12. R. O. Brown, Jr., P. L. Ward, G. Plafker, *Geological and Seismological Aspects of the Managua, Nicaragua, Earthquakes of December 23, 1972* (U.S. Geological Survey professional paper No. 838, Government Printing Office, Washington, D.C., 1973).
13. Committee on the Alaska Earthquake, Eds., *The Great Alaska Earthquake of 1964: Human Ecology* (National Academy of Sciences, Washington, D.C., 1970).
14. Such figures, based on census reports and housing count, would tend to undercount the poor.
15. *New York Times* (4 May 1973), p. 12.
16. G. W. Baker and D. W. Chapman, Eds., *Man and Society in Disaster* (Basic Books, New York, 1962); W. Form and S. Rosow, *Community in Disaster* (Harper & Row, New York, 1958); R. R. Dynes, J. E. Haas, E. L. Quarantelli, "Some preliminary observations on organizational responses in the emergency period after the Niigata, Japan, earthquake of June 16, 1964," Research report No. 11, Disaster Research Center, Ohio State University, December 1964; J. E. Haas and R. S. Ayre, *The Western Sicily Earthquake Disaster of 1968* (National Academy of Engineering, Washington, D.C., 1969); A. H. Barton, *Communities in Disaster* (Doubleday, New York, 1969); R. R. Dynes, *Organized Behavior in Disasters* (Heath, Lexington, Mass., 1970).
17. D. Yutzy and J. E. Haas, in *The Great Alaska Earthquake of 1964: Human Ecology*, Committee on the Alaska Earthquake, Eds. (National Academy of Sciences, Washington, D.C., 1970), pp. 90–95.
18. These and related issues are now being investigated with financial support provided by grant GI-39246 from the National Science Foundation.
19. National Oceanic and Atmospheric Administration, *A Study of Earthquake Losses in the San Francisco Bay Area: Data and Analysis* (report prepared for the Office of Emergency Preparedness, Washington, D.C., 1972). The damage estimates refer only to residential structures. Other estimates place the *total* loss at $11 to $25 billion.
20. State of California, Senate Bill No. 519, 21 November 1972.
21. Source of data: "Evaluación preliminar de daños a consequencia del terremoto de Managua—23 Deciembre 1972," emergency report prepared by a task force of persons in private enterprise and authorized by the Comité Nacional de Reconstrucción Económica.
22. Sources: author's evaluations of data in (*21*); Mercali intensity summarized from S. T. Algermissen, J. W. Dewey, C. Langer, W. Dillinger, "Managua, Nicaragua, earthquake of December 23, 1972: Location, focal mechanism, and intensity distribution," paper presented at the annual meeting of the Seismological Society of America, Golden, Colorado, 16 May 1973.
23. Sources: "San Fernando earthquake, February 9, 1971," report of the Los Angeles County Earthquake Committee, 1971; "The San Fernando earthquake of February 9, 1971, and public policy," report of the Special Subcommittee of the Joint Committee on Seismic Safety, California Legislature, 1972.
24. The following persons provided significant assistance to our research activity: Ing. Carl Ahlers, Lic. William Baéz, Fundación Nicaragüense de Desarrollo; George Baker, National Science Foundation; Ernest Barbour, U.S. Agency for International Development; Gary Bergholdt, Instituto Centroamerica de Administración de Empresas; Carlos H. Canales, Ministry of Health and Hospitals; Edgar Chamarro C.; Arq. Eduardo Chamarro C.; Ing. Filadelfo Chamarro C.; William Dalton, U.S. Agency for International Development; Orlando Espinosa B., Ministry of Labor; Ing. Alfonso Guerrero, Empresa Nacional de Luz y Fuerza; Pdr. Ramiro Guerrero, University of Central America; Janice Hutton, University of Colorado; Verona Norton; Cap. Ortegaray, La Guardia Nacional; Doña María Elena de Porras, Emergency Relief Committee; Carlos Ramón Romero, Ministry of Health and Hospitals; Ing. Cristóbal Rugama Nuñez, Ministry of Public Works; Renée Spinosa, Caritas; Harry Strachen, Instituto Centroamerica de Administración de Empresas. The authors alone are responsible for any omissions or errors in fact and interpretation.

The Relative Operating Characteristic in Psychology

A technique for isolating effects of response bias finds wide use in the study of perception and cognition.

John A. Swets

Psychological measurements of an individual's ability to make fine discriminations are often plagued by biasing factors that enter as he translates his covert discrimination into an overt report about it.

Reliable, valid measures are desired of an individual's ability to make a great variety of sensory discriminations, along dimensions such as brightness, hue, loudness, pitch, and the intensive and various qualitative attributes of taste and smell and touch. Sometimes the focus is on the organism's capacity for discrimination, as when the functioning of the sense organs is under study. At other times, interest centers upon the discriminability of the alternatives, as when the measures are used in the development of a product such as color film or tea.

Also sought are accurate measures of more complex perceptual discriminations. How well do individuals judge relative size, distance, direction, time, and motion? How noticeable is a given road sign, and how distinguishable are the signs that employ different combinations of shape, color, and notation to convey different meanings?

Further, it is important to develop unbiased measures of cognitive discriminations, such as those related to memory and conceptual judgment. Psychologists ask people to distinguish objects they have seen before from objects they have not, perhaps nonsense syllables or advertisements; to tell from an article's title, descriptors, or abstract whether it is relevant or irrelevant to a particular need for scientific information; to say whether a given opinion is representative of source A or of source B; and so on.

The translation of covert discrimination into overt report is not direct and simple, according to psychological theory, either because the output of the discrimination process is not definite or because judgmental considerations can override that output. In any case, an inherent ambiguity makes an individual's report prone to influence by such factors as his expectations and motivations or, more specifically, by such factors as probabilities and utilities. Thus: The immediate evidence may favor alternative A, but alternative B is more probable on the whole, so I'll more likely be correct if I report B. Again: The evidence may favor A, but the penalty for incorrectly reporting A is relatively large (or the reward for correctly reporting B is relatively large), so I'd be wise to report B.

That probabilities and utilities influence outcomes of the important discriminations people are called upon to make is perfectly clear—as when the clinician reads an x-ray, when the pilot emerges from a low ceiling, or when the Food and Drug administrator suspects that a product is harmful. Less

The author is senior vice president of Bolt Beranek and Newman Inc., 50 Moulton Street, Cambridge, Massachusetts 02138.

clear, perhaps, is that these and similar biasing factors can play a large role in any discrimination problem, even those problems posed in the rarified atmosphere of the laboratory. A laboratory subject may have unrealistic notions about the prior probabilities of the alternatives presented to him, or about the sequential probabilities in random sequences of alternatives. One subject may not mind failing to notice a very small difference and feel foolish asserting a difference when there is none, while another subject may strive to detect the smallest possible difference and accept errors of commission as simply part of the game.

One bothersome effect of the biasing factors, of course, is the variability they introduce. When biases vary out of control, then measurements vary for no apparent reason—from one subject to the next, from one day to the next, from one laboratory to the next. Worse, however, is the potential that biases contribute for misinterpretation. As I shall show, the effects of biasing factors on the report have often been viewed as properties of the discrimination process, with the result that incorrect conclusions have been drawn about the nature of perception and cognition and have been held for long periods.

Psychologists have sought for more than a century to devise measurement procedures that minimize the extent of bias, and, indeed, one procedure more than a century old is largely successful in this respect. The procedure is to present two alternatives at a time, with the assurance that one is A and the other is B, and to ask which is which. Under this scheme, probabilities and utilities are essentially symmetrical. However, it is often desirable, and sometimes necessary, to present just one alternative at a time. In these cases, one must let the biases play and then try to remove their effects by later analysis.

An analytical technique developed in recent years does the trick fairly well. It distills and quantifies, collectively, the various factors that bias an individual's report and leaves a relatively pure measure of discrimination. It amounts, quite simply, to plotting the data in the form of what I call here the relative operating characteristic (ROC). The ROC is a curve whose overall location corresponds to a particular degree of discrimination, while the position of any point along the curve represents a particular degree of bias in the report. This provision for two independent measures contrasts the ROC with measurement techniques available earlier, in which discrimination and report bias are confounded in a single free parameter.

The ROC originated in the concept of the operating characteristic, as developed in the statistics of testing hypotheses. This concept was refined—transformed into the relative operating characteristic—in the context of electronic signal detection. I shall trace this development, but I shall first consider how psychology arrived at the point of being ready to accept the ROC. This entails mainly a consideration of psychophysics, a discipline whose beginnings laid the foundation of experimental and quantitative psychology.

After placing the ROC in historical perspective, I describe how to work with it. In speaking to certain practical questions, such as estimation procedures, I also try to indicate ways in which the ROC analysis falls short of perfectly accomplishing the tasks set for it. Last, I review a broad range of applications in psychology, emphasizing those outside of psychophysics, which were accomplished largely during the last 5 years.

Psychophysics

One of the first people to tackle the problem of obtaining precise measures of discrimination was Gustav Theodor Fechner (1801 to 1887). Though a physicist, physiologist, doctor of medicine, poet, aestheticist, and satirist through 70 years of intellectual productivity, he aspired most consistently to be a philosopher. His aim was to overthrow materialism, and he conceived psychophysics to help in this aim by showing empirically the relationship between mind and body. He developed psychophysics as the measurement of attributes of sensation (intensity, quality, duration, extent) and the correlation of these measurements with physical measurements of the stimuli. His *Elemente der Psychophysik* was published in 1860 and translated into English to celebrate its centennial (*1*).

Fechner brought together and further developed what are still the basic psychophysical methods. He used them to measure both the just noticeable difference between two stimuli (otherwise called the difference limen or difference threshold) and the stimulus just noticeable (the absolute limen or threshold). Whereas he sought to obtain in this way the unit and the natural origin of the psychological continuum, I shall not be concerned here with scaling stimuli, but with absolute and difference measures as they are useful individually, depending on the problem. In current terminology, the absolute case is called "detection" and the difference case is called "recognition"; detection is the special case of recognition where one of the two stimuli to be discriminated is the null stimulus.

As Fechner put it, the first problem psychophysical methods confront is "the great variability of sensitivity due to individual differences, time, and innumerable internal and external conditions" (*1*, p. 44). As a matter of course, the methods do so by replication. Each stimulus value may be presented to each subject hundreds of times in order to obtain a relatively stable estimate of the proportion of positive responses. By a positive response, I mean either "yes, I recognize stimulus A [as opposed to B]," in the single-stimulus, or yes-no, forms of the various methods; or "stimulus A is greater than stimulus B," in what have come to be called the paired-comparison, or forced-choice, forms.

Fechner plotted the proportion of positive responses against a physical measure of stimulus strength or stimulus difference to get the psychometric function. This function ordinarily takes the form of an ogive, as shown in Fig. 1a. This form is consistent with a constant sensory effect of a given stimulus and a bell-shaped distribution over time of an assumed physiological threshold, as well as with a fixed physiological threshold and a bell-shaped distribution of the sensory effect of repetitions of a given stimulus (Fig. 1b). Fechner extracted one number from the psychometric function—either a measure of central tendency (usually the median) or a measure of dispersion (a transformation of the standard deviation)—to represent the keenness of discrimination, in particular, to designate the average magnitude of stimulus or stimulus difference needed to exceed the physiological threshold. This number was expressed in units of the physical measure and was taken as a stimulus

threshold, which is a statistical construct, as contrasted with the hypothesized physiological threshold. The one number, unfortunately, particularly in the single-stimulus methods, is subject to wide variation with variation in the judgmental factors that intervene between discrimination and response.

Louis Leon Thurstone (1887 to 1955) made the next great advance in the study of discrimination. Bringing the tradition of psychometrics together with psychophysics, he showed how Fechner's methods could be used to quantify psychological attributes of stimuli not readily susceptible to physical measurement—how they could be used, for example, to assess the excellence of handwriting. Thurstone stressed the variable psychological magnitude (sensory effect) of repetitions of a given stimulus and ignored the concept of a physiological threshold. He also emphasized the importance of the paired-comparison procedure in reducing distortions of response frequencies and concentrated on difference (recognition), as opposed to absolute (detection), measurements.

Thurstone's approach to measuring discrimination was outlined in his basic work in 1927 (*2*). His model begins with the assumption of overlapping distributions of the psychological magnitudes of two similar stimuli, shown in Fig. 2, the starting point of every other model I consider here. The model proceeds with some very specific assumptions (also characteristic of most of the more recent models), including normality of the distributions, zero correlation between stimuli, and equal standard deviations. From the proportion of times stimulus B is judged greater than stimulus A, along with a table of areas under the normal curve, one determines the difference between the means of the two distributions. This difference, denoted d in Fig. 2, is expressed in units of the standard deviation.

Thurstone could get by with the single parameter because he supposed, with justification, that the paired-comparison procedure comes close to eliminating judgmental biases. He assumed that subjective probabilities and utilities were symmetrical, that the observer would select B as his response whenever the psychological magnitude of stimulus B exceeded that of stimulus A, and vice versa.

This assumption of symmetry can be simply related to the essence of the ROC analysis, if one considers the

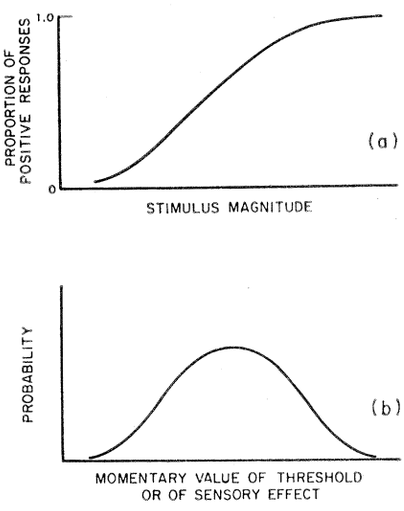

Fig. 1. (a) The psychometric function—the proportion of positive responses as a function of stimulus magnitude; (b) the mechanism assumed to underlie the psychometric function—a temporal variation either in a sensory threshold or in the sensory effect of a given stimulus (*1*).

implications of the assumption for the single-stimulus procedure. First, define the judgmental or response bias in terms of the decision criterion: the decision criterion is a cutoff point (c) along the axis of sensory effects (x) such that values of x above c lead to response B, while values of x below c lead to response A. Then note that Thurstone's assumption of symmetry is equivalent to assuming a decision criterion (represented by the dashed line in Fig. 2) located at the point where the two distributions cross. The ROC analysis, on the other hand, allows the observer to locate his decision criterion anywhere throughout the entire range of x and extracts a measure of discrimination, essentially the one I

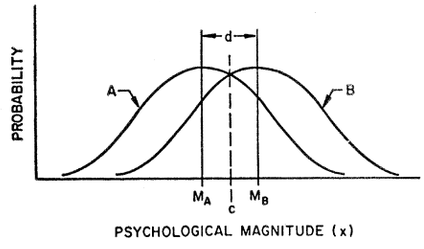

Fig. 2. Hypothetical distributions of the psychological magnitudes of two confusable stimuli, A and B (*2*). The distance between their means, d, can be inferred from the observer's judgments and is a measure of discriminability. The dashed line, c, represents a symmetrical decision criterion, in anticipation of the ROC analysis, which emphasizes a variable criterion.

have denoted d in Thurstone's model, that is independent of the location of the criterion. The ROC technique thus salvages the single-stimulus procedure, in which a symmetrical or otherwise reliable criterion is highly unlikely. The ROC technique also yields a second measure, that of the location of the decision criterion.

The next step in psychophysics was taken in the 1940's and is typified in the work of H. Richard Blackwell (*3*). Blackwell advocated a procedure akin to the paired-comparison procedure, which he called the forced-choice procedure, but it is his application of some of Thurstone's thinking to the single-stimulus (yes-no) procedure that is of interest here.

Blackwell focused on the detection problem, in which one of the two stimuli considered is the null stimulus. By the time he began his work, however, developments in electronics and physiology had changed the conception of the null stimulus from being nothing, or a blank presentation, to being a stimulus in fact. In particular, the new view was that the variability inherent in the environment and in the observer—what we have come to call "noise"—would produce sensorineural activity that could be confused with the sensory effect of the stimulus to be detected. So one of Thurstone's (normal, uncorrelated, constant-variance) distributions, the one with the lower mean, could represent noise alone, while the other could represent noise plus "signal." Noise and signal effects can be plotted on the same axis because they are, by definition, qualitatively the same. The minimal background of noise is created by uncontrollable events outside and inside the observer; the level of noise can be raised by introducing a masking stimulus background—for example, an illuminated screen when the signal is a brief spot of light or a hissing sound when the signal is a brief tone.

Although Blackwell realized that noise could interfere with detection of the signal, he made a further assumption that essentially eliminated the possibility of noise being mistaken for a signal. He assumed the existence of a criterion for a positive response at a level such that the observer would rarely be misled by the noise into reporting a signal (see dashed line in Fig. 3). With such a criterion, fixed over time, the number of false-positive responses ("yes" responses to noise alone) that could be attributed to the

noise, when it momentarily reached a high level, would be negligible. Such a criterion reminds one of the familiar rule in statistical testing for rejecting the null hypothesis at, say, the 1 percent level of confidence.

Blackwell acknowledged that biasing factors might favor a positive response, that the observer might say "yes" on some trials even though the sensory effect during the trial period failed to exceed the criterion (although he ignored the possibility that the observer might say "no" when the sensory effect did exceed the criterion). Blackwell assumed, however, that values of sensory effect below the fixed criterion were indistinguishable—as if this criterion were a physiological threshold—so that the observer could only be guessing that a signal was present when one of these values occurred and would then be correct only by chance. He therefore applied a correction for chance success, according to which the proportion of false-positive responses is taken as an index of the amount by which the proportion of correct-positive responses is inflated, so that by a subtractive procedure the proportion of "true" positive responses is obtained. A correction for chance success was used by others in sensory studies; indeed, it is familiar throughout psychology, having a long history in, for example, studies of recognition memory. In the form of the correction Blackwell adopted, the proportion of correct-positive responses minus the proportion of false-positive responses is divided by 1 minus the proportion of false-positive responses.

Given the proportion of so-called true responses for each stimulus magnitude, it was a short step to plotting the psychometric function and taking the stimulus magnitude corresponding to a response proportion of 0.50 as the stimulus threshold, a one-parameter measure of discrimination.

It can be shown that use of the chance, or guessing, correction assumes that all psychometric functions based on raw proportions, whatever the proportion of guesses or false-positive responses, will correct to a single true curve, as represented in Fig. 4. To state it another way, the chance correction assumes statistical independence of false-positive and true-positive responses. The next pertinent development in psychophysics was the empirical finding that this assumption is not justified, a finding that served to discredit the notion of a

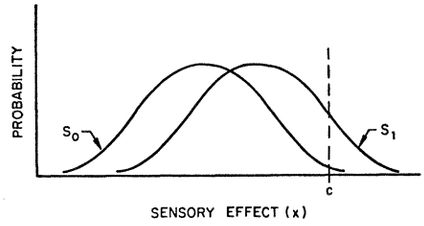

Fig. 3. A possible revision of Thurstone's recognition model to represent detection. Sensory effect is assumed to vary according to the left-hand distribution when the null stimulus, S_0, or noise alone, is present and according to the right-hand distribution, S_1, when a given signal is added to the noise. The criterion for a positive response, c, is assumed to be fixed at such a point that it is rarely exceeded by noise alone, with no discrimination possible below that point; therefore, positive responses to the null stimulus can be considered random guesses (*3*).

fixed criterion for response (or a physiological threshold) located at the upper end of the noise distribution and to undermine the associated measure of discrimination.

Before turning to those empirical results, a few more words about the concept of the null stimulus. In the view of classical psychophysics, the observer should report his *sensations*; to base his reports on the *stimulus*—for example, to let stimulus probabilities and stimulus-response utilities affect the report—is to commit the "stimulus error." In that context, presentation of the null stimulus is a "catch trial": If the observer is caught in a false-positive response, he is admonished to pay better attention to his task. The null stimulus is presented infrequently, and the false-positive responses are not counted (*4*).

An opposing view, the so-called objective view, was advanced about the turn of the century, adopted by

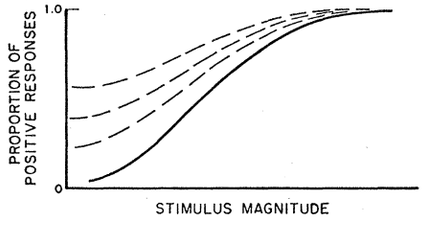

Fig. 4. The correction for chance success implies that psychometric functions obtained with different proportions of false-positive responses (dashed lines) will all correct to one "true" function (solid line). The correction simply normalizes any curve obtained, reducing to zero the proportion of positive responses made to zero stimulus magnitude.

Thurstone and to a lesser extent by Blackwell, and embraced with a vengeance in the context of the ROC analysis. The observer is expected to commit the stimulus error. The probability of a false-positive response must then be carefully estimated, preferably on the basis of as many trials as the probability of a correct-positive response. Rather than subjecting these proportions to the correction for chance, which assumes a decision criterion fixed at a particular value, the ROC analysis uses them to determine the location at that time of a variable criterion.

Substantive support for the ROC approach was supplied by three empirical studies conducted independently in the early 1950's by Moncrieff Smith and Edna A. Wilson, William A. Munson and John E. Karlin, and Wilson P. Tanner, Jr., and me. In each study, data were obtained from observers using different decision criteria in yes-no detection tasks (*5*).

Smith and Wilson, working at the Lincoln Laboratory of the Massachusetts Institute of Technology, were studying the gains in detection performance that were expected to accrue from using teams of observers rather than individual observers. The basis for the study was the assumption that temporal variations in the sensitivity of individuals are less than perfectly correlated, and therefore if one observer were momentarily insensitive another might detect the signal. In their analysis, Smith and Wilson varied the number of individual positive responses that they would take as representing a positive response by the team and noted corresponding differences in the numbers of false-positive responses issued by the team. Recognizing that the number of team false positives would depend also on individual tendencies toward false positives, they instructed some observers to be "conservative" and others to be "liberal" in deciding to report a signal; still others were to respond according to a four-category scale of certainty that a signal existed.

Munson and Karlin, at the Bell Telephone Laboratories, were examining some concepts derived from the communications theory developed by Claude Shannon of that same laboratory. Measuring the rate of information transmitted by their observers' detection judgments required a good estimate of the probability of a positive response to the null stimulus. Their

data showed individual differences among their observers that they described by the terms "safe," "objective," and "risky."

Tanner and I, at the University of Michigan, had studied sensory psychology with Blackwell and were associated with the laboratory in which fellow graduate students Wesley W. Peterson and Theodore G. Birdsall were applying statistical decision theory to radar detection problems and were developing the ROC analysis. We encouraged our observers to vary the proportion of false positives from one group of trials to another by varying the a priori probability of signal presentation and the values and costs (in cents) associated with correct and incorrect "yes" and "no" responses. We also required observers to set several criteria simultaneously and report according to a rating scale.

The three sets of experiments—two in audition, one in vision—showed that the correction for chance did not map all psychometric functions, with different proportions of false-positive responses, onto the same curve. Stimulus thresholds decreased as false positives increased. Corrected, "true" proportions of positive responses at each signal level were highly correlated with proportions of false-positive responses. Evidently, the observers did not produce more positive responses by guessing, by responding "yes" to a random selection among indistinguishable sensory effects that fell beneath a fixed criterion or sensory threshold, but rather by setting a lower criterion. It was clearly courting trouble, then, to extract one of the traditional measures of discrimination without regard to the variable criterion.

At the same time, it was clear that means existed for calibrating any criterion an observer might adopt. What has come to be the preferred measure, the likelihood ratio, was proving useful in a related problem in the field of mathematical statistics. A brief review will give the highlights of that development.

Statistical Theory

The problem faced in testing statistical hypotheses, or in making statistical decisions, is usually represented pictorially in much the same way that Thurstone and Blackwell represented

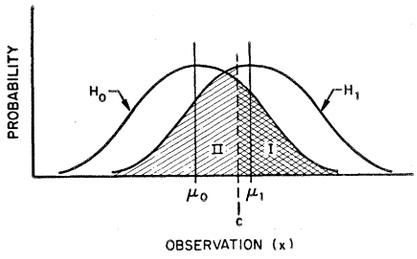

Fig. 5. Population distributions in statistical theory: H_0, the null hypothesis, asserts that the population mean, μ, equals μ_0; H_1, an alternative hypothesis, asserts that $\mu = \mu_1$. The area under H_0 to the right of the decision criterion c represents the probability of a type I error; the area under H_1 to the left of c represents the probability of a type II error.

the discrimination problem. Figure 5 shows the familiar overlapping, bell-shaped distributions, here representing sampling distributions of test statistics. The left distribution represents the null hypothesis, H_0, and the right one represents an alternative hypothesis, H_1. H_0 might assert, for example, that the mean of a population, μ, is equal to some value, μ_0, while H_1 asserts that μ is equal to some other value, μ_1. On the basis of an observation x, one or the other of the hypotheses is accepted.

The construction of a statistical test is equivalent to dividing the x axis into two regions; that is, setting a decision cutoff, or criterion (c), such that sample values of x less than c lead to acceptance of H_0 and sample values of x greater than c lead to acceptance of H_1. Where the criterion is set will

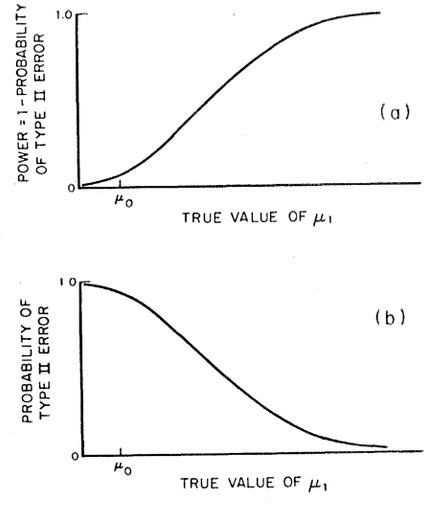

Fig. 6. (a) The power function of a statistical test. (b) The inverse of the power function, the operating characteristic.

determine the relative probabilities of the two possible types of errors: type I errors, which consist in accepting H_1 when H_0 is true, and type II errors, which consist in accepting H_0 when H_1 is true. In general, one wants to adjust these error probabilities in accordance with the relative costs of the two kinds of error, but he must choose among several different rules for doing so.

General principles governing such rules were advanced by Jerzy Neyman and Egon Sharpe Pearson in 1933 (*6*). The particular rule associated with them, and the most familiar rule in statistics, is to fix the probability of a type I error arbitrarily (at a significance level, or confidence level, usually .05 or .01) and then to choose the criterion in such a way as to minimize the probability of a type II error. They showed that the best such test is defined in terms of the likelihood ratio, which is the ratio at any value of x of the ordinate of the H_1 distribution to the ordinate of the H_0 distribution. One accepts H_1 whenever the likelihood ratio exceeds some number c, where c is chosen to produce the desired probability of a type I error.

Ordinarily, instead of considering the probability of a type II error, the focus is on 1 minus that probability, or the probability of accepting H_1 when it is true, called the "power" of the test. Under the Neyman-Pearson rule, then, one fixes the probability of a type I error and chooses the likelihood ratio equal to c in order to maximize the power of the test. When H_0 is tested against several alternatives instead of just one, the power function of the test can be represented as in Fig. 6a. Note that this function is essentially the same as the psychometric function defined by Fechner and that the Neyman-Pearson decision rule was assumed in Blackwell's model.

The operating characteristic, as defined in statistics, is simply 1 minus the power function, as shown in Fig. 6b. The ROC is a graphic way of comparing two operating characteristics—the one just defined and another, rarely seen, if ever, that shows the variation in the probability of a type I error with a fixed probability of a type II error. The ROC gives the two types of errors equal status and shows how they covary as the criterion changes for any given difference between the means of the two hypotheses.

The advance in statistical decision theory that brings one to the present time, although probably understood by Neyman and Pearson, was made by Abraham Wald in the 1940's (*7*). Wald showed that several quite different decision rules—such as maximizing the proportion of correct decisions, maximizing the expected value of a decision, and maximizing the minimum payoff—are unified by means of the likelihood ratio. He made it clear that one construct would handle many of the decision rules an observer might adopt, as well as any of the many criteria that one of those rules might dictate.

Detection Theory

The detection of electromagnetic signals in the presence of noise was seen in the early 1940's to be a problem of testing statistical hypotheses. Noise alone was identified with the null hypothesis, H_0, while noise plus a signal was associated with the alternative hypothesis, H_1. The concern then for radar signals highlighted the importance of a variable decision criterion and the possibility of various decision rules. In the radar context, type I errors are "false alarms" and type II errors are "misses," and whereas both are pretty clearly bad in a defensive situation, their relative cost varies widely with different threats and available reactions to a threat.

The unification of several decision rules by the likelihood ratio was described in two presentations at the 1954 Symposium on Information Theory (sponsored by the Institute of Radio Engineers at the Massachusetts Institute of Technology)—one by David Van Meter and David Middleton of Harvard University and another by Wesley W. Peterson, Theodore G. Birdsall, and William C. Fox of the University of Michigan. Discussion following the coincidence revealed that the Harvard theorists had read Wald, while the Michigan theorists had developed the idea independently. It was this unification of several decision rules that established the generality of the ROC analysis.

The ROC analysis first appeared in the literature in the transactions of that symposium (in papers by the two groups of authors just mentioned and in a paper by Tanner and me), although Peterson and Birdsall had presented it a year earlier in a technical report (*8*). So it is fair to say, from the vantage point of psychology, that Peterson and Birdsall showed us how to plot the data.

The ROC is a plot of 1 minus the probability of a type II error (which equals "power" in statistics and the probability of a "hit" in the detection context) against the probability of a type I error (or "false alarm"), as the decision criterion varies, with the difference between the means of the two hypothetical distributions as the parameter. This difference between the distributions' means is essentially the d of Fig. 2 (representing Thurstone) and the $\mu_1 - \mu_0$ of Figs. 5 and 6 (representing statistical tests). When Peterson and Birdsall used normal distributions of equal variance, which they derived for certain kinds of signal and noise, they used the symbol d to denote the difference between the means in units of the variance. When Tanner and I assumed distributions of that form, we used the symbol d' to denote the difference between the means in units of the standard deviation ($d' = \sqrt{d}$). Other assumptions and no assumption are possible, and require other notations, but the measure d' is the one used most often in psychology.

Figure 7 shows a family of ROC curves based on normal, equal-variance distributions (solid lines). Note that vertical and horizontal cuts through the curves yield the two kinds of operating characteristic mentioned earlier: the vertical cut gives the probability of a type II error for fixed probability of a type I error, the garden variety operating characteristic; the horizontal cut gives the unfamiliar reverse of that operating characteristic. It is because the ROC is a comparison of two operating characteristics that I use the term "relative" operating characteristic, according to a suggestion by Birdsall. Originally, serving to confuse the ROC and the OC (operating characteristic) in the detection context, the R stood for "receiver." That terminology stemmed from the broader perspective of communications, which views detection as part of the reception process. Sometimes, according to a suggestion by R. Duncan Luce, the ROC's footing in statistics is ignored in psychological usage and the ROC is called an "isosensitivity curve" (*9*). This term might be preferable except for the fact that the ROC is also applied to problem areas in psychology not usually thought of in terms of sensitivity (to memory, for example), and the proliferation of terms like "isomnemonic curve" obscures the identity of the single, underlying technique.

The ROC analysis thus gives a measure of discrimination that is independent of the location of the decision criterion and is presumably uncontaminated by the processes, such as expectation and motivation, that affect the response. At the same time, the

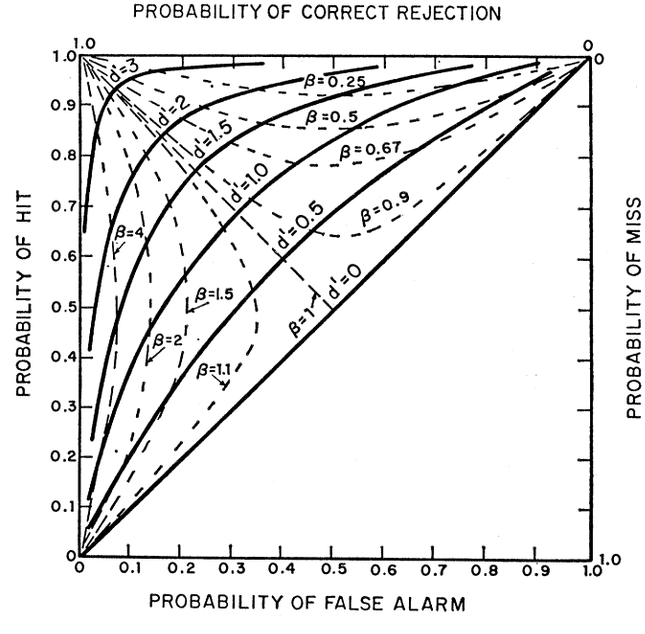

Fig. 7. A family of theoretical ROC curves based on normal, equal-variance distributions, with the parameter d' (solid lines). Also shown is a family of theoretical isobias curves, with the parameter β (dashed lines). The quantities shown on the left and lower coordinates are the two quantities ordinarily used in ROC analysis; the quantities shown on the right and upper coordinates are added here to point out that they are complements of the other two, respectively.

ROC analysis provides a measure of the net effect of processes that influence response—specifically, the location of the decision criterion—at any given time. This measure, called β in the ROC context, is the value of the likelihood ratio at which the criterion has to be set to yield a particular point on a given ROC curve. It can be shown that the measure β equals the slope at which the given ROC curve passes through that particular point. The dashed lines in Fig. 7 are curves of constant β, or isobias curves.

We will proceed shortly to consider some computational details. Note simply now that, if one assumes normal, equal-variance distributions, then certain response proportions plotted as a single point in the ROC space yield independent measures of discrimination (d') and extra-discrimination effects (β). Ordinarily, the decision criterion is manipulated from one group of trials to another, or a rating technique is used to the same effect, in order to obtain better definition of the curve.

In leaving this brief treatment of detection theory, I should observe that the theory has been developed to specify a variety of forms of ROC curves for various kinds of signals and noises and to specify ideal detection performance for various kinds of signals and noises. Although I do not treat the topic here, mathematical models of ideal observers—which show in the limit how d' or a similar measure varies with a physical measure of signal-to-noise ratio—have been used as normative models in sensory psychology, in an attempt to determine what sort of information the human observer extracts from the stimulus (*10*).

Computational Procedures

Data collected for a given location of the decision criterion yield a 2-by-2 contingency table of stimuli and responses. I refer to the stimuli as S_1 and S_2 in the recognition case, and as S_0 and S_1 in the detection case (where S_0 is the null stimulus); and, similarly, to the responses as R_1 and R_2 or R_0 and R_1. Although one of my major purposes is to relate the ROC analysis to a broad class of perceptual and cognitive discrimination problems, it will be simplest to use the terminology of the detection problem throughout the following discussion of computational procedures.

	S_0	S_1
R_0	90	20
R_1	10	80
	100	100

Fig. 8. A contingency table of stimulus and response. The detection notation is used: S_0 for the null stimulus and S_1 for the stimulus to be detected; R_0 for the "no" response and R_1 for the "yes" response.

An example of a contingency table for the detection case is shown in Fig. 8. These frequency data give estimates of the conditional probability of a false alarm

$$P(R_1|S_0) = 10/100 = .10$$

and of the conditional probability of a hit

$$P(R_1|S_1) = 80/100 = .80$$

(The other two conditional probabilities implied by the table, "misses" and "correct rejections," are their complements.)

The straightforward way to calculate d' and β from these two probabilities is by means of a table of normal curve functions. The false-alarm probability of .10 indicates that the criterion is 1.28 standard deviations above the mean of the S_0 distribution, and the hit probability of .80 indicates that the criterion is 0.84 standard deviations below the mean of the S_1 distribution. The value of d' is the sum

$$1.28 + 0.84 = 2.12$$

The measure β is the ratio of the ordinate of the S_1 distribution to the ordinate of the S_0 distribution at the criterion setting

$$0.28/0.18 = 1.55$$

Tables that give d' and β directly for any pair of false alarm and hit proportions are also available (*11*, *12*).

When data are available for several criterion locations, either because the observer varied his criterion from one group of trials to another or because he reported by means of a rating scale, one may use graphs to estimate d'. The theoretical curves of Fig. 7 are straight lines of unit slope when plotted on probability scales—that is, on coordinate scales that space linearly the normal deviates—as shown in Fig. 9. Thus, one can fit several points by a straight line of unit slope and get d', by subtracting the normal-deviate value corresponding to the hit proportion from the normal-deviate value corresponding to the false-alarm proportion, at any point along that line (*13*).

Not all data are fitted well by a straight line of unit slope, of course, and this fact presents a complication. A departure from linearity violates the normality assumption, and nonunit slope violates the equal-variance assumption; in fact, for normal distributions, the slope equals the ratio of the standard deviation of S_0 to the standard deviation of S_1. As it happens, the linearity condition is usually met, and the question is what to do about nonunit slope. It is apparent that the measure d' is everywhere different along a line of nonunit slope, and so does not provide the necessary invariance.

There are three or four ways to contend with this problem. A direct reaction is to use two parameters to represent the ROC curve; for example, (i) the difference between the means of the two supposedly normal distributions, with the standard deviation of the S_0 distribution as the unit, and (ii) the slope of the ROC. An alternative is to use a one-parameter description that ignores slope information; for example, the value of d' at the negative diagonal or the perpendicular distance from the center of the ROC space to the ROC curve, each of which enlists a unit based on the standard deviations of both distributions.

Another way to achieve a one-parameter representation is to assume distributions that predict nonunit slopes—specifically, distributions that predict how the slope will vary with the difference between the means. One can assume, for example, normal distributions with standard deviations that are constant fractions of the means (although then the decision axis cannot be monotonically related to likelihood ratio). One might otherwise assume Poisson, exponential, gamma, or Rayleigh distributions, in each case distributions whose variances are direct functions of their means. This class of alternatives predicts slopes less than unity, slopes that decrease with increasing discrimination—a prediction reasonably in accord with data. In

some discrimination tasks a rationale exists for selecting one of these distributions.

At the other extreme, one might want to assume nothing at all about the distributions underlying the ROC. The proportion of the area of the entire ROC space that lies beneath the ROC curve is a distribution-free measure of sensitivity. It is equal to the probability of a correct choice in a two-alternative, forced-choice task no matter what distributions exist. This and the other measures mentioned in the preceding paragraphs have been described in more detail elsewhere (*10, 11, 14*). The main point for present purposes is that, although the simple measure d' will not do the whole job, other measures of discrimination that are independent of report bias can be extracted from the ROC.

While on the subject of alternative measures of discrimination, let me note that measures of criterion location other than the likelihood ratio are also available. The distance in standard-deviation units from the mean of the S_0 distribution to the criterion location is one such measure, with certain advantages and disadvantages relative to the likelihood ratio (*15*). A nonparametric index of response bias is also available (*16*).

The fact that detection theory predicts ROC curves quite different in form from the ones that can be derived from the chance correction formula, with which the psychological detection theory was first compared, may be the reason that some investigators have tried to infer too much about underlying processes from the form of empirical ROC curves. The basic thought to keep in mind in this regard is that the ROC form reflects only the differences between the two underlying distributions; it does not imply anything about the form of the distributions individually. Indeed, most unimodal distributions will produce an ROC curve that is very nearly a straight line on probability scales; even a rectangular S_0 distribution and a ramp S_1 distribution lead to a linear ROC on those scales (*17*). Moreover, assumptions about the decision process enter in a critical way. Thus, for example, if the variance of the decision criterion is large, the slope of the linear ROC will approach unity, even if the standard deviation of S_1 is much larger than the standard deviation of S_0 (*18*). Again, with certain parameters assumed for the decision process, a two-state discrimination process can produce as smooth an ROC curve as that produced by a discrimination process that yields a continuous output (*19*). It may also be noted in this connection that the technique used to calculate an ROC curve from rating data is cumulative and therefore forces a monotonic, increasing curve.

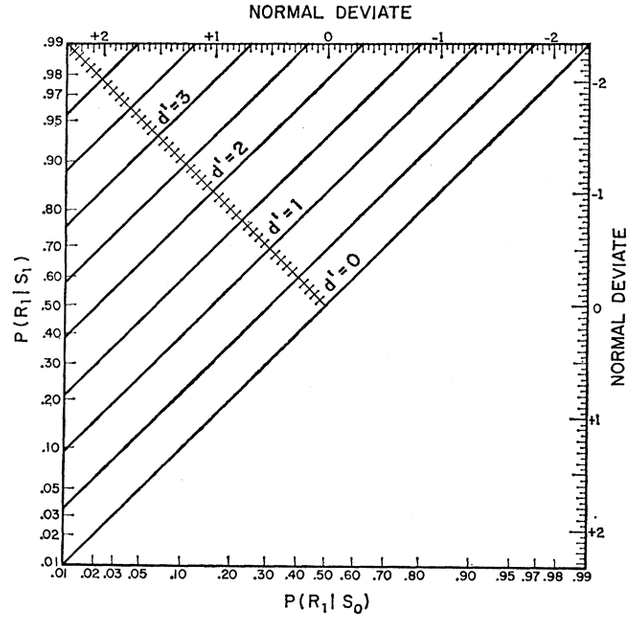

Fig. 9. The theoretical ROC curves of Fig. 7 plotted on probability scales (left and bottom) and on linear normal-deviate scales (top and right).

When the ROC analysis was first devised, data points were fitted by eye, for lack of a known alternative, and this procedure is still probably adequate for many applications. However, several more objective estimation procedures have been devised (*20*). Significance tests for observed differences in d' have also been developed (*21*), and the sampling variability of the value of d' at the negative diagonal and of the area under the ROC curve have been examined (*22*).

Applications

The ROC confers the ability to measure covert discriminations in single-stimulus tasks in a relatively pure form —to measure these discriminations, at least to first order, unconfounded by the biasing factors that tended to distort the overt report in the measurement procedures previously available. How has psychology profited?

Published studies in which the ROC has been used fall into a dozen or so substantive areas. In some of these areas the value has primarily been more reliable and valid measurements; I will treat those areas here only by reference. In other areas the ROC has led to substantially revised interpretations; I will discuss briefly a few examples: sensory functions, vigilance, perceptual selectivity, and memory.

The basic applications of the ROC to sensory processes have been presented in a collection of articles (*11*) and a systematic textbook (*10*). These volumes report studies of detection and recognition processes in several sensory modalities and support the following conclusions.

1) Empirical ROC curves can be reliably obtained by manipulating signal probability or the values and costs of the various stimulus-response outcomes; by verbal instructions to adopt, say, a "strict," "medium strict," "medium," "medium lax," or "lax" criterion; or by instructions to use those criteria simultaneously as the boundaries of rating categories.

2) Measures of sensitivity that do not isolate effects of changes in the decision criterion ignore a substantial source of variation.

3) The ROC analysis rescues an important test method, previously suspect because of its great susceptibility to biasing effects—the single-stimulus method.

4) Of sensitivity measures extant, only those associated with the ROC model give reasonably consistent results across yes-no, rating, and forced-choice procedures.

5) When the stimulus is measured

in terms prescribed by detection theory, the functional relationship obtained between d' and stimulus magnitude is practically the same from one laboratory to another.

6) Some prominent conceptions of the sensory threshold are incorrect, and the theories and test methods that depend on them are invalid.

Vigilance is the practical detection problem—the observation period is long and signals occur infrequently. The first studies, around 1950, showed that the probability of a correct detection drops off noticeably in only a half hour or so of observation (*23*). What seemed at the time to be a rapid decrement in performance was surprising, for the subject was not asked to work very hard. Hundreds of studies conducted since, with a great variety of stimulus displays, have shown the same sort of decrement. They have examined a host of psychological, physiological, situational, and environmental variables thought to affect alertness, including work-rest cycle, intersignal interval, irrelevant stimulation, incentives, knowledge of results, introversion-extroversion, temperature, drugs, age, and sex (*24, 25*). At least five theories have been proposed to account for the apparent decrement in sensitivity (*26*).

About 10 years ago, several investigators, led by James P. Egan (*27*), began to question the assumption that a sensitivity decrement occurs in these vigilance tests. The ROC analysis makes it clear that the probability of a hit could decline without implying a decrease in sensitivity; if the proportion of false alarms also dropped, sensitivity might remain constant while the decision criterion changed. A progressively more conservative decision criterion might come about as the result of a decreasing expectation of signal occurrence (for example, when a naive subject experiences a high signal probability in training sessions and then a low signal probability in test sessions), or it might result from a motivational change (for example, if the perceived value of a hit were to decrease over time relative to the perceived cost of a false alarm).

The ROC analysis has now been employed in some 30 studies of vigilance (*28*). What they add up to is that with almost all stimulus displays, including those used in the earliest experiments, the sole change over time is in the decision criterion. Sensitivity remains constant. Alertness, apparently, remains essentially constant. With a few stimulus displays—specifically, visual displays with undefined trials or with trials occurring at a rate greater than about one per second—changes in both d' and β are observed: sensitivity decreases and the criterion becomes more stringent. In short, a sensitivity decrement in vigilance tests is uncommon, and when it occurs it is smaller than originally supposed. The practical problem is training the observer to hold a constant decision criterion. One study has shown that the observer will hold a constant decision criterion if the values of correct responses and the costs of incorrect responses are well defined (*29*).

It was in the late 1940's that several theorists in the field of perception emphasized the perceiver's contribution to what he perceives, a contribution that stems from inner states such as needs, emotions, and values. They were dubbed "new look" theorists, and the term "old look" was then applied to the theorists, primarily Gestaltists, who concentrated on stimulus determinants and denied effects of experience.

Several experiments, most of them using words as stimuli, were purported to show perceptual selectivity. Differences in measured thresholds of different classes of words were attributed to mechanisms of perceptual vigilance, sensitization, and perceptual defense. However, most of these results were soon accounted for in terms of differences in the commonality of the words employed, the "word-frequency effect." At first, word frequency was viewed as affecting perception per se, as effecting a selective intake of information. Then much converging evidence showed instead that stimulus frequency affects the response system.

At the same time that differences in thresholds were examined, several experimenters sought to demonstrate subthreshold, or "subliminal," perception related to inner states of the observer. Several lines of criticism were applied to these experiments, including the criticism that many of the results showed only that discrimination was possible when identification was not, or that forced-choice thresholds are typically lower than yes-no thresholds.

The ROC analysis came into contact with developments in this area at many points. For one reason, many of the experiments used the correction for chance and the associated concept of threshold. For another reason, the notion of relative willingness to respond was put forth in connection with taboo-word experiments. The ROC analysis was used to explain that yes-no thresholds are higher than forced-choice thresholds because stringent yes-no criteria are encouraged and, of course, to suggest that the presumed threshold was actually a response criterion (*30*).

Granting that stimulus probability affects the response system, which brings one back to the old look, the question now being addressed is whether the mechanism is a guessing process or a variable criterion. At the moment, contradictions are flying back and forth, and there is some doubt that the most familiar experimental paradigm can distinguish the two kinds of mechanism (*31*). After the most extensive analysis to date, including the results of a new experimental paradigm, Donald E. Broadbent argues quite persuasively that the mechanism is a variable criterion (*25*). His experimental results, incidentally, indicate that d' is *lower* for common words than for uncommon words.

In a recognition memory task, the subject is asked to say whether each of a series of items was presented before ("old") or not ("new"). Following Egan again (*10, 32*), one may refer to the picture of the two distributions and a variable criterion and identify the new items with the left-hand (noise) distribution and the old items with the right-hand (signal) distribution. The assumption is that all items fall along a continuum of memory strength, at a location affected by acquisition and forgetting. This identification should allow phenomena of memory to be separated from biases associated with the response.

Several presumed memory effects have recently been shown by ROC analysis to be response effects only: that is, d' remains constant, while β changes. One such effect is the better recall of more common words; in fact, here again there is evidence that d' is lower for common than for uncommon words (*25, 33*). A related finding is that familiar associations are no better recalled than unfamiliar ones when unbiased measures are used (*34*). The typical increasing rate of

false-positive responses in a continuous recognition task has been shown to be a reflection of criterion change rather than a buildup of proactive interference (*35*). Various amounts of learning interpolated between acquisition and recall have an effect on β but not on d'; this result is contrary to the extinction hypothesis of retroactive inhibition and indicates that generalized response competition is responsible for the criterion change (*36*). Although changes in semantic or association context from acquisition to recall have been reported to reduce recognition accuracy, with the reduction explained in terms of multiple representations in memory, a new study shows these changes to affect only the response process (*37*). Another study uses the ROC to assert an equality of females and males in recognition memory, even though they may differ with respect to response bias (*38*).

Other memory studies show a change in both d' and β, thereby demonstrating that the ROC analysis is required. Meaningfulness in a paired-associate, short-term memory task was found to affect memory and response bias (*39*). In that study and in another (*40*), both d' and β were correlated with serial position. Finally, increasing the similarity of distracting items to target items raises the criterion while lowering d' (*41*). Two articles have reviewed many of these studies in more detail and have considered some of the theoretical issues involved (*15, 17*).

The ROC analysis has also been applied in the areas of attention (*25*), imagery (*42*), learning (*43*), conceptual judgment (*44*), personality (*45*), reaction time (*46*), manual control (*47*), and speech (*10*). Rats, pigeons, goldfish, and monkeys have produced exceptionally neat ROC curves, by rating and yes-no responses, with variations in signal probability or with differential reinforcement (*48*). Lee B. Lusted has applied the ROC to medical decisions, particularly in radiology (*49*). A recent finding in physiological psychology is that the amplitude of a particular component of evoked cortical potentials increases monotonically with increasing strictness of the decision criterion, whether the criterion is manipulated by varying the signal probability or by varying the values and costs of the possible stimulus-response outcomes. This component was always present when the observer correctly reported a signal to exist, but was never present when the response was a miss or a false alarm (*50*). The ROC has also been used to evaluate the effectiveness of information retrieval systems (*51*).

Summary

The clinician looking, listening, or feeling for signs of a disease may far prefer a false alarm to a miss, particularly if the disease is serious and contagious. On the other hand, he may believe that the available therapy is marginally effective, expensive, and debilitating. The pilot seeing the landing lights only when they are a few yards away may decide that his plane is adequately aligned with the runway if he is alone and familiar with that plight. He may be more inclined to circle the field before another try at landing if he has many passengers and recent memory of another plane crashing under those circumstances. The Food and Drug administrator suspecting botulism in a canned food may not want to accept even a remote threat to the public health. But he may be less clearly biased if a recent false alarm has cost a canning company millions of dollars and left some damaged reputations. The making of almost any fine discrimination is beset with such considerations of probability and utility, which are extraneous and potentially confounding when one is attempting to measure the acuity of discrimination per se.

The ROC is an analytical technique, with origins in statistical decision theory and electronic detection theory, that quite effectively isolates the effects of the observer's response bias, or decision criterion, in the study of discrimination behavior. This capability, pursued through a century of psychological testing, provides a relatively pure measure of the discriminability of different stimuli and of the capacity of organisms to discriminate. The ROC also treats quantitatively the response, or decision, aspects of choice behavior. The decision parameter can then be functionally related to the probabilities of the stimulus alternatives and to the utilities of the various stimulus-response pairs, or to the observer's expectations and motivations. In separating and quantifying discrimination and decision processes, the ROC promises a more reliable and valid solution to some practical problems and enhances our understanding of the perceptual and cognitive phenomena that depend directly on these fundamental processes. In several problem areas in psychology, effects that were supposed to reflect properties of the discrimination process have been shown by the ROC analysis to reflect instead properties of the decision process.

References and Notes

1. G. T. Fechner, *Elemente der Psychophysik* (Breitkopf & Hartel, Leipzig, 1860). English translation of volume 1 by H. E. Adler, *Elements of Psychophysics*, D. H. Howes and E. G. Boring, Eds. (Holt, Rinehart & Winston, New York, 1966). Reviewed by J. A. Swets, *Science* **154**, 1532 (1966).
2. L. L. Thurstone, *Psychol. Rev.* **34**, 273 (1927); *Am. J. Psychol.* **38**, 368 (1927).
3. H. R. Blackwell, *J. Exp. Psychol.* **44**, 306 (1952); *J. Opt. Soc. Am.* **53**, 129 (1963).
4. E. G. Boring, *Am. J. Psychol.* **32**, 449 (1921).
5. M. Smith and E. A. Wilson, *Psychol. Monogr.* **67**, No. 9 (1953); W. A. Munson and J. E. Karlin, *J. Acoust. Soc. Am.* **26**, 542 (1954); W. P. Tanner, Jr., and J. A. Swets, *Psychol. Rev.* **61**, 401 (1954).
6. J. Neyman and E. S. Pearson, *Phil. Trans. Roy. Soc. London Ser. A Math. Phys. Sci.* **231**, 289 (1933).
7. A. Wald, *Statistical Decision Functions* (Wiley, New York, 1950).
8. W. W. Peterson, T. G. Birdsall, W. C. Fox, *Trans. IRE Prof. Group Inf. Theory* **PGIT-4**, 171 (1954); D. Van Meter and D. Middleton, *ibid.*, p. 119; W. P. Tanner, Jr., and J. A. Swets, *ibid.*, p. 213; W. W. Peterson and T. G. Birdsall, *The Theory of Signal Detectability* (technical report No. 13, Electronic Defense Group, University of Michigan, Ann Arbor, 1953).
9. R. D. Luce, in *Handbook of Mathematical Psychology*, R. D. Luce, R. R. Bush, E. Galanter, Eds. (Wiley, New York, 1963), pp. 103–189.
10. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966). A reprint of this book, with a supplementary bibliography of publications from 1967 to 1973, is in press (Krieger, Huntington, N.Y.).
11. J. A. Swets, Ed., *Signal Detection and Recognition by Human Observers* (Wiley, New York, 1964).
12. D. R. Freeman, *Tables of d' and β* (technical report No. APU 529/64, Applied Psychology Research Unit, Cambridge, England, 1964); L. Hochhaus, *Psychol. Bull.* **77**, 375 (1972).
13. Codex graph paper No. 41,453 runs from .01 to .99 on the probability scales, which is adequate for psychological applications, and gives the normal-deviate scales as well. Kueffel and Esser No. 47 8062 is the other probability-by-probability graph paper available; it is scaled from .0001 to .9999 and does not give the normal-deviate scales.
14. A. I. Schulman and R. R. Mitchell, *J. Acoust. Soc. Am.* **40**, 473 (1966); J. Markowitz and J. A. Swets, *Percept. Psychophys.* **2**, 91 (1967); B. C. Brookes, *J. Doc.* **24**, 41 (1968); S. E. Robertson, *ibid.* **25**, 1 (1969); *ibid.*, p. 93.
15. W. P. Banks, *Psychol. Bull.* **74**, 81 (1970).
16. W. Hodos, *ibid.*, p. 351; J. B. Grier, *ibid.* **75**, 424 (1971).
17. R. S. Lockhart and B. B. Murdock, Jr., *ibid.*, p. 100.
18. W. Wickelgren, *J. Math. Psychol.* **5**, 102 (1968).
19. D. Krantz, *Psychol. Rev.* **76**, 308 (1969).
20. J. C. Ogilvie and C. D. Creelman, *J. Math. Psychol.* **5**, 377 (1968); D. D. Dorfman and E. Alf, Jr., *Psychometrika* **33**, 117 (1968); *J. Math. Psychol.* **6**, 487 (1969); I. G. Abramson

and H. Levitt, *ibid.*, p. 391; D. R. Grey and B. J. T. Morgan, *ibid.* **9**, 128 (1972).
21. V. Gourevitch and E. Galanter, *Psychometrika* **32**, 25 (1967); L. A. Marascuilo, *ibid.* **35**, 237 (1970).
22. I. Pollack and R. Hsieh, *Psychol. Bull.* **71**, 161 (1969).
23. N. H. Mackworth, "Research on the measurement of human performance" (Medical Research Council Special Report Series No. 268, His Majesty's Stationery Office, London, 1950).
24. D. N. Buckner and J. J. McGrath, Eds., *Vigilance: A Symposium* (McGraw-Hill, New York, 1963); J. F. Mackworth, *Vigilance and Habituation: A Neurophysiological Approach* (Penguin, Middlesex, England, 1969).
25. D. G. Broadbent, *Decision and Stress* (Academic Press, London, 1971).
26. J. P. Frankmann and J. A. Adams, *Psychol. Bull.* **59**, 257 (1962).
27. J. P. Egan, G. Z. Greenberg, A. I. Schulman, *J. Acoust. Soc. Am.* **33**, 993 (1961).
28. About half of these studies were published early enough to be discussed in reviews by Broadbent (*25*), by J. A. Swets and A. B. Kristofferson [*Annu. Rev. Psychol.* **21**, 339 (1970)], and by J. F. Mackworth [*Vigilance and Attention: A Signal Detection Approach* (Penguin, Middlesex, England, 1970)].
29. P. Lucas, *J. Acoust. Soc. Am.* **42**, 158 (1967).
30. J. A. Swets, W. P. Tanner, Jr., T. G. Birdsall, *The Evidence for a Decision-Making Theory of Visual Detection* (technical report No. 40, Electronic Defense Group, University of Michigan, Ann Arbor, 1955); I. Goldiamond, *Psychol. Bull.* **55**, 373 (1958); J. Pierce, *ibid.* **60**, 391 (1963).
31. D. E. Broadbent, *Psychol. Rev.* **74**, 1 (1967); J. Catlin, *ibid.* **76**, 504 (1969); L. H. Nakatani, *ibid.* **77**, 574 (1970); M. Treisman, *ibid.* **78**, 420 (1971); J. R. Frederiksen, *ibid.*, p. 409; L. H. Nakatani, *ibid.* **80**, 195 (1973); J. Catlin, *ibid.*, p. 412.
32. J. P. Egan, *Recognition, Memory, and the Operating Characteristic* (technical note AFCRC-TN-58-51, Hearing and Communication Laboratory, Indiana University, Bloomington, 1958).
33. J. D. Ingleby, thesis, Cambridge University (1969).
34. D. McNicol and L. A. Ryder, *J. Exp. Psychol.* **90**, 81 (1971).
35. W. Donaldson and B. B. Murdock, Jr., *ibid.* **76**, 325 (1968).
36. W. P. Banks, *ibid.* **82**, 216 (1969).
37. F. DaPolito, D. Barker, J. Wiant, *Psychonomic Sci. Sect. Hum. Exp. Psychol.* **24**, 180 (1971).
38. M. Barr-Brown and M. J. White, *ibid.* **25**, 75 (1971).
39. G. A. Raser, *J. Exp. Psychol.* **84**, 173 (1970).
40. B. B. Murdock, Jr., *ibid.* **76** (Suppl. 4, part 2) 1 (1968).
41. G. Mandler, Z. Pearlstone, H. S. Koopmans, *J. Verb. Learning Verb. Behav.* **8**, 410 (1969).
42. S. J. Segal and V. Fusella, *J. Exp. Psychol.* **83**, 458 (1970); *Psychonomic Sci. Sect. Hum. Exp. Psychol.* **24**, 55 (1971).
43. G. R. Grice, *Psychol. Rev.* **75**, 359 (1968).
44. Z. J. Ulehla, L. Canges, F. Wackwitz, *Psychonomic Sci. Sect. Hum. Exp. Psychol.* **8**, 221 (1967); *ibid.*, p. 223.
45. R. H. Price, *Psychol. Bull.* **66**, 55 (1966); W. C. Clark, *ibid.* **65**, 358 (1966).
46. R. Pike, *Psychol. Rev.* **80**, 53 (1973); R. G. Swensson, *Percept. Psychophys.* **12**(1A), 16 (1972).
47. H. S. Cohen and W. R. Ferrell, *IEEE Trans. Man-Mach. Syst.* **10**, 41 (1969).
48. D. S. Blough, *Science* **158**, 940 (1967); J. A. Nevin, in *Animal Psychophysics*, W. C. Stebbins, Ed. (Appleton-Century-Crofts, New York, 1970); D. Yager and I. Duncan, *Percept. Psychophys.* **9**(3B), 353 (1971); M. Terman and J. S. Terman, *ibid.* **11**(6), 428 (1972); A. A. Wright, *Vision Res.* **12**, 1447 (1972); B. M. Clopton, *J. Exp. Anal. Behav.* **17**, 473 (1972); T. F. Elsmore, *ibid.* **18**, 465 (1972); W. Hodos and J. C. Bonbright, Jr., *ibid.*, p. 471.
49. L. B. Lusted, *Science* **171**, 1217 (1971); *Introduction to Medical Decision-Making* (Thomas, Springfield, Ill., 1968); in *Computer Diagnosis and Diagnostic Methods*, J. A. Jacquez, Ed. (Thomas, Springfield, Ill., 1972).
50. D. Paul and S. Sutton, *Science* **177**, 362 (1972); K. C. Squires, S. A. Hillyard, P. H. Lindsay, *Percept. Psychophys.* **13**, 25 (1973).
51. J. A. Swets, *Science* **141**, 245 (1963); *Am. Doc.* **20**, 72 (1969).

NEWS AND COMMENT

Sloan-Kettering: The Trials of an Apricot Pit—1973

These are bad times for reason, all around. Suddenly, all of the major ills are being coped with by acupuncture. If not acupuncture, it is apricot pits. . . . —LEWIS THOMAS, president, Memorial Sloan-Kettering Cancer Center, in an address delivered 11 October 1973.

At the Memorial Sloan-Kettering Cancer Center on the upper east side of Manhattan, some perfectly respectable scientists are taking a new look at some thoroughly unrespectable cancer remedies. Inevitably, they are generating a fair amount of controversy in the process.

One of the unorthodox remedies Sloan-Kettering researchers are evaluating—and one that has caused them considerable embarrassment recently—is a drug called Laetrile. Laetrile, known chemically as amygdalin, is derived from apricot pits. According to its proponents, who are legion, Laetrile often cures cancer. And, they claim, in those cases in which it fails to actually cure, it gives terminal cancer patients a sense of well-being and surcease from pain that allows them to live out their days in relative peace. According to its detractors, who also are legion, Laetrile does nothing of the sort. In the eyes of the National Cancer Institute, the Food and Drug Administration, and the American Cancer Society, Laetrile therapists are quacks.

And, Sloan-Kettering's new president, Lewis Thomas, shares the view that much of what has been claimed in the name of Laetrile goes beyond the bounds of reason. But the institute's searching look at Laetrile is another story altogether.

Preliminary results of one Sloan-Kettering study suggest that Laetrile might actually have some anticancer activity in mice. Understandably, that study is provocative. The fact that it was meant to be kept confidential and that it came to light through a leak adds a touch of intrigue to the drama.

The story apparently began about 2 years ago, when investment banker Benno C. Schmidt, who is also on the board of Sloan-Kettering, became President Nixon's number one adviser in the national war against cancer. There are a lot of people in this country who believe in Laetrile. Many of them buy it for themselves or their dying friends or relatives on the black market. Many go to Tijuana to get it at a clinic operated by a pathologist named Ernesto Contreras. These people began writing Schmidt letters.

"Since I've been chairman of the President's cancer panel, I've had literally hundreds of letters about Laetrile. Some people ask me whether it is any good. Others flatly state that it cures. A great many say that, in any case, it alleviates pain. When I answer these people and tell them that Laetrile has no effect, I would like to be able to do so with some conviction," Schmidt said in a conversation with *Science*. His curiosity piqued, he began asking questions.

He took it up with the National Cancer Institute. People there told him they had looked into the matter long since and found no basis for any claims that Laetrile is good for fighting cancer. The American Cancer Society, which lists Laetrile in its book, *Unproven Methods of Cancer Treatment*, concurs. Schmidt asked a couple of leading cancer scientists what they knew about Laetrile. They, too, told him it has no value. But when he asked for evidence, he recalls, "I couldn't get anybody to show me his work."

The research that has been done on Laetrile by so-called reputable scientists