Psychological Bulletin
1986, Vol. 99, No. 1, 100–117

Indices of Discrimination or Diagnostic Accuracy: Their ROCs and Implied Models

John A. Swets
BBN Laboratories Incorporated
Cambridge, Massachusetts

Tasks in which an observation is the basis for discriminating between two confusable alternatives are used widely in psychological experiments. Similar tasks occur routinely in many practical settings in which the objective is a diagnosis of some kind. Several indices have been proposed to quantify the accuracy of discrimination, whether the focus is on an observer's capacity or skill, on the usefulness of tools designed to aid an observer, or on the capability of a fully automated device. The suggestion treated here is that candidate indices be evaluated by calculating their relative operating characteristics (ROCs). The form of an index's ROC identifies the model of the discrimination process that is implied by the index, and that theoretical form can be compared with the form of empirical ROCs. If an index and its model yield a grossly different form of ROC than is observed in the data, then the model is invalid and the index will be unreliable. Most existing indices imply invalid models. A few indices are suitable; one is recommended.

Subjects in experiments on perception, learning, memory, and cognition are often required to make a series of fine discriminations. In a common method, a single stimulus is presented on each trial and the subject indicates which of two similar stimuli it is, or from which of two similar categories of stimuli it was drawn. In addition, in several practical settings, professional diagnosticians and prognosticators must say time and again which of two conditions, confusable at the moment of decision, exists or will exist. Among them are physicians, nondestructive testers, product inspectors, process-plant supervisors, weather forecasters, mineralogists, stockbrokers, librarians, survey researchers, and admissions officers. There is interest in knowing both how accurately the experimental subjects and professionals perform and how accurately their various tools perform, and a dozen or more indices of discrimination accuracy are in common use. In this article I cover a way of discriminating among those indices that permits sifting the ones that are valid and reliable from the ones that are not. This proposed touchstone for indices is the *relative* (or *receiver*) *operating characteristic* (ROC).

In this article I argue that there is no model-free approach to confusion data, and specify the models implied by several common indices. Many of the points I make may be familiar to experimental psychologists from previous discussions of signal detection theory, but they are generalized now to provide a theoretical overview of questions usually addressed heuristically, and with uneven success. The package is presented as a useful contribution to other fields and to those who have avoided the indices of detection theory in favor of indices presumed to make fewer or weaker assumptions.

The path of this article is not simple and quick, but the outcome is quite manageable. A half-dozen indices imply threshold models, which are clearly at odds with existing data. These indices are hence subject to unnecessary unreliability (instability, imprecision), and so, as a mnemonic device, might be given a near-failing grade of "D." Four indices are consistent with variable-criterion models, which are in much better agreement with the data. However, they assume as fixed a certain parameter that the data tell us must be free, and hence they may be given a "C." Lastly, a few indices drawn from the class of variable-criterion models accommodate the free parameter mentioned. They are the best available, but because improvements might still be made, they could be given a "B."

With the single-stimulus method, in which either Alternative A or Alternative B is presented, analysis of data in terms of the ROC can provide a relatively pure index of discrimination capacity, or accuracy. In particular, an ROC index may be largely unaffected by the discriminator's criterion for choosing, say, Alternative A (or, as in the terminology of the threshold models, by the discriminator's bias toward the choice of A). Data show that the decision criterion (or response bias) is necessarily involved in the single-stimulus method, and varies both from one person to another and within a person over time. It, and its vari-

Theodore G. Birdsall, in the Electrical Engineering Department of The University of Michigan, first taught me about ROCs when he invented them; he has continued to share his thoughts and help me refine mine, and did so with this article. Charles E. Metz, of The University of Chicago's Radiology Department, contributed substantially as the structure of this article developed by describing his mathematical insights and providing derivations and proofs. Ian B. Mason, my correspondent in the Australian Bureau of Meteorology, helped advance some of the ideas treated here and extended my view of their generality. David J. Getty responded to soundings in the next office and assisted with mathematics. Others whose earlier work and comments on a draft contributed to this article are David M. Green, R. Duncan Luce, and J. E. K. Smith. As a referee, Neil Macmillan made helpful suggestions for exposition.

This article was prepared in part during a sabbatical leave granted under the Science Development Program of BBN Laboratories Incorporated. Additional support was obtained from a contract with The University of Chicago, under National Cancer Institute Grant R01-CA-33261.

Correspondence concerning this article should be addressed to John A. Swets, BBN Laboratories Incorporated, 10 Moulton Street, Cambridge, Massachusetts 02238.

ation, will confound indices of accuracy unless steps are taken to isolate it.

The ROC is a graph of the functional relation between the proportion of times that Alternative A is chosen when it occurs and the proportion of times that Alternative A is chosen when Alternative B occurs—as the decision criterion or response bias varies. In signal detection theory, the first quantity is termed the *hit rate* and the second, the *false-alarm rate.* (Also indicating the frequent asymmetry between Alternatives A and B are the corresponding terms "true-positive ratio" and "false-positive ratio.") The two quantities in question vary together from low to high as the criterion for choosing Alternative A is made more lenient (or the bias toward the choice of A becomes stronger)—and, thus, for any particular degree of accuracy, an ROC curve is traced from left to right and low to high. Figure 1 shows an example implied by a particular model (discussed in the section on *Equal-variance, normal PDFs*). In general, an index of discrimination that specifies the locus of such a curve, rather than a single point on it, reflects all possible decision criteria or response biases, and hence is independent of any one (see, e.g., Swets, 1973).

Many accuracy indices are calculated from a single ROC point, in disregard of the full curve that results from variation in the decision criterion or in the response bias. However, I suggest here that a candidate index may be evaluated by plotting the family of ROCs that it implies. Strictly, one plots the isopleths, at various values of the index, on the ROC graph. An isopleth, or curve connecting points at which the index has a constant value, is the ROC implied by the index for that value. The gist of this article is that an index is valid, and is likely to be reliable, only if its implied ROCs have the same form as the empirical ROCs found for the discrimination problem (observer, task, setting) in question.

Fine structure is added to that theme because an index's ROCs can reveal several of its properties. They may show, for example, that the index violates a basic measurement purpose by giving the same value to better-than-chance and poorer-than-chance performances. They can disclose that the index depends on factors irrelevant to discrimination per se, not only the observer's tendency to choose one or the other alternative, but also the relative frequency of occurrence of the two alternatives. Further, ROCs may show that two apparently different indices are practically the same.

Principally, an index's ROCs can identify certain fundamental assumptions that use of the index makes about the nature of the discrimination process. In effect, they specify the model implied by the index for that process. One example of an assumption or model is that the representations (observations or samples) of the alternatives on which choices are based have just a few values (or states). The opposing assumption is that the values of representations vary continuously over a wide range. A second example is that certain states of the representations, bounded by a fixed threshold, lead directly to the choice of a given alternative. The opposing assumption is that the observer can choose which values will lead to which choice, or can relocate a decision criterion at will. A final example is that the basic statistics of the variable representations are the same across all tasks and discriminators. The opposing assumption is that those statistics can vary in some specified manner.

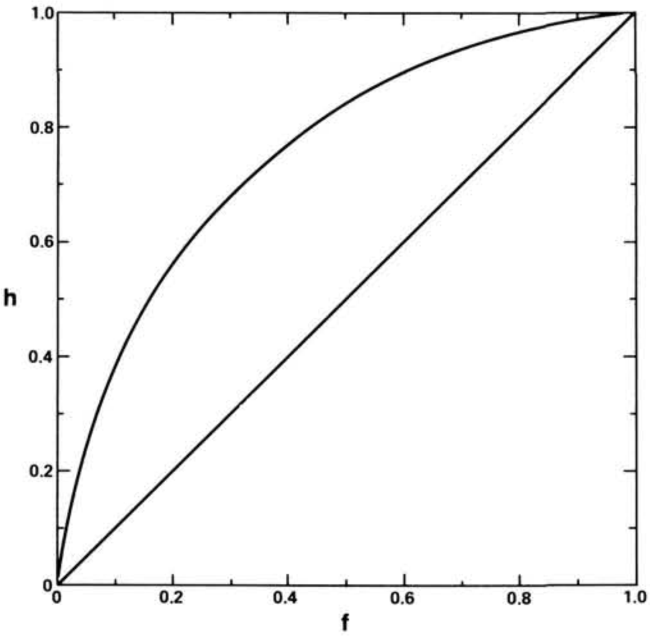

Figure 1. An illustrative ROC (relative operating characteristic), showing the conditional probability of a hit (h) as a function of that of a false alarm (f), as the decision criterion varies.

One might imagine that different models are appropriate for different tasks or observers. However, the working hypothesis is advanced here that all discrimination data will agree best with a model that includes: (a) continuous representations of the alternatives, (b) a decision criterion controlled by the observer, and (c) a particular free parameter of the statistics of the representation values; namely, the relative variance of the distributions of representation values that are associated with the two alternatives. If this hypothesis is true, then using an index that implies another model requires specific justification, conceivably on some pragmatic basis.

My plan is first to set forth the formal description of discrimination performance in the 2 × 2 contingency table, or confusion matrix, and to define various accuracy indices in terms of the quantities in that table. Then I review ROC theory; I discuss general ways of generating ROCs of various forms for the class of continuous, variable-criterion models; the ROCs of four indices consistent with such a model, with fixed and equal variances of the two distributions, are presented in that context. Third, I present the ROCs for six indices that imply one or another threshold model and discuss these models. Fourth, I review ROC practice; I point to illustrative empirical ROCs from several areas of psychology and from other fields, shown in a companion paper (Swets, in press), and summarize their general features. Lastly, I point out the implications of empirical ROCs for the validity of the various models, and review indices that are appropriate to the form of empirical ROCs.

Formal Description of Discrimination Performance

For present purposes, the relevant data from a two-alternative, single-stimulus discrimination task are fully contained in a

2×2 contingency table (Table 1). Occurrences of the two alternatives are denoted by A and B and the corresponding choices are denoted by $\tilde{A}$ and $\tilde{B}$. The cell entries (a, b, c, and d) indicate the frequencies of the four possible conjunctions of occurrence and choice. The column sums give the frequencies of occurrences and the row sums give the frequencies of choices. The total sample size (N) appears as the overall sum.

Table 1
Formal Description of Discrimination Performance

	Occurrence		Sum of row frequencies
Choice	A	B	
$\tilde{A}$	a	b	$a + b$
$\tilde{B}$	c	d	$c + d$
Sum of column frequencies	$a + c$	$b + d$	$N = a + b + c + d$

The relative frequency of a conjunction can be taken as an estimate of the *joint* probability of its two elements: for example, a/N is an estimate of $P(A \cdot \tilde{A})$. Dividing a cell frequency by its column sum yields a ratio that is an estimate of a *conditional* probability, specifically, the probability of a choice conditional on an occurrence. For example, $a/(a + c)$ is an estimate of $P(\tilde{A}|A)$. The latter probability and $b/(b + d)$, or $P(\tilde{A}|B)$, contain all of the information in the table, because the other two conditional probabilities are their complements.

As I mentioned, the two probabilities just listed are the coordinates of the ROC graph. Given that the two alternatives are often the presence and absence of something (e.g., a weak spot of light or a weak tone in an experiment on sensory capacity—and disease, rain, or oil, in practice), the terms *hit* and *false alarm* are often used, and I use h and f to stand for these conditional probabilities. Another variable of importance is the probability of occurrence of Alternative A ("something" or "signal"), namely $(a + c)/N$, here denoted by s.

Definitions of Various Indices

Tables 2 and 3 define illustrative indices as proposed in various fields, first in terms of the frequencies of the 2×2 table (Table 1) and then in terms of the two or three main probabilities derived from them: h, f, and (in some cases) s. The derivations of the second definition of each index are not given here, but in general are obtained by substitution according to equalities of the following sort: $a/(a + c) = h$, $(a + c)/N = s$, $a = hsN$; $b/(b + d) = f$, $(b + d)/N = 1 - s$, and $b = f(1 - s)N$. Tables 2 and 3 also give the formulas for the ROCs that may be obtained by rearranging the (second) definitions in terms of h, f, and (in some cases) s. I examine later the forms of the ROCs specified by these formulas.

Table 2 lists six indices that imply fixed-threshold models. Table 3 lists four indices that are consistent with variable-criterion models that have fixed distributional parameters, that is, equal variances. The indices identified in this article as preferred to either kind just mentioned are defined later, after the form of empirical ROCs is adduced.

Consider first Table 2. The first two indices listed are forms of "corrected hit probability." Both indices focus on h, but attempt to correct it for any spurious component that may be induced by a tendency toward false alarms, which is estimated by f. The first index, designated H_C, subtracts f from h, and then divides by $(1 - f)$ to normalize the range of the corrected value (e.g., Blackwell, 1963; Fisk & Schneider, 1984). It has been used primarily in studies of sensory functions. The second, H'_C, corrects simply by subtracting f from h (e.g., Gillund & Shiffrin, 1984; Woodworth, 1938). In psychology this index is associated primarily with studies of recognition memory, and it is also prominent in weather forecasting (e.g., Hanssen & Kuipers, 1965) and medical diagnosis (e.g., Galen & Gambino, 1975; Youden, 1950).

"Percentage correct" is the name usually given the overall percentage of correct choices of either alternative. It is listed in Table 2 in the more convenient form of "proportion correct" (PC). Proposed at least a century ago (Finley, 1884) to evaluate the accuracy of tornado prediction, it is still popular in many fields, including weather forecasting (e.g., Brier & Allen, 1952; Ramage, 1982) and medical diagnosis; in fact, in medical diagnosis, it is often taken as synonymous with accuracy (see Metz, 1978).

A contemporary of Finley (Gilbert, 1885) pointed out the dependency of PC on s and, indeed, the fact that PC could be as high as s or $(1 - s)$ by chance, without discrimination. Now, in weather forecasting, indices that measure the extent to which discrimination exceeds chance performance are generally called *skill scores*. An example is given by the index designated Z (Woodcock, 1976) listed in the table. As shown later, a higher value of h relative to f must be achieved to attain a given value of Z as s departs from the symmetrical .50. Several other indices proposed in weather forecasting have been analyzed in ROC terms by Mason (1982), including indices proposed by Heidke (1926), Vernon (1953), Appleman (1959), Schrank (1960), Bermowitz and Zurndorfer (1979), and Rousseau (1980).

A statistic used as a measure of observer agreement in clinical medicine (Landis & Koch, 1977), called the kappa statistic (K), is another form of chance-corrected counterpart to PC, and is listed fifth in Table 2.

The final index considered in Table 2 is another measure of association in statistics. Phi, equal to $(\chi^2/N)^{1/2}$, is called the fourfold point coefficient and the root–mean-square contingency (Hays, 1973). It (or its square) has been used as a discrimination accuracy index in experimental psychology (Wellman, 1977; see also Nelson, 1984), weather forecasting (Pickup, 1982); and nondestructive testing (see Swets, 1983a, 1983b).

Table 3 lists four indices that have been, or can be, associated with variable-criterion models. They are all consistent with the assumption (as will be seen later) that the distributions of observation values stemming from the two alternatives to be discriminated are of equal variance—a restrictive assumption not supported by data.

The first index listed is the first index defined in the psychological application of signal detection theory, the detectability index d' (Tanner & Swets, 1954). It is defined in terms of integrals of normal (Gaussian) distributions and is given in terms of the normal deviate, or z score; d' is the z score corresponding to f minus the z score corresponding to h.

The next index, η, is defined in Luce's (1959, 1963) general theory of choice. Both η and the next measure, the log odds ratio

Table 2
Definitions and ROC Formulas for Indices That Imply a Threshold Model

Index name and symbol	Definitions and ROC formula
1. Corrected hit probability, H_C	$H_C = [a/(a + c)] - [b/(b + d)]/[1 - [b/(b + d)]]$ $= (h - f)/(1 - f)$ $h = H_C + f(1 - H_C)$
2. Corrected hit probability, H'_C	$H'_C = (ad - bc)/(a + c)(b + d)$ $= h - f$ $h = H'_C + f$
3. Proportion correct, PC	$\mathrm{PC} = (a + d)/N$ $= (1 - s)(1 - f) + sh$ $h = [\mathrm{PC} - (1 - s)(s - f)]/s$
4. Skill test, Z	$Z = 4\,(ad - bc)/N^2$ $= 4s\,(1 - s)(h - f)$ $h = f + [Z/4s(1 - s)]$
5. Kappa statistic, K	$K = \dfrac{2(ad - bc)}{2(ad - bc) + N(b + c)}$ $= \dfrac{2s(1 - s)\,[h - f]}{(1 - 2s)[hs + f(1 - s)] + s}$ $h = \dfrac{f(1 - s)[1 - (1 - K)(1 - 2s)] + sK}{s[1 + (1 - K)(1 - 2s)]}$
6. Phi coefficient, ϕ	$\phi = (ad - bc)/[(a + c)(b + d)(a + b)(c + d)]^{1/2}$ $= \dfrac{[(1 - s)s]^{1/2}\,[h - f]}{[\{sh + (1 - s)f\}\{1 - [sh + (1 - s)f]\}]^{1/2}}$ $h = \{\sigma^2 + 2(1 - s)(1 - \phi^2)f$ $+ \phi[\phi^2 + 4(1 - s)(1 - f)/s]^{1/2}\}/\{2[(1 - s) + s\phi^2]\}$

Note. ROC = relative operating characteristic. The numbers 1–6 order these equations in sequence with those presented in the text. The first two equations for each index are different, but equivalent, definitions; the third equation is the ROC formula for that index.

(LOR), are similar to d', but depend on logistic, rather than normal or Gaussian, distributions (thus permitting an explicit writing of the ROC function, as shown).

The LOR was described by Goodman (1970) and is used extensively in biostatistics (see Gart, 1971). The odds are those of a correct choice (*ad*) relative to an incorrect choice (*bc*). As the two ROC equations show, $\eta^2 = e^{-\mathrm{LOR}}$.

Lastly, the measure Q was defined by Yule (1912). For two alternatives, Goodman and Kruskal's (1954) gamma measure is equal to Q. It was recently proposed as an accuracy index in psychology by Nelson (1984). Nelson contrasted PC, (see Table 2), d', and Q, and advocated Q over d' primarily on the basis that it was thought to make weaker assumptions. As I show later, and as Table 3 indicates by the common form of the ROC formulas for η, LOR, and Q, the latter index is also consistent with (though it need not assume) logistic distributions in detection theory.[1]

[1] In treating the related indices Q, η, and LOR (Table 3) as well as ϕ (Table 2), I essentially include almost all standard measures of association as derived in statistical theory. Bishop, Fienberg, and Holland (1975) point out that in 2 × 2 tables, almost all such measures reduce either to functions of the cross-product ratio (and are independent of marginal totals), as are Q, η, and LOR, or to functions of the correlation coefficient (and are sensitive to marginal totals), as is ϕ. The kappa statistic (Table 2) is a measure of agreement described by Cohen (1960) as a special case of association for larger tables, and as essentially equivalent to association in 2 × 2 tables (again, sensitive to marginal totals).

ROC Theory

Basic Model

The ROC graph was designed in the context of the theory of signal detectability by Peterson, Birdsall, and Fox (1954) to provide an index of accuracy consistent with their basic model of the detection process. They saw the detection task as one of discriminating occurrences of "signal plus noise" (*sn*) from occurrences of "noise alone" (*n*). Given that noise is a random variable, the two alternatives can be considered as statistical hypotheses.

The theory of statistical decision, or of testing statistical hypotheses (e.g., Wald, 1950), is the basis for a model that provides an accuracy index that is independent both of the probability of occurrence of the two alternatives (s and $1 - s$) and of the discriminator's tendency to favor the choice of one or the other alternative. Neither variable, the detection theorists suggested, is usefully or properly regarded as part of the process of discrimination per se and neither should therefore influence an index of discrimination capacity or accuracy. Because they are variables, an accuracy index tacitly dependent on them would be imprecise at best. An accuracy index that incorporated an explicit and monitored dependence on them would imply a model of the choice process as well as of the discrimination process and would need to be validated by data.

The detection-theory model is depicted in Figure 2. The horizontal axis is akin to the decision variable of statistical theory: The variable x is a measure of the strength of the observation, or the magnitude of a sample statistic. The vertical axis is probability density. The left-hand distribution is the probability density function (PDF) of x for n, analogous to the null hypothesis, H_0. The right-hand distribution is the PDF for *sn*, analogous to the other hypothesis under test, H_1. The degree of overlap of the two distributions determines the confusability of the two alternatives, reflected in the figure by the difference between the distribution means, and denoted by θ. Thurstone (1927) called these distributions "discriminal dispersions" in psychometric theory: They acknowledge the idea that representations of a given alternative, either psychologically in perception or cognition or more generally as samples of any kind, vary from one occurrence of the alternative to another, and can be considered to lie along a single dimension.[2]

Table 3
Definitions and ROC Formulas for Indices Consistent With a Variable-Criterion Model

Index name and symbol	Definitions and ROC formula
7. Detectability, d'	$d' = z_{b/(b+d)} - z_{a/(a+c)}$ $= z_f - z_h$ $z_h = z_f - d'$
8. Choice-theory measure, η	$\eta = (bc/ad)^{1/2}$ $= \{[f(1-h)]/[h(1-f)]\}^{1/2}$ $h = f/[f + \eta^2(1-f)]$
9. Log odds ratio, LOR	$\mathrm{LOR} = ln(ad/bc)$ $= ln[h(1-f)/f(1-h)]$ $h = f/[f + e^{-\mathrm{LOR}}(1-f)]$
10. Yule's Q	$Q = (ad - bc)/(ad + bc)$ $= (h - f)/(h - 2fh + f)$ $h = f/\left[f + \frac{1-Q}{1+Q}(1-f)\right]$

Note. ROC = relative operating characteristic. The numbers 7–10 order these equations in sequence with those presented in the text. The first two equations for each index are different, but equivalent, definitions; the third equation is the ROC formula for that index.

A critical value of x, or the decision criterion (x_c), separates the values of x that lead to the choice of *sn* (i.e., $x \geq x_c$) from those that lead to the choice of n (i.e., $x < x_c$). The particular value of x_c selected depends in theory on the probability s and on the benefits and costs of the four possible choice outcomes.

The area under the *sn* distribution to the right of x_c (hatched) equals the probability h, and the area under the n distribution to the right of x_c (crosshatched) equals f. The ROC is traced from left to right on the graph of h versus f as x_c moves from right to left. This graph and an illustrative ROC are shown in Figure 1. The ROC shown in Figure 1 is that indicated by the specific model and specific discriminability portrayed in Figure 2: namely, Gaussian or normal distributions of equal variance separated by a particular value of θ. Larger θs lead to ROCs that are higher on the graph, and vice versa. Discrimination accuracy is at the chance level when the ROC follows the positive diagonal (the distributions overlap completely), and is perfect when the ROC follows the left-hand and top axes (the distributions do not overlap at all). Appropriate indices are discussed later, but note that accuracy independent of the criterion for choice can be indexed either by a theoretical parameter related to the PDFs that might underlie an ROC, such as θ, or by some other measure of the locus of the ROC, perhaps one empirically based, such as the proportion of the graph's area lying beneath the ROC.

ROC theory as applied in psychology is discussed in detail elsewhere (Green & Swets, 1966/1974; Swets, 1973), and its applications in other fields have also been summarized (Swets & Green, 1978). Note that within the framework of detection theory, ROCs can be generated in two main ways. (a) One can assume PDFs, on the decision variable, of one or another specific form. As indicated previously, the PDF form determines the ROC form. (b) One can assume some ROC form directly, without recourse to PDFs, as specified by an algebraic formula that relates h to f. Relevant examples in each category are described next.

ROCs Generated by PDFs

Equal-variance, normal PDFs. The equal-variance, normal PDFs of Figure 2 were the first considered in detection theory, and represent a signal with all of its parameters (i.e., frequency,

[2] Multidimensional representations can be mapped onto the single dimension of likelihood ratio, the ratio of the ordinate of the PDF for *sn* to the ordinate of the PDF for *n* (Green & Swets, 1966/1974; Peterson, Birdsall, & Fox, 1954).

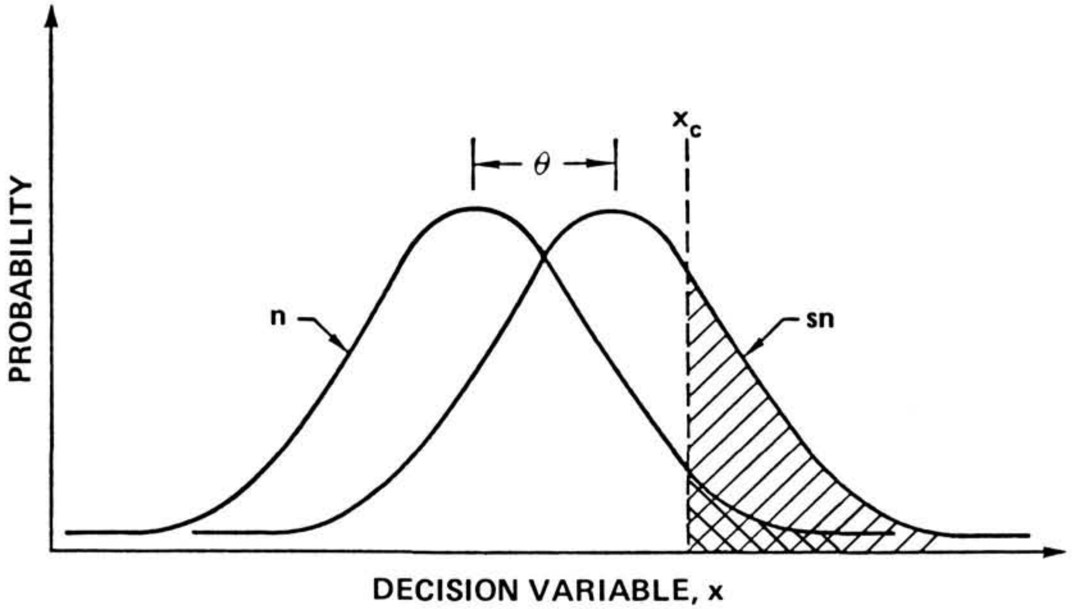

Figure 2. Detection-theory model of the discrimination process (see text). (*x*: a measure of the strength of the observation; *n*: the distribution of observations [or probability density function] that arises from the noise-alone events; *sn*: same, but for the signal-plus-noise event; x_c: a critical value of *x*; and θ: the difference between the means of the two distributions.)

phase, amplitude, starting time, and duration) exactly known by the observer. The corresponding accuracy index was taken as

$$d' = (m_{sn} - m_n)/\sigma_n, \qquad (11)$$

which expresses the difference between the means (*m*) of the two PDFs in terms of the standard deviation (square root of variance) of the PDF for the *n* (or, equivalently, for the *sn*) alternative. Figure 3a shows illustrative ROCs for equal-variance, normal PDFs; as in later figures, the form of the PDFs is shown in the inset at lower right.

Equal-variance, logistic PDFs. The logistic PDF is similar to the normal (Gaussian) PDF (e.g., Bush, 1963; Laming, 1973), and logistic PDFs of equal variance for *n* and *sn* yield ROCs similar to those of equal-variance, normal PDFs (Luce, 1963). Figure 3b shows such PDFs and illustrative ROCs. Compared with the normal PDF, the logistic PDF is slightly taller, narrower through the midsection, and wider at the tails. The accuracy index shown, η, is that defined in Luce's choice theory (Equation 8, Table 3).

As indicated, the LOR index, though defined outside of detection theory and the concept of the ROC, is based on the logistic distribution (and, in detection theory, would be derived from equal-variance, logistic PDFs), and yields ROCs identical in form to those of η as shown in Figure 3b. I show later, when dicussing algebraically defined ROCs, that *Q*, also defined without regard to the ROC concept or underlying distributions, yields ROCs identical to those of η and LOR. Their ROC formulas, as noted in Table 3 (Equations 8–10), are of the same form. The ROCs of the LOR and *Q*, with representative index values, are shown in Figure 4.

ROCs on a binormal graph. The use of a different scale on the axes of the ROC graph is convenient for fitting empirical ROCs and for comparing theoretical ROCs. The scale used most often is one on which the spacing of the probabilities is transformed so that their corresponding normal deviates, or *z* scores, are linearly spaced. A graph with such a scale on both axes is called a *binormal* graph. On such a graph, ROCs for constant *d'* are straight lines with slope of 1, as Figure 5a shows (Peterson et al., 1954).

Logistic PDFs yield straight-line ROCs on what Birdsall (1966) called a "lor–lor" graph, where, as with the LOR index, *lor* stands for the natural logarithm of the odds ratio. However, given the similarity of the normal and logistic ROCs apparent in Figure 3, one would suspect that logistic-based ROCs are not far from straight lines on a binormal graph. Following Birdsall (1966), logistic ROCs (with the values of η shown in Figure 3) are shown on a binormal graph in Figure 5b, where a slight bow can be seen. Figures 3 and 5 indicate that the logistic and normal forms of ROC could be distinguished in data only with exceptionally reliable data, based on a large number of observations, and could not be distinguished for most practical purposes.

Unequal-variance, normal PDFs. The early detection experiments with human observers produced some ROCs like those of Figure 3, but most were not symmetrical about the negative diagonal; they rose more steeply from the origin and then bent more sharply toward the upper corner. This effect at first seemed to be more pronounced in data as the signal strength was increased, that is, for progressively higher ROCs (Swets, Tanner, & Birdsall, 1961). Such ROCs would arise from normal PDFs having a larger variance for *sn* than for *n*, and such that the difference in variance increased with ($m_{sn} - m_n$). ROCs based on a mean-to-sigma ratio of 4.0 (i.e., assuming $[m_{sn} - m_n]/[\sigma_{sn} - \sigma_n] = \Delta m/\Delta\sigma = 4.0$) agreed reasonably well with the early data (Green & Swets, 1966/1974).

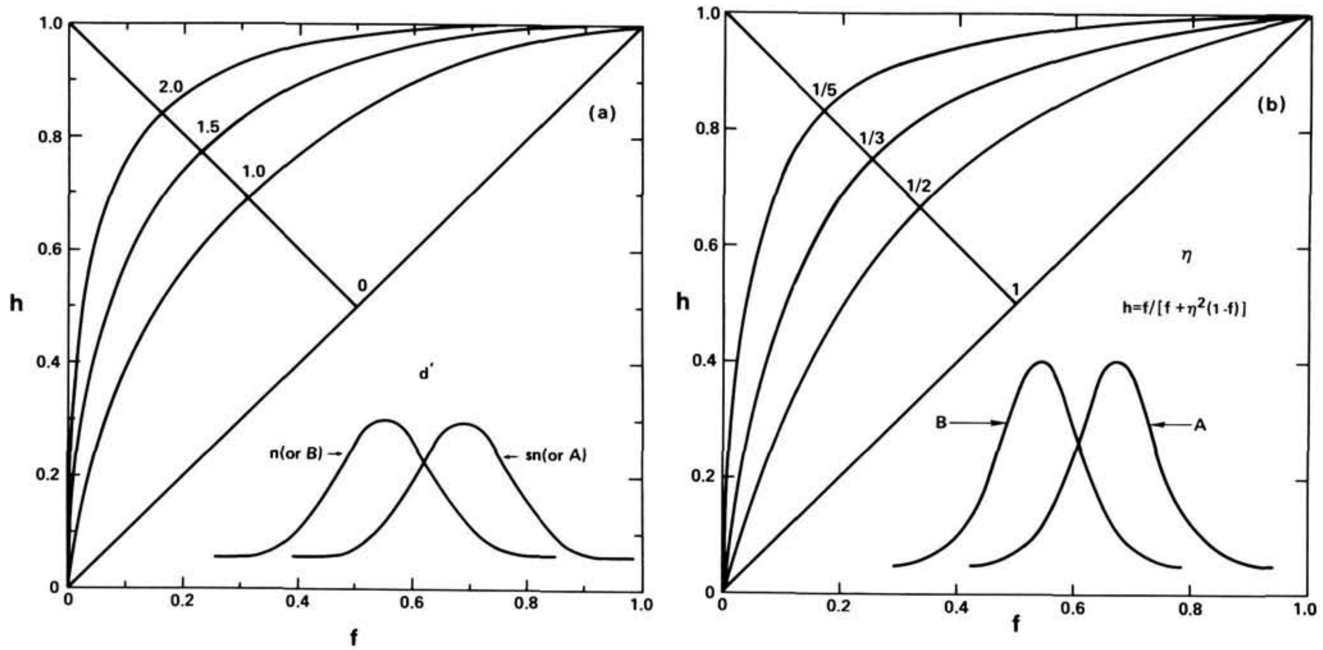

Figure 3. a: relative operating characteristics (ROCs) for equal-variance, normal probability distributions. b: ROCs for equal-variance, logistic probability distributions. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; *n*: the distribution of observations [or probability density function] that arises from the noise-alone event; *sn*: same, but for the signal-plus-noise event; *A* and *B*: more general designations of the two events; *d'*: an accuracy index based on equal-variance, normal distributions; and *η*: an accuracy index based on equal-variance, logistic distributions.)

Illustrative ROCs based on that ratio, and illustrative PDFs, are shown on ordinary scales in Figure 6a. On a binormal graph, ROCs with a fixed mean-to-sigma ratio are straight lines with a progressively shallower slope as they rise on the graph, as shown in Figure 6b. Given that such ROCs cannot be indexed by d' (Equation 11), because d' is not constant along them, Figure 6 shows Δm as an index, which was an early attempt to handle ROCs consistent with unequal-variance, normal distributions. The definition of Δm is analogous to d',

$$\Delta m = (m_{sn} - m_n)/\sigma_n, \qquad (12)$$

and can be used along with the slope of an empirical ROC in a two-parameter description of data (Green & Swets, 1966/1974).

Other PDFs for asymmetrical ROCs. Though unequal-variance, normal (or logistic) ROCs cannot be indexed by a single quantity unless some further assumption is made, such as that the mean-to-sigma ratio is fixed, some other forms of PDF give rise to ROCs similar to those of Figure 6, and can be indexed by a quantity that is constant along a given ROC without concern for relative variances. Some of these were defined in detection theory as suitable for a signal having one or more parameters (e.g., phase) known only statistically to an observer (Peterson et al., 1954). Included are the Rayleigh PDF (Jeffress, 1964), chi-square and noncentral chi-square PDFs (Birdsall & Lamphiear, 1960), and exponential PDFs (Green & Swets, 1966/1974).

Algebraic ROCs

Power ROC. ROCs having the general pattern of those in Figure 6 are specified by the formula for a power function: $h = f^k$, where $k < 1$ (Egan, Greenberg, & Schulman, 1961). A few power curves are shown in Figure 7a, which indicates how the value of k can serve as an accuracy index. The PDFs inset, which produce power ROCs, are exponentials (Green & Swets, 1966/1974). These ROCs, of course, are straight lines on log–log scales. Again, however, they are quite close to straight lines on normal–normal scales, as Figure 7b shows.

Conic ROCs. Arcs of circles, ellipses, parabolas, and hyperbolas can serve as algebraically specified ROCs (Birdsall, 1966). Of these, the hyperbola is singled out here, for reasons stated next.

Algebraic ROCs and Related PDFs

As mentioned previously, exponential PDFs generate power ROCs, though these ROCs are simply identified directly by their algebraic formula. In the same vein, Birdsall (1966) observed that logistic PDFs generate ROCs that are rectangular hyperbolas. Thus, the ROC specified by the equation for a hyperbola was shown in Figures 3b and 4a, and both η (Equation 8) and LOR (Equation 9) are associated with hyperbolas. Not originally derived from logistic PDFs (nor dependent on them), but implying hyperbolic ROCs, is the index Q (Equation 10). I show that LOR, η, and Q specify hyperbolic ROCs in the Appendix.

Relation of d', η, LOR, and Q

The indices η, LOR, and Q are related by the following equations defining LOR and Q in terms of η:

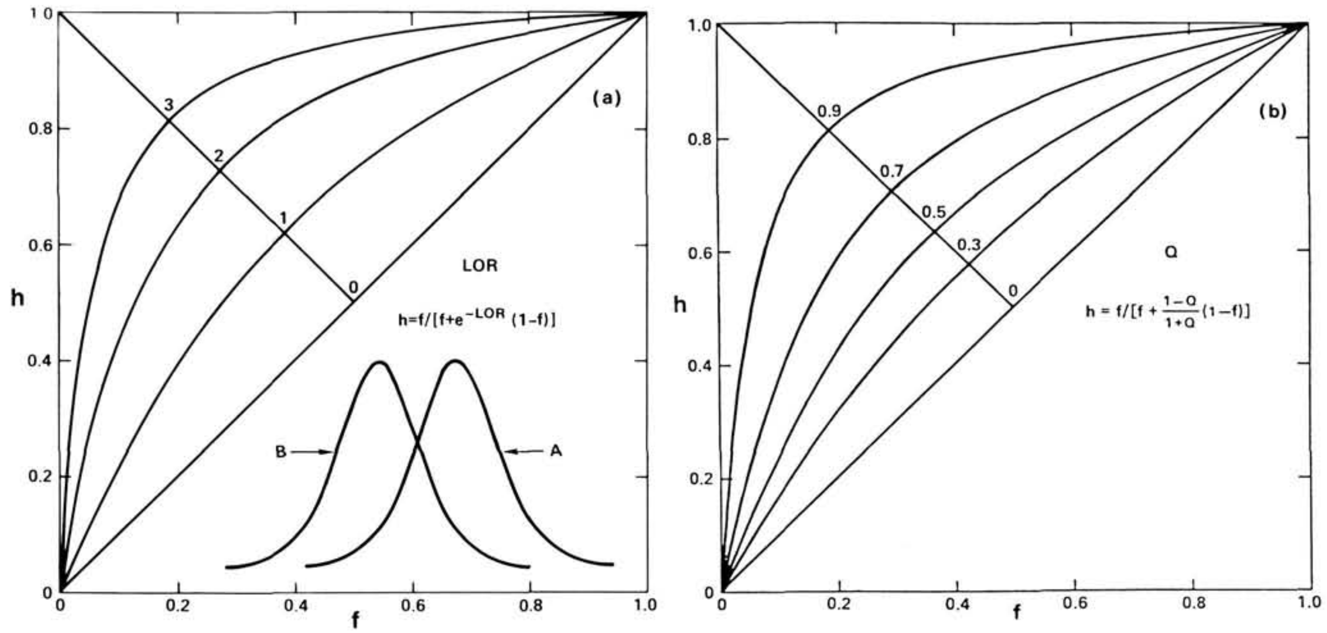

Figure 4. a: relative operating characteristics (ROCs) for the log odds ratio (LOR) index. b: ROCs for Yule's Q index. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; *A* and *B*: distributions of observations [or probability density functions] that arise from the two possible events; and LOR and *Q*: accuracy indices that correspond in detection theory to equal-variance, logistic distributions.)

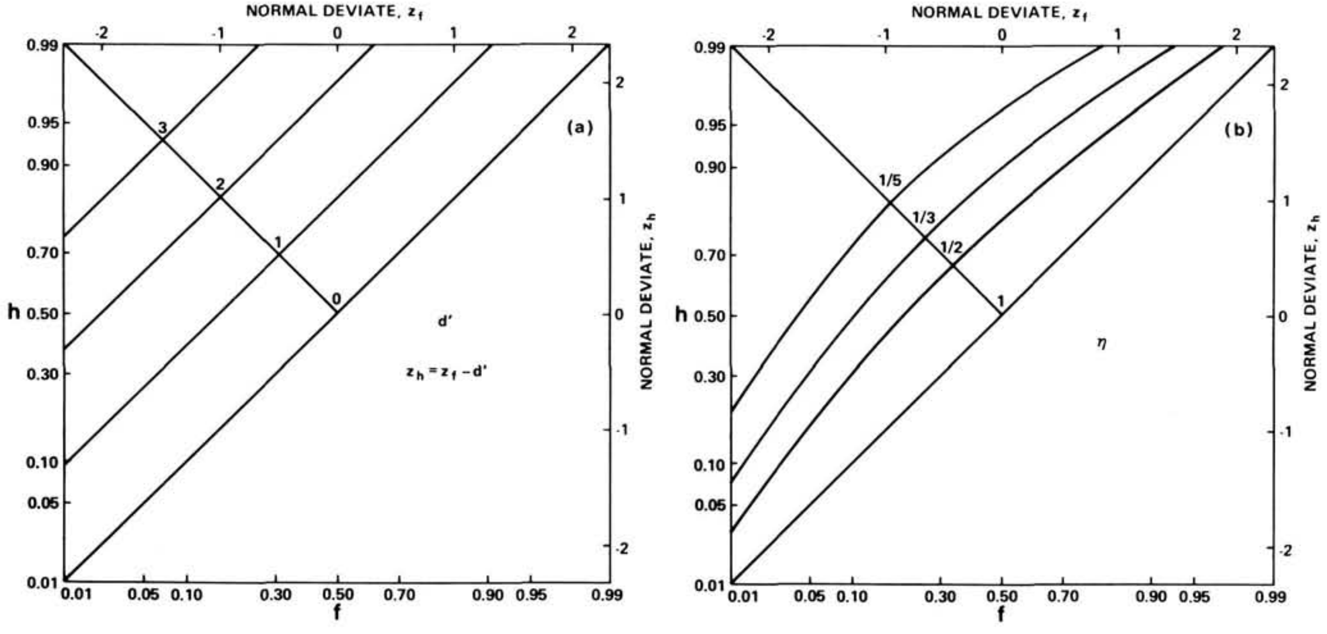

Figure 5. a: relative operating characteristics (ROCs) for equal-variance, normal distributions on a binormal graph. b: ROCs for equal-variance, logistic distributions on a binormal graph. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; z_h and z_f: normal-deviate values of *h* and *f*; *d'*: an accuracy index based on equal-variance, normal distributions; and η: an accuracy index based on equal-variance, logistic distributions.)

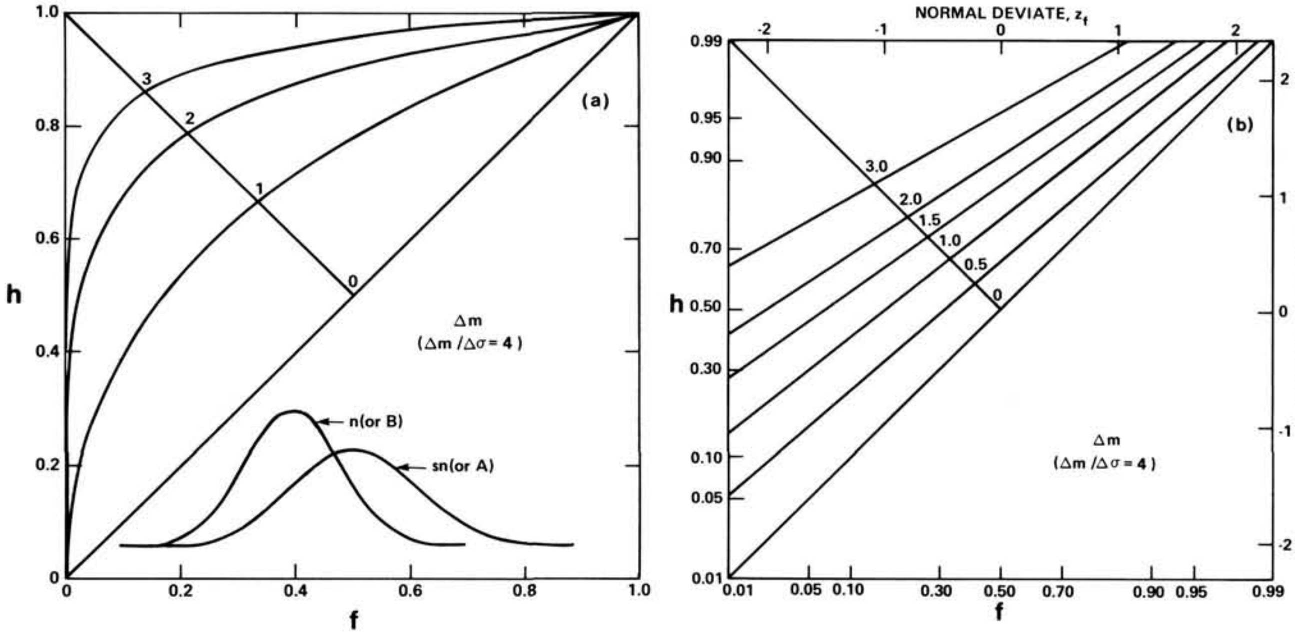

Figure 6. a: relative operating characteristics (ROCs) for unequal-variance, normal probability distributions, with a mean-to-sigma ratio of 4.0, on ordinary scales. b: the same ROCs but on a binormal graph. (*h*: conditional probability of a hit; *f*; conditional probability of a false alarm; *n* (or *B*) and *sn* (or *A*): distributions of observations [or probability density functions] that arise from the two events; Δ*m*; an accuracy index based on unequal-variance, normal distributions; *m*: the mean of a distribution; and *σ*: the standard deviation of a distribution.)

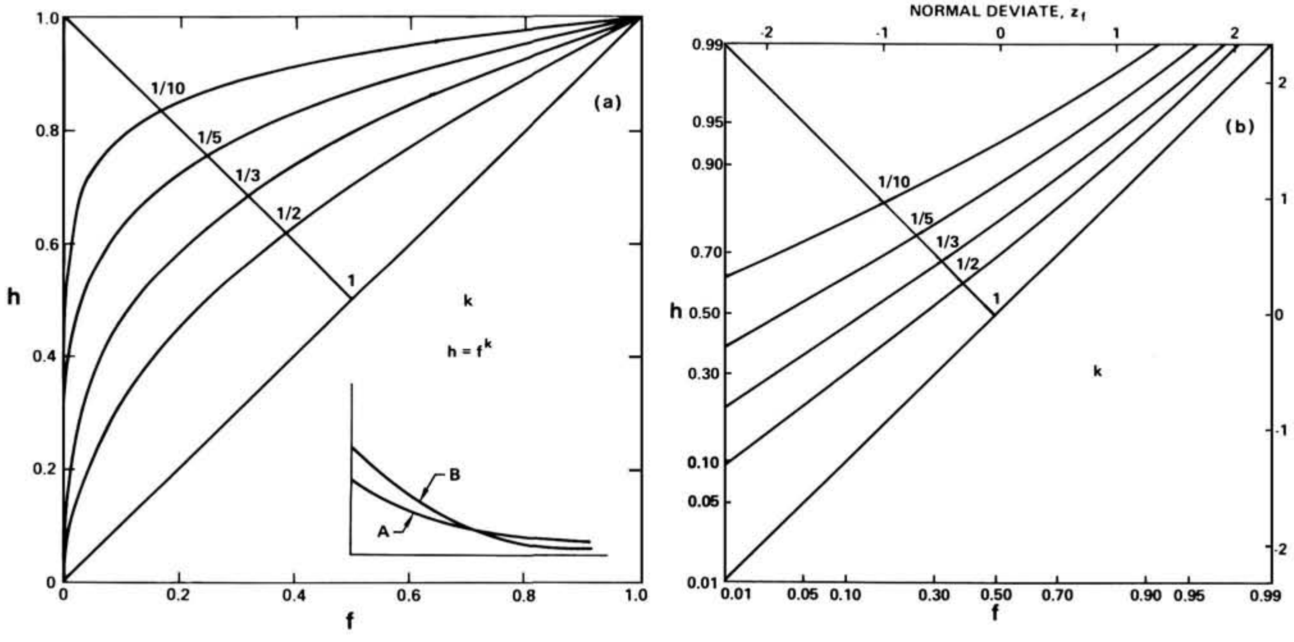

Figure 7. a: relative operating characteristics (ROCs) for exponential distributions, in the form of a power function, on ordinary scales. b: the same ROCs on a binormal graph. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; *A* and *B*: distributions of observations [or probability density functions] that arise from the two events; and *k*: an accuracy index based on exponential distributions.)

$$\text{LOR} = ln(1/\eta^2), \quad (13)$$

and

$$Q = (1 - \eta^2)/(1 + \eta^2). \quad (14)$$

As developed by C. E. Metz (personal communication, 1984), there is a one-to-one correspondence between those three indices and the value of d' taken at the negative diagonal of the ROC graph, termed d'_e. Following Luce's (1963) derivation of

$$\eta \approx \exp[-(2/\pi)^{1/2}d'_e], \quad (15)$$

and Equations 13 and 14, we have

$$\text{LOR} \approx 2(2/\pi)^{1/2}d'_e, \quad (16)$$

and

$$Q \approx \tanh[(2/\pi)^{1/2}d'_e]. \quad (17)$$

These approximations can be shown to be quite good for $d'_e < 2$ and can be extended by a correction term to higher values. They are reasonably good for a fairly substantial range of ROC points centered at the negative diagonal; for example, they are good to within 2% for ROC points having f values within ±.10 of its value at the negative diagonal. The need for approximate equations stems from the nature of the definition of d', but corresponding values of the four indices, according to exact relations, could be tabled. The key exact relation is

$$\eta = \{1/[\Phi(d'_e/2)]\} - 1. \quad (18)$$

To review a bit of history, Luce (1963) showed that η corresponds to a distance between the means of logistic PDFs in signal detection theory, via the relation $\eta = \exp[-(\Delta m)/2]$, where Δm is defined as in Equation 12. The logistic function was demonstrated to be "very similar" to the normal function (Bush, 1963, p. 448; Laming, 1973, p. 23), differing at most by "less than two parts in the hundred" (Luce, 1959, p. 55). The ROCs of η were observed to have "substantially the same" form as those of d' (Luce, 1963, p. 131). Ogilvie and Creelman (1968) observed that LOR is approximately 1.64 times d'_e. C. E. Metz and I. B. Mason (independent personal communications, 1984) pointed out that LOR and Q are transformations of η. Edwards (1963) had shown that measures of association drawn from a 2 × 2 table that are independent of the marginal frequencies should logically be some function of the cross ratio, that is, η^2; he specifically qualified LOR and Q (and disqualified ϕ, defined in Equation 6.) Birdsall (1966) showed that logistic PDFs generate hyperbolic ROCs.

Birdsall (1966) pointed out that hyperbolic ROCs are difficult to distinguish from the ROCs of d' in a figure reproduced here as Figure 8a. It shows an equal-variance, normal ROC as a curve ($d' \approx 1.5$), along with selected points from the hyperbolic ROC that fits it best. As he put it, "If one's definition of 'close' means 'as points appear when plotted on linear [scales],' then these two [PDFs] yield ROC curves that are quite close" (1966, p. 178).

Birdsall observed, however, that as a second-order effect, η, LOR, and Q do diverge from d' at the edges of the plot. Ordinarily, too few observations are made to define such extreme probabilities with adequate reliability, but that may not always be the case. So let us examine some representative values of Q and d' for values of $f < .10$. Figure 8b shows Q as a function of f for three values of d' (solid lines). Between $f = .01$ and .10, the value of Q for constant d' varies by about 4% for $d' = 2.0$, 17% for $d' = 1.0$, and 26% for $d' = 0.5$. At such values of f, especially at low index values, Q and d' cannot be used interchangeably.[3]

Figure 8b also helps demonstrate a point made earlier. It shows the contour of values of f taken at the negative diagonal of the ROC graph for various ROCs (broken line). It can be seen that Q (and, hence, LOR and η) vary by only a few percent from d' over a range in f of 0.20 or more, centered about the negative diagonal. This graph, then, supports the statement that the approximations given above for the correspondence of η, LOR, and Q to d'_e are quite good for a fairly large range of off-diagonal ROC points.

In summary, ROC analysis shows that the four indices have a kinship in theory and indicates that under most conditions, they would lead to the same conclusions in practice. Surely an advocacy of one over another is considerably enlightened by a comparison of their ROCs. That is particularly true for Nelson's recent advocacy of Q for research on "feeling of knowing," inasmuch as researchers in that field have acknowledged the problems of the variable decision criterion (Nelson, 1984, pp. 117–119, 121–122, 125–126).

Note also that the same or similar conclusions that would usually be indicated by the four indices may be faulty because of the straitjacket these indices put on the form of a ROC. For example, consider that if the top curve of Figure 6b represented an observed ROC, calculated values of d' would vary by about 100% along that curve (from about 2.7 at the left to about 1.3 at the right), and would vary by about 20%–25% in a range of $f \pm .10$ about the negative diagonal, where $f = .14$ (i.e., from about 2.0 at $f = .04$ to about 2.5 at $f = .24$); all of which is unnecessary error of measurement.

Concept of the Regular ROC

In making the transition from indices associated with the class of continuous, variable-criterion models to those associated with threshold models, it will be helpful to have in mind the concept of the "regular" ROC. Almost any form of ROC can be generated within signal detection theory: even the forms implied by threshold models, as we shall see, follow from assuming rectangular PDFs. However, what can be thought of as canonical, or classic, detection theory contains an assumption that leads to regular ROCs. This assumption is that any value of the observation or decision variable can arise from either alternative (n or sn); in other words, that noise is thoroughly noise, perturbing observation values throughout their range.

Given enough observations, this assumption implies that $f = 0$ will be attained only when $h = 0$, and that $h = 1$ will be attained only when $f = 1$. A regular ROC is thus interior to the unit-square ROC graph (ordinary scales) except at the chance points ($f = 0$, $h = 0$) and ($f = 1$, $h = 1$). Nonregular ROCs are those permitting "singular" detection, that is, ROC points having $h > 0$ for $f = 0$ or $h = 1$ for $f < 1$. On ordinary scales, such

[3] I. B. Mason (personal communication, 1985) pointed out the importance of the divergence of Q and d' in weather forecasting, where the usual decision variable, posterior probability, serves to expand the left side of the graph as compared with Figure 8b. That expanded region is of detailed interest in that field, in which a distinction is made routinely among posterior probabilities of 0, .02, and .05 (see, e.g., Murphy, 1977).

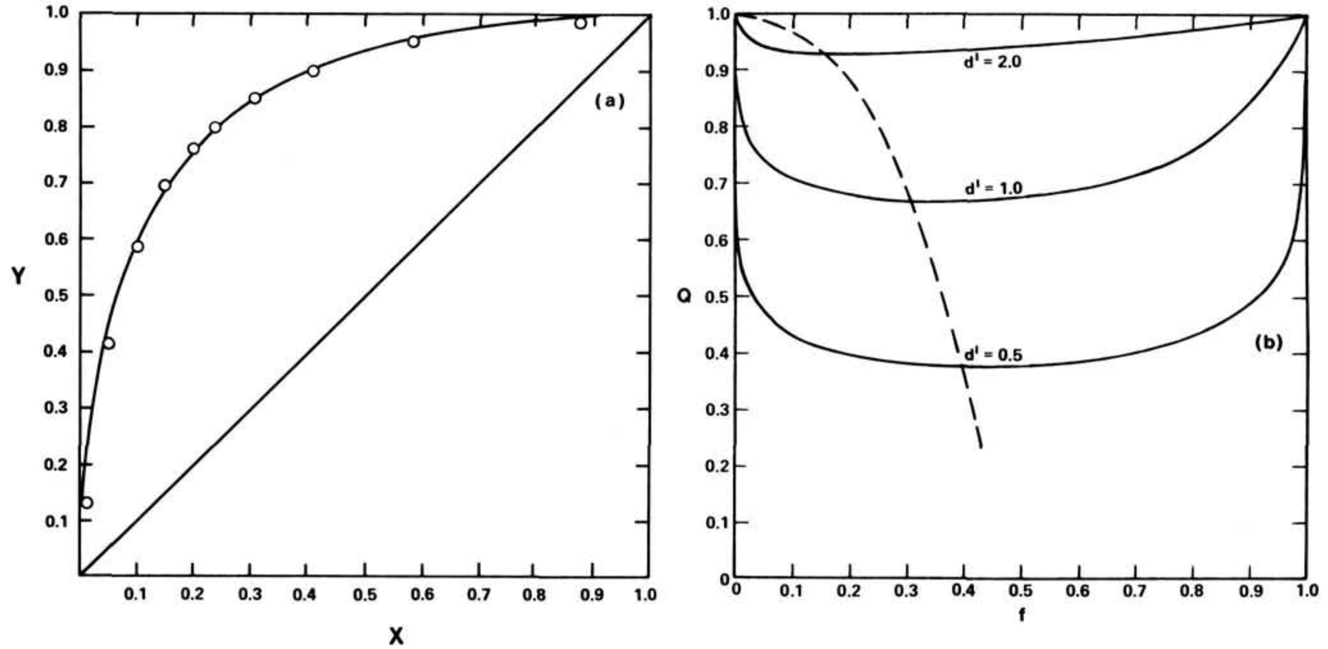

Figure 8. a: comparison of a hyperbolic relative operating characteristic (ROC), as generated by equal-variance, logistic probability distributions (open circles) and an ROC based on equal-variance, normal probability distributions (curve). (Reproduced from Birdsall, 1966.) b: *Q* as a function of *f* for three values of *d'* (solid lines) and the contour of values of *f* at the negative diagonal of the ROC graph (broken line). (*Y*: conditional probability of a hit; *X* or *f*: conditional probability of a false alarm; *Q*: an accuracy index that corresponds in detection theory to equal-variance, logistic distributions; and *d'*: an accuracy index based on equal-variance, normal distributions.)

ROCs will cross the left-hand axis above the lower left-hand corner or the top axis short of the upper right-hand corner. The theoretical ROCs seen so far are regular; the theoretical ROCs seen next are nonregular. Data, as I show elsewhere (Swets, in press) and characterize here, yield regular ROCs.[4]

Indices Implying a Threshold Model

The six indices remaining to be considered, those in Table 2 that were anticipated to imply a threshold model, can be handled with fewer subtleties than the four from Table 3 discussed previously. For one thing, their ROCs look very different from the ones already discussed, so that contrast is easy. For another, because their ROCs all look very different from data, I do not go to great lengths to compare and contrast them with each other. One of these indices implies a high-threshold model; the other five imply a double-threshold model.

Note first that the ROC formulas for the first five indices in Table 2 show *h* as a linear function of *f*. Hence, their ROCs will be straight lines on *ordinary* scales. Thus they are nonregular. The ROCs of the sixth index listed are nearly straight and are also nonregular. For the most part, the models implied by these indices assume that observation values have just two or three states. Such models might include underlying PDFs that are truncated or, alternatively, a threshold on the observation variable, such that values on a given side of the threshold are indistinguishable from each other. I present the six ROCs next, followed by a brief characterization of the models.

Corrected Hit Probability, H_C

The ROCs for H_C, the first of two chance-corrected versions of *h*, are shown in Figure 9a. They were derived by Tanner and Swets (1954) as representing both the particular correction and the so-called high-threshold model advocated, for example, by Blackwell (1963). They intersect the left-hand axis and the upper right-hand corner. Craig (1979) reported that an index used to assess inspector accuracy in industrial monitoring also leads to the ROC of the high-threshold model.

Corrected Hit Probability, H'_C

The ROCs for H'_C, the simpler chance-corrected version of *h*, appear in Figure 9b. They were first drawn by Egan (See Green & Swets, 1966/1974). They cross both the left-hand and upper axis. Craig (1979) pointed out that the nonparametric measure of the area beneath an ROC that is constructed by connecting

[4] I borrow "regular" from Birdsall (1966) who meant something slightly more specific by it. His definition is essentially the same as Egan's (1975) definition of a "proper" ROC. A *proper* ROC is based on likelihood ratio as the decision variable (see Footnote 2). Such an ROC will have two other properties of a regular ROC: (a) It will be complete; that is, for each value of the horizontal axis there is one value of the vertical axis. (b) It will be convex; that is, it will be on or above the line segments connecting any two points the discriminator can produce. In general, these ROCs will have a monotonically decreasing slope from the point (0, 0) to the point (1, 1).

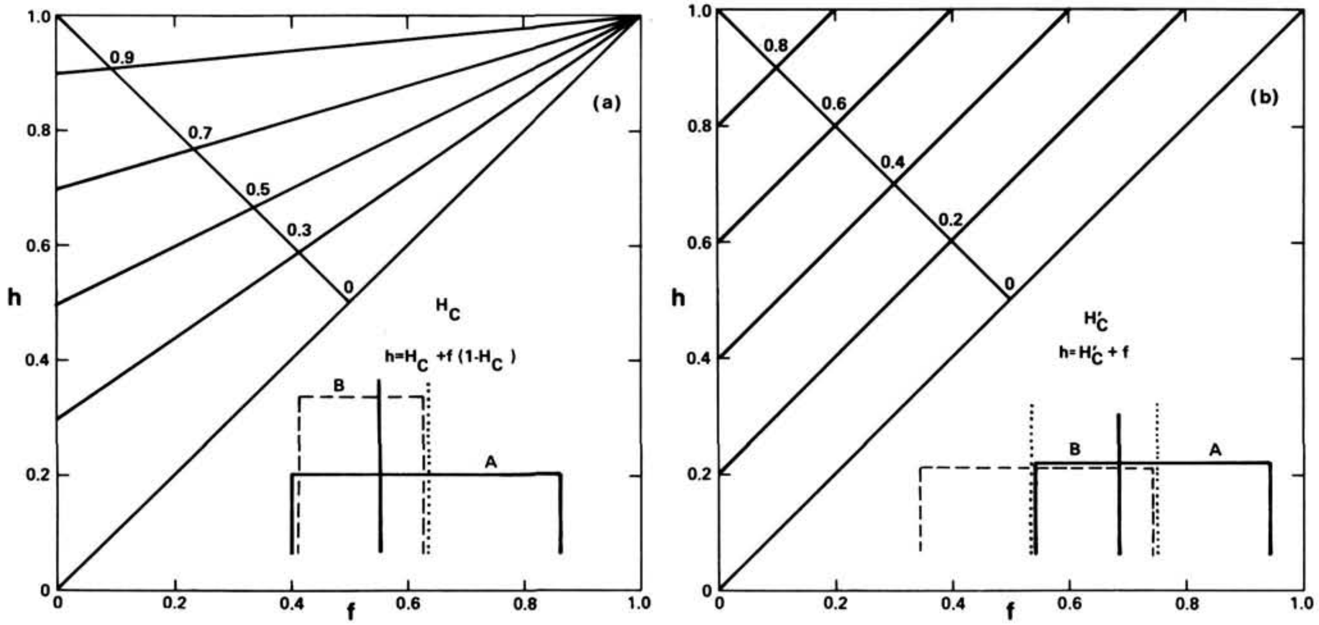

Figure 9. a: relative operating characteristics (ROCs) for the H_C index on ordinary scales. b: ROCs for the H'_C index on ordinary scales. (h: conditional probability of a hit; f: conditional probability of a false alarm; A and B: distributions of observations that arise from the two events; H_C: an accuracy index that implies a high-threshold model; and H'_C: an accuracy index that implies a double-threshold model.)

a single observed point and the corners (0, 0) and (1, 1), a special case of the general area measure discussed by Green and Swets (1966/1974), also leads to the theoretical ROC of the double-threshold model.

Percentage (Proportion) Correct, PC

The ROCs for PC are shown in Figure 10a. The three values of s included (.25, .50, .75) indicate the strong dependence of PC on that variable. For $s = .50$, these ROCs have the same form as those of H'_C, as Macmillan and Kaplan also showed (1985).

In general, the slope of the ROCs for PC is $(1 - s)/s$. The ROCs of different slopes rotate about their intersection with the negative diagonal. All points along an ROC that intersects the negative diagonal at a given point are assigned the same index value, which is equal to the value of h at the negative diagonal. The reader might note, for example, the broad locus of ROC points assigned PC = .70, at the three values of s shown: thus, at $f = .40$, h could be 1.0, .80, or .73; at $f = .10$, h could be .63, .50, or .10. For values of $s \neq .5$, some ROCs will cross the positive diagonal, with the result that some better-than-chance performances are given the same index value as some poorer-than-chance performances.

Skill Test, Z

ROCs for the index Z, calculated by Mason (1982), are shown in Figure 10b. They have the same form as those of H'_C and of PC for $s = .5$. As the curves for $s = .70$ and .30 show, Z depends on s in that better performance in terms of h and f is required to achieve a given index value if the probabilities of the two alternatives differ.

Kappa Statistic, K

The use of K was suggested as a possible accuracy index by G. Koch (personal communication, 1984), and my calculations of its ROCs are shown in Figure 11a. Again, slopes of 1 occur for $s = .50$, and, indeed, for $s = .50$, $K = H'_C$. As with PC, the slope varies with s. All points on the upper three ROCs yield the same value of K (.50) and so do all points on the lower three ROCs (.20).

Phi Coefficient, ϕ

In Figure 11b, it can be seen that ROCs for ϕ (discussed by Swets, 1983a, 1983b) share with H'_C, Z, and K the intersection of both axes other than at the corners. They differ in having a slight curvature. Only the ROCs for $s = .5$ are shown; as with PC and K, ROCs for ϕ tilt away from symmetry about the negative diagonal for other values of s. N. Macmillan (personal communication, 1985) pointed out that a nonparametric area index based on a single observed ROC point, as proposed by Pollack and Norman (1964), denoted A', yields ROCs that closely resemble those of ϕ for $s = .5$ (see their Figure 2 or Macmillan and Kaplan's [1985] Figure 11). C. E. Metz (personal communication, 1985) observed that both of these ROCs are arcs of an ellipse.

Threshold Models

As mentioned previously, the model for H_C is a high-threshold (two-state) model. As indicated in the inset to Figure 9a, representations of Alternative A (distributed according to the solid line) may fall above or below the threshold (shown as a dotted vertical line). The threshold is "high" in that it is never exceeded by representations of Alternative B (distributed according to the

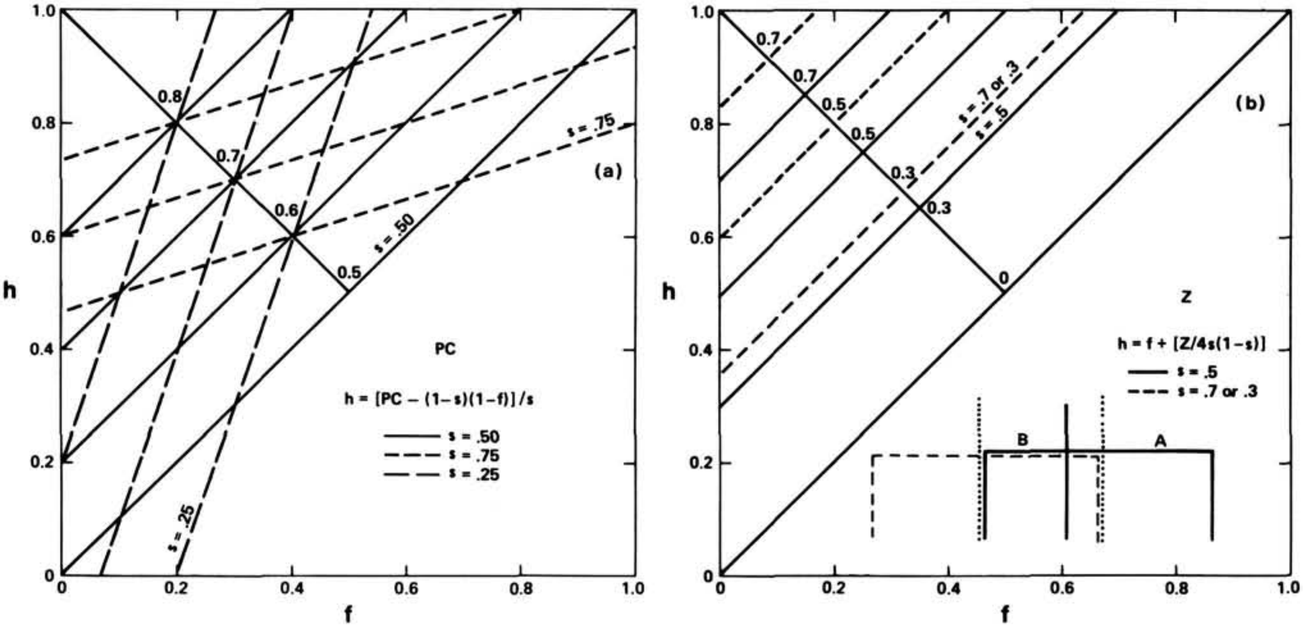

Figure 10. a: relative operating characteristics (ROCs) for the PC index on ordinary scales. b: ROCs for the *Z* index on ordinary scales. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; *s*: probability of occurrence of one event; *A* and *B*: distributions of observations that arise from the two events; PC: the accuracy index that is the percentage of correct responses; and *Z*: an accuracy index, as defined in the figure, used in weather forecasting.)

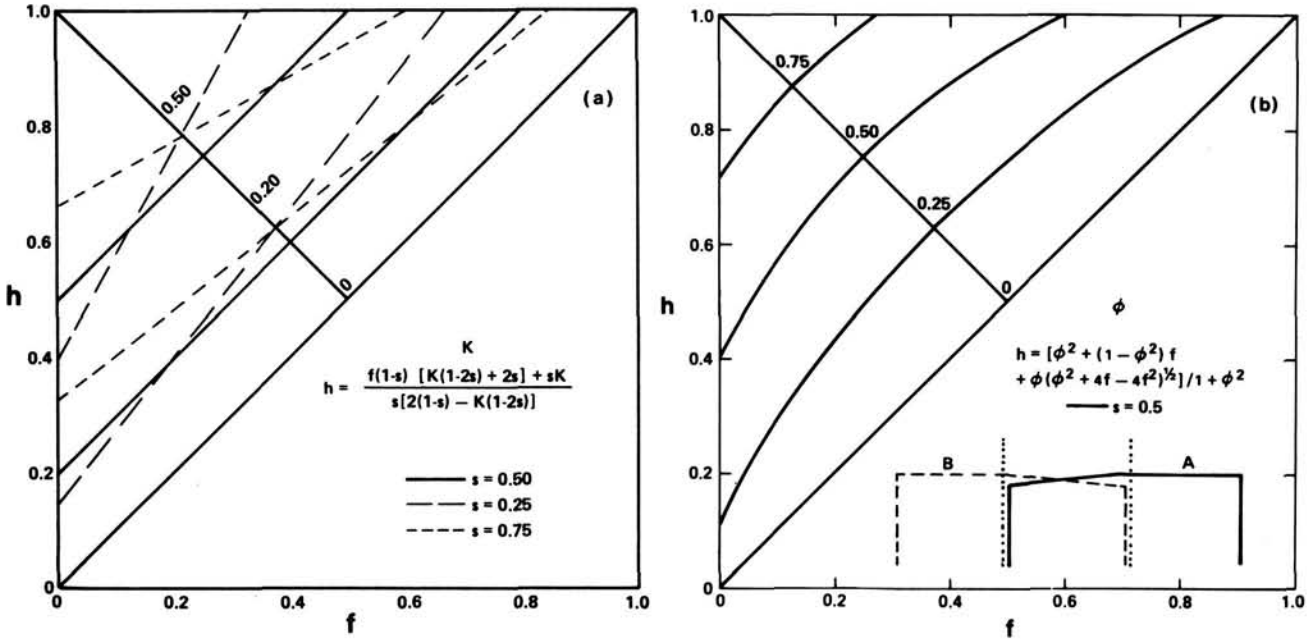

Figure 11. a: relative operating characteristics (ROCs) for the kappa index on ordinary scales. b: ROCs for the ϕ index on ordinary scales. (*h*: conditional probability of a hit; *f*: conditional probability of a false alarm; *s*: probability of occurrence of one event; *A* and *B*: distributions of observations that arise from the two events; *K*: the accuracy index supplied by the kappa statistic; and ϕ: the accuracy index supplied by the phi coefficient.)

dashed line, slightly offset). Values above the threshold are indistinguishable from one another, as are values beneath it. The value of h at threshold, that is, at $f = 0$, is the "true" h and is inflated according to a chance mechanism when $f > 0$. That is, the observer who desires a higher h than that given at $f = 0$ must respond "Alternative A" to a random selection of values of the decision variable x that fall beneath the threshold. One may picture such an observer in detection-theory terms as setting a decision criterion somewhere beneath the threshold; the solid vertical line represents a criterion set so that about one-third of the observations beneath the threshold receive the "Alternative A" response. Of course, the x-axis in this picture is not a continuum; the probabilities associated with the subareas of the rectangles can be viewed as massed at appropriate points along the x-axis.

A double-threshold model is implied by H'_C, PC, Z, K, and ϕ. All but ϕ correspond to a three-state model: Two thresholds define three categories of representations of the alternatives such that the values within each category are indistinguishable (Swets, 1961). This model can be related to uniform PDFs as in Figures 9b and 10b. Some representations, arising only from A, are above a high threshold (dotted vertical line on right) and lead correctly to the choice of A; others, arising only from B, are beneath a low threshold (dotted vertical line on left) and lead correctly to the choice of B. Representations in a third category arise from A or B, fall between the two thresholds, lead to either choice, and are correct or incorrect strictly by chance. In detection-theory terms, the solid vertical line represents a decision criterion that assigns about one-third of the indeterminate representations to Alternative A. (Again, however, the x-axis is not a continuum.) The model for ϕ is similar, but the representations in the middle category are distinguishable from each other to a slight degree (see Fig. 11b, inset) and thus permit slightly better than chance behavior. The ROCs in Figures 9–11 (for $s = .5$) are symmetrical about the negative diagonal, as they would be for symmetrically placed thresholds, and as implied by the various indices. More detailed discussions of the models (except that of ϕ) are given by Green and Swets (1966/1974).

Empirical ROCs and Suitable Indices

Characterization of Empirical ROCs

A companion paper (Swets, in press) includes a large number of empirical ROCs drawn from psychological experiments on sensory capacity, memory, cognition, and learning, as well as from other fields, including aptitude testing, polygraph lie detection, weather forecasting, information retrieval, and medical diagnosis. They strongly support the conclusions that empirical ROCs are fitted well on a binormal graph by straight lines of varying slope. Predominantly, the observed binormal slopes are between 0.7 and 1.0—some a little lower (to about 0.5) or higher (to about 1.5). The implication of the straight line is that empirical ROCs are regular, as that term is defined above. The implication of the variation in binormal slope is that a free slope parameter is required to fit data. A corollary of the observed variation in binormal slope is that any index defined in terms of a single 2 × 2 table will vary along empirical ROCs that have a different form (or different binormal slope) than the fixed form (and binormal slope) that is assumed by the index.

Index Variation

I illustrated earlier how much d' (and, by implication, η, LOR, and Q) could vary along an ROC of binormal slope other than 1. Now, viewing the ROCs of threshold models on a binormal graph shows that they can easily be distinguished from straight lines and that their indices vary along any empirical ROC. Figure 12a shows the ROCs for the high-threshold model and the index H_C; Figure 12b shows ROCs for the double-threshold models, corresponding to H'_C and Z, and also to PC, K, and approximately to ϕ for equal-probability alternatives ($s = .50$); it gives index values for H'_C. None of these curves would come close to meeting a chi-square criterion for a fit by a straight line.

The dashed lines in the figures bound the space of observed ROCs (binormal slope of 1.5 in Figure 12a and 0.5 in 12b) and illustrate the variation in threshold-model index values that could be assigned to a discriminator with fixed capacity to discriminate, that is, with fixed accuracy. The value of H_C assigned could vary from near 0 (chance performance) to near 1.0 (perfect performance). It varies from about 0.5 to 0.9 within a range of $f \approx .20 \pm .10$. Similarly, the indices H'_C, Z, PC, K, and ϕ could vary considerably for fixed accuracy. In the illustration, H'_C varies from about 0.8 to 0.4 in the range of f from about .1 to .5.

The use of one of the indices considered so far appears to need specific justification on conceptual and empirical grounds. An article on memory by Gillund and Shiffrin (1984) provides an example of one kind of empirical justification that might be attempted. They reported that findings and conclusions based on H'_C were unchanged by an analysis in terms of "d' measures" (p. 5). The foregoing analysis suggests that such could be the case for ROC points lying quite close to the negative diagonal of the graph, as would result from a symmetrical decision criterion—that is, one yielding equal error rates. That would probably not be the case for a nonsymmetrical criterion, for example, one set to yield a given value of f, say, .05. For an ROC point along a vertical line at that value, H'_C and d' diverge considerably from their values at the negative diagonal.

Nelson (1984) based one justification for Q on a reliability argument; specifically, he implied that Q has a smaller error variance for a given number of observations than either d' or an ROC area index of the kind defined in the next section as suitable to binormal ROCs of varying slope. However, establishing the facts in this matter may not be simple. According to an analysis suggested by C. E. Metz (personal communication, 1985), the relative error for d'_e is smaller than that for Q up to $d'_e \approx 1.4$ and $Q \approx .8$, and then is larger; in addition, the relative error for the area index A_z is always smaller than that for Q, ranging from 0.004 of the relative error for Q at $Q = .01$, up to 0.567 of the relative error for Q at $Q = .99$.[5] A fairer comparison, though, as

[5] These calculations are based on the following two equations for relative error:

$$\frac{\sigma_{d'_e}}{d'_e} = \frac{\sigma_Q}{Q} \cdot \frac{Q}{1 - Q^2} \cdot \frac{1}{ln\left(\frac{1 + Q}{1 - Q}\right)}, \qquad 22$$

and

$$\frac{\sigma_{A_z}}{A_z} = \frac{\sigma_Q}{Q} \cdot \left\{ \frac{1}{4\sqrt{2}} \cdot \frac{Q}{1 - Q^2} \cdot \frac{e}{\Phi\left(\frac{\sqrt{\pi}}{4} \, l\eta\left[\frac{1 + Q}{1 - Q}\right]\right)} \right\}; \qquad 23$$

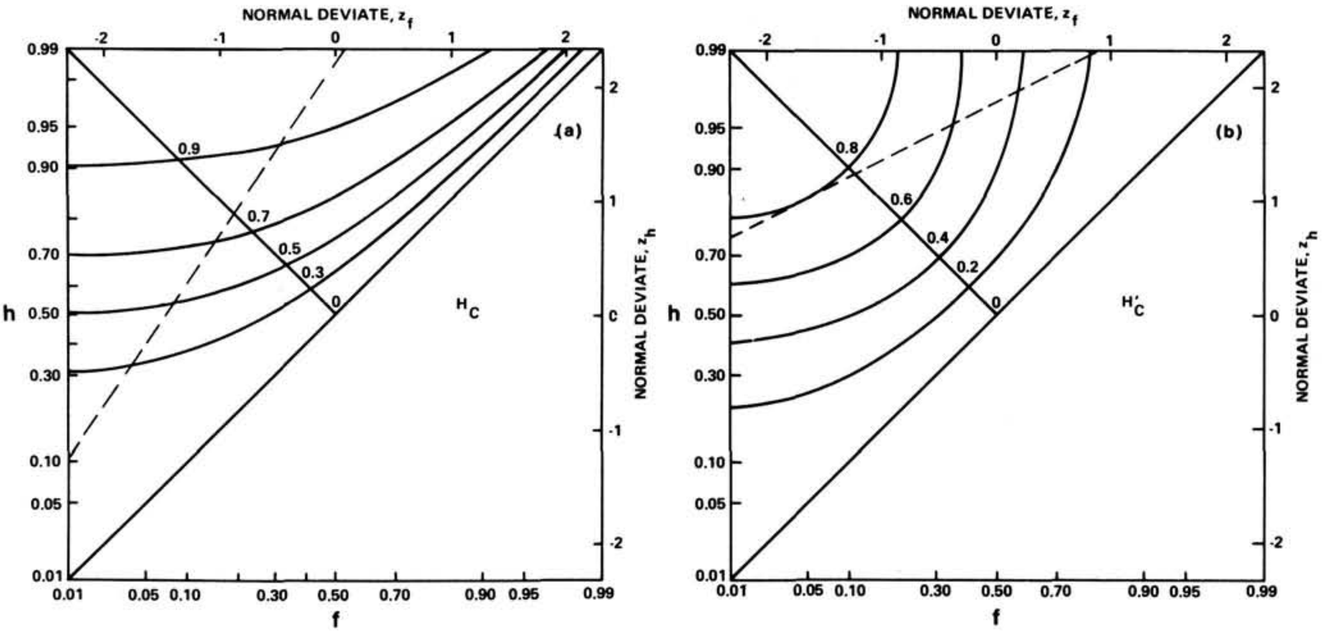

Figure 12. a: relative operating characteristics (ROCs) for the high-threshold, two-state model on a binormal graph. b: ROCs for the double-threshold, three-state model on a binormal graph. (h: conditional probability of a hit; f: conditional probability of a false alarm; z_h and z_f: normal-deviate values of h and f; H_C: an accuracy index associated with the high-threshold model; and H'_C: an accuracy index associated with the double-threshold model.)

Metz observed, would be of A_z (which ranges from .5 to 1.0) and d'_e (which ranges from 0 to infinity) in relation to Q after they are normalized to the range of Q from 0.0 to 1.0. Pursuing that trail indicates that the relative error in A_z is smaller than that of Q up to $Q \approx .94$, and then is larger; but it must be appreciated that certain relations involved in the derivation are accurate (within 5%) only for $Q < .90$. Normalizing d'_e relative to Q, given that d'_e ranges from 0 to infinity, is also problematic. In total, there seems to be no justification for Q in terms of a greater reliability.

Suitable Indices

Figure 13 shows ROC slopes of 1 and 0.7 on a binormal graph and indicates some quantities that may be considered in defining indices that are reasonably appropriate for nonunit slopes. The index Δm, defined in Equation 12, is the distance in z units from the origin (at $z_h = z_f = 0$) to the intersection of the ROC and the axis at $z_h = 0$. Ordinarily, one reports the slope along with Δm and together they serve more as an economical description of data than as an accuracy index. The index d'_e, also mentioned earlier, equals $z_f - z_h$ at the intersection of the ROC and the negative diagonal. In Figure 13, the value of d'_e indicated is $0.6 - (-0.6) = 1.2$. Green and Swets (1966/1974) showed

$$d'_e = 2\Delta m[\text{slope}/(1 + \text{slope})]. \qquad (19)$$

The quantity $z(A)$ is the perpendicular distance, in z units, from the origin of the graph to an ROC (Simpson & Fitter, 1973). It had earlier been defined (with other notation) in psychology (Schulman & Mitchell, 1966) and information retrieval (Brookes, 1968). It can be shown that

$$z(A) = \text{slope}(\Delta m)/(1 + \text{slope}^2)^{1/2}. \qquad (20)$$

For ROCs of slope = 1,

$$d' = d'_e = 2^{1/2}z(A). \qquad (21)$$

The quantity $z(A)$ has a variance that is familiar in statistics and more tractable than that of d'_e (Brookes, 1968).

I would recommend the index called A_z, which is now probably the most widely used of the indices suitable to ROCs of varying slope (Swets & Pickett, 1982). In terms of the quantity just defined, it is the (tabled) area under the cumulative normal function up to the normal-deviate value equal to $z(A)$. However, A_z is better thought of as the proportion of the area of the ROC graph that lies beneath an ROC (on ordinary scales) that is assumed to be a straight line on a binormal graph. This index runs from .5, at the diagonal of the ROC graph representing chance performance, up to 1.0 for perfect performance; that is, when an ROC point is observed in the upper left-hand corner. Illustrations of the use of A_z are available (e.g., Swets et al., 1979).[6]

which, in turn, are based on the relations:

$$A_z = \Phi(d'_e/\sqrt{2}) \doteq \Phi\left(-\frac{\sqrt{\pi}}{2}\ln\eta\right),$$

where

$$\eta = \sqrt{\frac{1-Q}{1+Q}}.$$

[6] Perhaps contrary to appearances, A_z does not assume normal distributions, but rather any form of distribution that can be transformed monotonically to the normal. A_z and, more generally, the linear fit on a

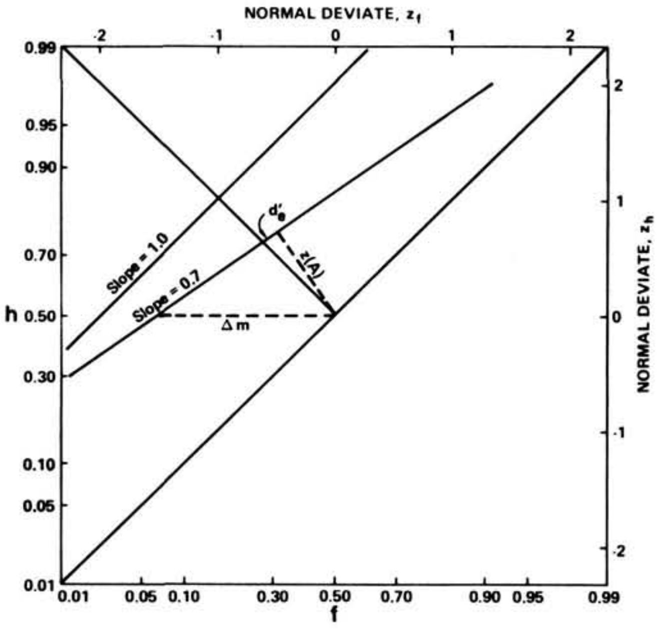

Figure 13. Relative operating characteristics (ROCs) of two illustrative slopes on a binormal graph, and some quantities used to define acceptable indices. (h: conditional probability of a hit; f: conditional probability of a false alarm; z_h and z_f: normal-deviate values of h and f; Δm and d'_e: accuracy indices based on unequal-variance, normal distributions; and $z(A)$: an accuracy index defined as the perpendicular distance from the origin of the binormal graph to the ROC.)

A computer program developed by Dorfman and Alf (1969) and revised by Dorfman (Swets & Pickett, 1982) provides a maximum-likelihood fit to ROC data obtained by the rating method, and estimates of the indices Δm, d'_e, $2^{1/2}z(A) = d_a$, and A_z, along with their variances. A listing of this program was given by Swets and Pickett (1982), who also described the construction of ROC curves by the rating method.

Though the linearity of empirical ROCs on a binormal graph is a robust finding, one might prefer an index that is not parameterized in terms of any underlying distributions. The index here termed *PA*, for "proportion of area," is the proportion of area in the unit-square ROC graph (ordinary scales) beneath the observed points when they and the (0, 0) and (1, 1) corners are connected by straight lines (Green & Swets, 1966/1974). Like A_z, PA runs from .5 to 1.0. If PA is taken as the area under a continuous curve (vs. the series of linear segments), it will, of course, be slightly larger. This larger value can be shown to equal the proportion of correct choices in a two-alternative, "forced-choice" task—that is, when both Alternatives A and B are presented on each trial and the observer says which is which. This equality holds for any and all forms of assumed PDFs (Green & Swets 1966/1974). The area index defined as A_z above, which depends on a line fitted to the observed data points (on a binormal graph), has an advantage over PA determined by linear segments (on ordinary scales) in being less dependent on the spread of the points.

I have assumed that one can usually obtain a sufficient number of data points (say, 4 or more) to define an ROC, and the force of this article is that one should do that if possible. When it is not possible, as perhaps in sensory testing of young children, the choice would seem to be among the indices of Table 3—d', η, LOR, and Q—as at least being independent of the relative frequencies, s, and implying a linear (or near-linear) binormal ROC. The use of one of those indices would be supported by a finding of nearly equal error proportions.

Summary

The two-alternative single-stimulus discrimination task appears in many guises. Several indices of discrimination accuracy have been defined in terms of a single 2 × 2 data table. However, a discriminator under fixed conditions, and with fixed capacity or accuracy, can (and usually will) generate other 2 × 2 data tables that, for each of these indices, lead to different values. Similarly, two discriminators in a given setting having the same intrinsic accuracy may produce different tables, and different values on each index; in addition, two discriminators with unequal accuracies may produce the same table, and the same value on each index. The source of this variation is variation in the decision criterion used for choosing one alternative over the other, which the discriminator can usually select at will, and is independent of discrimination capacity. In short, indices defined in terms of a single 2 × 2 table confound discrimination capacity and decision criterion. Hence, investigators should use an index of this sort only when a very rough estimate of accuracy is adequate for the measurement problem in question.

On considering such an index, one can be informed in several ways by the ROCs that it implies. The ROC is a graph that shows how the 2 × 2 data tables can vary for any constant value of an index. It can reveal an index's violation of basic measurement purpose, such as giving the same value to performances of obviously different accuracies, for example, performances better and poorer than chance. It can point out that an index depends, inappropriately, on the relative frequencies of the two alternatives. It can disclose that two apparently different indices lead to the same result. Primarily, the ROC will reveal what an index has laid on as assumptions about the discrimination process, to the point of specifying the index's general model of the process.

Several representative indices I examined imply threshold models and produce what are defined as nonregular ROCs, and are thus inconsistent with available data, which show regular ROCs across a wide variety of discrimination tasks and settings. Other representative indices imply variable-criterion models and produce regular ROCs, but ROCs that are fixed across all conditions whereas empirical ROCs vary in a particular way. In general, discrimination accuracy is most reliably determined when several different 2 × 2 data tables are collected—or, better, one 2 × r table based on r confidence ratings—so that a full ROC is obtained, and so that an index can be calculated that depends on the locus of the full, measured ROC. Such an index is not confounded by the decision criterion, nor by the relative frequency of the two alternatives. It is consistent with the fun-

binormal graph, make a particular assumption about the (observable) functional form of the ROC, and not about the (usually unobservable) forms of underlying distributions. A_z is parameterized in terms of an effective pair of normal distributions only as a convenient convention.

damental properties of the discrimination process, and accommodates the particular variation in ROCs that occurs across observers, tasks, and settings.

References

Appleman, H. S. (1959). A fallacy in the use of skill scores. *Bulletin of the American Meteorological Society, 41,* 64–67.

Bermowitz, R. J., & Zurndorfer, E. A. (1979). Automated guidance for predicting quantitative precipitation. *Monthly Weather Review, 107,* 122–128.

Birdsall, T. G. (1966). The theory of signal detectability: ROC curves and their character. *Dissertation Abstracts International, 28,* 1B.

Birdsall, T. G., Lamphiear, D. E. (1960). *Approximations to the noncentral chi-square distribution with applications to signal detection models.* (Tech. Rep. No. 101). Ann Arbor: University of Michigan, Electronic Defense Group.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53,* 129–160.

Brier, G. W., & Allen, R. A. (1952). Verification of weather forecasts. In T. F. Malone (Ed.), *Compendium of Meteorology* (pp. 841–848). Boston: American Meteorological Society.

Brookes, B. C. (1968). The measures of information retrieval effectiveness proposed by Swets. *Journal of Documentation, 24,* 41–54.

Bush, R. R. (1963). Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 429–469). New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

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

Dorfman, D. D., & Alf, E., Jr. (1969). Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology, 6,* 487–496.

Edwards, A. W. F. (1963). The measure of association in a 2 × 2 table. *Journal of the Royal Statistical Society, Series A, 25,* 109–114.

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

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *Journal of the Acoustical Society of America, 33,* 993–1007.

Finley, J. P. (1884). Tornado predictions. *American Meteorological Journal, 1,* 5–88.

Fisk, A. D., & Schneider, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 181–197.

Galen, R. S., & Gambino, S. R. (1975). *Beyond normality: The predictive value and efficiency of medical diagnoses.* New York: Wiley.

Gart, J. J. (1971). Comparison of proportions. *Review of the International Statistical Institute, 39,* 148–169.

Gilbert, G. K. (1885). Finley's tornado predictions. *American Meteorological Journal, 1,* 167–172.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association, 45,* 226–256.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–764.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics* (Reprint). Huntington, NY: Krieger. (Original work published 1966)

Hanssen, A. W., & Kuipers, W. J. A. (1965). On the relationship between frequency of rain and various meteorological parameters. *Royal Netherlands Meteorological Institute, Mededelingen en Verhandelingen, 81,* 2–15.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd. ed). New York: Holt, Rinehart & Winston.

Heidke, P. (1926). Berechnung des Erfolges und der Guete der Windstarkevohersagen im Sturmwarnungsdienst. *Geografiska Annaler, 8,* 301–349.

Jeffress, L. A. (1964). Stimulus-oriented approach to detection. *Journal of the Acoustical Society of America, 36,* 766–774.

Laming, D. (1973). *Mathematical psychology.* London: Academic Press.

Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98,* 185–199.

Mason, I. (1982). *On scores for yes/no forecasts.* Preprints of papers delivered at the Ninth Conference on Weather Forecasting and Analysis, the American Meteorological Society (pp. 169–174). Seattle, Washington.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298.

Murphy, A. H. (1977). The value of climatological, categorical, and probabilistic forecasts in the cost–loss ratio situation. *Monthly Weather Review, 105,* 803–816.

Nelson, T. O. (1984). A comparison of current measures of accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum likelihood estimation of ROC curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory, 4,* 171–212. (Reprinted in R. D. Luce, R. R. Bush, & E. Galanter [Eds.]. [1963]. *Readings in mathematical psychology* [pp. 167–211]. New York: Wiley)

Pickup, M. N. (1982). A consideration of the effect of 500 mb cyclonicity on the success of some thunderstorm forecasting techniques. *Meteorological Magazine, 111,* 87–97.

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

Ramage, C. S. (1982). Have precipitation forecasts improved? *Bulletin of the American Meteorological Society, 63,* 739–743.

Rousseau, D. (1980). *A new skill score for the evaluation of yes/no forecasts.* Preprints of papers delivered at the World Meteorological Organization Symposium on Probabilistic and Statistical Methods in Weather Forecasting (pp. 167–174). Nice, France.

Schrank, W. R. (1960). A solution to the problem of evaluating forecast techniques. *Bulletin of the American Meteorological Society, 42,* 277–280.

Schulman, A. I., & Mitchell, R. R. (1966). Operating characteristics from yes-no and forced-choice procedures. *Journal of the Acoustical Society of America, 40,* 473–477.

Simpson, A. J., & Fitter, J. J. (1973). What is the best index of detectability? *Psychological Bulletin, 80,* 481–488.

Swets, J. A. (1961). Is there a sensory threshold? *Science, 134,* 168–177.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182,* 990–1000.

Swets, J. A. (1983a). Assessment of nondestructive-testing systems—Part

I: The relationship of true and false detections. *Materials Evaluation, 41,* 1294–1298.

Swets, J. A. (1983b). Assessment of nondestructive testing systems—Part II: Indices of performance. *Materials Evaluation, 41,* 1299–1303.

Swets, J. A. (in press). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin.*

Swets, J. A., & Green, D. M. (1978). Applications of signal detection theory. In H. L. Pick, Jr., H. W. Leibowitz, J. E. Singer, A. Steinschneider, & H. W. Stevenson (Eds.), *Psychology: From research to practice* (pp. 311–331). New York: Plenum Press.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., & Freeman, B. A. (1979). Assessment of diagnostic technologies. *Science, 205,* 753–759.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340.

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Vernon, E. M. (1953). A new concept of skill score for rating quantitative forecasts. *Monthly Weather Review, 81,* 326–329.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A development study of memory monitoring. *Child Development, 48,* 13–21.

Woodcock, F. (1976). The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review, 104,* 1209–1214.

Woodworth, R. S. (1938). *Experimental psychology.* New York: Holt, Rinehart & Winston.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3,* 32–35.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 579–642.

Appendix

Hyperbolic Form of ROCs for Three Indices

The ROCs associated with the indices η, Q, and LOR, and with equal-variance logistic PDFs in detection theory, are demonstrated here to be hyperbolas, according to a proof by C. E. Metz (personal communication, 1984).

The common functional form of these ROCs (as shown in Table 3) is

$$y = \frac{x}{x + c(1 - x)} = \frac{x}{c + (1 - c)x},$$

where c is a nonnegative constant (<1) specifying the ROC; a smaller c implies a higher ROC. These ROCs have maximum slope, $1/c$, at (0, 0) and minimum slope, c, at (1, 1).

If the coordinates are shifted from the x, y system to a u, v system having the center of the hyperbola at its origin, in particular, such that

$$u = x + \frac{c}{1 - c} \quad \text{and} \quad v = y - \frac{1}{1 - c},$$

so that the origin of the u, v system is $\left(\frac{c}{1 - c}\right)$ to the left of and $\left(\frac{c}{1 - c}\right)$ above the (0, 1) point in the x, y system, then

$$x = u - \frac{c}{1 - c} \quad \text{and} \quad y = v + \frac{1}{1 - c}.$$

Hence, in the u, v coordinates, the equation of the ROC curve is:

$$v + \frac{1}{1 - c} = \frac{\left[u - \frac{c}{1 - c}\right]}{c + (1 - c)\left[u - \frac{c}{1 - c}\right]} = \frac{u - \frac{c}{1 - c}}{(1 - c)u},$$

or, after rearrangement,

$$uv = \frac{-c}{(1 - c)^2} = \text{constant}.$$

Thus, the curve is hyperbolic, and, in particular, is a rectangular hyperbola with the u, v axes as asymptotes.

Received May 10, 1984
Revision received March 11, 1985 ■