Annu. Rev. Psychol. 1999. 50:215–41

QUANTIFYING THE INFORMATION VALUE OF CLINICAL ASSESSMENTS WITH SIGNAL DETECTION THEORY

Richard M. McFall and Teresa A. Treat

Department of Psychology, Indiana University, Bloomington, Indiana 47405;
e-mail: mcfall@indiana.edu; ttreat@indiana.edu

KEY WORDS: ROC, Bayesian, probability theory, base rates, cutoff value

ABSTRACT

The aim of clinical assessment is to gather data that allow us to reduce uncertainty regarding the probabilities of events. This is a Bayesian view of assessment that is consistent with the well-known concept of incremental validity. Conventional approaches to evaluating the accuracy of assessment methods are confounded by the choice of cutting points, by the base rates of the events, and by the assessment goal (e.g. nomothetic vs idiographic predictions). Clinical assessors need a common metric for quantifying the information value of assessment data, independent of the cutting points, base rates, or particular application. Signal detection theory (SDT) provides such a metric. We review SDT's history, concepts, and methods and provide examples of its application to a variety of assessment problems.

CONTENTS

0084-6570/99/0201-0215$08.00

FUNDAMENTALS REVISITED

Psychologists too often make the mistake of equating clinical assessment with the administration of tests or interviews to clinical patients for purposes of arriving at individual diagnoses, predicting outcomes, planning interventions, or tracking therapeutic changes. These applied activities may be the most visible part of clinical assessment, but they are only the exposed tip of the whole clinical assessment enterprise, which really is much broader, deeper, and more complicated than these surface activities reveal. Beneath every clinical application of a valid psychological test lies an extensive foundation of scientific theory, empirical research, and quantitative modeling. To the extent that psychologists neglect this foundation, construing clinical assessment narrowly as the clinical application of psychological tests, they are impeding scientific advances in clinical assessment.

This review critically examines the current status and future prospects of clinical assessment, broadly defined. The review focuses on the aims, concepts, methods, and evaluative criteria that underlie the clinical assessment enterprise in general. It is not about specific tests; it is about the functions of clinical assessment, the standards by which methods can be evaluated, and the most promising approaches to achieving the broad goals of clinical assessment. Of course, to the extent that the review helps strengthen the foundations of clinical assessment, it also—as a byproduct—should have practical implications for the more applied aspects of clinical assessment.

Let us start by recapping the basics: The purpose of all psychological assessment is to gather data that provide information regarding specific theoretical questions. This seemingly simple sentence is saturated with important implications. First, the sentence highlights the essential link between theory and assessment (McFall 1993). All assessments are driven by questions; these questions, in turn, always reflect the assessor's theoretical preconceptions, hunches, and assumptions, whether formal or informal, explicit or implicit (Popper 1962). To be useful, an assessment must be tailored to the specific questions that gave rise to it in the first place; its value is determined entirely by its ability to illuminate these questions (Meehl 1971).

Second, the ties between theory and assessment are bidirectional. The ability of an assessment to shed light on a theory is constrained by the validity of the underlying assumptions and constructs of the theory it is attempting to illuminate (Popper 1962). No assessment is atheoretical or assumption-free. Just as water cannot rise naturally above its source, no assessment can be more valid than the theoretical conceptions and assumptions from which it springs.

Third, the term information, in this sentence, has a very specific meaning. It is defined as "the reduction of uncertainty," which is a relativistic concept referring to the relative increment in predictive accuracy, or the relative decrease in predictive error, that is yielded by data (Gigerenzer & Murray 1987). Thus, data reduce uncertainty: They have information value or are illuminating to the degree that they allow us to predict or control events with greater accuracy or with less error than we could have done without them (Mischel 1968).

Fourth, this conception of information is fundamentally quantitative (Gigerenzer & Murray 1987, Meehl & Rosen 1955). The information value of assessment data is represented by a scaled numerical value corresponding to the magnitude of the quantitative difference between the predictive accuracy of our prior model (i.e. the accuracy achieved without the data) and the predictive accuracy of our posterior model (i.e. the accuracy achieved after adjusting the model to reflect the new data). This quantitative view of assessment information suggests a Bayesian epistemology both conceptually and computationally. We will elaborate these Bayesian connections below, but for now, the spirit of this epistemology is reflected in the familiar assessment concept of incremental validity.[1]

Probability Theory

Before proceeding to the main level of our review, we must lay one final block in our conceptual foundations. The cornerstone of psychological assessment is probability theory.[2] Contemporary scientific theories increasingly assume that events in nature including human behavior are probabilistic, or stochastic, rather than deterministic (Gigerenzer et al 1989). Events are determined, in part, by the chance confluence of many other events that are, themselves, unpredictable, random, or the result of chance. Thus, not all assessment variability (error, uncertainty) is due to the inadequacies of our theories and measures; some simply is due to the fact that the events we are attempting to assess and predict are inherently probabilistic.

This means that we cannot expect nature to be so well-behaved as to allow us to predict single events with certainty; instead, the best we can hope for is to identify an array of possible outcomes and to estimate the relative likelihood of 1each. From this probabilistic perspective, the idealized goal of traditional assessment predicting unique, remote events with precision is fanciful, reflecting our naivete and/or hubris. A more realistic assessment goal would be to estimate with incremental accuracy the relative probabilities of the array of possible outcomes for an event. Useful assessments provide data with infor-

[1]We have traced the origins of the concept of incremental validity to Meehl (1959).

[2]Probability theory, as discussed in this chapter, is not synonymous with the concept of probability as taught in the usual psychology statistics courses.

mation value, data that improve the relative accuracy of our probability estimates.

The mechanics of empirically generating a normative probability distribution are straightforward. First, we must impose some category structure on nature, segmenting our target event into two or more mutually exclusive and exhaustive classes of outcomes. Our choice of category structures is guided by our theoretical preconceptions, by our assessment questions, and by practical and methodological considerations. It is up to us to decide, for example, whether to employ a dichotomous category structure (e.g. yes-no, success-failure) or a finer-grained structure (e.g. ratings on a 5-point scale, scores on a 100-point scale, age in years). This choice invariably involves trade-offs. On one hand, simple category structures require smaller samples to fill them, and tend to be more reliable, more readily analyzed and interpreted, and easier to work with. On the other hand, finer-grained structures tend to capture and retain more information. Once we have decided on our category structure, we must observe a suitably large and representative sample of actual outcomes, tabulating and summing the frequencies for each of our categories. These raw frequency data then are normalized by transforming them into proportions; each category total is divided by the grand total of observations. The resulting proportions represent the empirically observed relative probabilities of the categorical events.

Thus, empirically derived probability distributions actually are quantitative records of historical events. As clinical assessors, however, we usually are more interested in predicting the future than in recounting the past. So why should we care about historical probabilities? Because we assume that such historical probability distributions provide the best estimates of the future probabilities for maximally similar events assessed under maximally similar circumstances. The assumption that past probabilities are predictive of future probabilities seems to be a huge leap of faith. Fortunately, however, this idea has been studied extensively by probability theorists, actuaries, mathematicians, philosophers, and scientists for more than three centuries, and consistently has shown itself to be a robust and extremely useful basis for predicting all sorts of events—even random or chance events. Although it seldom yields perfect predictions, no other approach does as well. Note that this assumption cannot be proven, so it must remain an assumption. No matter how many times past probabilities provide useful estimates of future probabilities, we never can be certain that they will do so the next time. This uncertainty notwithstanding, probability theory is the cornerstone of all clinical assessment and prediction.

Conditional Probabilities

We have asserted that historical probability distributions provide the best estimates of the future probabilities for maximally similar events assessed under

maximally similar circumstances. Imbedded in this assertion is an important caveat: As the variability of the circumstances surrounding our assessment of past and future events increases, the accuracy of our probability estimates is likely to decrease. For example, if we used a sample consisting primarily of men to estimate the probability distribution of body weights for a sample consisting primarily of women, our estimates would be unacceptably inaccurate. Fortunately, we can reduce the threat that varying conditions pose to our probabilistic predictions by identifying the specific conditions that systematically affect the outcomes, and incorporating these variables into our probability models.

Until now, we have described only probability models based on a simple, one-dimensional array of categories. To deal with the added complexity of other factors that systematically affect the variability of our probability estimates, we need multidimensional models, which yield joint probability distributions. The simplest form of a joint distribution is a two-dimensional, two-category model, as represented by a 2 x 2 contingency table.[3] The horizontal axis of such a table, for example, might represent the dimension of body weight with two categories (e.g. light = 150 lbs or below; heavy = more than 150 lbs); the vertical axis might represent the two-category dimension of gender. Each observation in a sample would be tabulated in one of the table's four cells (e.g. heavy-male; light-female). To normalize this two-dimensional table of joint frequencies, the observed frequency for each cell would be divided by the total number of observations.

With multidimensional probability models, we not only can compute joint probabilities, but also can compute conditional probabilities, which are even more useful. In our simple two-dimensional table, for example, we can generate separate estimates—one for men, one for women—of the relative probabilities of the two weight categories (light, heavy). These conditional probabilities are computed by dividing each cell total by the marginal total for each level of the conditional variable (e.g. for each sex separately). Conditional probabilities provide more accurate estimates because, in effect, they yield separate estimates of probability distributions for each level of conditions suspected of systematically affecting the outcome.

Clinical psychologists' interest in conditional probabilities is not unique. Virtually all scientific research and theory involves probabilistic models of conditional relations among variables (along with hypothetical explanations for these relations). The careers of epidemiologists, insurance actuaries, business executives, and casino operators, for example, all hinge on an ability to

[3]There may be more than two dimensions, of course, and each dimension may be divided into more than two categories; however, a model's complexity increases rapidly as the number of joint probability cells increases multiplicatively.

estimate conditional probabilities with reasonable accuracy. Laypeople, too, have a vital interest in estimating conditional probabilities; almost every life decision involves a subjective appraisal of the likely outcomes (risks and gains) of different choices. No doubt the popular appeal of astrology and numerology stems from their illusory reduction of uncertainty. The horoscopes in daily newspapers, after all, are nothing but tables of conditional probability statements (e.g. if you are a Leo, you can expect this event; if you are a Capricorn, you can expect that event, and so forth).

The distinguishing characteristics of scientific clinical psychology46s interest in conditional probabilities are (*a*) its use of a scientific epistemology and quantitative methods to build and test theoretical models of conditional probabilities, and (*b*) its focus on a specific content area: psychopathology (the assessment, prediction, prevention, amelioration, and explanation of abnormal human behavior). Clinical psychologists and meteorologists employ similar scientific methods, for example, but focus on different content. Clinical psychologists and fortune tellers share some content interests, but employ different methods.

In their scientific pursuit of psychopathology, clinical psychologists engage in a wide variety of tasks, all involving probability theory. They might work, for example, on the development of a classification system for abnormal behaviors, searching for conditional probabilities related to clusters of symptom patterns, common causes, expected course, or treatment response. Once they have settled on a working classification system (e.g. DSM-IV), they might search for diagnostic signs to help them diagnose, differentiate, refine, or explain the categories. In an effort to predict, prevent, or explain disorders, they might search for antecedent characteristics associated with increased risks of specific pathology. Alternatively, they might search within diagnostic groups for client characteristics that predict differential responses to different treatments.

We could go on describing various content questions that might drive the research programs and assessment interests of clinical scientists, but this sample is sufficient to help us make two important points: First, all of these examples—if performed properly—involve the clinical assessment and multidimensional quantitative modeling of conditional probabilities. Second, all of these are nomothetic activities. That is, they all represent attempts to build general models with which to increase the accuracy of predictions for groups of individuals. In general, they all address questions of the following type: p(S|D). That is, what is the probability (p) of a diagnostic sign (S), given membership in a diagnostic category (D)?

To illustrate, imagine a hypothetical study in which the Revised Hamilton Rating Scale for Depression (HRSD-R) (Riskind et al 1987) is administered to two groups, one composed of patients known to be clinically depressed, the

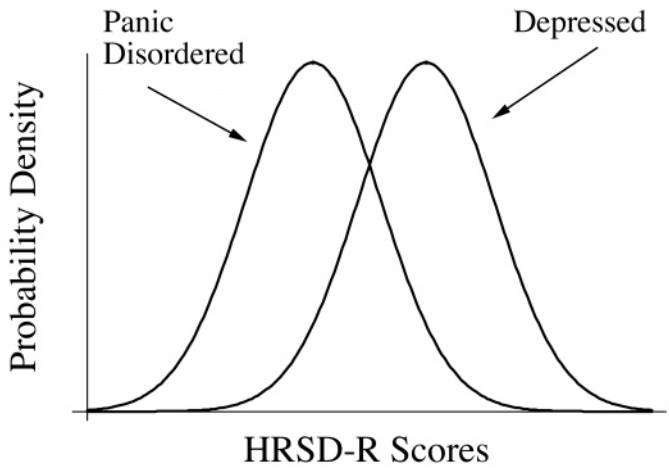

Figure 1 Hypothetical distributions of scores on HRSD-R for persons receiving panic disorder and major depression diagnoses. Prevalence of depression is assumed to be 0.5.

other made up of patients known to be suffering from panic disorder. The study's aim is to evaluate how well HRSD-R ratings differentiate depressed patients from panic patients.[4] Fictitious idealized results are displayed in Figure 1. We see a separate (conditional) probability distribution of HRSD-R ratings for each patient group. We also see that the mean of the depressed group's distribution is higher than the mean of the panic group. Now, how might we quantify and evaluate the information value of these data?

If we were to analyze these data using traditional Fisherian statistical methods (Gigerenzer et al 1989), we would test the null hypothesis; that is, we would ask, "Can we reject the hypothesis that the means of these two distributions are not significantly different from one another?" This is an odd, double-negative question. Even before collecting the data, we know that the likelihood of rejecting the null hypothesis for a given absolute difference (no matter how trivially small) between sample means increases as the sample size increases (Cohen 1994, Loftus 1996, Meehl 1978). Besides, rejecting the null hypothesis tells us nothing about which of the many plausible rival hypotheses might be supported by the results. Furthermore, traditional statistical tests shed little light on the question that gave rise to the study in the first place. They don't tell us how useful the HRSD-R is for differentiating between depressed and panic patients. Neither do correlation-based methods, as favored in the Neyman-

[4]We're ignoring for the moment another major issue—the criterion problem. That is, how does the investigator "know" for certain the patients' "true" diagnoses. We will return to it later

Pearson tradition (Gigerenzer et al 1989), although they shed some light on our research question by telling us the strength of the association between HRSD-R scores and group membership. Still, we need better methods of quantifying the information value of such data.

Cutting Scores, Base Rates, and Inverse Probabilities

Over 40 years ago, Meehl & Rosen (1955) identified three classical problems that further complicate the task of evaluating the information value of data sets. The first of these was the problem of choosing the optimal cutting score (cut-point) for differentiating between two groups. This problem is illustrated in Figure 2, where three possible cutting scores (*A* = liberal; *B* = moderate; *C* = conservative) have been applied to our fictitious HRSD-R data. Setting the cutting score at point *A* correctly identifies most of the depressed patients as depressed, but also misidentifies a high percentage of panic patients as depressed. Thus, cut-point *A* shows what epidemiologists call good sensitivity (i.e. a high true positive rate, or the proportion of depressed patients classified as depressed), but also shows poor specificity (i.e. a low true negative rate, or the proportion of panic patients classified as not depressed). Shifting the cutting score to point *C* results in the correct identification of most panic patients as not depressed, but also misidentifies a high percentage of depressed patients as not depressed. Thus, cut-point *C* shows good specificity, but poor sensitivity. This example illustrates the inevitable trade-off between sensitivity and specificity as a function of changes in the cutoff value.

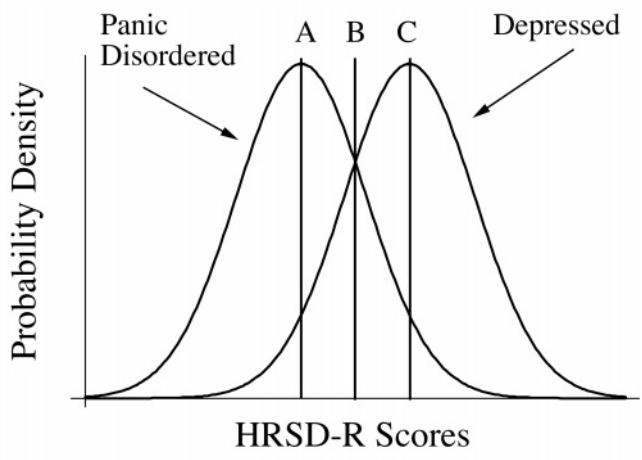

Figure 2 Hypothetical distributions of scores on HRSD-R for persons receiving panic disorder and major depression diagnoses. Prevalence of depression is assumed to be 0.5. Liberal (*A*), moderate (*B*), and conservative (*C*) cutoff values are shown.

The second problem identified by Meehl & Rosen (1955) was the base-rate problem. In short, the discriminatory power of a particular measure will vary as a function of the base rate of the variable being predicted (e.g. depression) in the population being assessed. In our imaginary data set, for example, HRSD-R ratings were obtained from equal numbers of depressed and panic patients, so the relative density of cases under the two curves was equal. Thus, the base rate, or prevalence, of depression in this fictitious study was 0.5. Suppose instead that the base rate, or prevalence, of depression had been 0.1. The hypothetical distributions of HRSD-R scores for depressed and panic patients have been redrawn in Figure 3 to illustrate the effects of these altered base rates. We can see, for example, how the problem of choosing a cutting score is compounded by the base rate problem. For a fixed cut-point (say, point *B*), the sensitivity and specificity indices will not change (i.e. the proportion of depressed persons classified as depressed will remain constant), but the practical utility of the measure will change as a result of changes in the base rate, or prevalence, of the disorder. The 9:1 ratio of panic patients to depressed patients shown in Figure 3 means that where the two distributions overlap, classification errors are far more frequent for panic patients than for depressed patients.

The third problem identified by Meehl & Rosen (1955) (see also Meehl 1973) was the logical fallacy of using nomothetic, or normative, probability distributions to make idiographic decisions and predictions. This is called the

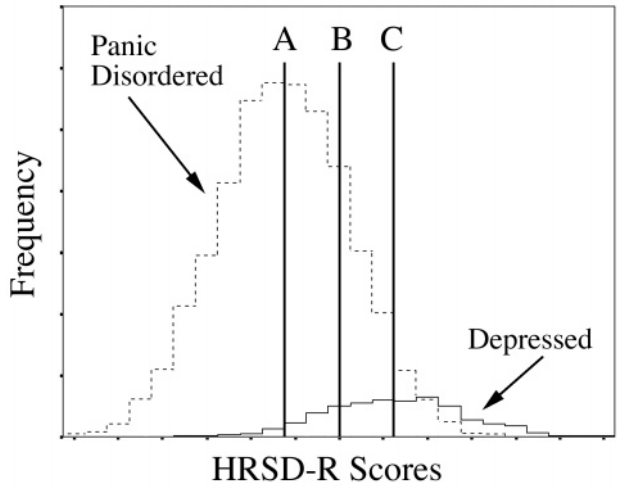

Figure 3 Hypothetical distributions of scores on HRSD-R for persons receiving panic disorder and major depression diagnoses. Prevalence of depression is assumed to be 0.1. Liberal (*A*), moderate (*B*), and conservative (*C*) cutoff values are shown.

inverse probability problem. The problem arises when clinical assessors confuse two types of probability: p($S|D$), the probability of a particular score on a diagnostic test, given membership in a diagnostic group; and p($D|S$), the probability of being a member of a diagnostic group, given a particular score on the diagnostic test. The inverse probability problem interacts with the base rate problem. When the base rate, or prevalence, of a disorder in the sample is exactly 0.5, then p($S|D$) = p($D|S$). When the prevalence is not 0.5, however, these two probabilities are not equal. The more asymmetrical the prevalence rates, the greater the inequality. Meehl and Rosen (1955) showed how Bayes' Theorem (Bayes 1763; see Gigerenzer & Murray 1987) solves the problem by controlling for base rates while using normative probability distributions to estimate inverse probabilities. Bayes' Theorem is as follows:

$$\text{p(Hypothesis} \mid \text{Data)} = \frac{\text{p(Hyp)} * \text{p(Data} \mid \text{Hyp)}}{\text{p(Data)}}$$

or, using our notation,

$$\text{p}(D|S) = \frac{\text{p}(D) * \text{p}(S|D)}{\text{p}(S)}.$$

The Bayesian approach addresses two of Meehl & Rosen's problems (base rates and inverse probabilities), but not the third (cutting scores). Selecting optimal cutting scores always requires subjective judgments regarding how best to resolve the inevitable trade-offs between sensitivity and specificity, or between the relative costs and benefits of different types of classification errors. Because cutoff decisions never can be value free, there is no magic formula for finding an absolute, all-purpose, optimal cutoff. Every formula for optimal cutoffs is based on hidden assumptions and values. Given a data set and population base rate, for example, we might choose a cutting score that maximizes overall percent correct; however, this choice assumes that the optimal solution should assign equal weights to the two types of error (i.e. false positives and false negatives). Often the costs of these two errors are unequal, however. To prevent airplane terrorism, for instance, society tolerates a very high false positive rates and treats everyone at airports as potential terrorists because society places a greater value on ensuring the highest possible true positive rate, i.e. catching the rare terrorist.

The choice of cutting scores also is influenced by the personal biases of the individuals involved. Suppose we asked three psychologists (A, B, and C) to view videotaped HRSD-R interviews of a mixed sample of depressed and panic patients, and to make a dichotomous diagnostic judgment (depressed | not depressed) for each patient. Although the three psychologists view the same interview data, their judgments are likely to differ. Such interjudge differences arise from two sources. The judges may differ in their perception of the diag-

nostic information contained in the videotapes. But even if they were equally perceptive of the presence of diagnostic cues, their judgments still could differ as a result of the cutting score, or decision criterion, that they selected for calling a patient depressed. Figures 2 and 3 depict our three judges (with 0.5 and 0.1 base rates, respectively). Judge A has been very liberal in setting the criterion for discriminating between normal and abnormal; judge B has employed a moderate criterion; and judge C has drawn the line very conservatively. Here we see how three equally perceptive judges might produce three different diagnostic results as a function of their individual biases in selecting the criterion, or cutting score, for calling a patient depressed.

Wanted: A Common Metric for Information Value

Paradoxically, Meehl & Rosen's (1955) paper on the problems of cutting scores, base rates, and inverse probabilities, with its focus on Bayes' Theorem, has been cited widely for nearly half a century, yet it has had surprisingly little impact on actual practice in clinical assessment. Over the years, major texts on clinical assessment (e.g. Mischel 1968, Wiggins 1973) have reiterated the problems and reemphasized the importance of a Bayesian solution, but with minimal added impact. This is puzzling. Clinical assessors should have been attracted to Bayes' Theorem on epistemological grounds, if not on methodological grounds. As we noted at the outset, the concept of incremental validity is central to clinical assessment and prediction. Bayes' Theorem provides a quantitative method of iteratively incrementing the accuracy of probability estimates by systematically using the information contained in each new batch of data to transform the prior model into a more precise posterior model (Schmitt 1969). In this respect, the Bayesian approach provides a solid foundation for a more rigorously quantitative approach to clinical assessment.

Unfortunately, however, the Bayesian approach provides an incomplete solution to the clinical assessor's needs. It still does not provide a common metric with which to quantify the information value of assessment data that is independent of changes in cutoffs and prevalence rates. Such a standard scale is essential if clinical assessors wish to compare the incremental validity, or relative utility, of different assessment methods. Without such a metric, assessors will have little choice but to continue using current, inadequate strategies, typically evaluating the statistical significance, relative to the null hypothesis, of the difference between group means (disordered vs normal, treated vs control). But if research in clinical assessment is to build cumulative knowledge, assessors must be able to quantify and compare the results of assessment methods across studies, populations, and conditions. Scientific progress in clinical assessment will be stymied until assessors find a way of doing this (e.g. see Meehl 1973, 1978).

A metric for the information value of data yielded by an assessment method should be a property of the method alone, not the prevalence of the disorder in the sample to which the method was applied or the decision biases or criterion choices of the assessors using the method. Although the differential information value provided by various assessment methods may guide our selection of an assessment method, practical application of the method for assessment or prediction purposes hinges in part upon these latter factors, such as prevalence and cutoff scores. For an assessment method with a specific information value, for example, we might ask, "How does the practical utility of this assessment method vary as a function of prevalence rates or selection criteria?" But examination and quantification of the practical utility of an assessment method in a particular context require separate metrics that should be independent of the metric defining the information value of the method.

In sum, future advances in clinical assessment await the development of a common metric for quantifying assessment information. We believe that signal detection theory (SDT) provides such a metric. In the next section, we outline the history, concepts, and methods of SDT. Then we review recent examples of SDT's application to a cross-section of assessment problems. We conclude with a discussion of SDT's possible limitations and its potential contributions to clinical assessment.

SIGNAL DETECTION THEORY: MEASURING MEASURES

Background

SDT's history is not the story of a theory evolving smoothly and continuously over time; rather, it is the story of a conceptual framework being reborn periodically, each incarnation more elaborated and refined than the last, but with little apparent memory of former lives (see Ashby 1992; Gigerenzer & Murray 1987; Gigerenzer et al 1989; Link 1994; Murray 1993). Historians trace the roots of contemporary SDT to Neyman & Pearson's (1933) work on hypothesis testing and statistical inference (e.g. Gigerenzer et al 1989), but the underlying probabilistic concepts can be traced back chronologically, if not genealogically, more than 200 years. The concepts were central, for example, in the prepsychology contributions of Bayes and Gauss. The concepts reappeared at psychology's beginning in the work of Fechner, who extended Gauss's "true score plus error" model to the psychophysics of sensory perception (Link 1994, Murray 1993). The concepts resurfaced again during the first half of this century in Thurstone's (1927) pioneering work on the Law of Comparative Judgment, a unidimensional probabilistic scaling model (Murray 1993). They then played a featured role in the work of Neyman & Pearson, as noted. Around midcentury they reappeared in the guise of information theory, which

included the work of engineers and physicists such as Shannon & Weaver (1949), Peterson, Birdsall, and Wiener (see Macmillan & Creelman 1991 and Pierce 1980) , as well as psychologists [e.g. Tanner, Green, and Swets (Swets 1973)]. Today, the concepts are prominent in the work of cognitive scientists such as Luce, Townsend, Ashby, Ennis, MacKay, and Zinnes (see Ashby 1992). SDT's concepts and methods also are being adopted with increasing frequency by scientists in other fields that place a premium on minimizing error through the development of assessment, prediction, and decision systems with high accuracy and discriminatory power—fields such as information retrieval, aptitude testing, psychiatric diagnosis, and medical detection and decision-making (Murphy et al 1987, Swets 1996).

It is beyond the scope of this review to provide a detailed and comprehensive summary of SDT. Our aims here are (*a*) to provide a clear overview of SDT that is more conceptual than mathematical; (*b*) to highlight SDT's relevance and potential contributions to clinical assessment in psychology; and (*c*) to stimulate and entice readers into pursuing the topic further on their own, in greater depth. Throughout, we try to supply readers with linked pointers to key resources. In the end, we hope to convince readers that there no longer is any excuse for continuing to conduct business as usual, now that SDT provides the necessary tools for comparing and evaluating the information value of our clinical assessment methods. SDT-based indices represent a clear and significant advance over traditional accuracy indices such as sensitivity, specificity, and predictive power (Hsiao et al 1989; Metz 1978; Murphy et al 1987). Moreover, the practical application of SDT methods has been enhanced by the development of methods to help determine optimal cutoff scores and to examine the influence of prevalence, or base rates, on both the selection of cutoff scores and the accuracy of estimates (Metz et al 1973, Somoza & Mossman 1991, Somoza et al 1989).

Overview of SDT

The aim of diagnostic assessment systems is to discriminate between two mutually exclusive states, such as the presence or absence of a signal. For instance, a radiologist uses X-rays to help decide whether a tumor is present or absent; a psychologist uses an HRSD-R score to help decide whether to diagnose a patient as suffering from depression or panic disorder. SDT methods partition the variability in the data produced by such diagnostic systems into two independent components: perceptual and decisional. The perceptual index is a measure of diagnostic accuracy; that is, it represents quantitatively how well the system discriminates between the two possible states. The decisional index, in contrast, represents quantitatively the position of the cutoff score, or criterion, employed to arrive at the discriminations, e.g. whether the criterion was liberal or conservative.

By providing separate indices of these perceptual and decisional components, SDT offers a significant improvement over other, more traditional methods of assessing the accuracy of diagnostic systems (e.g. percent correct; sensitivity; specificity; positive predictive power; negative predictive power). All the traditional indices confound the contributions of these two components, thereby yielding estimates of diagnostic accuracy, or discriminatory power, that are influenced by the diagnostic system's criterion for discrimination. For reasons discussed in the previous section, this is not good.

To illustrate, suppose we wanted to evaluate the accuracy of the Bulimia Inventory-Revised (BULIT-R) (Thelen et al 1996), a self-report questionnaire designed to assist in diagnosing bulimia nervosa. The gold standard, or presumably true diagnoses, against which we evaluate the accuracy of the BULIT-R, will be the diagnostic decisions of experts who conducted extensive diagnostic interviews.[5] Table 1 summarizes the relevant frequency data in a two-dimensional contingency table, where one dimension classifies cases (bulimia present or absent) based on a conventional BULIT-R cutting score of 104 or above. The second dimension classifies cases based on the experts' true diagnosis (bulimia present or absent). Displayed in the cells are raw frequencies and both traditional and SDT cell labels. Hit, false alarm, miss, and correct rejection rates are identical to true positive, false positive, false negative, and true negative rates; each is equal to the cell frequency divided by the corresponding column marginal frequency. For example, the miss rate (the probability of a true bulimia present case being identified by the BULIT-R as a bulimia absent case) is 2/23, or 0.087.

We could use the data in Table 1 to compute several traditional indices of the BULIT-R's discriminatory power, or its ability to predict true diagnoses: (*a*) Percent correct is the sum of hits and correct rejections divided by the overall sample size (140/147 = 0.952); (*b*) Sensitivity is the hit rate (or true positive rate; 21/23 = 0.913); (*c*) specificity is the correct rejection rate (or true negative rate; 119/124 = 0.960); (*d*) positive predictive power (PPP) and negative predictive power (NPP) are computed using Bayes' Theorem, as discussed before. The calculation of PPP = 0.808 is illustrated below (NPP = 0.983).

p(bulimia | score of 104+) = p(bulimia)*p(score of 104+ | bulimia)/p(score of 104+) = 0.1565*0.9130/0.1768 = 0.1429/0.1768 =0.808.

PPP and NPP also may be obtained more easily by dividing the hit or correct rejection cell frequency, respectively, by the corresponding row marginal.

[5]Meehl (1959) pointed out the silliness of using tests for the sole purpose of predicting the opinions of psychological experts. We use this gold standard because we presume that the experts' opinions represent some true state of nature. Ultimately, the validity of the gold standard must be demonstrated somehow.

Table 1 Frequencies of hits (true positives), misses (false negatives), false alarms (false positives), and correct rejections (true negatives) using a cutoff score of 104 on bulimia inventory—revised (BULIT-R) to diagnose the presence or absence of bulimia nervosa. [Adapted from Thelen et al (1996).]

	True diagnosis		
Diagnosis based upon BULIT-R	Bulimia present	Bulimia absent	Row totals
Bulimia present (score 104+)	21 hits or true positives	5 false alarms or false positives	26
Bulimia absent (score <104)	2 misses or false negatives	119 correct rejections or true negatives	121
Column totals	23	124	147

It is interesting to note that sensitivity, specificity, PPP, and NPP all are conditional probabilities. Sensitivity and specificity are conditional on the true diagnosis (i.e. they summarize the column data), whereas PPP and NPP are conditional on the diagnostic system's classification (i.e. they summarize the row data).

Each of these indices summarizes the BULIT-R's discriminatory power in this unique situation, as long as the prevalence and cutoff values are fixed. The values of all of these indices will change, however, if different cutoff values are used. In addition, the values of PPP and NPP will change if the prevalence, or base rate, changes. This is because PPP and NPP, unlike the other indices, include prevalence information in their formulas. Because the optimal cutoff value will vary, in part, as a function of prevalence, the other indices also are influenced indirectly by prevalence. Thus, all these common accuracy indices are unsatisfactory; none provides a unique measure of accuracy that is independent of cutoff (criterion, decision bias) and base rate (prevalence). None of these indices can serve as a common metric for comparing the information value and discriminatory power of the data from assessment methods.[6]

In contrast to these conventional indices of accuracy, which are unacceptable because they confound decisional and perceptual contributions to performance, SDT provides an estimate of diagnostic accuracy that is not confounded by changing cutoff values or prevalence rates. SDT estimates accuracy by analyzing the receiver (or relative) operating characteristic (ROC). Engineers originally developed ROC analysis to quantify how well an electronic receiver detects electronic signals in the presence of noise; ROC analysis ac-

[6]Other common reliability or concordance indices, such as kappa and the Y and Q statistics, also fluctuate as a function of prevalence, and thus are inadequate indices of information value (see Langenbucher et al 1996; Swets 1996, Ch. 3).

quired its name from its application to radar detection problems during World War II (Pierce 1980).

ROC analysis yields a quantitative index of accuracy corresponding to what we have been calling the information value of the data. The axes of an ROC plot are the hit and false alarm rates, and each point on an ROC curve corresponds to a pair of hit and false alarm rates that result from use of a specific cutoff value.[7] In more traditional language, the ROC curve is a plot of sensitivity against 1–specificity at all possible cutoff values. Figure 4 presents several idealized ROC curves on a single plot. As one moves along a specific ROC curve from the lower left corner (where false alarm and hit rates both are 0.0), to the upper right corner (where false alarm and hit rates both are 1.0), the cutoff changes from maximally conservative to maximally liberal. Traditional accuracy indices would vary widely as a function of such marked variability in cutoff values, but the area under the ROC curve (AUC) quantifies the information value of the assessment method independently of the cutoff value. Alternative indices in the SDT family, such as d' and d'_e, quantify the distance between the means of the two underlying distributions in standard deviation units. Both indices assume that the underlying distributions are normal, and d' also assumes that the variances of the distributions are homogeneous. Presently, AUC is the preferred SDT accuracy index because nonparametric procedures are available for estimating AUC and many users prefer a proportion-based, rather than a distance-based, measure of accuracy (Macmillan & Creelman 1991; Swets 1996).

An ROC curve that lies on the main diagonal indicates that the diagnostic system is operating at the level of chance, because the hit and false alarm rates are equal across the range of possible cutoff values. Chance performance corresponds to an AUC of 0.5. As the information value of the diagnostic system increases, the distance of the observed ROC curve from the chance line increases. Several ROC curves and their corresponding AUC values are depicted in Figure 4. The values for AUC can range from 0.0 (when the ROC curve passes from the lower left corner through the lower right corner to the upper right corner) to 1.0 (when the ROC curve passes from the lower left corner through the upper left corner to the upper right corner). AUC also has a readily interpretable probabilistic meaning: It corresponds to the probability that a randomly selected pair of observations drawn from the two underlying distributions will be ranked (and thus classified) correctly (Green & Swets 1974,

[7]In some presentations, the ROC curve is a plot of z-transformed hit and false alarm rate pairs [z(FAR), z(HR)]. In this coordinate system, the ROC curve becomes a straight line with a slope of 1.0 when the underlying distributions are normal and have homogeneous variances, whereas the ROC curve becomes a straight line with a slope other than 1.0 when the underlying distributions are normal but show nonhomogeneous variances (Swets 1996).

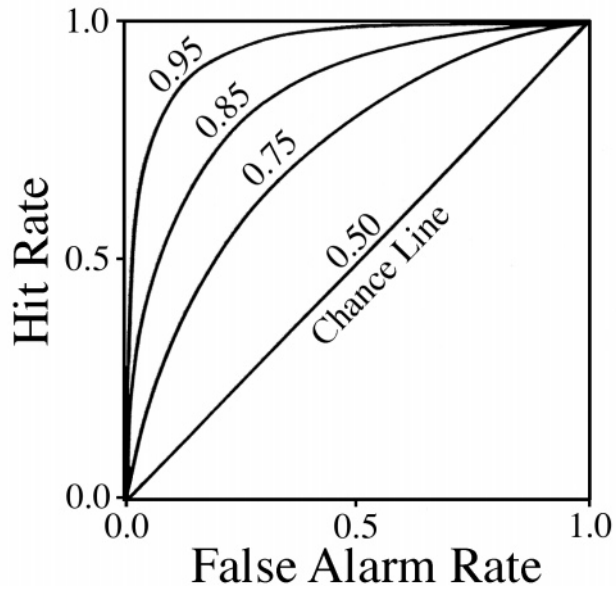

Figure 4 Idealized receiver operating characteristic (ROC) curves associated with various Area Under Curve (AUC) values. [Adapted from Swets (1988).]

Hanley & McNeil 1982). Thus, the value 1-AUC intuitively reflects the degree of overlap between the two distributions.

ROC curves can be generated in a variety of ways. First, multiple pairs of hit and false alarm rates can be calculated from a single data set by varying the cutoff. Second, the assessment method may be used repeatedly with different decision criteria employed on each occasion (i.e. from conservative to liberal). Each occasion provides a unique set of hit and false alarm rates. Third, a rating scale method may be used, in which raters not only classify the stimulus into one of two categories, but also indicate their confidence level for the accuracy of their classification, typically on a five-point scale. In this case, multiple pairs of hit and false alarm rates can be obtained by treating each confidence level as a separate cutoff value (Macmillan & Creelman 1991).

Earlier, we summarized a study by Thelen et al (1996) that examined the discriminatory power of the BULIT-R for the diagnosis of bulimia. Thelen et al presented the hit (0.808) and false alarm (0.017) rates for a BULIT-R cutoff score of 104, thus providing the coordinates for a single point on an ROC curve. However, we cannot compute an ROC curve from a single data point. This is a common limitation of published data, restricting our ability to perform ROC analyses on data from multiple studies to compare the information value of different methods.

To illustrate the benefits of ROC analyses of information value, therefore, we now turn to a study by Somoza et al (1994) that actually used SDT to examine the differential discriminatory power of six self-report measures (three for mood, three for anxiety) for the diagnosis of major depression and panic disorder. (We have been using this study as a model for our hypothetical examples up to now.) Figure 5 presents the ROC curves for the Revised Hamilton Psychiatric Rating Scale for Depression (HRSD-R), the Beck Depression Inventory (BDI) (Beck & Steer 1987), and the Beck Anxiety Inventory (BAI) (Beck et al 1988). To obtain an ROC curve for each self-report measure, Somoza et al calculated the hit and false alarm rates resulting from multiple cutoff values and used ROC programs developed by Metz et al (1973) to fit smoothed curves to these values. For the BAI, for example, they used cutoff values of 7, 13, 17, 24, and 32; these values are indicated by filled circles and are labeled *A* through *E*, respectively. Cutoff value *A* corresponds here to a very liberal criterion, which results in a substantial hit rate, but also a high false alarm rate. In contrast, the conservative cutoff value *E* results in a very low false alarm rate, but also an unimpressive hit rate. The AUC values for the HRSD-R, BDI, and BAI were .0896, 0.816, and 0.696, respectively. Further statistical analyses demonstrated that the HRSD-R does a significantly better job than the remaining five measures of discriminating between persons diagnosed with depression and panic disorder, regardless of whether a liberal, moderate, or conservative cutoff value is used. This example highlights the utility of the AUC index

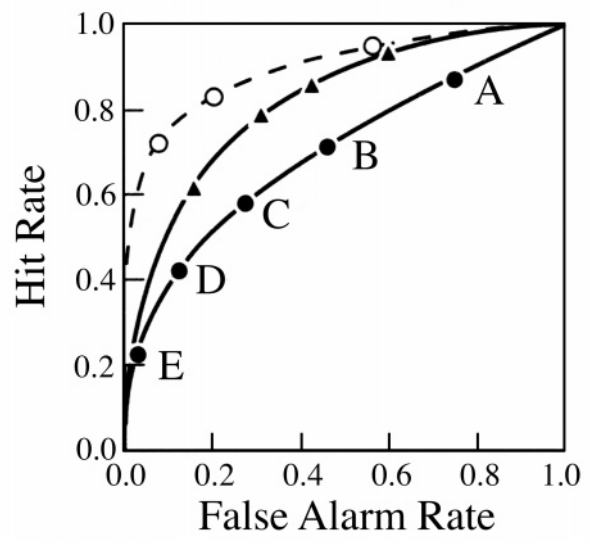

Figure 5 ROC curves for HRSD-R (*dashed curve, open circles*), BDI (*solid curve, solid triangles*), and BAI (*solid curve, solid circles*). [Adapted from Somoza et al (1994).]

as a common metric for quantifying information value independent of prevalence rates and cutoff values.

Of course, ROC models, like any mathematical model, are constrained by assumptions and limited in scope. ROC curves and AUC values typically are estimated using parametric methods only when the underlying distributions are normal and show homogeneous variances. Fortunately, parametric estimation appears to be robust to violations of these assumptions (Hanley 1988), and nonparametric estimation methods also are available when either or both of these assumptions are violated (Hanley & McNeil 1982). Both parametric and nonparametric methods allow the user to compare AUC values either to chance performance values (AUC = 0.5) or to the maximum AUC value attainable, given the prevalence rate of the phenomenon. AUC values for different diagnostic systems also can be compared statistically. Presently, ROC analysis is applicable only to unidimensional classifications into dichotomous categories, although diagnosticians often are called upon to make multidimensional classifications into more than two discrete categories. Fortunately, well-developed multidimensional generalizations of SDT exist (Ashby & Townsend 1986, Kadlec & Townsend 1992) that may be amenable to ROC analysis, and Scurfield (1996) recently generalized ROC analysis to unidimensional classifications into three or more categories.

Selection of Cutoff Values and Effects of Prevalence

Although ROC analysis provides an index of information value independent of cutoff value and prevalence, it neither provides the optimal cutoff value nor illustrates how prevalence affects cutoff selection (Hsiao et al 1989; Mossman & Somoza 1989; Murphy et al 1987; Somoza et al 1994). Selection of an optimal cutoff value necessarily involves specification of a function to be maximized. Thus, there is no true and unique optimal cutoff value. Because the usefulness of a diagnostic test in a practical setting is a function of the hit rate, false alarm rate, and prevalence of the phenomenon, researchers must consider all three factors when choosing a cutoff. The indices of percent correct, hit frequency, sensitivity, and specificity do not reflect all three factors.

There are two common approaches to selecting optimal cutoff values that incorporate these three factors in their criterion function. Meehl & Rosen (1955) and Somoza & Mossman (1991), among others, have advocated the use of an approach that combines an SDT analysis with utility-based decision theory (see also Metz 1978; Somoza et al 1989). This approach allows the user to place a differential value upon (i.e. to specify the differential utility of) hits (*H*), false alarms (*FA*), correct rejections (*CR*), and misses (*M*). Frequently, the user does not value these four possible outcomes equally because of their differential implications. As summarized in the following equation, the overall utility of a

specific cutoff value is a function of the hit and false alarm rates (HR and FAR) that result from a given cutoff value and a prevalence estimate (Pr):

$$U_{overall} = (Pr)(HR)(U_H) + (Pr)(1\text{-}HR)(U_M) + (1\text{-}Pr)(FAR)(U_{FA}) + (1\text{-}Pr)(1\text{-}FAR)(U_{CR}). \quad (2)$$

This utility approach has been criticized because it requires the user to specify quantitatively the utilities of the four outcomes, even though these often are thought of qualitatively. A user may view hits as more important than correct rejections, for example, but struggle to specify precisely how much more important hits are. Fortunately, it is possible for the user to specify a range of utility estimates rather than precise utility estimates (see Somoza & Mossman 1991). It also is important to reiterate that there is no absolute optimal cutoff value, apart from assumptions and criteria specifying the meaning of optimal. Ultimately, users have no option but to pay their money and make their choice.

To finesse the use of subjective utilities, Metz et al (1973) proposed that an information theory (Shannon & Weaver 1949) analysis of the ROC curve provides a natural criterion (information maximum, or I_{max}) for the selection of an optimal cutoff value (see also Mossman & Somoza 1989; Somoza et al 1989, 1994). According to Metz et al's approach, information refers to the reduction of uncertainty about the true classification of a person that results from administering the diagnostic measure (i.e. the difference between the prior and posterior uncertainties).

Inspection of the criterion functions specified by decision theorists and information theorists reveals that both incorporate the false alarm rate, the hit rate, and the prevalence rate, but I_{max} maximizes information, whereas $U_{overall}$ maximizes overall utility. Interestingly, $U_{overall}$ is simply a general case of I_{max}, because I_{max} simply provides an alternative specification of the utilities of the four outcomes; the two formulas are equivalent if U_H is constrained to be $\log_2(HR/B)$, U_M is constrained to be $\log_2[(1\text{-}HR)/(1\text{-}B)]$, U_{FA} is constrained to be $\log_2(FAR/B)$, and U_{CR} is constrained to be $\log_2[(1\text{-}FAR)/(1\text{-}B)]$. Thus, whereas METZ et al's (1973) approach to criterion selection sidesteps the necessity of the researcher's explicitly specifying the outcome utilities, the assumptions underlying the information theory formulation of the criterion function nonetheless exert an implicit influence on the criterion selection. There may be advantages to explicit specification of the outcomes' utilities, or at least to weighting the I_{max} terms by the more traditional utility estimates.

Regardless of the approach taken to specification of the criterion function, the user proceeds by calculating the value of the function for a wide range of cutoff values and prevalence rates (as the latter will influence the optimality of varying cutoff values). The maxima in the resulting three-dimensional topography correspond to the optimal cutoff values, given the criterion function specification. Various qualitative characteristics of this topography, such as

the number of maxima and the steepness of their surrounding areas, may provide helpful indicators of the measure's robustness under suboptimal conditions. It is important to note here that neither $U_{overall}$ nor I_{max} are indices of information value as we have defined it. Both utility and information vary as a function of prevalence and the cutoff value, so they are not equivalent to the AUC index of information value.

Somoza et al (1994), in their analysis of the relative discriminatory power of mood and anxiety measures for the diagnosis of major depression and panic disorder, used the I_{max} criterion function to select the optimal cutoff values for each of the six measures for varying prevalence rates. To illustrate the impact of the cutoff score and prevalence on I_{max} for each of the six measures, Somoza et al present three two-dimensional figures (I_{max} by cutoff score, I_{max} by prevalence, and prevalence by cutoff score). We discuss only the HRSD-R results below. The steepness of the criterion function around the maximal cutoff score in their I_{max} by cutoff figure indicates that the practical utility of the HRSD-R decreases rapidly as the cutoff deviates from its optimal value. It also is interesting to note in their I_{max} by prevalence figure that minimal values of the criterion function are much more likely when the prevalence rate is extreme, whereas maximal values are more likely when the prevalence rate is nearer 0.5. Thus, these figures illustrate quantitatively what Meehl & Rosen (1955) described long ago. Ideally, of course, we would like to depict I_{max} for the HRSD-R as a function of cutoff score and prevalence simultaneously, in a three-dimensional rendition of their three separate two-dimensional figures, as cutoff values and prevalence exert interactive influences as well as independent influences on the value of the criterion function.

Specific SDT Applications

Swets (1988, 1996) summarized the use of ROC methods to evaluate the diagnostic performance of assessments in various fields outside of clinical psychology. In clinical medicine, for example, ROC methods are used to quantify both the discriminatory power of medical imaging techniques (for the detection of pathology) and the decision criteria used by individual interpreters. In the field of aptitude testing, ROC analyses are used to evaluate the validity of various aptitude indices for predicting dichotomous outcomes, such as satisfactory or unsatisfactory school or work performance, regardless of whether the criterion used to distinguish the two outcomes is conservative, moderate, or liberal. ROC methods also are used to evaluate the performance of various information retrieval methods independent of the criteria used for inclusion of information.

A systematic search of the empirical literature in clinical psychological assessment, however, yielded surprisingly few published examples of SDT's ap-

plication to real-world clinical problems. Thus, the aim of this section is not to provide a detailed and exhaustive summary of all studies that have used SDT; rather, the aim is simply to give readers a sense of the range or diversity of problems that can be analyzed by SDT methods.

ROC methods have been used most extensively to evaluate the utility of laboratory tests or questionnaires for discriminating between diagnostic classes or between disordered and nondisordered persons. Several studies of this type are noteworthy for their methodological rigor and conceptual clarity. Our explication of SDT's theoretical and methodological foundations has drawn heavily from one leading example from this class of applications—Somoza et al's (1994) use of SDT to quantify and compare the relative ability of three depression measures to differentiate between samples of depressed patients and panic disorder patients. In that same paper, they also describe the use of SDT to quantify and compare the ability of three anxiety measures to differentiate between the same samples of diagnostic groups. In an earlier study, Mossman & Somoza (1989) used ROC methods to evaluate the literature on the utility of the dexamethasone suppression test (DST) for discriminating between depressed and nondepressed persons. Although AUC indices suggested moderate discriminatory power for the DST across studies, the optimal cutoff (as assessed by the I_{max} criterion function) varied widely across studies and as a function of prevalence, demonstrating clearly the nonexistence of a context-free optimal cutoff value. Similarly, Battaglia & Perna (1995) used ROC analyses to contrast the discriminatory power of two laboratory assessments of panic disorder and provided optimal cutoffs (using the I_{max} criterion function) for their particular prevalence rates (although not for others). Finally, Somoza & Mossman (1991) quantified the adequacy of REM latency for discriminating between depressed and nondepressed persons, and illustrated the use of utility-based decision theory for selecting optimal cutoff values as a function of prevalence.

Although diagnostic status has been the primary criterion variable investigated using ROC methods, several researchers have used, or have suggested using, ROC methods to evaluate clinical assessment and prediction across a wide range of criterion variables, including the presence of child maltreatment (Camasso & Jagannathan 1995), the likelihood of suicide attempts (Erdman et al 1987), decisions about whether to remove a child from a home (Dalgleish 1988), risk of future disorder (Olin et al 1995), the presence of violence (Mossman 1994), violence recidivism (Rice & Harris 1995), malingering (Mossman & Hart 1996), treatment relapse (Marder et al 1991), and treatment response (Ackerman et al 1996). In each instance, ROC methods provided an improved estimate of predictive power, relative to traditional methods, that resulted from their independence from cutoff values and their attention to the impact of prevalence on the optimal cutoff values.

Future Directions

SDT is a theory-based method of quantifying the performance of diagnostic systems. We are aware of no competing methods that are as well developed, powerful, promising, or enduring. Indeed, the probabilistic concepts underlying SDT are at least as old as psychology itself. Nevertheless, our review of the assessment literature uncovered surprisingly few empirical reports of clinical psychologists employing SDT. In contrast, we found that the method has been discovered by diagnosticians and decision makers in other fields, such as medicine, aptitude testing, and information retrieval systems. Given SDT's demonstrated value in these other fields, we found it all the more puzzling that clinical psychologists still have not adopted SDT as their primary method for evaluating and comparing competing clinical assessments (and interventions). At this point, we only can speculate about the possible reasons.

Two related factors may be (*a*) the quantitative demands of SDT, and (*b*) the conceptual demands of classical probability theory upon which it is based. These features may be intimidating to clinical psychologists whose quantitative training has been limited to psychology courses in traditional statistical methods. Even psychologists who study this approach may find it elusive. Sedlmeier (1997) reported that past efforts to teach Bayesian inference, for example, have achieved disappointing results; this way of thinking does not seem to stick, for some reason. This certainly is consistent with the apparent lack of practical impact that Meehl & Rosen's (1955) widely cited paper has had over the years. To overcome this problem, Sedlmeier (1997) has developed a computerized tutorial (BasicBayes) that has shown promise.

Another factor may be that SDT has not gone without criticism. Like all theories, it is based on assumptions that sometimes may not be appropriate. In its original form, for example, SDT assumed that the variable used to discriminate between two states (e.g. the test score used to distinguish diseased from healthy) was normally distributed within each state; however, this assumption is not always valid. As it turns out, investigators have explored the implications of non-normal distributions, and have found that AUC analyses tend to be robust, even when the normality assumption is violated (Hanley 1988). These same investigators also have introduced nonparametric methods of analyzing AUC that do not require this assumption, but still yield similar results (Hanley & McNeil 1982).

Yet another possible criticism of SDT is that it is limited to diagnostic prob0lems involving dichotomous decisions. This is a lesser concern than it might appear at first. Virtually any diagnostic task—whether dimensional or categorical—can be recast as a dichotomous problem. For analysis of those infrequent diagnostic problems that absolutely require more elaborate structures, more elaborate multidimensional modeling methods that generalize

unidimensional SDT have been developed (e.g. general recognition theory) (see Ashby & Townsend 1986; Kadlec & Townsend 1992; see also Macmillan & Creelman 1991 for information on a related method based upon choice theory).

Swets (1988) has identified several other possible threats to the reliability and validity of SDT as a method of quantifying the information value and accuracy of diagnostic tests. One is the so-called gold standard problem. If we cannot determine with certainty for every case in our sample the true state, that is, whether each case is positive or negative, then we cannot possibly expect SDT to provide a valid evaluation of a test's accuracy. For example, how can the discriminatory power of polygraph tests in real-world criminal cases be determined if the true guilt or innocence of each case is uncertain? Another problem arises when the assessment system and the determination of actual truth are not independent. For example, if the gold standard in criminal cases is defined by criminals' confessions, and the polygraph test is used to predict guilt or innocence, then the predictive system may contaminate the truth because confessions may be more likely after the polygraph has indicated guilty. It is also a problem when the procedures for determining the gold standard influence the selection of cases for the test sample. In general, methodological concerns about the representativeness of samples are just as critical to the evaluation of diagnostic test accuracy as they are in any other clinical research. Swets (1988) emphasized, however, that none of these problems is due to weaknesses in SDT; all stem from inadequacies in our tests and in our ability to determine truth.

Clinical psychology's failure to discover SDT over the years cannot be blamed on lack of access to the method. Several authors have made extensive efforts to promote an awareness and understanding of SDT's value as a method of evaluating diagnostic systems [e.g. see Meehl's collected works (1973); the collected papers of Swets (1996); the last in a series of seven papers by Somoza & Mossman in the *Journal of Neuropsychiatry and Clinical Neurosciences* (1992)]. Clinical assessors simply have not responded to these efforts with appropriate enthusiasm.

The time is ripe, however, for clinical psychologists at long last to acquire the requisite knowledge and skills to employ SDT methods (and their multidimensional cousins, when necessary) as the standard benchmark system for evaluating and comparing the incremental validity and accuracy of clinical assessment methods. For example, Langenbucher et al (1996) emphasized the importance of comparing empirically the classification results yielded by competing nosologies (e.g. DSM-IV vs ICD-10). For all the reasons outlined in this review, SDT is the obvious method of choice for such comparisons.

Clinical psychologists should be able to pursue on their own the use of SDT in clinical assessment. A number of excellent resources are available for this.

We recommend the following resources, which we have listed in order from the most accessible and general overviews to the most demanding theoretical and quantitative analyses: (*a*) Somoza and Mossman (1990—first in a series in the same journal); (*b*) Swets (1988); (*c*) Murphy et al (1987); (*d*) Hsiao et al (1989); (*e*) Mossman & Somoza (1989); (*f*) Metz (1978); (*g*) Swets (1996); (*h*) Macmillan & Creelman (1991).

We acknowledge that some clinical psychologists may find daunting the up-front investment required to retool as experts in signal detection theory, and to reconceptualize the clinical assessment enterprise from this new perspective. The long-term benefits of such an investment promise to be well worth these up-front costs. This review began with descriptions of conceptual and methodological problems that have stymied scientific progress in clinical assessment. The review then introduced signal detection theory as a solution to many of those problems, and gave specific examples of SDT's successful application to similar problems across a range of fields. We have given readers a bite of the SDT apple. It is our hope that, having tasted of this knowledge, readers will not find it easy to return to old, inferior ways of evaluating clinical assessment methods. For those willing to make the effort to learn and adopt SDT methods, the future of clinical assessment should look brighter.

Literature Cited

Ackerman DL, Greenland S, Bystritsky A. 1996. Use of receiver-operator characteristic (ROC) curve analysis to evaluate predictors of response to clomipramine therapy. *Psychopharmacol. Bull.* 32:157–65

Ashby FG, ed. 1992. *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Erlbaum

Ashby FG, Townsend JT. 1986. Varieties of perceptual independence. *Psychol. Rev.* 93:154–79

Battaglia M, Perna G. 1995. The 35% CO_2 challenge in panic disorder: optimization by receiver operating characteristic (ROC) analysis. *J. Psychiatr. Res.* 29:111–19

Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London* 53:370–418

Beck AT, Epstein N, Brown G, Steer RA. 1988. An inventory for measuring clinical anxiety: psychometric properties. *J. Consult. Clin. Psychol.* 56:893–97

Beck AT, Steer RA. 1987. *Manual for the Revised Beck Depression Inventory*. San Antonio, TX: Psychol. Corp.

Camasso MJ, Jagannathan R. 1995. Prediction accuracy of the Washington and Illinois risk assessment instruments: an application of receiver operating characteristic curve analysis. *Soc. Work Res.* 19:174–83

Cohen J. 1994. The earth is round ($p<.05$). *Am. Psychol.* 49:997–1003

Dalgleish LI. 1988. Decision making in child abuse cases: applications of social judgment theory and signal detection theory. In *Human Judgment: The SJT View*, ed. B Brehmer, CRB Joyce, 54:317–60. Amsterdam: Elsevier

Erdman HP, Greist JH, Gustafson DH, Taves JE, Klein MH. 1987. Suicide risk prediction by computer interview: a prospective study *J. Clin. Psychiatry* 48:464–67

Gigerenzer G, Murray DJ. 1987. *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge, UK: Cambridge Univ. Press

Green DM, Swets JA. 1974. *Signal Detection Theory and Psychophysics*. Huntington, NY: Krieger. 2nd ed. Reprint. New York: Wiley, 1966. 1st ed.

Hanley JA. 1988. The robustness of the "binormal" assumptions in fitting ROC curves. *Med. Decis. Mak.* 8:197–203

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29–36

Hsiao JK, Bartko JJ, Potter WZ. 1989. Diagnosing diagnoses. *Arch. Gen. Psychiatry* 46:664–67

Kadlec H, Townsend JT. 1992. Signal detection analyses of dimensional interactions. In *Multidimensional Models of Perception and Cognition*, ed. FG Ashby, pp. 181–227. Hillsdale, NJ: Erlbaum

Langenbucher J, Labouvie E, Morgenstern J. 1996. Measuring diagnostic agreement. *J. Consult. Clin. Psychol.* 64:1285–89

Link SW. 1994. Rediscovering the past: Gustav Fechner and signal detection theory. *Psychol. Sci.* 5:335–40

Loftus GR. 1996. Psychology will be a much better science when we change the way we analyze data. *Curr. Dir. Psychol. Sci.* 5: 161–71

Macmillan NA, Creelman CD. 1991. *Detection Theory: A User's Guide*. Cambridge, UK: Cambridge Univ. Press

Marder SR, Mintz J, Van Putten T, Lebell M, Wirshing WC, Johnston-Cronk K. 1991. Early prediction of relapse in schizophrenia: an application of receiver operating characteristic (ROC) methods. *Psychopharmacol. Bull.* 27:79–82

McFall R. 1993. The essential role of theory in psychological assessment. In *Improving Assessment in Rehabilitation and Health*, ed. RL Glueckauf, LB Sechrest, GR Bond, EC McDonel, pp. 11–32. Newbury Park, CA: Sage

Meehl PE. 1959. Some ruminations on the validation of clinical procedures. *Can. J. Psychol.* 13:102–28

Meehl PE. 1971. High school yearbooks: a reply to Schwarz. *J. Abnorm. Psychol.* 77: 143–48

Meehl PE. 1973. *Psychodiagnosis: Selected Papers*. New York: Norton

Meehl PE. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.* 46:806–34

Meehl PE, Rosen A. 1955. Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychol. Bull.* 52:194–216

Metz CE. 1978. Basic principles of ROC analysis. *Semin. Nucl. Med.* 8:283–98

Metz CE, Goodenough DJ, Rossmann K. 1973. Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology* 109:297–303

Mischel W. 1968. *Personality and Assessment.* New York: Wiley

Mossman D. 1994. Assessing predictions of violence: being accurate about accuracy. *J. Consult. Clin. Psychol.* 62:783–92

Mossman D, Hart KJ. 1996. Presenting evidence of malingering to courts: insights from decision theory. *Behav. Sci. Law* 14:271–91

Mossman D, Somoza E. 1989. Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. *Arch. Gen. Psychiatry* 46:653–60

Murphy JM, Berwick DM, Weinstein MC, Borus JF, Budman SH, Klerman GL. 1987. Performance of screening and diagnostic tests. *Arch. Gen. Psychiatry* 44: 550–55

Murray DJ. 1993. A perspective for viewing the history of psychophysics. *Behav. Brain Sci.* 16:115–86

Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London Ser. A* 231:289–337

Olin SS, John RS, Mednick SA. 1995. Assessing the predictive value of teacher reports in a high risk sample for schizophrenia: a ROC analysis. *Schizophr. Res.* 16:53–66

Pierce JR. 1980. *An Introduction to Information Theory: Symbols, Signals and Noise*. New York: Dover. 2nd rev. ed.

Popper K. 1962. *Conjectures and Refutations.* New York: Basic Books

Rice ME, Harris GT. 1995. Violent recidivism: assessing predictive validity *J. Consult. Clin. Psychol.* 63:737–48

Riskind JH, Beck AT, Brown G, Steer RA. 1987. Taking the measure of anxiety and depression: validity of the reconstructed Hamilton scales. *J. Nerv. Ment. Dis.* 175: 474–79

Schmitt SA. 1969. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Reading, MA: Addison-Wesley

Scurfield BK. 1996. Multiple-event forced-choice tasks in the theory of signal detectability. *J. Math. Psychol.* 40:253–96

Sedlmeier P. 1997. BasicBayes: a tutor system for simple Bayesian inference. *Behav. Res. Methods Instrum.* 29:328–36

Shannon CE, Weaver W. 1949. *The Mathematical Theory of Communication*. Urbana: Univ. Ill. Press

Somoza E, Mossman D. 1990. Introduction to neuropsychiatric decision making: binary diagnostic tests. *J. Neuropsychol. Clin. Neurosci.* 2:297–300

Somoza E, Mossman D. 1991. "Biological markers" and psychiatric diagnosis: risk-benefit balancing using ROC analysis. *Biol. Psychiatry* 29:811–26

Somoza E, Mossman D. 1992. Comparing diagnostic tests using information theory: the INFO-ROC technique. *J. Neuropsychol. Clin. Neurosci.* 4:214–19

Somoza E, Soutullo-Esperon L, Mossman D. 1989. Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. *Int. J. Biomed. Comput.* 24:153–89

Somoza E, Steer RA, Beck AT, Clark DA. 1994. Differentiating major depression and panic disorders by self-report and clinical rating scales: ROC analysis and information theory. *Behav. Res. Ther.* 32: 771–82

Swets JA. 1988. Measuring the accuracy of diagnostic systems. *Science* 240:1285–93

Swets JA. 1996. *Signal Detection Theory and ROC Analysis in Psychological Diagnostics: Collected Papers*. Mahwah, NJ: Erlbaum

Thelen MH, Mintz LB, Vander Wal JS. 1996. The bulimia test—revised: validation with DSM-IV criteria for bulimia nervosa. *Psychol. Assess.* 8:219–21

Thurstone LL. 1927. A law of comparative judgment. *Psychol. Rev.* 34:273–86

Wiggins JS. 1973. *Personality and Prediction: Principles of Personality Assessment.* Reading, MA: Addison-Wesley