Signal Detection Theory – Alternatives

J.D. Balakrishnan and J. A. MacDonald
Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, USA

1. INTRODUCTION

In *Signal Detection Theory*, we discussed the rational and the basic concepts behind the "formal model" approach to human performance assessment and then briefly sketched out the assumptions and methods associated with the field's most popular formal measurement system, signal detection theory. Over the past several decades, the experimental methods and statistics associated with the signal detection theory framework have become so deeply ensconced in the measurement literature that relatively few theorists would see any serious reason to question them. Competing views have also been developed, however, and should not be ignored merely because they are less familiar. After all, the validity of signal detection theory has never been established, and from time to time some cracks in its foundation have been discovered. In fact, here we review some very recent empirical results that purportedly show that the signal detection theory measures have actually grossly misled human factors researchers about the nature and limits of human performance. First we review the experimental data that make signal detection theory seem so compelling and then consider the strength of this evidence and some alternative interpretations.

2. EVIDENCE IN SUPPORT OF SIGNAL DETECTION THEORY

For most human performance researchers, the important question about signal detection theory is not whether it is a valid theory or not, but whether it is valid enough for their purposes. Typically, the performance statistics are recruited to identify changes in sensitivity and response bias, not their absolute values. The signal detection theory measures are trusted in this role because the way they change under different experimental conditions is usually predictable and consistent with the theory. Most importantly, perhaps, the theory correctly predicts the effects of changing the base rates or payoffs in a yes–no detection task: increasing the frequency of the signal trials almost invariably increases both the hit and the false alarm rates. Stronger manipulations lead to stronger effects of the same kind. Similarly, the sensitivity indices (e.g. d' and area under the ROC curve) seem relatively unaffected by base rates and payoffs, while the bias measure b is not: it increases and decreases appropriately when the base rates or payoffs are manipulated, indicating that the operator's decision-making strategy is rational. Other kinds of evidence seem to justify the 'technical" assumptions of the theory while rejecting the underlying assumptions of many other statistics (Swets 1986b). Empirical z-ROC curves, for example, usually are almost perfectly linear, consistent with the distributional assumptions (normality) of signal detection theory.

Using these statistics for more fine-grained analyses of operator performance also leads to a plausible and cohesive account of human behavior. For example, studies have consistently shown that the operator shifts the detection criterion when the base rates are changed, but the size of the shift is presumably insufficient to cause the decision process to be "optimal" (i.e. to maximize the percentage of correct decisions). In other words, the decision process is "conservative" (e.g. Creelman and Donaldson 1968, Macmillan and Creelman 1990). In laboratory studies of watchkeeping (i.e. the vigilance paradigm), the measures have been used to "establish" that changes in the detection rate are due to changes in the operator's willingness to make a detection response under some conditions (e.g. relatively slow paced detection tasks) and losses of sensitivity under others (e.g. relatively fast paced detection tasks with a memory load; Parasuraman 1979). These accounts are interesting and plausible, and certainly do not raise any special concerns about the validity of the model that gives rise to them, even though it could easily have been otherwise. If the model was untenable, one might expect a more inconsistent or a more confusing pattern of results.

3. ALTERNATIVE INTERPRETATIONS OF THE DATA

Although the classical findings undoubtedly do tell us something important about human performance, they can also be interpreted in other ways, some of which are at least as plausible as the signal detection theory explanation. Response time models, for example, can reproduce all of these properties of the data, even though they represent the decision-making processes in a profoundly different manner (e.g. Townsend and Ashby 1983, Luce 1986). Instead of adjusting a detection criterion, these models assume that the operator accumulates information until enough evidence has been collected to justify one of the two possible responses. Because the decision is reached only after the accumulated information crosses a boundary, the number of samples (the encoding time) will depend on the quality or strength of the evidence as it is collected. If the operator receives weak information early on, for example, s/he will wait longer before responding so that additional information can be obtained. In signal detection theory, the decision-making process only plays a role after the encoding process is completed. Thus, this "static" model assumes that the amount of information collected is entirely independent of the information quality.

In some circumstances, such as clinical or medical diagnoses, the "fixed sample" assumption of signal detection theory is reasonable, or perhaps even necessarily true if the amount of physical evidence available to the decision-maker is not under the decision-maker's control. In many other situations, however, the decision to stop collecting new information and formulate a response is a crucial aspect of the operator's decision-making process, which may itself be biased in some way. In the real world detection problems associated with watchkeeping, for example, the operator must respond quickly as well as accurately to changes in the status of the system. Attempting to decide more quickly whether a signal or a non-signal event has taken place will generally increase both the false alarm and the miss rate (the so-called "speed–accuracy trade-off"), causing d' to decrease. Such biases might show up in mean response times of the operator or in the estimates of the parameters of a dynamic model, but would be invisible to signal detection theory. The difference between tasks that lend themselves to a signal detection theory analysis and those that do not is not always recognized: many researchers

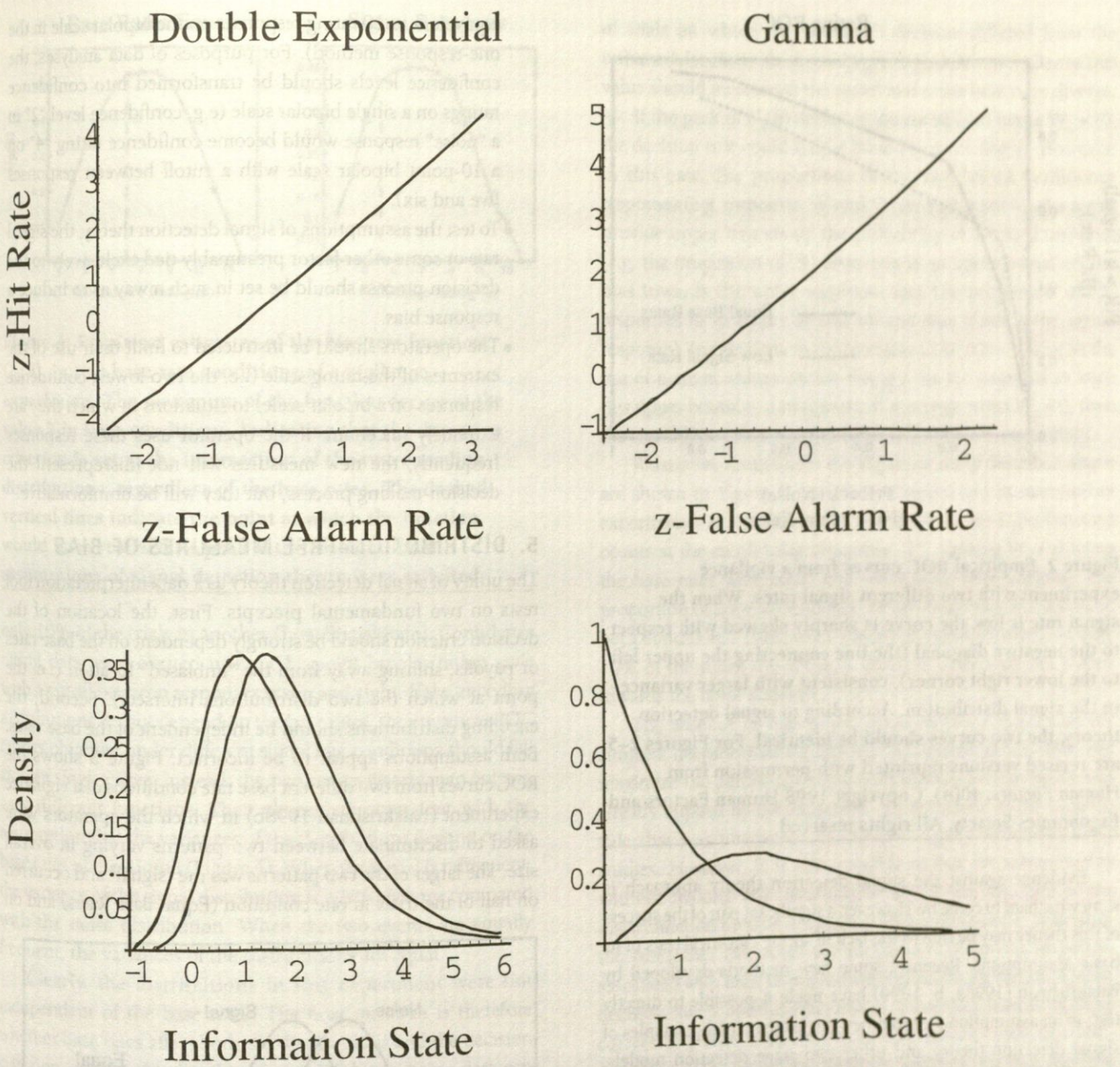

Figure 1. *z*-ROC curves obtained when pairs of hit and false alarms rate from non-normal distributions are converted to *z* scores from the standard normal distribution. The two "double exponential" distributions in the lower left panel have equal variance but are positively skewed (e.g. Luce 1986: 508). The "gamma" distributions are unequal variance and positively skewed (e.g. Luce 1986: 507). Many other distributions also predict quasi-linear *z*-ROC curves.

use *d'* or similar measures even when the operators are under time pressure.

The evidence purportedly supporting the normality assumptions of the signal detection theory model — i.e. the shape of the empirical ROC curve — is also very weak. Many other distribution models predict a virtually linear *z*-ROC curve. In fact, the small deviations from linearity in empirical *z*-ROC curves are actually larger than the deviations predicted by these alternative distributions. Examples of the *z*-ROC curve predictions of two other distributions, vastly different from the normal (and not "transformable to the normal"), are shown in Figure 2.

4. OTHER APPROACHES

Apart from the response time models, very few of the alternatives to signal detection theory were specifically developed to repair any specific or known deficiencies of *d'*. Instead, they were proposed as alternatives that might be more reliable if complete list of these alternative indices would be quite long (>20). Fourteen sensitivity indices are reviewed in Balakrishnan (1998a). Five different bias measures were recently compared by See *et al.* (1997). Not all of these measures were derived from an explicit description of the decision-making process, making it more difficult to evaluate them. However, the supposedly non-parametric indices make some testable predictions about the shape of the ROC curve and other predictions related to the effects of base rates and payoffs (e.g. Swets 1986a, Macmillan and Creelman 1996). In his review of 10 different sensitivity indices, Swets (1986a) concluded that a variable criterion measure that assumes normal distributions with different variance (area under the normal ROC curve) was the most viable index. Similarly, See *et al.* (1997) recommended a variant of the criterion value from signal detection theory over several supposedly non-parametric bias indices.

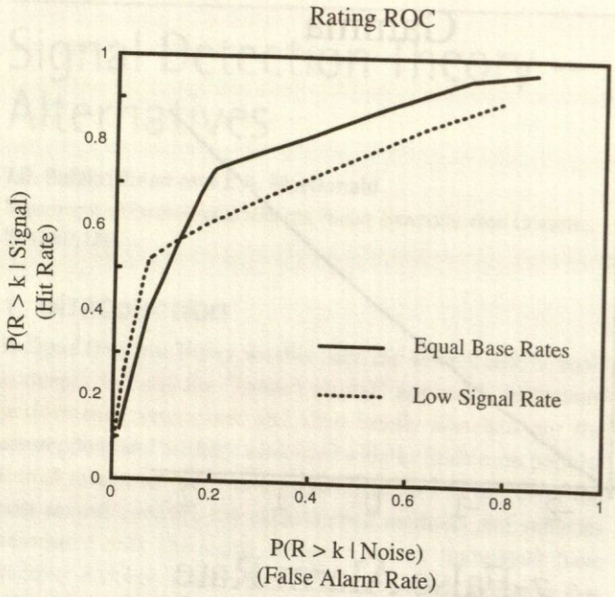

Figure 2. Empirical ROC curves from a vigilance experiment with two different signal rates. When the signal rate is low, the curve is sharply skewed with respect to the negative diagonal (the line connecting the upper left to the lower right corner), consistent with larger variance in the signal distribution. According to signal detection theory, the two curves should be identical. For Figures 2–5 are revised versions reprinted with permission from *Human Factors*, 40(4). Copyright 1998 Human Factors and Ergonomics Society. All rights reserved.

Evidence against the signal detection theory approach is somewhat hard to come by. However, a significant part of the success of this theory may be due to the lack of strong empirical tests of its basic assumptions. Recently, some new methods developed by Balakrishnan (1998a, b, 1999) have made it possible to directly test, in an assumption-free manner, the fundamental principles of signal detection theory and other two stage detection models, including the assumption that (1) response biases exist, (2) the decision process tends to be conservative and (3) the encoding and decision-making processes are independent. To avoid the assumptions required by *d'* and b, these tests make use of some extra information about encoding and decision-making contained in confidence ratings data. In many respects, these new methods are a relatively straightforward extension of the ratings paradigm developed earlier by signal detection theorists. However, they differ from the traditional signal detection theory methods in three ways:

- A cutoff between the "noise" and "signal" responses must be explicitly defined on a bipolar confidence rating scale. For example, if the rating scale has 10 values numbered from 1 to 10, rating responses five and six would be labeled "lowest confidence noise" and "lowest confidence signal" responses respectively. The two extremes of the scale (responses 1 and 10) would represent "highest confidence noise" and "highest confidence signal" responses respectively. Alternatively, the researcher may elicit first the yes–no detection response and then a "confidence level" response. For each confidence level in this two-response method, there is a corresponding confidence rating in the one-response method (e.g. five levels of confidence in the two-response method would be equivalent to 10 rating responses on the bipolar scale in the one-response method). For purposes of data analyses, the confidence levels should be transformed into confidence ratings on a single bipolar scale (e.g. confidence level "2" in a "noise" response would become confidence rating "4" on a 10-point bipolar scale with a cutoff between responses five and six).
- To test the assumptions of signal detection theory, the signal rate or some other factor presumably tied exclusively to the decision process should be set in such a way as to induce a response bias.
- The operators should be instructed to limit their use of the extremes of the rating scale (i.e. the two lowest confidence responses on a bipolar scale) to situations in which they are extremely uncertain. If the operator uses these responses frequently, the new measures will not misrepresent the decision-making process, but they will be uninformative.

5. DISTRIBUTION-FREE MEASURES OF BIAS

The utility of signal detection theory as a data interpretation tool rests on two fundamental precepts. First, the location of the decision criterion should be strongly dependent on the base rates or payoffs, shifting away from the "unbiased" location (i.e. the point at which the two distributions intersect). Second, the encoding distributions should be independent of the base rates. Both assumptions appear to be incorrect. Figure 3 shows the ROC curves from two different base rate conditions of a vigilance experiment (Balakrishnan 1998b) in which the operators were asked to discriminate between two patterns varying in overall size. The larger of the two patterns was the "signal" and occurred on half of the trials in one condition (Equal Base Rates) and on

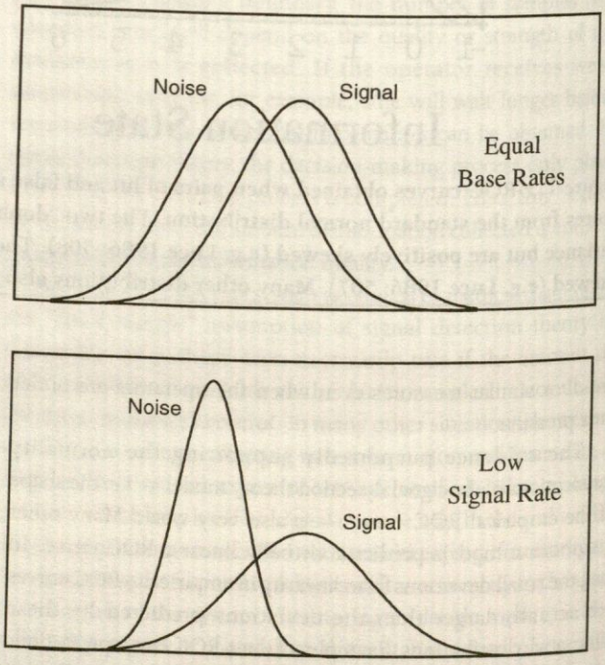

Figure 3. One possible distribution model that could account for the ROC curve data in Figure 2. Decreasing the signal rate increases the variance of the signal distribution while decreasing the variance of the noise distribution.

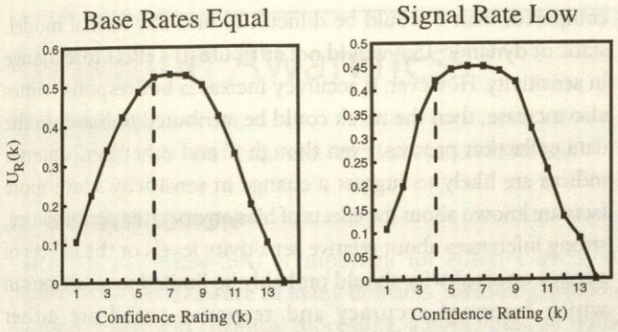

Figure 4. Empirical estimates of the bias test function, $U_R(k)$, in two base rate conditions of a vigilance experiment. The maximum of the function occurs at the value 7 in both conditions, indicating that the detection criterion is set at the intersection of the two encoding distributions, regardless of the base rates. The dashed vertical lines indicate the point at which the function would have reached its peak if the distributional assumptions of signal detection theory were satisfied.

only 10% of the trials in another (Low Signal Rate). Confidence rating data were elicited using a 14-point bipolar rating scale with a cutoff between responses seven and eight. If the encoding distributions do not depend on the base rates, the empirical ROC curves obtained under different signal rate conditions should fall along a single curve. Instead, the two curves clearly map out two very different functions. Their shapes are consistent with the assumption that the variances of the distributions depend on the base rates of the stimuli (Figure 4). When the signal is infrequent, the variance of the signal distribution is relatively large compared with the noise distribution. When the two events are equally frequent, the variances of the distributions are equal.

Clearly, the distributions in this experiment were not independent of the base rates. The next question is therefore whether base rates affect both the distributions and the decision criterion, or just the distributions. Confidence rating data can also be used to determine whether there is any shift of the criterion, without involving any assumptions about the shapes of the encoding distributions. The test is based on the difference between the cumulative relative frequency histograms of the operator's confidence rating responses,

$$U_R(k) = F_N(k) - F_S(k),$$

where $F_N(k)$ and $F_S(k)$ are the proportions of rating responses less than or equal to k on noise and on signal trials respectively, and the argument k is the rating value on a bipolar scale with a cutoff at k^*. If this function is decreasing for any k associated with a noise response ($k \leq k^*$), or increasing for any k associated with a signal response ($k > k^*$), then the decision rule is biased (i.e. the criterion is not set at the point where the two encoding distributions intersect). The total proportion of these "biased rating responses" in the data, or W_p, is an estimate of the proportion of times the participant makes a biased response. For example, if the rating scale has 10 values with a cutoff between responses five and six, and $U_R(k)$ reaches it maximum value at rating response three, then the total proportion of four and five responses during the experiment is the (estimated) proportion of trials on which the operator's decision differed from the unbiased decision rule. According to signal detection theory, this value should increase as the signal and noise base rates diverge.

If the peak of $U_R(k)$ occurs at the cutoff (and hence $W_p = 0$), the decision rule could still be biased to some degree. However, in this case, the proportions of the two lowest confidence responses (e.g. responses "4" and "5" on a 10-point bipolar scale) provide upper bounds on the probability of a biased response (e.g. the proportion of "4" responses is an upper bound on the bias towards the noise response, and the proportion of "5" responses is an upper bound on the bias towards the signal response). Instructions to the operator to be conservative in the use of extreme values on the rating scale are intended to keep this upper bound to a minimum. If it is large when $W_p = 0$, then the test will be relatively uninformative (but not misleading).

Illustrative results from the vigilance study described above are shown in Figure 5. In these examples and in many other experiments in our laboratory, the peak of the $U_R(k)$ function occurs at the cutoff value (response "7"), causing $W_p = 0$ when the base rates were equal and when they were unequal. The proportion of lowest confidence "noise" responses was < 0.01 in both conditions. Thus, even when the signal occurred on only 10% of the trials, the subjects' decision rules were not biased towards the noise response.

The absence of bias in the decision rule implies that the subject's decision-making strategy is suboptimal (i.e. does not maximize the percentage correct decisions). Suboptimality was already implied by the supposed "conservatism" of the decision rule that was uncovered by classical signal detection theory studies. However, it is also possible to test for suboptimality without making any assumptions about the structure of the discrimination process or the encoding distributions. To do this, the researcher calculates the proportion of correct responses associated with each rating response given by the operator. If the decision rule is optimal (maximizes % correct), then all of these correct response proportions will be greater than one half, regardless of the base rates. If any are less than one-half, then the decision rule is suboptimal. For example, considering only the trials of the experiment on which the operator chose rating response "2" on the 10-point bipolar rating scale (i.e. the "noise

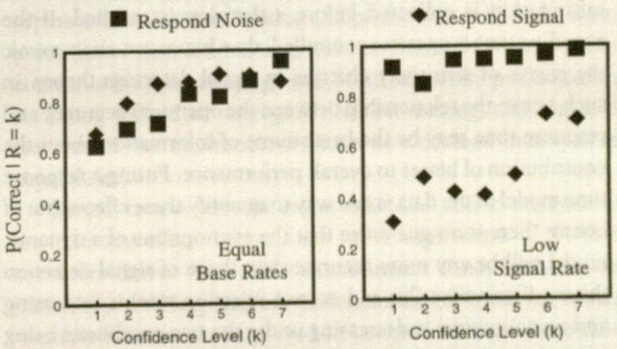

Figure 5. Proportion correct conditioned on the discrimination response and confidence level for two base rate conditions of a vigilance experiment. Any value < 0.5 indicates that overall proportion correct could be improved by changing the decision rule (i.e. the decision rule is suboptimal).

responses" given at confidence level 4), more than half of these responses should have occurred on "noise" trials. If this is not true, then the operator could improve his/her performance by simply switching from a "noise" to a "signal" response whenever s/he would normally make this "2" response.

Examples of this optimality test are shown in Figure 6. Notice that for several of the lower confidence "signal" responses, the subjects are more often incorrect than correct. Reversing these "suboptimal" responses to "noise" responses "corrects" for the suboptimality of the decision rule, providing a measure of the performance level that the subjects could have achieved if their decision rules were optimal. This *post-hoc* correction process has some important potential applications in the workplace: in principle, a trainer can make use of this information to give the trainees focused feedback that will allow them to optimize their decision-making strategies.

6. IMPLICATIONS FOR PERFORMANCE ASSESSMENT

Unless and until some other interpretation of these new empirical tests can be found that is consistent with the basic tenets of signal detection theory, the legitimacy of the *d'* and b analysis of discrimination performance is seriously open to question. Apparently, correcting for the effects of bias on the performance of a human operator is not merely a matter of adjusting for the value of a decision criterion, but instead involves explaining how and why biases affect the two distributions that describe the operator's information states. Area under the ROC curve, *d'* and other indices associated with signal detection theory may still provide some useful information about the operator's behavior, but it is not clear how this information should be interpreted. Similarly, most of the other indices developed to complement or replace the signal detection theory measures also rely heavily on the assumption (implicitly or explicitly stated) that the two encoding distributions are not affected by response bias.

Correcting for any suboptimality in the decision rule using the methods described above will make it possible to compare different operators or systems without the results being confounded by biases in the operator's decision rule. However, this method does not control for biases in the decision processes that influence other aspects of behavior, including how much information is collected before a decision is reached. If the encoding time is operator controlled, then biases can easily mimic the effects of sensitivity changes in signal detection theory. In such a case, the relationship between the operator's accuracy and response time may be the best source of information about the contribution of biases to overall performance. Fitting a response time model to the data is one way to quantify these effects, but of course there is no guarantee that the assumptions of a dynamic model will be any more accurate than those of signal detection theory. If accuracy (hit and correct rejection rate) is increasing and response time is decreasing under the two conditions being compared, than it would be difficult to find any formal model, static or dynamic, that would not attribute this effect to a change in sensitivity. However, if accuracy increases but response times also increase, then the result could be attributed to biases in the data collection process, even though *d'* and other performance indices are likely to suggest a change in sensitivity. Until more facts are known about the effects of bias on operator performance, strong inferences about relative sensitivity levels or the effects of a factor on sensitivity should probably be limited to situations in which response accuracy and response speed are either independent or positively covarying.

ACKNOWLEDGEMENTS

NASA Dryden Flight Research Center Grant NCC2-374 supported the authors.

REFERENCES

BALAKRISHNAN, J.D., 1998a, Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, 3, 68–90.

BALAKRISHNAN, J.D., 1998b, Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, 40, 601–23.

BALAKRISHNAN, J.D., 1999, Decision processes in discrimination: fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance* (in press).

CREELMAN, C.D. and DONALDSON, W., 1968, ROC curves for discrimination of linear extent. *Journal of Experimental Psychology*, 77, 514–16.

LUCE, R.D., 1986, *Response Times: Their Role in Inferring Elementary Mental Organization* (New York: Oxford University Press), 3, 164–70.

MACMILLAN, N.A. and CREELMAN, C.D., 1990, Response bias: characteristics of detection theory, threshold theory, and "non-parametric" indexes. *Psychological Bulletin*, 107, 401–13.

MACMILLAN, N.A. and CREELMAN, C.D., 1996, Triangles in ROC space: history and theory of "non-parametric" measures of sensitivity and response bias. *Psychonomic Bulletin and Review*,

PARASURAMAN, R., 1979, Memory load and event rate control sensitivity decrements in sustained attention. *Science*, 205, 924–7.

SEE, J.E., WARM, J.S., DEMBER, W.N. and HOWE, S.R., 1997, Vigilance and signal detection theory: an empirical evaluation of five measures of response bias. *Human Factors*, 39, 14–29.

SWETS, J.A., 1986a, Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, 99, 110–17.

SWETS, J.A., 1986b, Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin*, 99, 181–98.

TOWNSEND, J.T. and ASHBY, F.G., 1983, *Stochastic Modeling of Elementary Psychological processes* (New York: Cambridge University Press)