Psychological Bulletin
1986, Vol. 99, No. 2, 181–198

Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance

John A. Swets
BBN Laboratories Incorporated, Cambridge, Massachusetts

A sample of empirical relative operating characteristics (ROCs) is presented, drawn both from discrimination tasks in experimental psychology and from diagnostic tasks in several practical fields. These illustrative ROCs are seen to be fitted well by a straight line, of varying slope, on a binormal graph. This result has fundamental implications for models and indices of performance. The form of empirical ROCs is consistent with one version of the variable-criterion model provided by signal detection theory, and is inconsistent both with other versions of that model and with the main threshold models. That form supports the use of certain indices of discrimination accuracy derived from detection theory, and substantiates the potential unreliability of others of that heritage and, more importantly, the unreliability of the several accuracy indices in common use that can be shown to imply one or another threshold model. The preferred detection-theory indices of accuracy are ones that accommodate the varying slope of empirical ROCs. These indices are effectively independent of the decision criterion, which can be indexed separately. Measuring both accuracy and the decision criterion appropriately enables one to delineate which of these dependent variables is affected by which independent variables in psychological studies, and to assess the efficacy as well as the accuracy of diagnostic systems.

A companion article derives the form of the *relative* (or *receiver) operating characteristic* (ROC) that is algebraically implied by each of a dozen or so commonly used indices of discrimination accuracy, and identifies the models of the discrimination process that are implied by the main categories of those forms (Swets, 1986). In this article I present a broad sample of empirical ROCs for comparison with the theoretical forms. They are drawn from discrimination tasks used in the psychology of perception, learning, memory, and cognition, and from several practical fields in which a discrimination, or diagnosis, is made in the interest of prediction, selection, or corrective action. The fields included are medical imaging, information retrieval, weather forecasting, aptitude testing, and polygraph lie detection.

Precís of ROC Form, Indices, and Models

ROC Form

The ROC, in a sentence, is a graph showing the conditional probability of choosing Alternative A when that alternative occurs (here denoted by h, for "hit") plotted against the conditional probability of choosing A when Alternative B occurs (here denoted by f, for "false alarm"). Both h and f increase as the tendency to choose Alternative A increases, or as the criterion for choosing A becomes more lenient.

The form of an ROC is best visualized on a "binormal" graph—a graph in which the usual probability coordinates are rescaled so that their corresponding normal-deviate values are linearly spaced, as in Figure 1. On such a graph, empirical ROCs are consistently fitted well by a straight line that varies in slope; the slopes are generally between 0.5 and 1.5, as indicated in Figure 1a by dashed lines. (Other details of the figure are discussed next.)

Indices and Models

The previous article (Swets, 1986) showed that the form of predicted ROCs serves to sort common accuracy indices and their implied models into three categories. The first category contains models and indices that predict linear, or effectively linear, binormal ROCs of slope = 1.0. The indices and models of the second category predict binormal ROCs that are distinctly curvilinear. Those in the third category predict, or permit, ROCs consistent with both the observed linearity and variable slope, and are thus preferred. The earlier article showed how indices in the first two categories are subject to considerable unreliability.

Indices in the first category include one of the several indices associated with signal detection theory, namely d' (Green & Swets, 1966/1974), and three indices nearly equivalent to d': Luce's (1959, 1963) η, the log-odds ratio (LOR; e.g., Goodman, 1970), and Yule's (1912) Q. The index d' predicts a linear ROC of slope = 1.0, as indicated by the solid lines in Figure 1a. The indices η, LOR, and Q imply a slightly curved ROC, as shown

This article was prepared in part during a sabbatical leave granted me under the Science Development Program of BBN Laboratories Incorporated. Additional support was obtained from a contract with the University of Chicago, under Public Health Service Grant R01 CA-33261 awarded by the National Cancer Institute, Department of Health and Human Services.

Individuals and institutions supplied data and gave permission to reproduce figures, both essential to this article. Ronald M. Pickett made helpful suggestions for exposition.

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

in Figure 1b, that is indistinguishable from a straight line of slope = 1.0 with ordinary amounts of data.

The model for d' is a *variable-criterion* model of the general sort considered in signal detection theory, but specifically one in which observations under each alternative have normal (Gaussian) distributions of equal variance. Two normal distributions (though with unequal variance) are shown at bottom right in Figure 1a, and are denoted n and sn for the "noise-alone" and "signal-plus-noise" alternatives. The criterion is symbolized in the figure by the vertical line, x_c; observations $x > x_c$ lead to the choice of sn and observations $x < x_c$ lead to the choice of n. A variable-criterion model consistent with η, LOR, and Q contains logistic distributions of equal variance, which are similar to the normal distribution, as shown at bottom right in Figure 1b.

The second category contains a variety of other common indices, including two versions of the "hit" probability corrected for chance success, here denoted by H_C and H'_C (where $H_C = [h - f]/[1 - f]$ and $H'_C = h - f$); percentage correct, PC; the kappa statistic, K, as a chance-corrected PC; and the fourfold point correlation coefficient, ϕ. The index H_C implies the curvilinear ROC of Figure 2a. The index H'_C leads to the curvilinear ROC of Figure 2b; as do PC and K when the alternatives to be discriminated are equally probable; and the ROC for ϕ for equal probabilities is practically indistinguishable from that of Figure

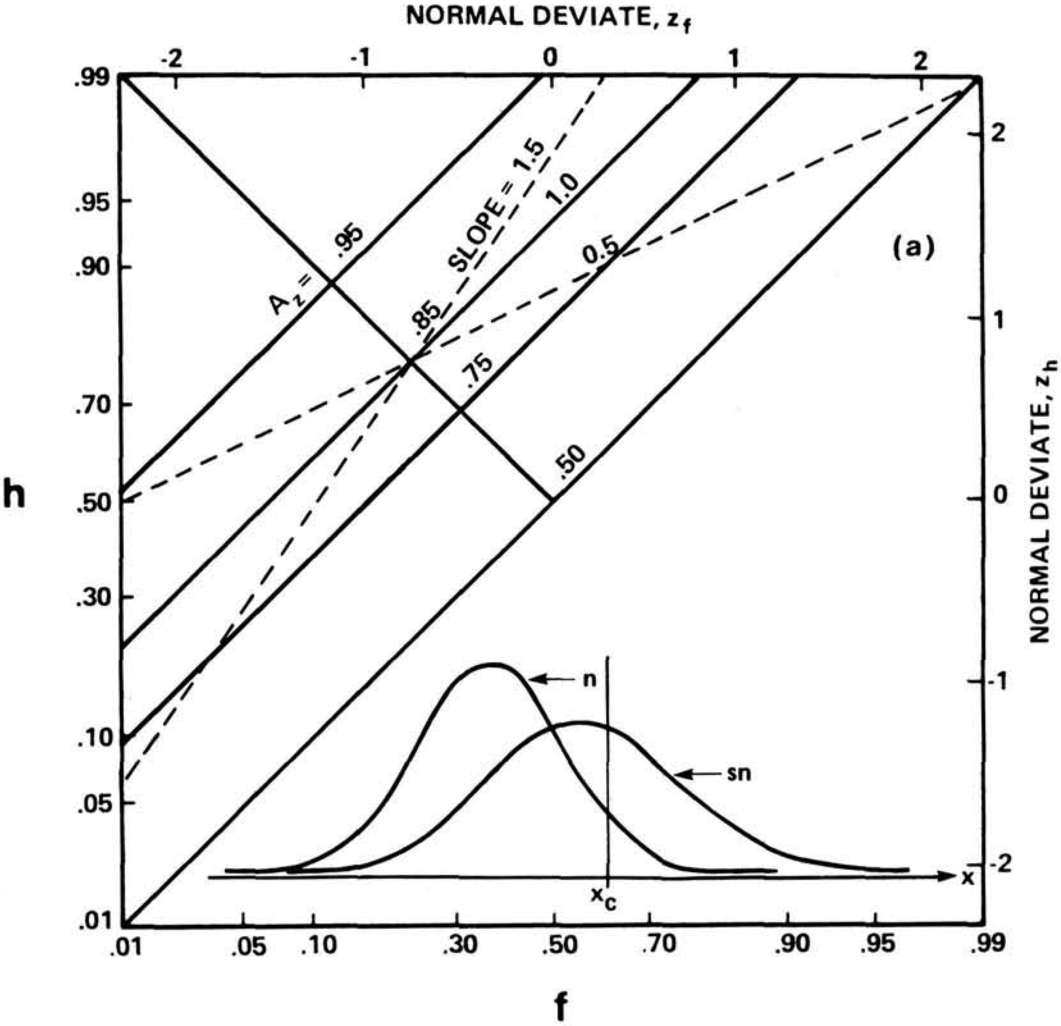

Figure 1. A binormal graph, on which probabilities (left-hand and bottom axes) are scaled so that the corresponding normal-deviate values are linearly scaled (right-hand and top axes). a: the solid lines of slope = 1.0 represent relative operating characteristics (ROCs) consistent with the d' index. (The index values shown are of A_z, the area under the binormal ROC [on ordinary scales]. The dashed lines, of slope ≠ 1.0, represent ROCs consistent with the A_z index, which might arise from distributions of observations of "noise alone" [n] and "signal plus noise" [sn] of unequal variance, as illustrated at bottom right; the distributions shown are normal. The slopes of 0.5 and 1.5 bound almost all empirical ROCs. h = probability of a "hit"; f = probability of a "false alarm.") b *(facing page):* ROCs implied by the η, log odds ratio (LOR), and Yule's Q indices, with index values of η. (Logistic distributions of equal variance, as illustrated at bottom right, produce ROCs having the form of those implied algebraically by η, LOR, and Q.)

2b. For the last three indices, unequal probabilities tilt the ROC away from symmetry about the minor diagonal.

The index H_C implies a *high-threshold* model and H'_C, PC, and ϕ imply a *double-threshold* model. In detection-theory terms, threshold models are based on uniform distributions, as shown in Figures 2a and 2b.

The third category contains a few nearly equivalent indices, including the perpendicular distance from the origin of the binormal ROC graph to the ROC, but perhaps the most common is the area under an ROC (on ordinary scales) that is assumed to be linear on a binormal graph, denoted A_z. The index A_z is consistent with a linear (binormal) ROC having a slope in the range of slopes found empirically. Swets and Pickett (1982) discuss the indices in this category and list a revised version of Dorfman and Alf's (1969) computer program for estimating A_z. Illustrative values of A_z are shown in Figure 1a; its values range from .50 at the positive diagonal, representing chance performance, to 1.00 for perfect discrimination. This index is equivalent to the percentage of correct responses made in a two-alternative, forced-choice test, that is, when a random draw from each of the *sn* and *n* distributions is compared on each trial (Green & Swets, 1966/1974).

The index A_z is associated with a variable-criterion model in which the underlying distributions can have unequal variances, as illustrated by the normal distributions at bottom right in Figure 1a. It should be noted that A_z does not assume normal distributions, but rather any form of distribution that can be transformed monotonically to the normal distribution. Thus A_z, and the binormal assumption more generally, make a particular assumption about the (observable) functional form of the ROC, and not about the (usually unobservable) forms of the underlying distributions. As discussed elsewhere, the forms of the underlying distributions imply a particular form of ROC, but, when the distributions are continuous (as in Figure 1) as opposed to uniform (as in Figure 2), the converse is not true. In general, the ROC reflects the difference between two distributions rather than the distributions themselves, and any monotonic transformation applied to two underlying distributions will result in the same ROC (Egan, 1975; Swets, Tanner, & Birdsall, 1961). Simply as a convenient convention, A_z is parameterized in terms of an ef-

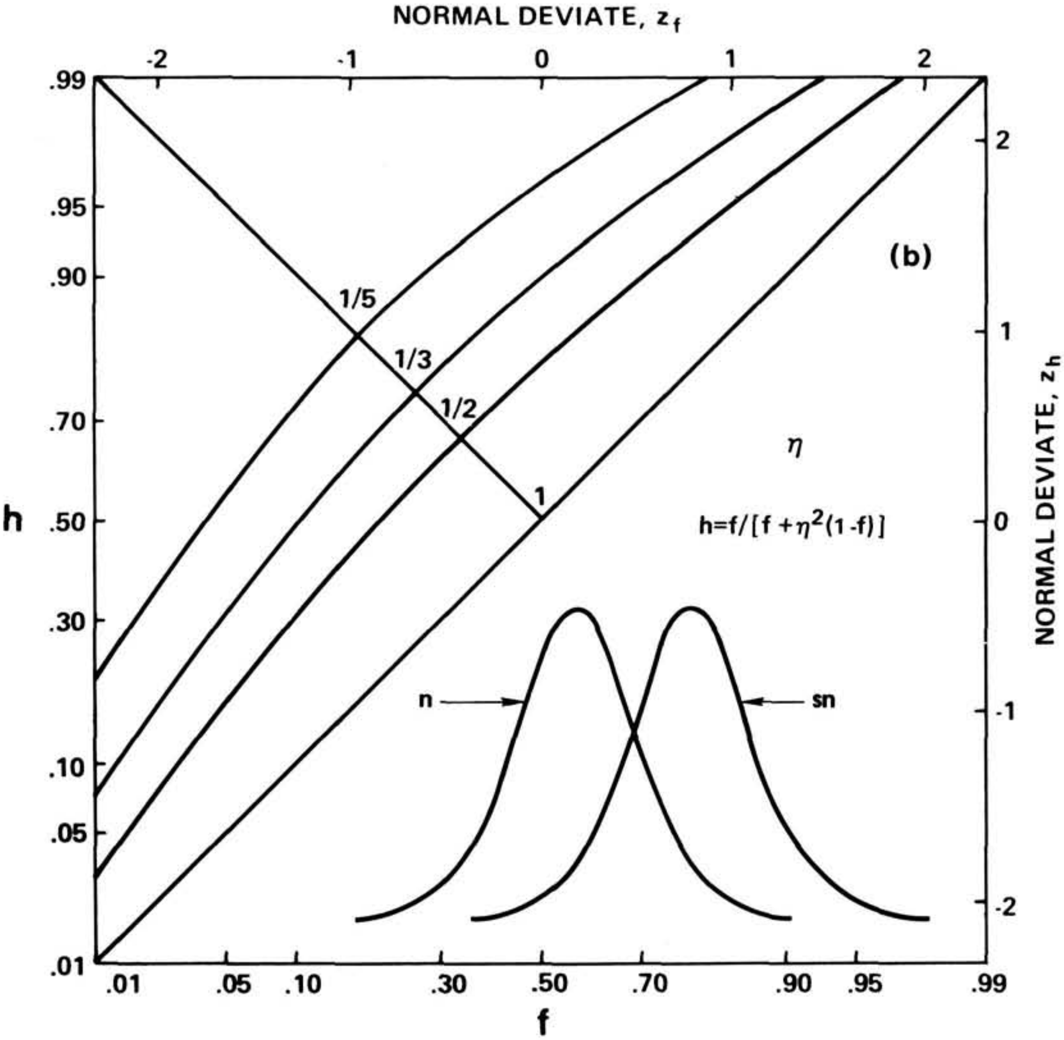

Figure 1. (continued)

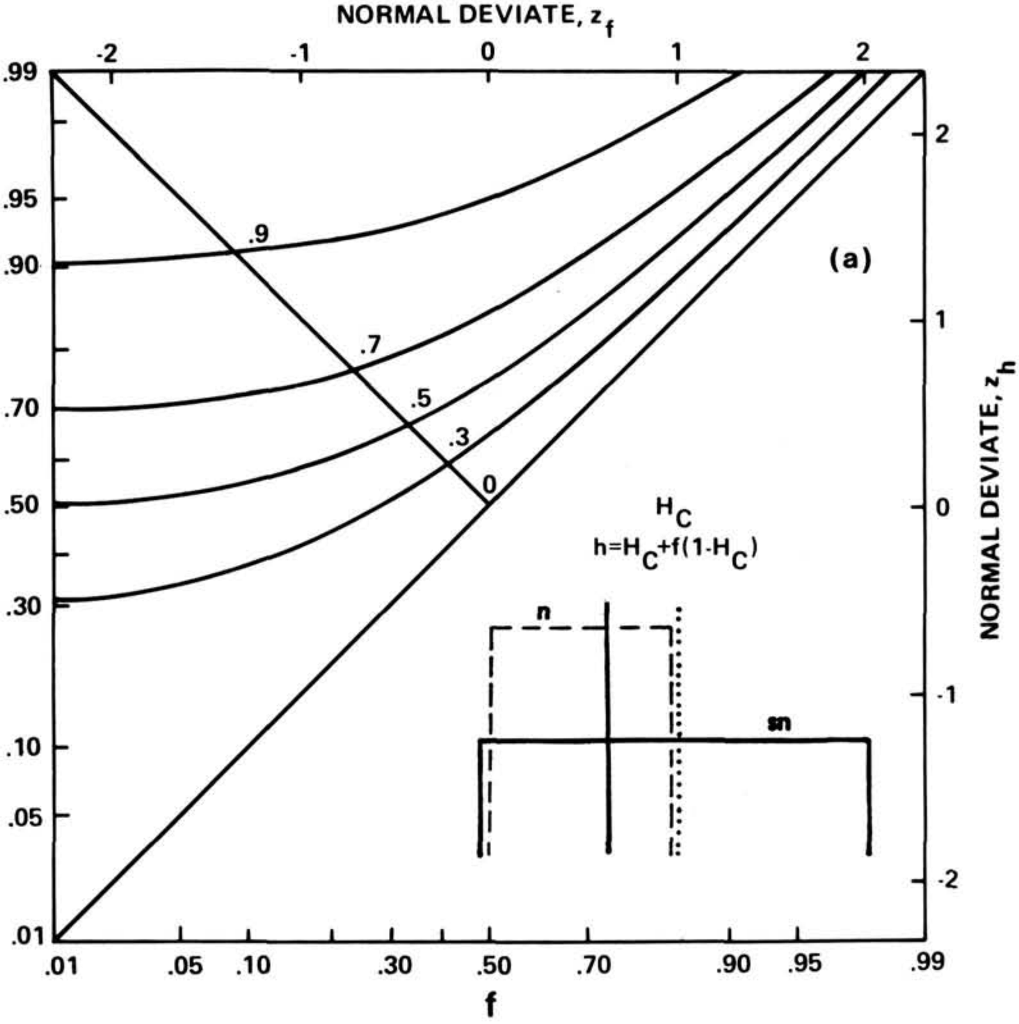

Figure 2. a: Theoretical relative operating characteristics (ROCs) of the high-threshold model on a binormal graph, labeled by the index associated with that model, H_c. (The uniform distributions show some observations of "signal plus noise" [*sn*] and none of "noise alone" [*n*] to exceed a high threshold symbolized by the dotted line; the solid vertical line connotes a variable criterion according to which, in this drawing, about one-third of the observations below the threshold lead to the choice of the *sn* alternative. h = probability of a "hit"; f = probability of a "false alarm.") b *(facing page):* Theoretical ROCs of the double-threshold model on a binormal graph, labeled by one of the indices associated with that model, H'_C. (The uniform distributions of *n* and *sn* are related to two thresholds [dotted lines]; in this illustration, a variable criterion [solid vertical line] is set so that about one-third of the observations falling between the thresholds lead to the choice of the *sn* alternative.)

fective pair of normal distributions, and then the binormal ROC slope consistent with A_z is equal to the ratio of standard deviations, σ_n/σ_{sn}.[1]

Scope and Procedure of This Article

The indices of the three categories delineated are the main ones used in experimental psychology and in practical fields like those mentioned previously. They include not only indices devised in one or more fields but essentially all measures of statistical association, inasmuch as the latter (for 2 × 2 tables) are usually functions either of the cross-product ratio (i.e., *ad*/*bc*, where those letters symbolize the cell entries)—as are η, LOR, and Q—or of the correlation coefficient, ϕ, which depends also on the marginal frequencies of the table (Bishop, Fienberg, & Holland, 1975).

After a few notes on the way ROCs for individual observers are obtained and then combined into average or typical ROCs,

[1] The unequal-variance model presents a problem for theory in that the observation or decision axis (labeled x in Figure 1a) is not monotonic with the likelihood ratio, that is, with the ratio of the ordinate of the *sn* distribution to the ordinate of the *n* distribution; in particular, the likelihood ratio is > 1.0 at both ends of this axis. As noted earlier (Laming, 1973; Swets, Tanner, & Birdsall, 1961), this model requires adding a substantive psychological assumption to the structure of statistical decision theory. One would prefer a better way to handle binormal ROC slopes unequal to 1.0, and better ways exist for certain fixed slopes (Swets, 1986),

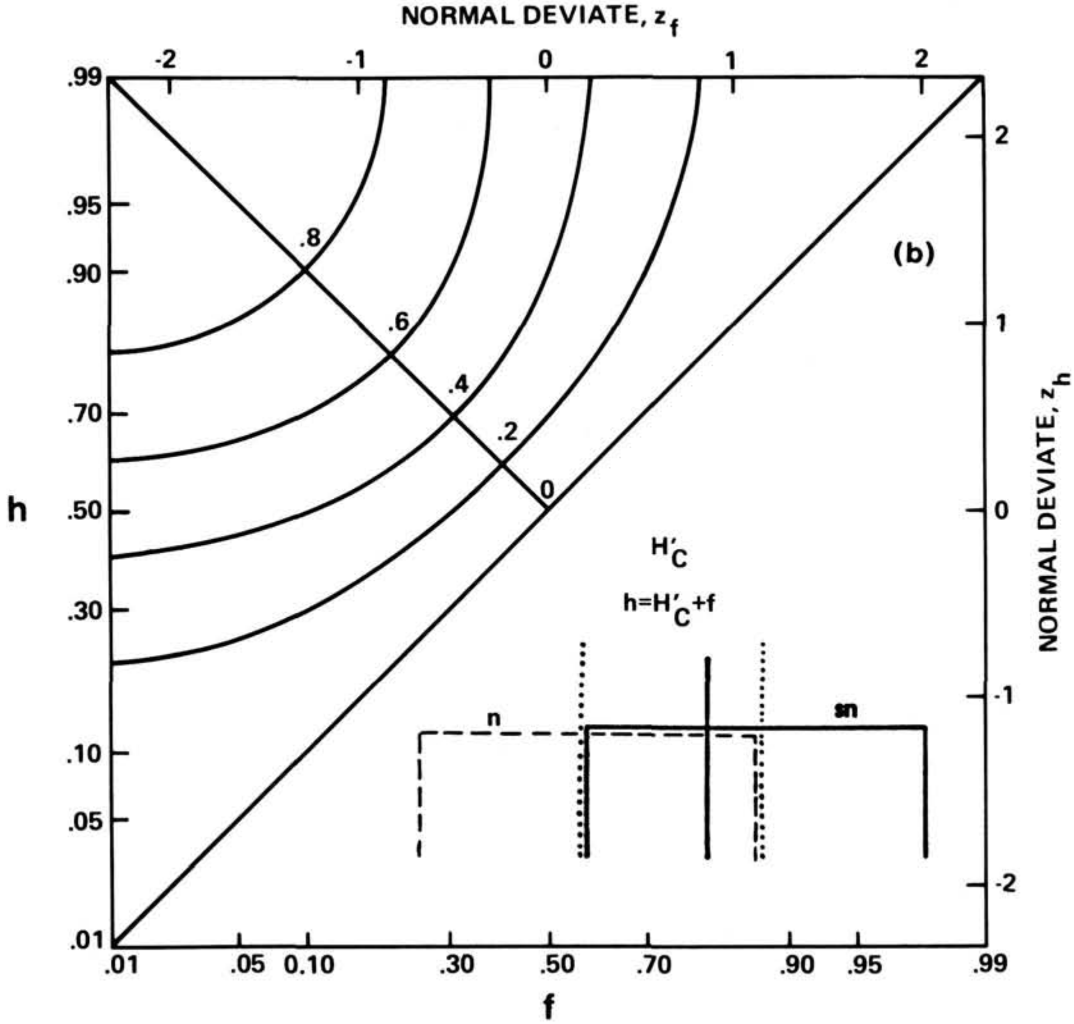

Figure 2. (continued)

I present the sample of empirical ROCs selected for this article. In the four practical fields other than medical imaging, they are illustrative of the only sets of ROCs of which I am aware. In that field and in experimental psychology, the main selection criterion was that they be based on a sufficient number of trials to show relatively little variability, in order to give a good look at their form. A second criterion, applied to ROCs in psychology, was that they represent the various types of discrimination tasks used in psychological experiments. As far as I know, no empirical ROCs support a form other than the one inferred here.

For the several ROCs shown, the interest lies first in the adequacy of a linear fit and then in the slope of the line. The goodness of fit is reported when available, that is, when the data have been submitted to a computer program, such as that described by Dorfman and Alf (1969), that makes a chi-square test of a maximum-likelihood estimation of a linear fit. The slope is reported in every case, as measured graphically from a visual fit when an objective fit was not made. The question of whether some pattern exists in the variation of slope that is observed is discussed briefly in a closing section.

All of the ROCs that follow are presented as being effectively linear and not in the least suggestive of one of the curvilinear forms of Figure 2. The visual effect is compelling enough, I believe, to permit us to forego a statistical comparison of each empirical ROC with each theoretical ROC.

Notes on Data Collection and Combination

The direct way to trace out an ROC is to have an observer adopt a different decision criterion, or response bias, from one group of trials to another, and so obtain several different points (Tanner & Swets, 1954). A more efficient way is to have the observer use a rating scale—say, a five-category scale of confidence that a particular alternative is the correct one—in a single group of trials (Swets et al., 1961). The procedure for obtaining simultaneously a number of different ROC points (one fewer

but the unequal-variance model seems to be the best available for treating variable slopes. The aberrations that occur in ROCs, if the decision is based on *x* rather than the likelihood ratio of *x*, are usually small and at the edges of the graph.

point than the number of categories) from rating data is detailed by Green and Swets (1966/1974). In brief, one assumes in analysis that observations that lead to responses of the highest category of confidence are those that meet the strictest decision criterion being used, and that observations leading to responses of either of the two highest categories are those that meet the next strictest criterion, and so on, cumulatively (when the lowest category is finally included, the ROC point calculated in the uninformative $h = f = 1.0$). Most of the data I present were obtained by some version of the rating-scale technique.

It is often desirable to portray a composite ROC based on several observers. Macmillan and Kaplan (1985) have published a general analysis of ways to obtain composite ROCs and support the pooling of rating data. Most of the ROCs I present are based on such a pooling.

Empirical ROCs in Experimental Psychology

The empirical ROCs displayed here are drawn from a range of tasks studied in experimental psychology—tasks focused on sensory functions or perception, memory, learning, and conceptual judgment.

Vision

The first rating ROCs obtained are shown in Figure 3 (Swets, Tanner, & Birdsall, 1955, 1961). Four observers used a 6-category rating scale for a signal that was a flash of a spot of light, and made nearly 1,200 observations. The figure is reproduced from Dorfman and Alf (1969); a later version of their program (Swets & Pickett, 1982) indicates that the probabilities associated with

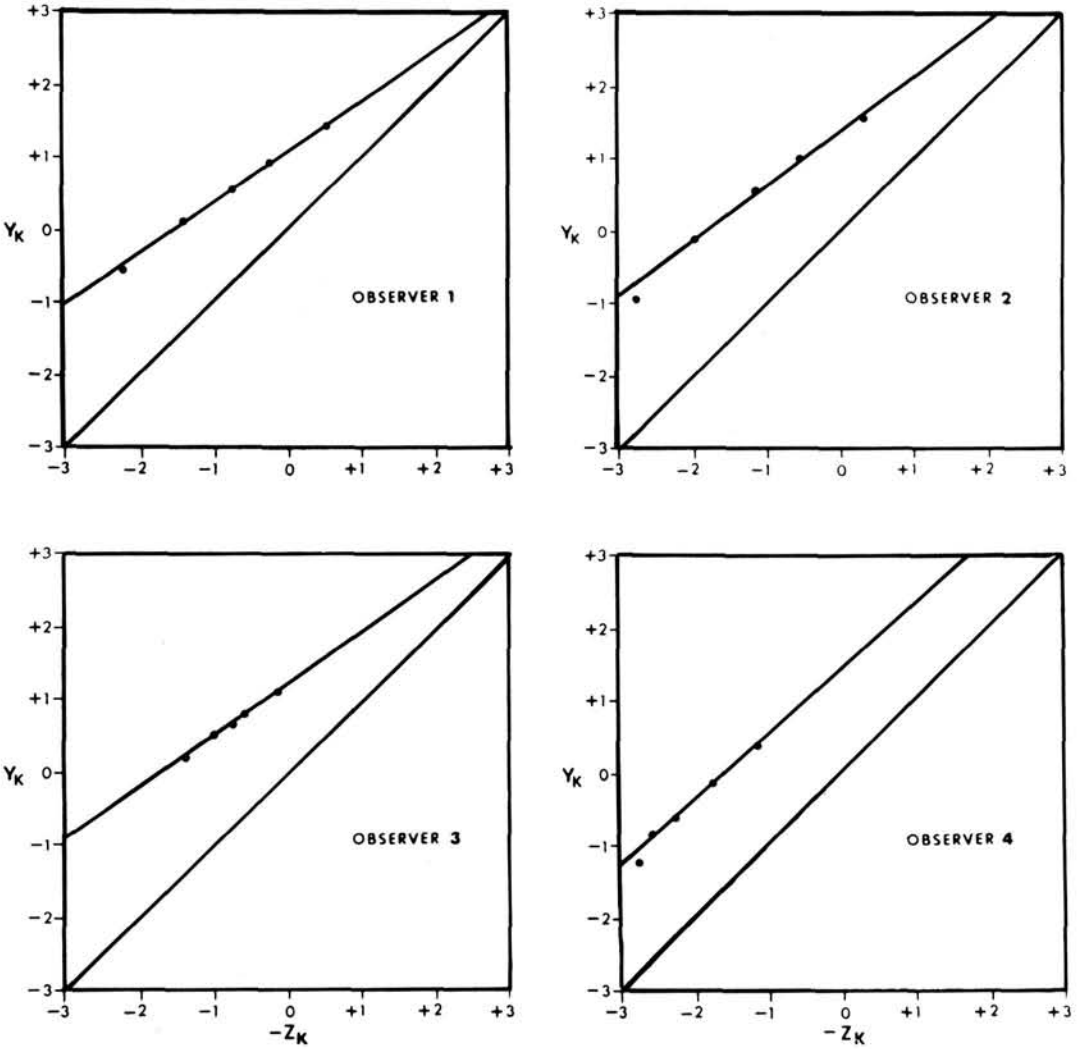

Figure 3. Relative operating characteristics (ROCs) for 4 observers in a vision experiment. (Y_K = normal-deviate value corresponding to the conditional probability of a "hit." $-Z_K$ = normal-deviate value corresponding to the conditional probability of a "false alarm." (From "Maximum Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals—Rating Method Data" by D. D. Dorfman and E. Alf, Jr., 1969, *Journal of Mathematical Psychology, 6,* p. 492. Copyright 1969 by Academic Press. Reprinted by permission.)

the chi-square tests of the linear fits are approximately .70, .05, .40, and .60 for the four observers, respectively, and that the group chi-square is not statistically significant (>.25). The four binormal slopes are 0.71, 0.74, 0.72, and 0.89.

Recognition Memory—Words

Some results of Egan's (1958) extension of the ROC to memory for words are reproduced in Figure 4. Subjects were given either one or two presentations of a list of 100 words and were later asked to rate their confidence (7-category scale) that each of a list of 200 words was on the first list; the 200-word list contained all of the words on the 100-word list. The labels on the graph's axes are the probabilities that "old" words (ordinate) and "new" words (abscissa) are said to be "old." The bottom ROC gives the combined result for the middle 50% of the 16 subjects given one presentation; the top ROC shows the same thing, but for two presentations. The binormal slopes of the two lines are 0.67 and 0.71.

Recognition Memory—Odors

Fifteen subjects of Rabin and Cain (1984) were exposed to 20 odors and were tested with those 20 embedded in a set of 40. They used a 10-category rating scale and were tested after intervals of 10 min, 1 day, and 7 days—each time with a different

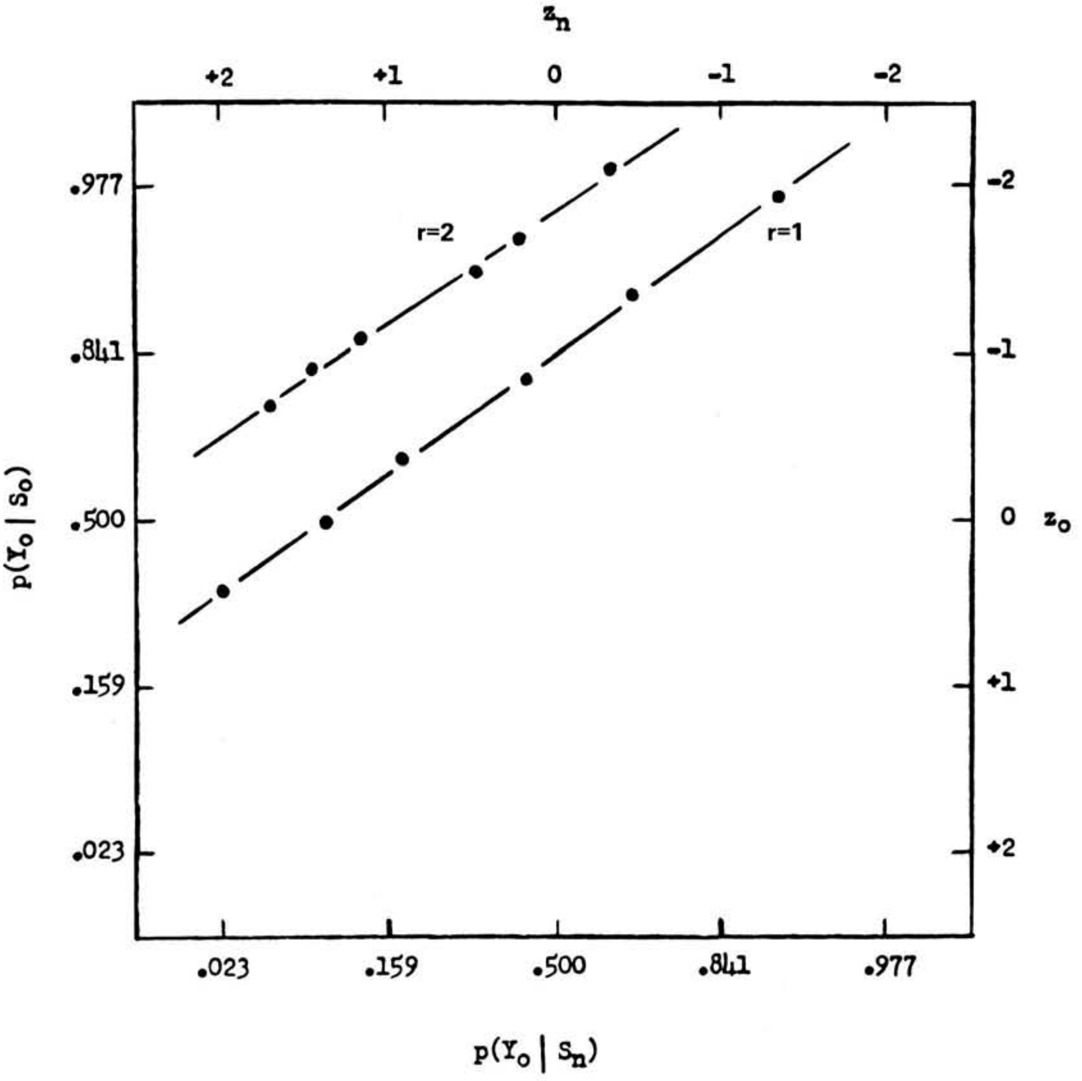

Figure 4. Two relative operating characteristics (ROCs), each based on 8 subjects, in an experiment on recognition memory for words. (r = number of repetitions of the list of "old" stimuli. S_n = presentation of a "new" stimulus. S_o = presentation of an "old" stimulus. Y_o = a "yes" response to an "old" stimulus, indicating that an old stimulus was called old. z_n = normal-deviate value corresponding to the conditional probability of a "false alarm," $P[Y_o|S_n]$. z_o = normal-deviate value corresponding to the conditional probability of a "hit," $p[Y_o|S_o]$. (From "Recognition Memory and the Operating Characteristic" by J. P. Egan, 1958, Technical Note, Indiana University, Hearing and Communication Laboratory, following p. 28. Reprinted by permission.)

set of 20 new odors. The pooled results are reproduced in Figure 5. The bottom two ROCs have slopes very near 1.0; the top ROC has a slope of about 1.1.

Animal Learning

Blough (1967) reinforced pigeons for pecking at a single wavelength and measured their (lower) rates of responding to nearby wavelengths. These response rates indicated the pigeons' certainty that the reinforced stimulus (signal) was present. Figure 6 shows quite good linear fits in the bottom two panels, for two birds, with slopes ranging mainly between 0.5 and 1.0 (with one at about 0.33).

Conceptual Judgment

Lee (1963) devised a discrimination task to make the observational continuum external to the observer, but we can view it as a cognitive task. Specifically, a dot was presented somewhere along the long dimension of a plain file card. The experimenter located two normal distributions along this dimension (not visible to the observer), randomly chose one on each trial, and then randomly chose a point from it to present to the observer. Information about the distributions was made available to the subject by means of feedback after each trial as to the source distribution of the point on that trial. In short, the subject was to build up an impression of experimental distributions similar to the picture at bottom right in Figure 1a (though Lee's distributions were of equal variance). The ROCs for two observers, based on an 8-category scale, are reproduced in Figure 7. Their slopes are about 1.05.

ROCs From Practical Fields

Considered next are illustrative ROC data from the fields of medical diagnosis (specifically, medical imaging), information

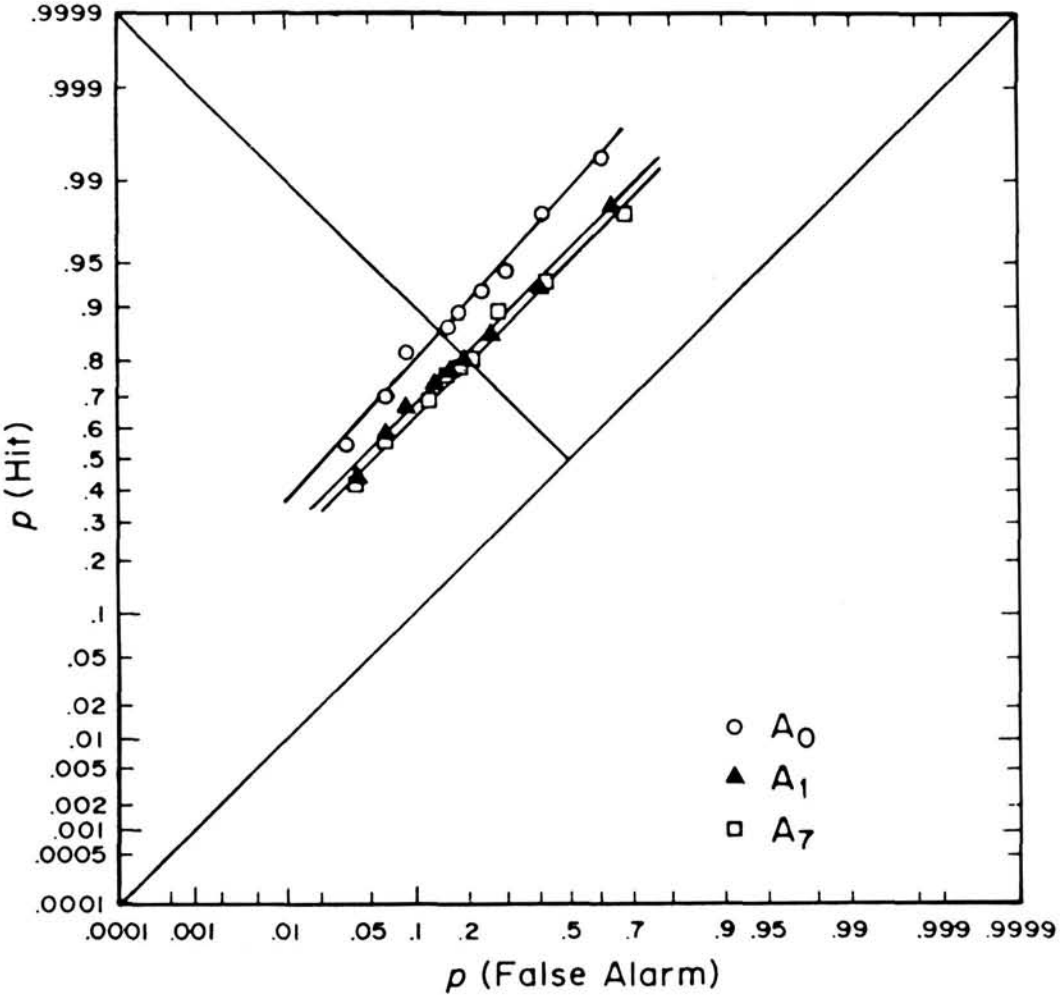

Figure 5. Three relative operating characteristics (ROCs), representing 15 subjects at three retention intervals, in an experiment on recognition memory for odors. (A_0, A_1, and A_7 = Group A, a group tested at all three intervals: 0 = 10 min; 1 = 1 day; and 7 = 7 days. From "Odor Recognition: Familiarity, Identifiability, and Encoding Consistency" by M. D. Rabin and W. S. Cain, 1984, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* p. 320. Copyright 1984 by the American Psychological Association. Reprinted by permission.)

retrieval, weather forecasting, aptitude testing, and polygraph lie detection. As in the previous section, the straight-line fit and the range of binormal slopes are of principal interest. Values of A_z are mentioned here when comparisons within a given field are of interest. An article in preparation will examine how well diagnostic systems of various kinds perform, and the relative ease and difficulty of obtaining good estimates of accuracy in different fields.

Medical Imaging

The ROCs shown in Figure 8 were obtained in an evaluation of computed tomography (CT) and radionuclide scans (RN) in the detection of brain lesions (Swets et al., 1979). One hundred thirty-six cases of patients subjected to both imaging modalities were interpreted by 12 radiologists (6 CT specialists and 6 RN specialists). The cases were selected on the basis of having adequate truth data (histological confirmation of 84 abnormal cases and 8-month follow-up of 52 normal cases) and to provide appropriate representation of lesion types and locations.

The ROCs shown were obtained via a 5-category rating scale of probability of abnormality. Each ROC is based on the pooled rating data of six radiologists. This method of combining data was defended on the grounds that the ROCs obtained from individual readers were fitted well by straight lines (chi-square

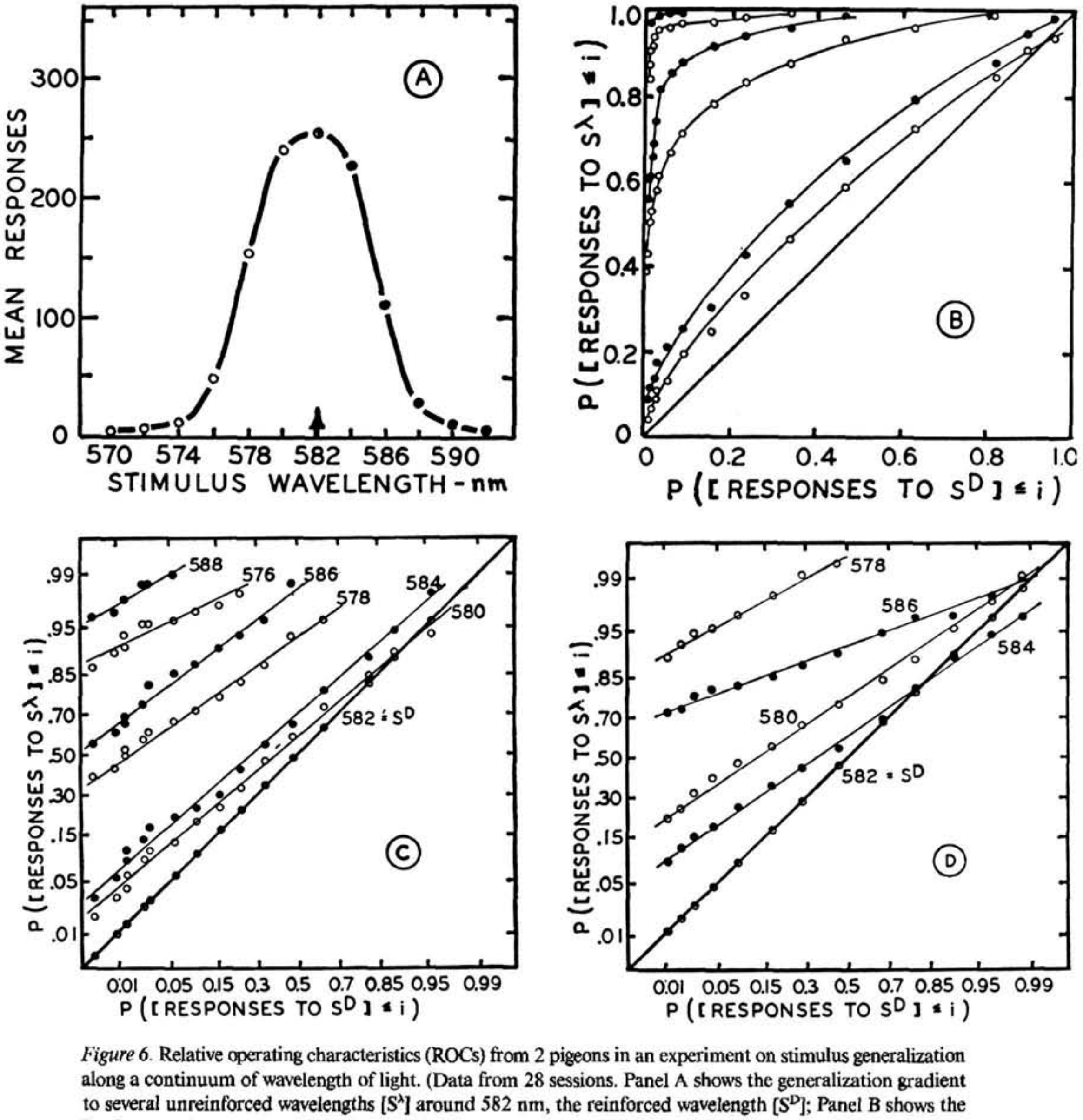

Figure 6. Relative operating characteristics (ROCs) from 2 pigeons in an experiment on stimulus generalization along a continuum of wavelength of light. (Data from 28 sessions. Panel A shows the generalization gradient to several unreinforced wavelengths [S^λ] around 582 nm, the reinforced wavelength [S^D]; Panel B shows the ROCs on ordinary scales, of the six stimuli nearest 582 nm, for 1 bird. Those ROCs are shown on a binormal graph in [C], and the ROCs of a second bird are shown in [D]. Axes in [B], [C], and [D] represent relative frequencies that a given number [*i*] of responses or fewer were made to the stimulus in question. From "Stimulus Generalization as Signal Detection in Pigeons" by D. S. Blough, 1967, *Science, 158,* p. 941. Copyright 1967 by the American Association for the Advancement of Science. Reprinted by permission.)

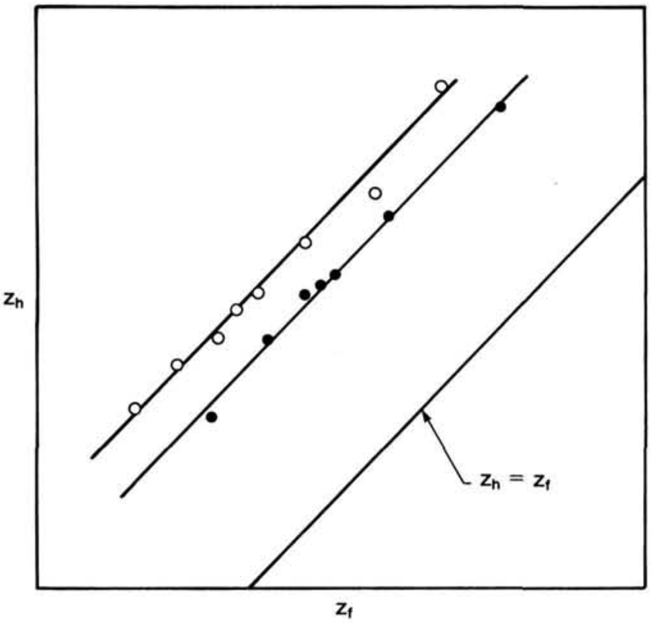

Figure 7. Relative operating characteristics (ROCs) for 2 observers in an experiment on conceptual judgment. (z_f = normal-deviate value corresponding to the conditional probability of a "false alarm." z_h = normal-deviate value corresponding to the conditional probability of a "hit." From "Choosing Among Confusably Distributed Stimuli With Specified Likelihood Ratios" by W. Lee, 1963, *Perceptual and Motor Skills, 16,* p. 230. Copyright 1963 by Southern Universities Press. Adapted by permission.)

analysis yielded $p > .20$ for 11 of the 12 readers) and showed little variation across readers in accuracy (A_z from .96 to .98 for CT and from .83 to .89 for RN). Both the pooling of rating data and the averages of A_z yielded $A_z = .97$ for CT and $A_z = .87$ for RN. Individual slopes averaged 0.70 for CT (range from 0.49 to 1.04), compared with the pooled slope of 0.61, and 0.52 for RN (range from 0.37 to 0.68), compared with the pooled slope of 0.51.

Figure 9 shows an ROC representing 10 cytotechnologists who viewed approximately 6,000 individual cell photomicrographs to discriminate between abnormal and normal cells in screening for cervical cancer (Bacus et al., 1984). True cell class was based on full case information and consensus among other cytotechnologists and pathologists. The observers classified the cells relative to 18 categories of both type and severity of abnormality; the experimenter ordered those categories according to the likelihood of abnormality to obtain the 17-point ROC shown. The relevant indices are $A_z = .87$, slope = 1.33. Computer-based, automated evaluation of the same slides—based on a multivariate Gaussian classification scheme and standard measurements of area, density, color, shape, and texture—yielded a very similar ROC, with $A_z = .90$ and slope = 1.21 (Bacus, 1982). (A review of medical studies reporting ROCs that were available at the time was given by Swets, 1979, who found values of A_z generally between .85 and .95).

Information Retrieval

Three major studies of information retrieval conducted in the mid-1960s were analyzed shortly thereafter in ROC terms (Swets, 1969). They were conducted at Harvard University (Salton & Lesk, 1966), at Cranfield, England (Cleverdon & Keen, 1966), and at Arthur D. Little, Inc. (Giuliano & Jones, 1966). The first and third studies were of computer-based systems; the second was of a traditional, manual library system. The computer system in the first study, for each query, examined every word of every document in a given collection, either of the full text or just the abstract; made associations of words in the document with words in the query by various techniques (word-stem match, synonym recognition, statistical word-word associations, etc.); and calculated the relevance of each document to the query. The ROCs were obtained by choosing various decision criteria (rating categories) on this relevance scale. Actual relevance and nonrelevance were determined by a panel of judges.

Figure 10 shows the results of six retrieval methods (as indicated in the figure) applied to one of the collections of documents used in the first study. For each method, 35 queries were directed to a collection of 82 documents. The six lines, fitted by eye, are reproduced on the full plot at the bottom of the figure. For present purposes, note that straight lines give a quite good fit (a possible staircase effect might be due to the small number of relevant documents per query) and that the slopes vary slightly from 1.0 (approximately 0.85 to 1.0). The scale along the negative diagonals in the graphs is of the accuracy index d' (which is suitable when the slope is 1.0). The six methods are seen to vary little about $d' = 1.0$, which corresponds to $A_z = .76$.

Other document collections and retrieval methods used in the three studies showed linear fits as good or better than those in Figure 10. All of them showed a similarly small effect of method within a given collection. Slopes typically range from about 1.0 to 1.3. Values of A_z range from .76 to .96 in the Harvard study, from .83 to .91 in the Cranfield study, and from .87 to .93 in the Arthur D. Little, Inc. study. (Overall, five of the six major conditions across studies produced values of A_z that are almost uniformly spread between .85 and .95.)

Weather Forecasting

Mason (1982) published some 20 ROCs for forecasting various types of weather, of which 6 are shown in Figure 11. They are all based on the probability reports generally issued by forecasters (usually in 13 categories for rain, and in smaller numbers of categories for other weather events). Whether the weather event in question actually occurred or not was determined according to procedures established in the weather forecasting field.

Figure 11a shows an ROC for rain based on some 17,000 reports at Chicago. The linear fit is very good; slope = 0.97, A_z = .85. Fits based on about 3,000 reports of individual forecasters were almost as good. Figure 11b refers to prediction of a minimum temperature < 28°F near Albuquerque; slope = 1.38, A_z = .89. Figure 11c refers to predictions of one or more tornadoes in areas delineated by the Severe Storms Forecast Center in Kansas City; based on about 90 reports, slope = 0.70, A_z = .77. (These figures are based on data published earlier by A. Murphy, who was instrumental in the move to probability forecasting, and R. L. Winkler [Murphy, 1977; Murphy & Winkler, 1977a, 1977b].) Figure 11d shows ROCs for fog-risk forecasts at the Canberra, Australia, airport issued 24, 18, or 12 hours earlier than the specified time, and based on over 300 reports. The A_z

rises from .72 to .76 to .81 as the time shortens; the corresponding slopes are 1.27, 1.22, and 1.0. (Mason's further analyses show A_z values for rain ranging across locations from .74 to .89. Predicting lightning and fog gave an A_z of about .75; predicting temperatures within intervals and tornadoes showed an A_z of about .70. So, the range over-all is approximately .70 to .90.)

Aptitude Testing

Historically, aptitude tests have predicted a continuous variable, principally graded school performance or rated job performance. In this case the product-moment correlation coefficient serves well as an index of the test's performance—that is, of its validity as a predictor—that is based on all available information. However, the predicted variable may be binary, as when students working under individually paced instruction complete the course or not, or when the rating of workers reduces to whether or not they can do the job well enough to stay on it, and then the accuracy of the test's discrimination is of interest. An accuracy index can serve as an alternate measure of validity.

To simulate such a binary outcome, I have analyzed data (kindly supplied to me by the Navy Personnel Research and Development Center) with a cut-score for success in a course of instruction set arbitrarily at the 50th percentile of the distribution of final grades. Nine ROC points were generated by taking deciles of aptitude-test scores.

Figure 12 shows the ability of the Armed Forces Qualification Test to predict such "pass-fail" performance in four Navy schools: (a) quartermaster, (b) signalman, (c) electrician's mate, and (d) mess management. The slopes vary from 0.86 to 0.96, and the values of A_z vary from .66 to .72. Based on a few hundreds of students, the linear fits are good: chi-square analysis of 12

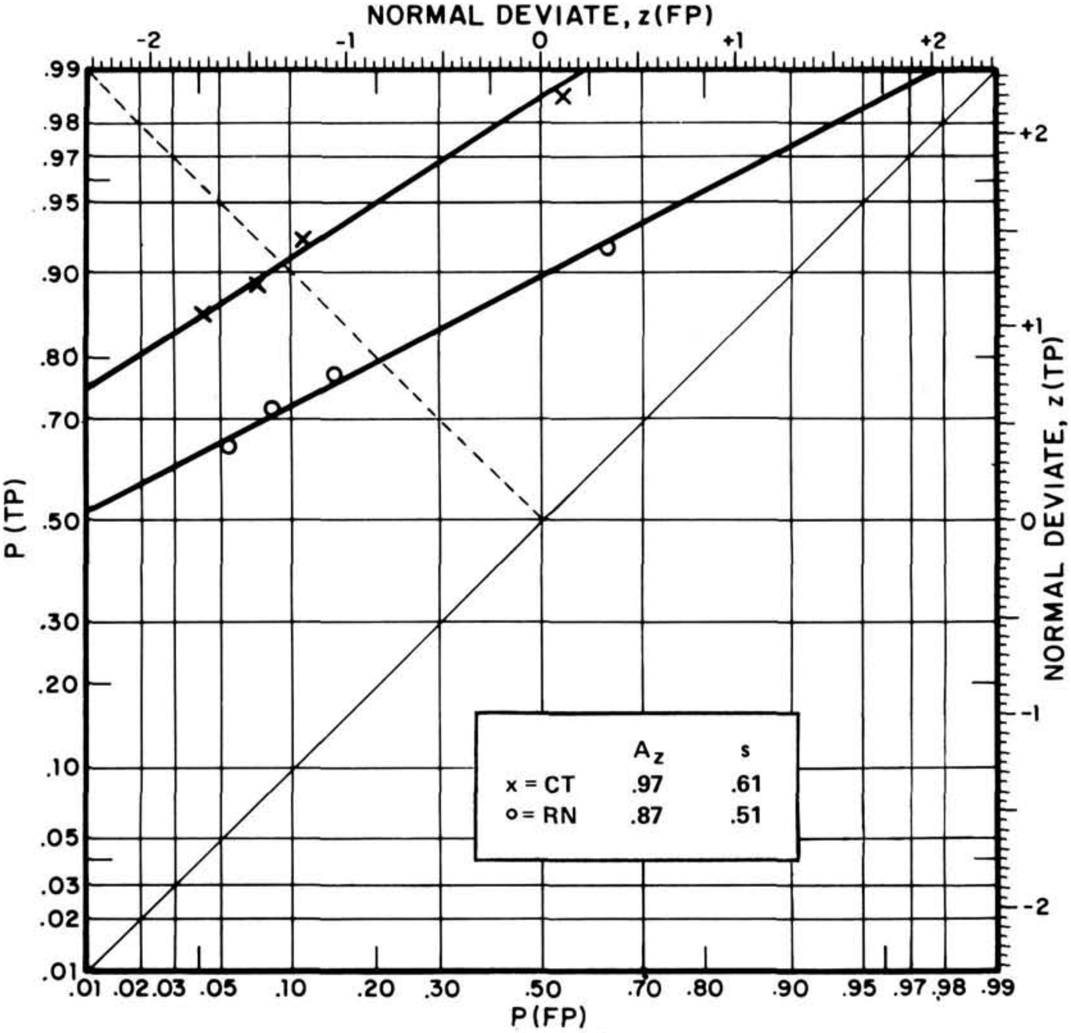

Figure 8. Relative operating characteristics (ROCs), each based on 6 observers, representing two image modalities in clinical medicine. (CT = computed tomography; RN = radionuclide scans. FP = false-positive response, or "false alarm." TP = true-positive response, or "hit." A_z = the area under the ROC plotted on ordinary scales. *s* = slope of the binormal ROC. From "Assessment of Diagnostic Technologies" by J. A. Swets, R. M. Pickett, S. F. Whitehead, D. J. Getty, J. A. Schnur, J. B. Swets, and B. A. Freeman, *Science, 205,* p. 757. Copyright 1979 by the American Association for the Advancement of Science. Reprinted by permission.)

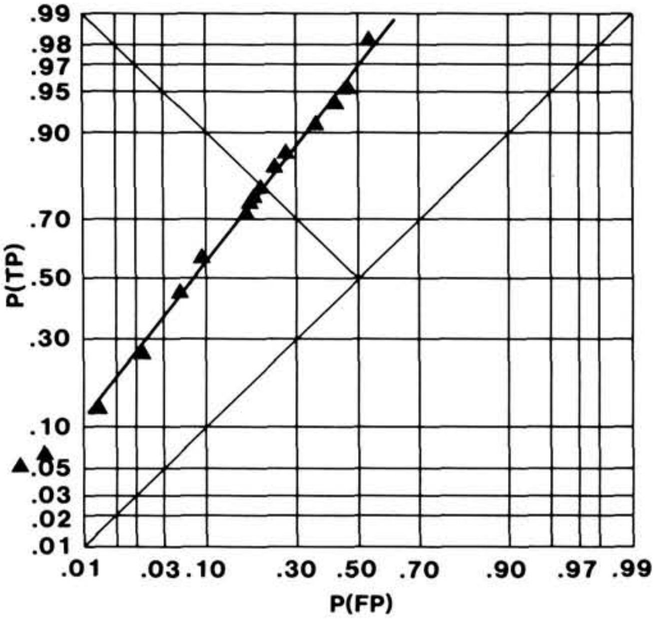

Figure 9. Relative operating characteristic (ROC) for 10 cytotechnologists screening slides for evidence of disease. (From "Malignant Cell Detection and Cervical Cancer Screening" by J. W. Bacus, E. L. Wiley, W. Galbraith, P. N. Marshall, G. D. Wilbanks, and R. S. Weinstein, 1984, *Analytical and Quantitative Cytology, 6,* p. 125. Copyright 1984 by The International Academy of Cytology. Reprinted by permission.)

linear fits (based on three different cut scores of final grades) gave probabilities ranging from .13 to .98.[2]

Polygraph Lie Detection

The open literature contains about 10 studies of the accuracy of polygraph lie detection in each of two main classes: field studies and laboratory, or analogue, studies. The former include various crimes and compare the polygraph examiners' decisions with judicial outcomes, panel decisions, or confessions. The latter are based on mock or role-playing crimes, so they have an advantage in the surety of "ground truth" and a disadvantage in the severity of the consequences. The field studies have been reviewed in the context of a detection-theory analysis by Ben-Shakhar, Lieblich, and Bar-Hillel (1982); both classes of studies were reviewed in an analysis for the federal Office of Technology Assessment (Saxe, Dougherty, & Cross, 1985).

One of the laboratory studies yielded a 7-point ROC; six of them can be analyzed (as I have) to provide 2-point ROCs, by virtue of including an "inconclusive" category along with "deception" and "no deception" categories. The field studies each yielded a single point in ROC space. Figure 13a shows the one full ROC available. In a study made by Szucko and Kleinmuntz (1981), 15 subjects carried out a mock crime and 15 did not. Their polygraph records were examined by six interpreters who judged the likelihood of deception, in response to each of three questions, on an 8-category scale. The six individual ROCs ranged in A_z from approximately .65 to .75; the pooled ROC shown has an A_z of about .75 and a slope of about 0.95. The straight-line fit is not good at the ends, which might result from the combination of a relatively low A_z and a small number of observations (45). Though, of course, not evidence for linear binormal ROCs, Figure 13b shows 2-point ROCs from six other laboratory studies. For them, A_z ranges from about .80 to .95 and the slopes range (with one exception) from about 0.75 to 1.3.

Possible Pattern of Slope Variation

My concern for empirical (binormal) ROC slopes has been to show that they vary considerably, enough to introduce considerable unreliability if they are assumed, by using a given index, to be fixed at some particular value. The question may then arise whether the observed slope variation has some pattern and whether some account of that pattern can be given. I believe that a partial pattern can be discerned and that some relevant theory exists, and so a few remarks on the topic may be heuristic.

In the original development of the signal detection theory that gave rise to the ROC (Peterson, Birdsall, & Fox, 1954), it was assumed that a signal is added at times to background noise to produce a distribution of observations of signal plus noise having a higher mean value than the distribution of observations of noise alone. Then, if the signal adds a constant, the two distributions would have the same variance. Given that the binormal ROC slope under the normal (Gaussian) model equals the ratio of standard deviations, σ_n/σ_{sn}, the ROC slope would be 1.0 in this case. The signal was assumed in theory to add a constant to the noise when all of its appropriate parameters—such as frequency, phase, amplitude, starting time, and duration—are fixed and known to the observer; this signal is "specified exactly" in the theory's terms.

If, on the other hand, one or more of the signal's parameters varies at random from one occurrence to another, the signal is only "specified statistically," and adds additional variance to that of the noise. Distributions other than the normal (e.g., exponential, Rayleigh) can be adduced to treat this case, or the normal model can be retained with the assumption that $\sigma_{sn} > \sigma_n$ and, therefore, that the ROC slope is <1.0 (Green & Swets, 1966/1974). Consistent with such a theory, early work on human sensory processes was taken to indicate slopes near 1.0 for pure-tone signals, specifiable as sine waves, and slopes <1.0 for auditory signals that are samples of white noise, visual signals that are flashes of white light, and other signals in which certain parameters (e.g., frequency or phase) are not specified exactly. The

[2] I should note that these values of A_z are depressed, as raw correlation coefficients on the same sample would be, because they are computed only on the candidates selected to take the course. In correlational terms, the effect is one of "restriction of range" of the two variables. Similarly, in discrimination terms, the spectrum of abilities considered is restricted by the lack of course data on unselected candidates, so that discrimination within the sample considered is more difficult. A correction for restriction of range is usually made for correlation coefficients, and a larger study of Navy data of the sort described here showed median uncorrected and corrected coefficients of .43 and .73, respectively (Swanson, 1979). I have not addressed the question of a similar correction for the A_z index, but Gray, Begg, and Greenes (1984) have provided an approach to the problem.

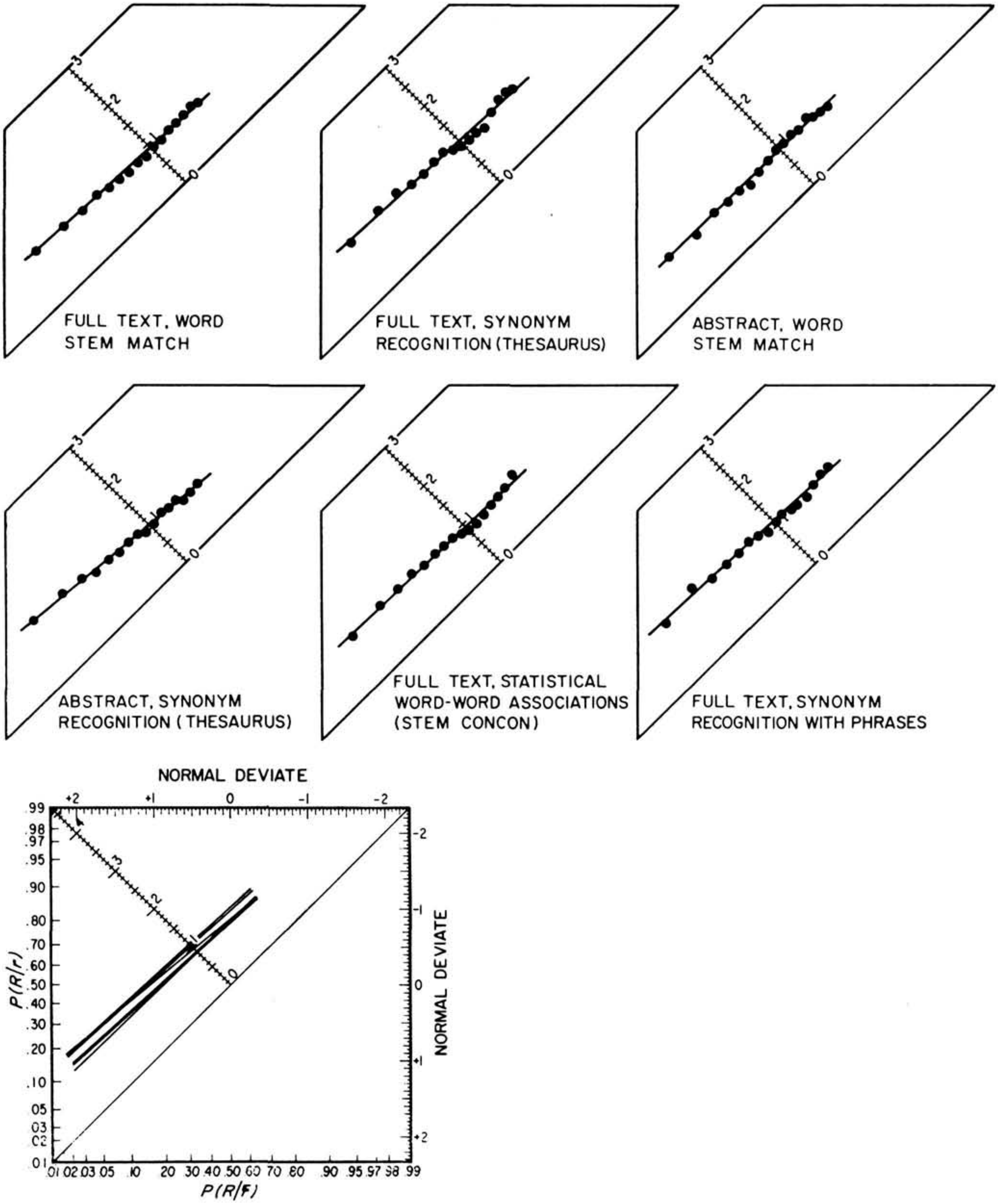

Figure 10. Relative operating characteristics (ROCs) representing six methods of information retrieval. (The 6 fitted lines of the upper panels are also shown in the bottom panel. *F* = the presentation of an irrelevant document. R = the response of "relevant." r = the presentation of an irrelevant document. From "Effectiveness of Information Retrieval Methods," by J. A. Swets, 1969, *American Documentation, 20,* p. 79. Copyright 1969 by John Wiley and Sons, Inc. Reprinted by permission.)

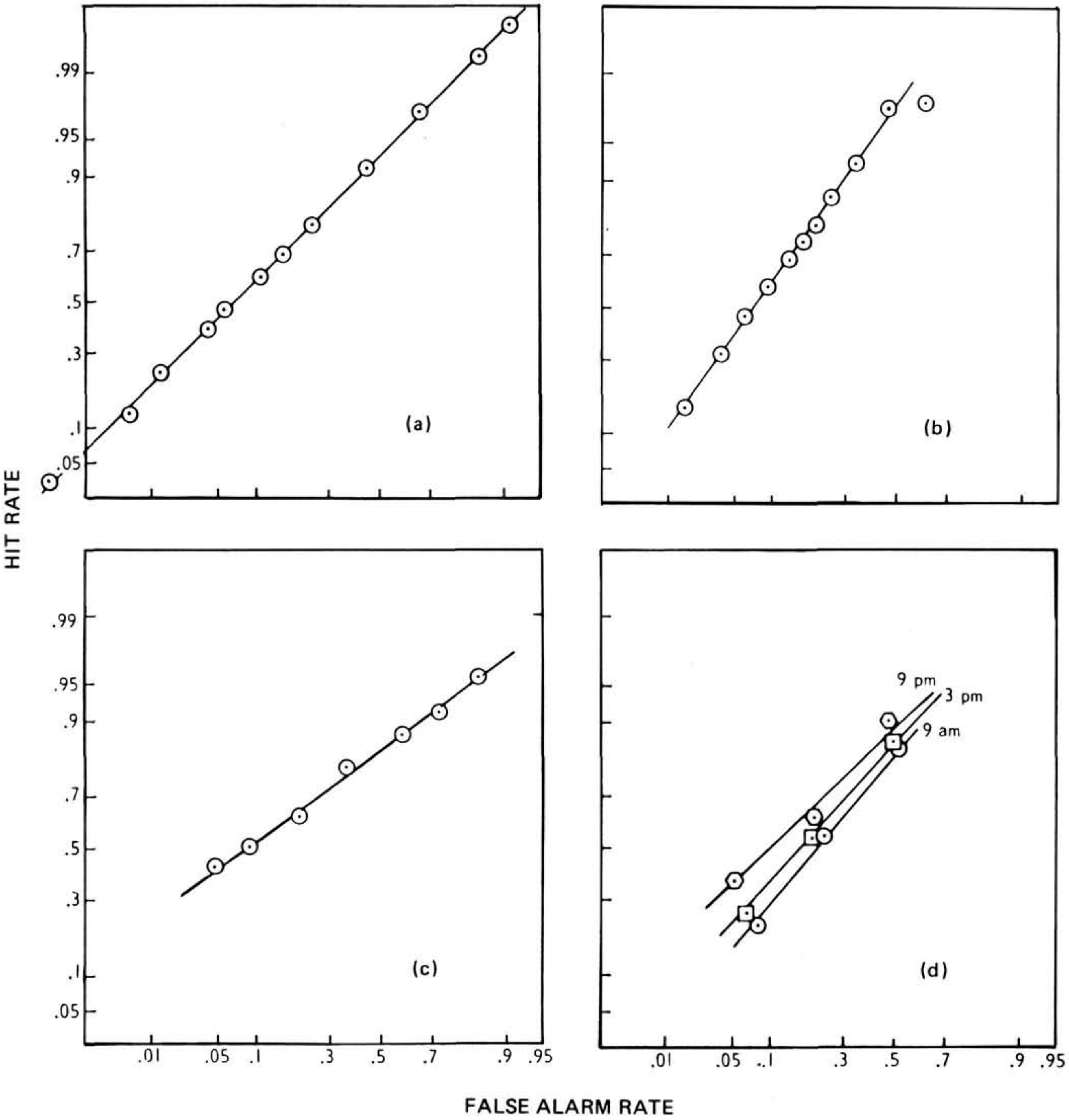

Figure 11. Relative operating characteristics (ROCs) for weather forecasting. (a: rain; b: minimum temperature; c: tornadoes; and d: fog. From "A Model for Assessment of Weather Forecasts" by I. Mason, 1982, *Australian Meteorological Magazine, 30,* pp. 296, 297, 299. Copyright 1982 by the Australian Government Publishing Service. Reprinted by permission.)

ROCs shown above for white visual signals have slopes mostly around 0.70.[3]

The first binormal slopes noted as quite consistently >1.0 are the ones for information retrieval, which range mostly from 1.0 to 1.3, as mentioned above. This result seems reasonable, inasmuch as the "signals" are then relevant documents and these signals are not in any sense added to "noise," or irrelevant documents. In this case, two separate kinds of stimuli, with some

[3] The idea that uncertainty about the signal adds variance to the *sn* distribution, beyond that of the *n* distribution, might suggest that binormal slopes < 1.0 will usually be observed whenever a brief signal is added to

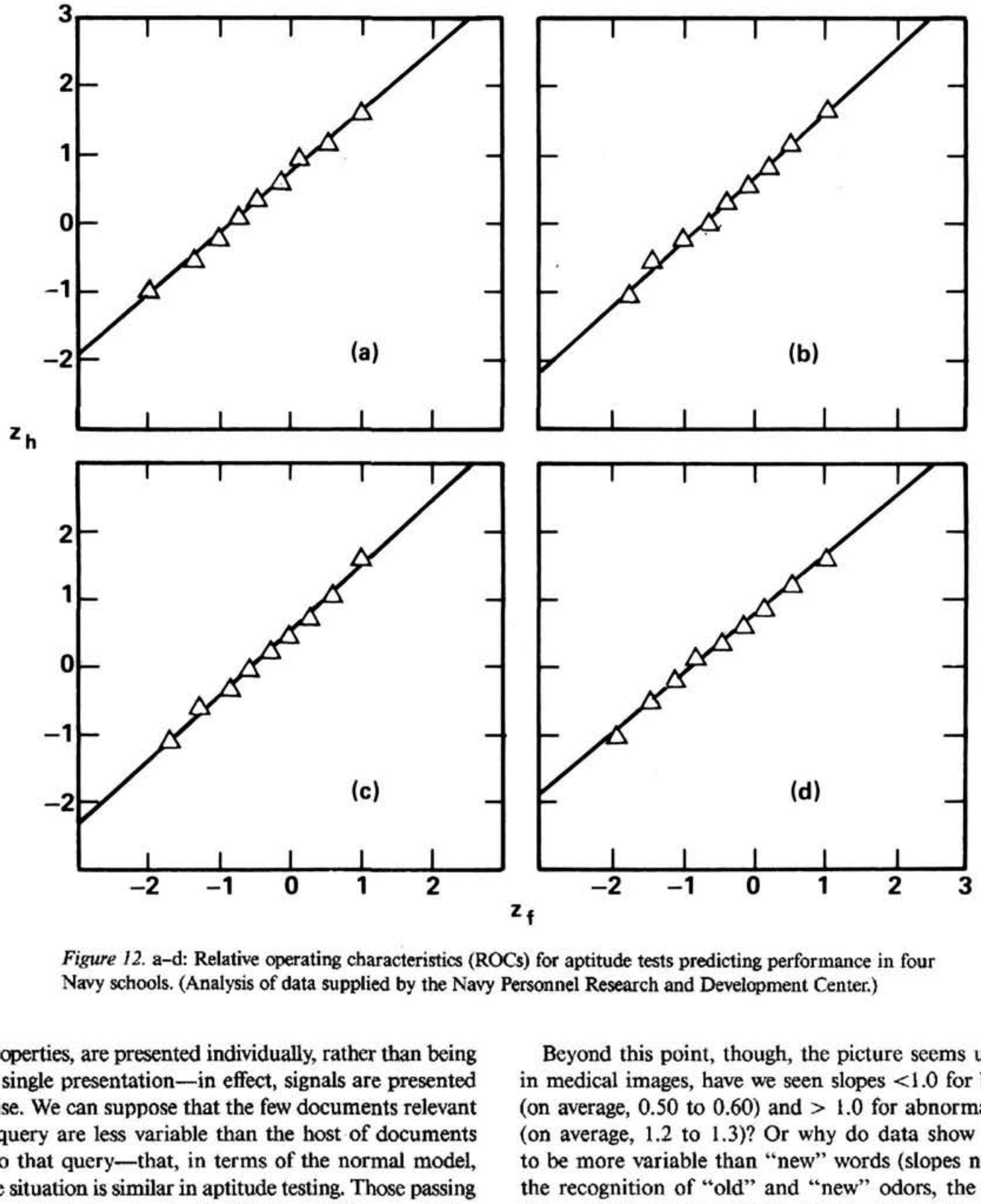

Figure 12. a–d: Relative operating characteristics (ROCs) for aptitude tests predicting performance in four Navy schools. (Analysis of data supplied by the Navy Personnel Research and Development Center.)

common properties, are presented individually, rather than being mixed in a single presentation—in effect, signals are presented without noise. We can suppose that the few documents relevant to a given query are less variable than the host of documents irrelevant to that query—that, in terms of the normal model, $\sigma_s < \sigma_n$. The situation is similar in aptitude testing. Those passing courses are not added to, or mixed with, those failing courses in a single presentation; the slopes <1.0 for aptitude tests are consistent with the finding that those passing are more variable (considering here only individuals meeting the test's criterion score).

Beyond this point, though, the picture seems unclear. Why, in medical images, have we seen slopes <1.0 for brain tumors (on average, 0.50 to 0.60) and > 1.0 for abnormal tissue cells (on average, 1.2 to 1.3)? Or why do data show "old" words to be more variable than "new" words (slopes near 0.70)? In the recognition of "old" and "new" odors, the present data show nearly equal variance (slopes near 1.0). Again, some weather events have shown slopes definitely <1.0, and others, slopes clearly >1.0.

In the net, several puzzles remain, and support at least the present use of an area index such as A_z. And, indeed, evidence suggests that different individuals can produce ROCs of different slopes under the same signal and noise conditions (e.g., Green & Swets, 1966/1974).[4]

a continuous noise background, because the observer will usually have some uncertainty about at least the location of the signal in time or space. A related suggestion would be that slopes near 1.0 will result when the two alternatives to be discriminated are alike with respect to most of their physical characteristics (e.g., both are brief tones, or lights, or samples of noise, and vary along a single dimension, such as frequency or amplitude). D. R. J. Laming and A. Craig (personal communications, 1985) have advanced this distinction as providing a good summary of existing data from sensory tasks.

[4] The present approach assumes that every point on an empirical ROC represents an observer or system operating at a constant accuracy, but there may be cases in which that assumption does not hold. For example, a human observer in a "yes–no" task might discriminate with greater accuracy at a lenient decision criterion, if that criterion is adopted because

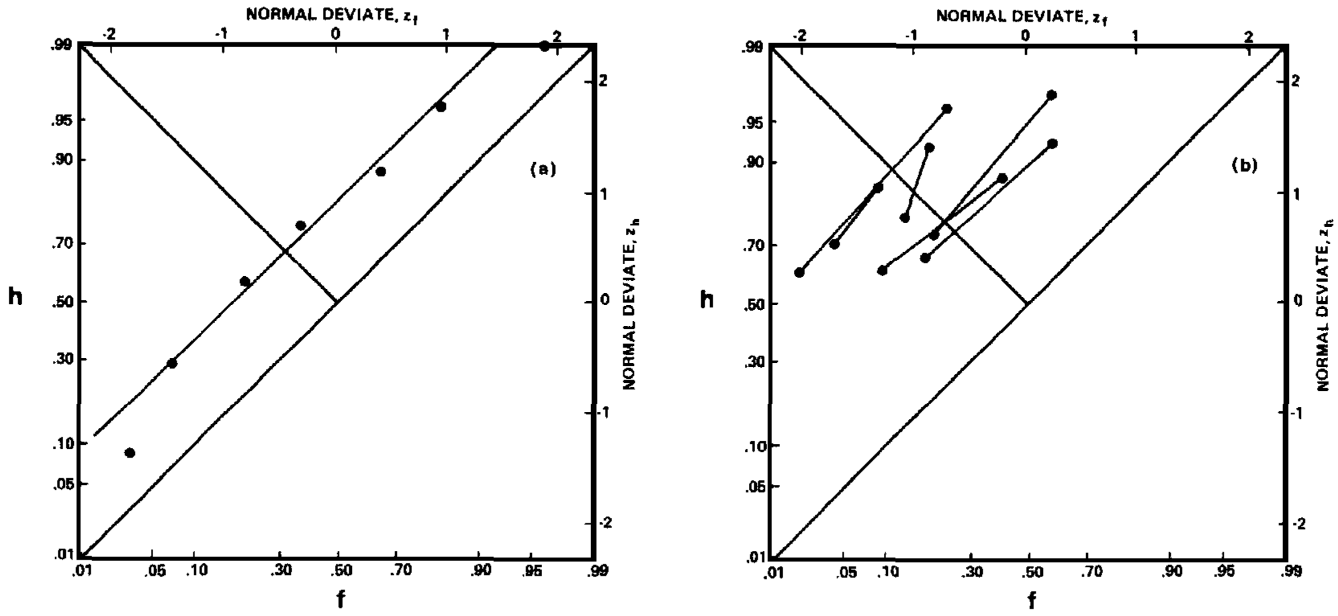

Figure 13. Relative operating characteristics (ROCs) for polygraph lie detection. a: for 6 examiners in a laboratory study. (h = probability of a "hit"; f = probability of a false alarm. Based on data from an unpublished doctoral thesis by J. J. Szucko summarized by Szucko & Kleinmuntz, 1981.) b: 2-point ROCs, derived from experiments using positive, negative, and inconclusive categories, conducted in 6 laboratories. (Based on data summarized by Saxe, Dougherty, & Cross, 1985.)

Conclusions

Empirical ROCs drawn from experimental psychology and several practical fields, representing available discrimination data, are fitted well on a binormal graph by straight lines of varying slope. This robust finding supports the use of the accuracy index A_z. It also supports the validity of a particular variable-criterion model of the discrimination process, one incorporating distributions of unequal variance. The indices d', η, LOR, and Q imply binormal ROCs that are linear or nearly linear but with a fixed slope = 1.0, and hence they do not agree sufficiently well with the data. By the same token, variable-criterion models that assume equal-variance distributions, which have been associated with d' and η and can be associated with LOR and Q, are of limited validity and utility. Several other common indices—including the chance-corrected hit probabilities, H_C and H'_C; percentage correct, PC; the kappa statistic, K; and the correlation coefficient, ϕ—imply binormal ROCs that are distinctly curvilinear and diverge considerably from the data. Therefore, their corresponding models, which are members of the class of threshold models, are invalid. The unreliability of the various indices that misrepresent empirical ROC form can be substantial: Index values can vary from low to high, by > 100%, when, in fact, accuracy is constant (Swets, 1986).

Discussion

Signal detection theory was originally developed as a mathematical theory for the process of detecting radar signals (Peterson et al., 1954), but was soon found useful in understanding the behavior of human observers of simple visual and auditory signals (Tanner & Swets, 1954; Tanner, Swets, & Green, 1956). The general applicability of the theory to human discrimination was indicated by its ability to treat empirical findings in recognition memory (Egan, 1958). Its applicability to humans, and to devices that aid or supplant humans, in practical discrimination or diagnostic tasks was suggested by analyses of information-retrieval systems (Swets, 1963, 1969). The common denominator in these tasks is, first, an observation process that lends varying degrees of assurance about the occurrence of the alternatives to be discriminated and, second, a desire to assign those varying degrees to one or the other alternative in some reasonable way.

In experimental psychology, the process of discrimination is of interest in its own right: Is it governed by a fixed, physiologically determined threshold or is it adaptive, via a variable decision criterion, to different conditions of expectancy and motivation?

the prior probability of a "signal" is high, than at a strict criterion, if that criterion is adopted because the signal probability is low; a possible mechanism for this effect is that the signal is better defined when presented relatively often (Markowitz & Swets, 1967). Knowing when the assumption of constancy holds and when it does not requires a knowledge of the distributions underlying the ROC. Thus, knowledge that the distributions are of equal variance would indicate varying accuracy when a binormal slope ≠ 1 is obtained. Unfortunately, evidence about the distributions tends to exist—for example, in machine-based systems for information retrieval or medical diagnosis, or in aptitude testing—when there is little reason to doubt the assumption of constant accuracy. In general, the working assumption of constant accuracy seems preferable to using an index that leads to the conclusion of inconstancy for all binormal ROC slopes other than the single slope that it implies. Moreover, the rating-scale technique is relatively insensitive to variables that might affect accuracy differently at different decision criteria—for example, prior probability of signal, and rewards and penalties for correct and incorrect responses.

Consistent support for the latter process unifies psychological conceptions of a broad range of behaviors and indicates the extensive role of cognitive factors in discrimination tasks. Also, in experimental psychology, the way in which discrimination acuity or capacity varies with independent variables can reveal something substantive about the nature of the particular mechanism of discrimination—be it perceptual, memorial, or cognitive. In practical fields, more emphasis is placed on the absolute level of discrimination acuity that is evidenced, which is of substantive interest in decisions about using, and attempting to improve on, a given technique or system for diagnosis (Swets & Pickett, 1982).

The bonus that the appropriate detection-theory model carries along is the ability it provides, via the ROC, to obtain a relatively pure index of discrimination capacity—one largely independent of the decision criterion or choice tendency—and also an index of the decision criterion that is operative in any given instance. Experimental psychology and practical fields thereby gain a valid and reliable index of discrimination capacity. Psychology, especially, acquires an ability to determine whether various variables that effect a change in performance do so by affecting discrimination acuity or the decision criterion (Swets, 1973). An example here is the finding that the declining hit rate observed in perceptual vigilance experiments is often the result of an increasingly strict criterion rather than of decreasing sensitivity (Parasuraman, 1984).

Practical fields need a criterion-free index of discrimination capacity when the criterion used with a given system varies widely over the different settings in which that system is used, and for which it is being evaluated. Thus, for example, the strictness of the criterion used with a particular imaging system in clinical medicine can be quite different in screening and referral settings, and the criterion used with a weather forecasting system will differ from one geographical region to another and from one user of forecast information to another.

Practical fields, moreover, acquire an ability from the ROC to assess the efficacy of a diagnostic system for a specific setting. In a given setting, one is fundamentally concerned with some measure of the system's utility, for example, its expected value or payoff, as determined by the probabilities of the various outcomes of the decision and by the benefits and costs of those outcomes. For specific settings in which the probabilities, benefits, and costs are stable and can be estimated, the emphasis is more on the payoff associated with a particular point on the ROC—that is, with a particular decision criterion—than on an index of the locus of all ROC points. For any of several decision rules that seek to maximize one or another quantity related to utility, one can calculate the optimal decision criterion, or operating point on the ROC (Green & Swets, 1966/1974; Swets & Pickett, 1982; Swets & Swets, 1979). And then, usually, the system can be adjusted to operate at or near that criterion or point. Because the binormal slopes of empirical ROCs vary widely from one instance to another, in a manner so far not predictable, both the calculation of, and adjustment to, the optimal criterion depend on having the empirical ROC in hand.

References

Bacus, J. W. (1982). *Application of digital image processing techniques to cytology automation* (Tech. Rep.) Chicago: Rush Presbyterian-St. Luke's Medical Center, Medical Automation Research Unit.

Bacus, J. W., Wiley, E. L., Galbraith, W., Marshall, P. N., Wilbanks, G. D., & Weinstein, R. S. (1984). Malignant cell detection and cervical cancer screening. *Analytical and Quantitative Cytology, 6,* 121–130.

Ben-Shakhar, G., Lieblich, I., & Bar-Hillel, Y. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. *Journal of Applied Psychology, 67,* 701–713.

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

Blough, D. S. (1967). Stimulus generalization as signal detection in pigeons. *Science, 158,* 940–941.

Cleverdon, C., & Keen, M. (1966). *Factors determining the performance of indexing systems: Test results* (Vol. 2). Cranfield, England: Association of Special Libraries and Information Bureaux.

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.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Technical Note). Indianapolis: Indiana University, Hearing and Communication Laboratory.

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

Giuliano, V. E., & Jones, P. E. (1966). *Study and test of a methodology for laboratory evaluation of message retrieval systems* (Interim Rep. No. ESD-TR-66-405). Cambridge, MA: Arthur D. Little.

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

Gray, R., Begg, C. B., & Greenes, R. A. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making, 4,* 151–164.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley. (Reprinted, 1974, Huntington, NY: Krieger)

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

Lee, W. (1963). Choosing among confusably distributed stimuli with specified likelihood ratios. *Perceptual and Motor Skills, 16,* 445–467.

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.

Markowitz, J., & Swets, J. A. (1967). Factors affecting the slope of empirical ROC curves: Comparison of binary and rating response. *Perception & Psychophysics, 2,* 91–100.

Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine, 30,* 291–303.

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

Murphy, A. H., & Winkler, R. L. (1977a). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society, 26c,* 41–47.

Murphy, A. H., & Winkler, R. L. (1977b). Probabilistic tornado forecasts: Some experimental results. *Preprints, Tenth Conference on Severe Local Storms* (pp. 403–409). Omaha, NE: American Meterological Society.

Parasuraman, R. (1984). Sustained attention in detection and discrimination. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 243–272). New York: Academic Press.

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.

Rabin, M. D., & Cain, W. S. (1984). Odor recognition: Familiarity, identifiability, and encoding consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 316–325.

Salton, G., & Lesk, M. (1966). *Information storage and retrieval* (Scientific Rep. No. 11). Ithaca, NY: Cornell University, Department of Computer Science.

Saxe, L., Dougherty, D., & Cross, T. (1985). The validity of polygraph testing. *American Psychologist, 40,* 355–366.

Swanson, L. (1979). Armed services vocational aptitude battery, Forms 6 and 7: Validation against school performance in Navy enlisted schools (July 1976–February 1978) (Tech. Rep. No. 80-1). San Diego, CA: Navy Personnel Research and Development Center.

Swets, J. A. (1963). Information retrieval systems. *Science, 141,* 245–250.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation, 20,* 72–89.

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

Swets, J. A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology, 14,* 109–121.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117.

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., & Swets, J. B. (1979). ROC approach to cost-benefit analysis. *IEEE Proceedings of the Sixth Conference on Computer Applications in Radiology* (pp. 203–206). Newport Beach, California. (Reprinted in K. L. Ripley & A. Murray [Eds.], *Introduction to automated arrhythmia detection* [pp. 57–60]. New York: IEEE Computer Society Press, 1980)

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1955). *The evidence for a decision-making theory of visual detection* (Tech. Rep. No. 40). Ann Arbor: University of Michigan, Electronic Defense Group.

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

Szucko, J. J., & Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist, 36,* 488–496.

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

Tanner, W. P., Jr., Swets, J. A., & Green, D. M. (1956). *Some general properties of the hearing mechanism* (Tech. Rep. No. 30). Ann Arbor: University of Michigan, Electronic Defense Group.

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

Received May 17, 1985
Revision received August 23, 1985 ■