HUMAN FACTORS, 1979, 21(1), 69—78

Nonparametric Measures of Sensory Efficiency for Sustained Monitoring Tasks

ANGUS CRAIG[1], *Medical Research Council Perceptual and Cognitive Performance Unit, Experimental Psychology Laboratory, University of Sussex, Brighton, England*

The validity of d', the signal detection theory measure of sensory efficiency, seems suspect in applications to vigilance and inspection, and it is suggested that the investigator should use an alternative, distribution-free measure instead. Three nonparametric measures of efficiency which seem particularly well suited to vigilance applications are considered. The measures, A_G (Green and Swets, 1966), A' (Pollack and Norman, 1964), and E (McCormack, 1961) are defined and illustrated, and an example is provided demonstrating their use and suitability in analyzing data from a sustained monitoring task. Norman's (1964) nonmetric procedure for comparing performance efficiency is also applied to the data and is shown to provide a useful means for distinguishing between measure-dependent and measure-independent inferences about changes in efficiency. The relative merits of the three nonparametric measures are discussed.

INTRODUCTION

Over the past decade, signal detection theory (Green and Swets, 1966) has emerged as the dominant theory for interpreting performance in prolonged vigilance and inspection tasks, and *d'*, the signal detection theory measure of sensitivity, has become increasingly popular as an index of the observer's sensory efficiency, his actual ability to discriminate between target and nontarget (or noise) events (Mackworth, 1970; Broadbent, 1971). But at the same time, there has also been a growing awareness that the *d'* statistic may suffer from problems of validity and reliability when derived from data obtained in such sustained monitoring tasks.

The validity issue relates to a central assumption which underlies the derivation of *d'*, namely, that the perceived effects of target and nontarget events are both normally distributed with equal variance, (the distributions differing only in their means by the quantity *d'*, expressed in unit normal deviates). When the equal-variance assumption is seen to be false, *d'*, is not a valid measure of the separation between the means of the distributions, and alternative parametric efficiency measures such as Δm and d'_e are used instead (see Green and Swets, 1966, pp. 96–98). But these measures depend on knowledge of the ratio of the variances and hence cannot be applied if the ratio is unknown, as, for example, where the basic datum from which the sensitivity measure is to be estimated consists of only a single pair of hit and false-alarm rates.

[1] Requests for reprints should be sent to Mr. Angus Craig, MRC Perceptual and Cognitive Performance Unit, Laboratory of Experimental Psychology, University of Sussex BN1 9QG, England.

© 1979, The Human Factors Society, Inc.
All rights reserved.

Various authors have indicated that the equal-variance assumption is particularly suspect in the typical vigilance task where no external feedback is provided, there is uncertainty regarding the target arrival time, and target occurrence is relatively rare (e.g., Mackworth, 1970; Taylor, 1967; Swets and Kristofferson, 1970). Unfortunately, in a great many cases, this disquiet is coupled with ignorance of the actual variance ratio, so that neither Δm nor d'_e can be used.

The problem is exacerbated by measurement unreliability, since, in within-session analyses of vigilance performance, d' is frequently estimated from a data base consisting of at most a few hundred trials, of which perhaps only 20 to 30 have included the target event, and over which the observer may occasionally fail to produce an error of commission or omission.

When similar problems arise in a psychophysical context, one can frequently point to an external criterion against which the competence of d' to describe the data may be assessed. For example, d' *should* increase when signal intensity is increased, but *should* remain constant when the signal level is unchanged. In vigilance, however, the external validating criterion is often lacking. It would be no more justifiable to consider d' acceptable because it failed to show a within-session decline, than it would be to accept it because it did show a decline. What *should* happen to d' is not known. Whenever such uncertainty exists, either within-session or between sessions, any doubts about the measurement validity of d' assume an even greater importance.

It has been suggested that these problems may be partially circumvented by the adoption of a nonparametric (distribution-free) measure of sensitivity (Swets and Kristofferson, 1970). It was the author's intention to consider in this paper the empirical merits of a number of such nonparametric alternatives to d', some of which may be viewed as theory-related measures, others merely as consistent with commonsense notions about operator efficiency. In fact only three measures will be dealt with here, since the other six in an original ensemble of nine measures failed to meet at least one of three criteria for selection.

CRITERIA FOR MEASURE SELECTION

(*a*) The first criterion is that the measure yields scores which preserve the ordering of performances about which categorical judgments of superiority or inferiority can be made, using the procedure outlined by Norman (1964). This approach is illustrated in Figure 1. It is assumed that the point *A*, representing an observed performance, lies on a curve which is the locus of all points representing performances equally sensitive to *A*. This curve is assumed to have one end point where the false alarm rate is zero, and another where the hit rate is unity. It is also assumed that the equal-sensitivity curve, relating hit rate to false alarm rate, is a

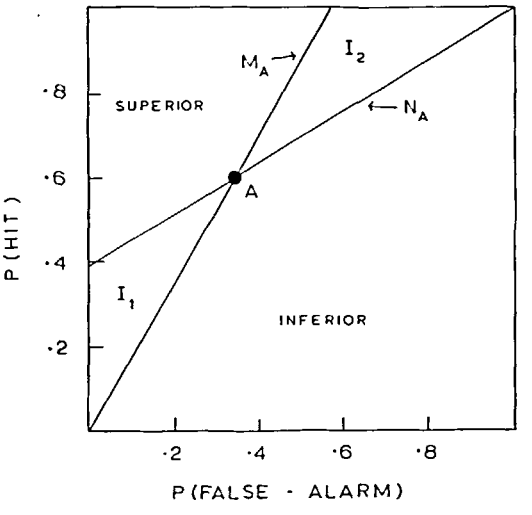


Figure 1. *Limits on "proper" ROCs passing through a yes—no operating point.*

monotonic increasing function, with non-increasing slope. The lines M_A and N_A represent the bounds on all such equal-sensitivity curves which pass through the point A.

Compared with performance A, a second performance B is categorically superior, and could not have been achieved by chance (e.g., by guessing), if it lies in the upper left region of the square. Similarly, B is categorically inferior to A if it lies in the lower right region. Otherwise, if B lies in I_1 or I_2, judgment is only determined if one specifies a particular measure of efficiency. (It will be shown later that Norman's procedure for comparison may prove sufficient without recourse to any further measurement).

(*b*) The second criterion is that the scores should be independent of response-bias, since otherwise the measure cannot be held solely to reflect the observer's ability to distinguish between the two events. In some instances, there are sufficient a priori grounds for rejecting a measure on the basis of this criterion; in other cases, the association with response-bias may not be obviously apparent, but the measure may nevertheless be rejected if an empirical association is found.

(*c*) Finally, it seems a reasonable requirement to demand that the measure should define a score even when there are no errors of omission or of commission. The examination of within-session data blocks from vigilance experiments frequently reveals that an operator has made no false alarms in one or more blocks, and this presents measurement problems if the investigator wishes to index efficiency by means of d'. Consequently, if one wishes to recommend an alternative measure which could be applied to this type of situation, it should satisfy this third criterion and provide a finite score, even with error-free data.

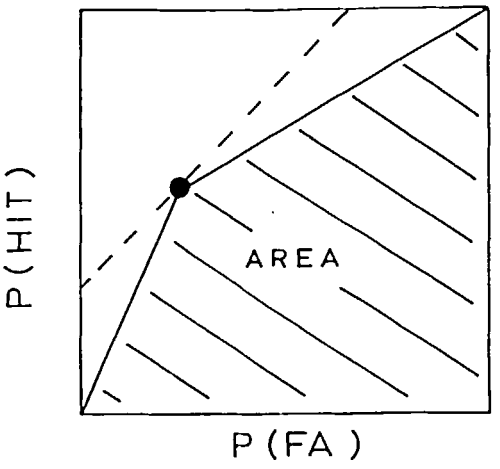
These three criteria impose sensible limits on the efficiency measures which may be considered. Norman's (1964) suggestion restricts one to measures whose implied equal-sensitivity loci are lines or curves for which hit rate, *P(H)*, is a monotonic increasing function of false alarm rate, *P(FA)*, while the other two criteria ensure that an efficiency score can always be determined and that it will not be confounded with changes in response bias. The six measures excluded from further consideration on these grounds were (1) the overall percentage of correct decisions made (an appropriately weighted sum of hit rate and correct rejection rate) which failed to meet the second criterion; (2) the proportional correctness of affirmative reports (i.e., the ratio of hits to hits plus false alarms) which might have proved useful for the temporally unstructured detection tasks but which also failed to meet the second criterion; (3) a measure derived from common sense notions of efficiency which turned out to be the inverse square root of the stimulus similarity index proposed by Luce (1959). Rather surprisingly, since Luce's measure is intended to be independent of response-bias, this measure also failed on criterion (*b*) in addition to failing on the third criterion; (4) the product of hit rate and correct rejection rate, which seemed an intuitively reasonable measure, but which failed to meet the first and second criteria; (5) the complement of the area of the region labelled "superior" in Figure 1 (i.e., the sum of areas A, B, and C shown for the A' measure in Figure 2 below) which failed on the second criteria; and (6) another area measure, the complement of the rectangle bounded by miss rate and false alarm rate which is formed in the upper left corner of the ROC space, which again failed on the second criterion, and which also failed to meet the first criterion.

THREE CANDIDATE MEASURES

For the measures actually considered here, the appropriate equations, graphic illustrations and implied equal-sensitivity loci (bro-

MEASURE 1

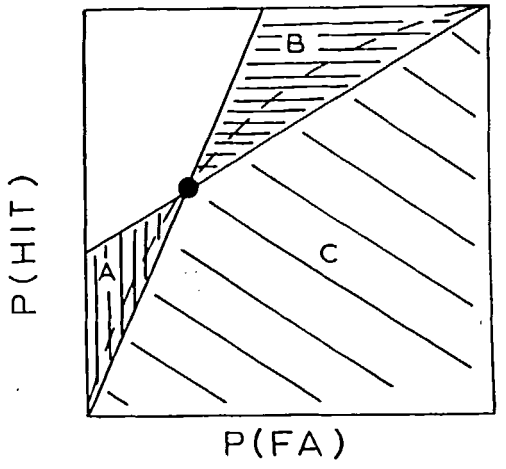
$$A_G = \text{AREA} = \frac{P(H) + [1 - P(FA)]}{2}$$



MEASURE 2

$$A' = \text{AREA} = C + \frac{A+B}{2}$$

$$= 1 - \tfrac{1}{4}\left[\frac{P(FA)}{P(H)} + \frac{[1 - P(H)]}{[1 - P(FA)]}\right]$$



MEASURE 3

$$E = 1 - \left[\frac{1 - P(H)}{1 - P(H).\,P(S) + P(FA).P(S) - P(FA)}\right]$$

Where $P(S)$ = signal probability
(i.e. batch quality)

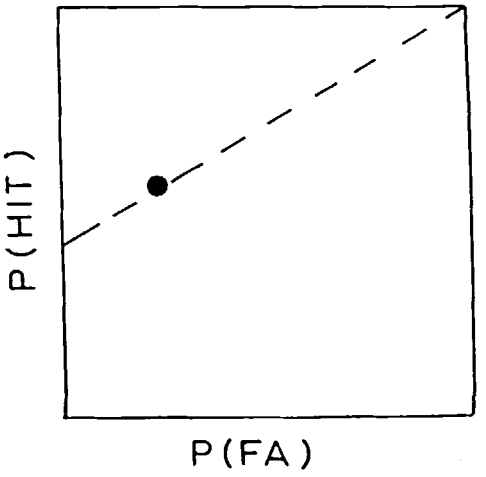


Figure 2. *Defining equations and implied equal-sensitivity contours for three distribution–free measures of sensory efficiency.*

ken lines) are presented in Figure 2. To avoid some confusion it might be noted that the areas shown for measures 1 and 2 represent the measures themselves and not the operation regions in the unit square mentioned by Norman (1964).

(1) This measure, A_G, is an estimate of the "area under the yes-no ROC" mentioned by Green and Swets (1966) and is of theoretical relevance in signal detection theory. Interestingly, the implied equal-sensitivity contour for A_G (the broken line in the first diagram in Figure 2) is a straight line at 45°, whereas signal detection theory implies that the equal-sensitivity contour (corresponding to constant d') is a curve of decreasing slope, similar in appearance to the broken curve shown in the central diagram in Figure 2. But a straight line at 45° is implied by the simple "threshold" decision model presented in Atkinson, Bower, and Crothers (1965) which provides a reasonable fit for at least some psychophysical data (Friedman, Carterette, Nakatani, and Alumada, 1978), so that A_G will also be directly related to this "threshold" model.

(2) This measure, A', is described by Pollack and Norman (1964) and derives from Norman's (1964) graphic representation of the upper and lower bounds on monotonic ROCs with nonincreasing slope. As Egan (1975) points out, all "proper" ROCs which pass through the operating point are contained within the limits of the areas A and B. The average of $A + B$ coupled with area C provides a closer estimate of the "area under the yes-no ROC" than does measure 1, which gives a minimal estimate only of this area.

(3) This measure reverts to a more commonsense notion about operator efficiency. It is the E measure defined by McCornack (1961) in his report on inspector accuracy. The quality of the batch (proportion of defective articles present) before and after inspection is compared, and the percentage improvement (or deterioration) achieved is then related to the maximum improvement which was possible. Clearly, this is an intuitively reasonable measure to use for vigilance and inspection tasks.

ASSESSMENT OF MEASURES

For illustrative purposes the data from a sustained monitoring study reported by Craig, Colquhoun, and Corcoran (1976) are used. This study, with its low target occurrence rate and moderate degree of discrimination difficulty, is reasonably representative of vigilance and inspection operations and produced typically low false alarm rates. Since individual hit rates and false alarm rates were presented by the authors for each of 18 subjects under each of the three display conditions investigated (auditory (A), visual (V), and combined audio-visual (AV) presentation) the data base thus provided is not meager, and enables reasonable comparisons of the measures to be made. An additional bonus is that individual differences between the subjects in the study produced a useful range of variation in the scores: hit rate ranged from approximately 20% to 90%, while false-alarm rate ranged from just over 1% to almost 11%.

As measures of response-bias, Craig, Colquhoun, and Corcoran (1976) reported both β, the likelihood-ratio criterion of signal detection theory, and P ("signal"), the probability of an affirmative report (irrespective of whether the report was correct or not). Table 1 represents the observed degree of association over all conditions, between each of the presently suggested efficiency measures and these response-bias measures. It seems safe to say that the efficiency scores are not confounded with response-bias, thereby satisfying the second criterion.

Table 1 also lists the correlation between the equal-variance d' estimates (read from

TABLE 1

Between Measure Correlations (Rho)

	P("signal")	*β*	*d'*
Measure 1 (A_G)	0.245	−0.130	0.909*
Measure 2 (A')	0.112	0.002	0.959*
Measure 3 (E)	0.259	−0.140	0.904*
d'	−0.104	0.218	
β	−0.947*		

*$p < 0.01$

Freeman's, 1973, tables of d' and β) and the nonparametric measures. These correlations were included in the table since the measures under discussion are intended as possible substitutes for d'. All of these empirically acceptable measures are very highly correlated with d', so that one might indeed regard each as a potentially useful substitute for that measure. This statement is in no way intended to imply that a high correlation with d' provides an additional criterion for measure selection. Instead, since it is the d' measure that is suspect, the obtained correlations may be regarded as providing empirical support for using the parametric d' measure to describe these particular data.

Table 2 lists the results of the between-condition comparisons based on each of the three nonparametric measures, and also on d'. All measures are in agreement regarding the comparisons A versus V and V versus AV. But there is no similar consensus about the two remaining comparisons, A versus AV and A, V versus AV. Measures 1 and 3 are in-agreement, but Measure 2 differs from them. The comparisons based on d' agree with those based on Measure 2. It is the author's surmise that these discrepancies arise because of the differences in form of the equal-sensitivity curves. Measures 1 and 3 both imply straight lines, whereas Measure 2 and d' imply curves with decreasing slope (see Figure 2). This possibly accounts for the agreement pairings and the differences between pairings, in the Table 2 comparisons A versus AV and A, V versus AV. (Note, however, that the disagreement between the pairings does not permit the inference that one pair of measures is superior to another. If the author's surmise is correct, then such an inference must rest on an additional presumption regarding the form of the equal-sensitivity curves.)

The results of these Table 2 comparisons, taken in conjunction with the correlations listed in Table 1, suggest that of the nonparametric measures considered here, Measure 2, which is Pollack and Norman's (1964) area measure, A', is the most suitable alternative to d'. Also, these results further attest to the adequacy of the d' measure to describe this particular set of data.

Measure 1, A_G, the simplest approximation to the area under the yes-no ROC curve (Green and Swets, 1966), is convenient and easy to use (for measurement purposes it is not even necessary to use the divisor, 2), and seems a useful alternative to d', as a measure of sensory efficiency, in respect of vigilance or inspection data.

The third measure, however, McCornack's (1961) E score, seems the most intriguing one of all. In practical applications, this measure is ideally suited for describing data from vigilance and inspection studies, because it comes closest to the definition of efficiency used by managers and the like who are re-

TABLE 2

Between Condition Comparisons (Wilcoxon)

Comparison	*A vs. V.*	*A vs. AV*	*V vs. AV*	*A, V[1] vs. AV*
Measure 1 (A_G)	A > V**	A < AV*	V < AV**	A, V < AV*
Measure 2 (A')	A > V**	n.s.	V < AV**	n.s.
Measure 3 (E)	A > V**	A < AV**	V < AV**	A, V < AV*
d'	A > V**	n.s.	V < AV**	n.s.

*$p < 0.05$
**$p < 0.01$
[1] comparing the better of A (auditory) or V (visual) with AV (combined audio-visual).

sponsible for real operational efficiency. For similar reasons it might also prove useful in theory-oriented research where its adoption as a nonparametric efficiency measure would render the description of research findings more readily intelligible to the less academically inclined managers.

Some comment is also in order regarding the d' measure itself. The finding that d' seemed an adequate descriptor for the data used in this study may be regarded as an indication, either that the underlying variances were in fact approximately equal or that d' is relatively robust to departures from equality. This could be explored usefully in future study by a Monte-Carlo simulation using signal and noise distributions with known variances and means. Measures could then be compared in relation to systematic departures from the equal-variance assumption. If, as is possible, the measures are differentially sensitive to these departures, this would provide a rational criterion for prefering one measure to another.

A NONMETRIC APPROACH

It was previously stated that Norman's (1964) procedure could also be used to compare performances. This is now demonstrated. The procedure has the distinct advantage of requiring fewer assumptions than any of the measures considered in the preceding section but suffers both the disadvantages of not producing a measure score and of producing a tendency toward more indeterminate judgments when bias shifts are present than when they are not. The latter disadvantage is particularly relevant to the analysis of within-session changes during vigilance where bias shifts are frequently observed (Mackworth, 1970). Richardson (1972) has further objected to the method on the grounds that it is a laborious one involving a graphic comparison for each data point. But, his criticism is hardly a scientific one; and, in any case, there is a less laborious way of comparing data points.

Consideration of Figure 1 indicates that if there exists a point B which is superior to point A, then the line M_B must be steeper than M_A, while N_B is less steep than N_A. The slope of lines M is defined by $P(H)/P(FA)$, while lines N have slope defined by $[1 - P(H)]/[1 - P(FA)]$, and these slopes are easily calculable from the basic data. The following categorical judgments may then be made:

A is superior to B if: $(M_A - M_B)$ is positive *and* $(N_A - N_B)$ is negative;

A is inferior to B if: $(M_A - M_B)$ is negative *and* $(N_A - N_B)$ is positive; otherwise, when slope differences M, N are both positive or both negative, or zero, no categorical judgment may be made. This procedure is simple and does not entail graphic comparisons. It was applied to the data reported by Craig, Colquhoun, and Corcoran (1976) with the results shown in Table 3.

From these data it can be seen that there remains some doubt only for the A versus AV and A, V versus AV comparisons with a bias only slightly favoring the superiority of AV and suggesting that a conclusion of "no difference" is more likely. For the comparisons A versus V and V versus AV, the nonmetric technique proves sufficient by itself for distinguishing between the efficiency levels, without recourse to any further computation (i.e., there is no need to derive an efficiency score). Despite these noted disagreements, the between-condition comparisons shown in Table 2 are all perfectly consistent with these conclusions drawn from Norman's (1964) technique, but, whereas, the significant differences obtained for A versus V and V versus AV are independent of the measure used, those obtained for A versus AV and A, V versus AV are clearly measure-dependent.

TABLE 3

Between Condition Comparisons Using Norman's Nonmetric Procedure

Comparison	*A vs. V*	*A vs. AV*	*V vs. AV*	*A, V vs. AV*
No. of categorical judgments possible	17	13	17	13
Minority judgments: number and favorable direction	1 (V)	3 (A)	1 (V)	4 (A,V)
Indeterminate	1	5	1	5

Permissible conclusions:

1. A superior to V ($p \leq 0.002$, binomial)
2. V inferior to AV ($p \leq 0.002$, binomial)

Tentative inferences:

1. Possibly A does not differ from AV
   for 3/13, $p > 0.05$, n.s.
   if 3/18, $p < 0.01$
   if 8/18, n.s.
2. Possibly A,V does not differ from AV
   for 4/13, $p > 0.20$, n.s.
   if 4/18, $p < 0.05$
   if 9/18, n.s.

CONCLUSIONS

The problems which beset the d' measure of efficiency, when estimated from vigilance data, may be circumvented by the adoption of any one of the three nonparametric measures considered in this report. Each of these nonparametric alternatives retains the desirable property of being free from the confounding effects of response-bias, and none is troubled by the presence of errorless data. In addition, each measure offers its own particular basis for recommendation. For example, if an investigator is persuaded, by considerations additional to those mentioned here, that the equal-sensitivity contour is indeed a curve with a monotonically declining slope (like the curve implied by signal detection theory), then he would be recommended to use Measure 2, A', which implies that shape. If he is not so persuaded about the shape of the curve, then he may elect to use Measure 1, A_G, because it is easy to calculate or because it relates to a theory of detection, or he may opt for Measure 3, E, for its direct, obvious interpretation. Further study of these measures, as in the suggested Monte-Carlo simulation, may well reveal additional bases for recommendation.

Since the same intended function is served by all three measures, and since the conclusions arrived at on the basis of one measure may not agree with those based on another, it would seem profitable to precede the measurement and scoring by a nonmetric classification of the data as demonstrated. Irrespective of the particular efficiency measure chosen, the nonmetric procedure would indicate whether any obtained differences were measure-dependent or not.

ACKNOWLEDGMENT

The author wishes to thank an anonymous referee for his helpful comments, and in particular for his suggestion regarding the Monte-Carlo simulation.

REFERENCES

Atkinson, R. C., Bower, G. H., and Crothers, E. J. *An introduction to mathematical learning theory.* New York: Wiley. 1965.

Broadbent, D. E. *Decision and stress.* London: Academic Press, 1971.

Craig, A., Colquhoun, W. P., and Corcoran, D. W. J. Combining evidence presented simultaneously to the eye and the ear: A comparison of some predictive models. *Perception & Psychophysics,* 1976, *19,* 473-484.

Egan, J. P. *Signal detection theory and ROC analysis.* New York: Academic Press, 1975.

Freeman, P. R. *Table of d' and β.* Cambridge: University Press, 1973.

Friedman, M. P., and Carterette, E. C., Nakatani, L., and Ahumada, A. Comparisons of some learning models for response bias in signal detection. *Perception & Psychophysics*, 1968, *3*, 5-11.

Green, D. M., and Swets, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

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

McCornack, R. L. Inspector accuracy: A study of the literature. Albuquerque: Sandia Corporation, SCTM 53-61 (14), 1961.

Mackworth, J. F. *Vigilance and attention*. Harmondsworth: Penguin Books, 1970.

Norman, D. A. A comparison of data obtained under different false-alarm rates. *Psychological Review*, 1964, *71*, 243-246.

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

Richardson, J. T. E. Non-parametric indexes of sensitivity and response bias. *Psychological Bulletin*, 1972, *78*, 429-432.

Swets, J. A. and Kristofferson, A. B. Attention. *Annual Review of Psychology*, 1970, *21*, 339-366.

Taylor, M. M. Detectability theory and the interpretation of vigilance data. *Acta Psychologica*, 1967, *27*, 390-399.