CHAPTER 37

MEASURING TEST PERFORMANCE WITH SIGNAL DETECTION THEORY TECHNIQUES

Teresa A. Treat and Richard J. Viken

The development and evaluation of assessment and prediction strategies designed to distinguish two mutually exclusive states are central enterprises in psychological science. For example, we might want to assess diagnostic status (present or absent) or child maltreatment (present or absent). Alternatively, we might be interested in predicting whether violence is likely or whether treatment relapse will occur. Once classic psychometric methods have been used to develop one or more assessment devices, the predictive or criterion validity of the measurement strategies must be evaluated (see Clark & Watson, 1995; Smith, 2005). Widely used indexes of test performance include Sensitivity (the proportion of positive cases correctly classified as positive), Specificity (the proportion of negative cases correctly classified as negative), Positive Predictive Power (the proportion of cases classified as positive who actually are positive), and Negative Predictive Power (the proportion of cases classified as negative who actually are negative). However, these indexes vary widely as a function of cutoff scores, the base rates (BRs) of the phenomenon of interest, and the costs and benefits associated with a particular assessment or prediction context, as we will see.

As a result, researchers increasingly are relying on the methods of signal detection theory, particularly receiver operating characteristic (ROC) analysis and utility-based decision theory approaches. ROC methods can be used to quantify and compare the discriminative power of measurement devices independently of cutoff scores, BRs, and costs and benefits. Decision theory methods then can be used to determine optimal cutoff scores for particular contexts, given specification of the BRs and the values placed on different kinds of correct and incorrect decisions.

After presenting background on the role of BRs in assessment and prediction as well as traditional accuracy indexes, we provide an overview of the use of ROC and decision-theory approaches for examination and enhancement of decision making in psychology. We close with recommendations regarding the reporting of the development of new measures, especially with regard to optimal cutoff values for a range of BRs and several common decision goals.

TRADITIONAL INDEXES OF TEST PERFORMANCE

In this chapter, we will illustrate issues in evaluating test performance with a data set from the National Comorbidity Survey Replication (NCS-R; Kessler & Merikangas, 2004; Kessler et al., 2004). The NCS-R is a nationally representative survey of English-speaking adults in the United States that was conducted from 2001 to 2003. *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV;* American Psychiatric Association, 1994) diagnoses were determined for 9,282 respondents by the World Mental Health Composite International Diagnostic Interview, a structured diagnostic interview administered by lay persons (see Kessler & Üstün, 2004). The K6 is six-item screening scale that was completed by 6,656 of the participants in the data set, which we use as our predictive test. The K6 contains questions

DOI: 10.1037/13619-038
APA Handbook of Research Methods in Psychology: Vol. 1. Foundations, Planning, Measures, and Psychometrics, H. Cooper (Editor-in-Chief)

about the frequency with which various aspects of psychological distress were experienced during the respondent's worst month emotionally in the past year (Kessler et al., 2002, 2003). Respondents indicated how often they felt worthless, depressed, restless, hopeless, nervous, and that everything was an effort. Responses were made on a five-point scale ranging from 1 = *all the time* to 5 = *none of the time.* Responses were summed to obtain a total score on both scales, with lower scores indicating greater psychological distress. The large size and the high quality of the NCS-R sample make it an excellent database to provide examples for the methods described in this chapter. The K6 data in the sample are not missing at random with respect to the full sample, so the results of the analyses in this paper should be considered illustrative only.

Evaluating the performance of the K6 screening scale requires selection of a "gold standard" of psychological distress. The gold standards used in test evaluation typically are higher quality or more expensive indicators of the phenomenon of interest, although they still may contain error (Swets, Dawes, & Monahan, 2000). In the current analyses, we used the presence or absence of the 12-month *DSM–IV* anxiety and mood disorders that were assessed for all respondents and included in the current public release of the NCS-R data set: agoraphobia with or without panic disorder, generalized anxiety disorder, panic disorder, specific phobia, social phobia, major depression, dysthymia, and bipolar I and II disorders. We considered participants to be positive for a disorder if they met criteria for at least one of these disorders. This chapter examines the extent to which the K6 scale provides far lower cost and less time-intensive assessments of psychological distress than the gold standard.

Figure 37.1 provides an initial look at the association between the interview-based diagnostic outcomes and the K6 screen. The two interview-based outcomes (Disorder Present or Absent) are depicted in the rows of the figure. The two K6-based predictions (Disorder Present or Absent) are depicted in the columns of the figure. Of the 6,656 subjects with

		K6-Based Classification		
		Disorder Present (K6 < 22)	Disorder Absent (K6 > 22)	
Interview-Based Classification ("Truth")	Disorder Present	Valid Positives (Hits) Frequency = 911 Percent = 13.7%	False Negatives (Misses) Frequency = 988 Percent = 14.8%	Base Rate Frequency = 1,899 Percent = 28.5%
	Disorder Absent	False Positives (False Alarms) Frequency = 509 Percent = 7.7%	Valid Negatives (Correct Rejections) Frequency = 4,248 Percent = 63.8%	100 – Base Rate Frequency = 4,757 Percent = 71.5%
		Selection Ratio Frequency = 1,420 Percent = 21.3%	100 – Selection Ratio Frequency = 5,236 Percent = 78.7%	Frequency = 6,656 Percent = 100%

FIGURE 37.1. Matrix of interview-based classifications by K6-based classifications for National Comorbidity Survey Replication data set. The percentages in the figure do not sum perfectly because of rounding issues.

data on both the screen and the interview, 1,899 (28.5%) met criteria for at least one of the disorders. The percentage (or proportion) of the sample meeting criteria for a disorder according to the gold-standard interview is referred to as the *base rate* (BR) or, in conventional clinical terms, the prevalence of the disorder. This BR is reflected in the BR entry to the far right in the Disorder Present row of the figure. With 28.5% of the sample meeting criteria for a disorder, this means that 71.5% (100 – BR) did not. All technical terms used in the chapter are listed chronologically with a brief definition in Exhibit 37.1.

The columns of Figure 37.1 depict the predictions made on the basis of the K6. In this example, we used a cutoff value on the K6 of 22: Anyone with a score of 22 or lower (recall that lower K6 scores indicate more distress) is predicted to have a disorder. This value was selected for illustrative purposes because it optimizes the percentage of correct classifications in the current sample. The last entry in the first column shows that 1,420 (21.3%) of K6 respondents were predicted to have a disorder. The percentage (or proportion) of a sample predicted to have the characteristic of interest is often called the Selection Ratio because in many practical applications of test prediction this group is being selected for further action. In the current context, for example, respondents scoring at or below 22 might receive further evaluation or referrals for treatment. The remaining 78.7% of the sample (100 – Selection Ratio) is predicted not to have a disorder.

The shaded cells of Figure 37.1 provide the core information about test performance. In this sample, 911 (13.7%) persons were predicted to have a disorder on the basis of the K6 *and* were found to have a disorder on the basis of the interview-based "truth." In clinical prediction contexts, such persons would be referred to as the Valid Positive, or True Positive, cases, and in the signal detection theory context, they would be referred to as Hits. The remaining 509 respondents who were predicted to have a disorder on the basis of the K6 did not, in truth, have a disorder. In clinical prediction contexts, these

Exhibit 37.1
Glossary of Technical Terminology

Base rate (or Prevalence): Percentage (or proportion) of cases identified by the gold standard as positive.
Cutoff (or Threshold): Value of assessment or prediction measure that distinguishes cases classified as positive and negative.
Selection ratio: Percentage (or proportion) of cases classified as positive.
Valid positives (or Hits): Cases identified by the gold standard as positive who are classified as positive.
Valid negatives (or Correct rejections): Cases identified by the gold standard as negative who are classified as negative.
False negatives (or Misses): Cases identified by the gold standard as positive who are classified as negative.
False positives (or False alarms): Cases identified by the gold standard as negative who are classified as positive.
Percent correct: Percentage of correctly classified cases.
Percent correct by chance: Percentage of cases that can be classified correctly by chance.
Predicting from the base rate: Predicting the more frequently occurring outcome for all cases.
Sensitivity: Proportion (or percentage) of positive cases correctly classified as positive.
Specificity: Proportion (or percentage) of negative cases correctly classified as negative.
Positive predictive power: Proportion of cases classified as positive who actually are positive.
Negative predictive power: Proportion of cases classified as negative who actually are negative.
Hit rate (or Valid positive rate): Probability of correctly classifying a positive case as positive.
False alarm rate (or False positive rate): Probability of correctly classifying a negative case as positive.
Area under the curve: The probability that a randomly selected pair of positive and negative cases will be ranked correctly by the assessment method.
Utility: Value placed on a specific decision-making outcome (i.e., user-perceived benefit or cost).
Overall utility: A utilities-weighted sum of the probabilities of the four decision-making outcomes.
Utility ratio: User-perceived relative importance of decisions about negative versus positive cases.
Information gain: The reduction of uncertainty about the true classification of a case that results from administering an assessment or prediction measure.

respondents would be referred to as the False Positive cases, and in signal detection theory contexts, they are referred to as False Alarms. Among those predicted to have no disorder on the basis of the K6, 988 (14.8%) were found actually to have a disorder present on the basis of the interview. In clinical prediction contexts, such persons are referred to as False Negatives, whereas in signal detection theory, they would be referred to as Misses. The 4,248 remaining people who were predicted to have no disorder on the basis of the K6 (63.8%) were indeed found to have no disorder as judged by the interview. These respondents are referred to as Valid Negative or True Negative cases in clinical prediction, and as Correct Rejections in signal detection theory.

One of the first things to evaluate in a table like this is the percentage of cases for which the K6-predicted classification was correct. There are two ways to be correct: Valid Positive and Valid Negative classifications. We can add the percentage of respondents in these two cells (13.7 and 63.8) to find that the K6-based prediction was correct 77.5% of the time. Although 77.5% accuracy sounds pretty good, it is important to compare percentage correct when using the K6 predictor to the percentage correct expected by chance. Conceptually, percentage correct by chance would be equivalent to making our predictions on the basis of a random process like a set of coin tosses rather than on the basis of the predictor. How often would we be correct if we randomly assigned 21.3% (the same percentage reflected in the Selection Ratio used for the K6) of participants to a prediction of Disorder Present? We can compute this expected percentage correct by considering the marginal percentages associated with each of the two ways of being correct. The percentage of Valid Positives expected by chance will be the product of the BR and Selection Ratio (i.e., 28.5% × 21.3% = 6.1%) because the BR and Selection Ratio are the marginal percentages associated with the Valid Positive cell in Figure 37.1. The percentage of Valid Negative cases expected by chance is the product of (100 – BR) and (100 – Selection Ratio; i.e., 71.5% × 78.7% = 56.2%). Summing these two ways of being correct (6.1 + 56.2) gives us an expected percent correct by chance of 62.3%. Thus, in our example, it does appear that the K6 is modestly more accurate than expected by chance (77.5% versus 62.3%). As first discussed by Meehl and Rosen (1955), the lower the BR is the more difficult it will be to make predictions that are better than chance. For instance, if the actual BR or prevalence of Disorder in the current example were 5% rather than 28.5%, then the expected percent correct by chance would be 75.9% [i.e., (5% × 21.3%) + (95% × 78.7%)]. Expected percent correct by chance is an important baseline against which to judge test performance.

There is another way to consider the effects of BR on our success in making accurate predictions on the basis of tests (Meehl & Rosen, 1955). As the BR of an outcome decreases, it becomes easier to obtain a high degree of accuracy just by predicting that no one will be in the affected group. In our current example, we would be accurate 71.5% of the time just by predicting that the disorder will never be present (in essence, setting the Selection Ratio to zero). We will always be wrong for respondents who do develop a disorder (the percentage of Valid Positive cases will be zero, because no positive cases are predicted to develop a disorder), but we will always be right for the far more numerous respondents who do not have a disorder (the percentage of Valid Negative cases will be 71.5, because all negative cases are predicted not to develop a disorder). This strategy is sometimes called *predicting from the base rate.* If the BR of the disorder were 5%, then we could achieve 95% accuracy just by predicting that no one will develop the disorder. It will be difficult to find a real predictor that can match the 95% accuracy obtained by predicting from the BR. This BR problem is a particular challenge in psychology, where many of the phenomena of interest have low BRs (e.g., violence, abuse, dementia, resilience), frequently prompting researchers to conduct studies with high-risk populations for which the BRs are higher.

The problem with measures of overall percentage correct (whether observed or expected by chance) is that they treat different kinds of correct predictions and different kinds of errors as though they are equal in importance. This assumption of equal importance will rarely be true. For instance, in a clinical prediction setting, because successfully

recognizing a disorder can lead to appropriate treatment, we might place high value on maximizing the percentage of Valid Positive cases and minimizing the percentage of False Negative cases. At the same time, we may view the cost of False Positives to be relatively low, consisting primarily of the time and expense it takes to follow up with our gold standard assessment. Once we begin placing different values on the four cells in the Prediction × Outcome matrix, overall percentage correct is no longer a good index of our success. In the example in Figure 37.1, 13.7% of the sample are Valid Positive cases, more than twice the 6.1% that would be expected by chance and much more than 0% we would identify by assuming that no one will have a disorder (i.e., setting the Selection Ratio to zero). If maximizing Valid Positives is important to us, then the test will do much better than the other strategies, because a far greater percentage of Valid Positive cases will be identified. Thus, even when low BRs make it difficult to achieve better than chance accuracy, or better accuracy than predicting from the BR, a test can still be useful if it can help us to exchange certain types of errors for others. In most assessment and prediction settings, it is the *profile* of correct predictions and errors that matters most, not overall percentage correct.

There are several indexes of test performance that recognize our interest in particular types of correct predictions and particular types of errors. *Sensitivity* (expressed as either a proportion or a percentage) is an index that focuses attention on our accuracy in correctly predicting disorder among those who have a disorder. It is based on the Disorder Present row of the matrix for the interview-based classification, and it is computed as the Valid Positive Percent / BR Percent. In the current example, Sensitivity is relatively low: less than half (48.1%) of the individuals who had a disorder according to the interview were predicted to have a disorder on the basis of a K6 score of 22 or below. *Specificity* (expressed as either a proportion or a percentage) is an index that focuses attention on our ability to avoid mistaken predictions of disorders among individuals in whom disorders are absent. Specificity is based on the Disorder Absent row of the matrix and is computed as Valid Negative Percent/(100 – BR Percent). With a cutoff of 22 or lower on the K6, Specificity in this example is very high at .892.

Although Sensitivity and Specificity reflect, in part, the accuracy of a predictive instrument, they are also strongly influenced by the cutoff or threshold at which we predict that a disorder will be present. If we were to increase the cutoff score in our example from a K6 score of 22 or less to a score of 26 or less (thereby including individuals with less severe K6 scores among those predicted to have a disorder), our Sensitivity will increase, and our Specificity will decrease. To see why this is so, consider Figure 37.2, which shows frequency distributions of K6 scores for the sample described in Figure 37.1, split into people with no diagnosis (white bars) and people with at least one mood or anxiety disorder (black bars). Consider the cutoff of 22 or lower, which is the basis of Figure 37.1. Clearly, most individuals without a diagnosis have scores higher than 22, which is reflected in the high Specificity of the K6 using this cutoff. Although individuals with diagnoses predominate in the part of the distribution at or below 22, it is obvious that about half of diagnosed individuals actually have K6 scores higher than 22. This is reflected in the relatively low Sensitivity of the K6 at this cutoff level. If we were to raise our cutoff to 26 (moving to the right on the *x*-axis and including individuals who show less extreme responses to the K6), we would increase Sensitivity from .481 to .758. The reason is obvious in Figure 37.2. By moving our cutoff to the right, we include all of the additional diagnosed individuals with scores between 22 and 26 on the K6. These individuals, who previously were False Negatives, now are converted to Valid Positives. But this increase in true positives comes at a cost, because setting the cutoff K6 score to 26 means that we are also including many individuals who do not have diagnoses. Indeed, most of the people added by moving our cutoff from 22 to 26 are not diagnosed. Those individuals were previously Valid Negatives and are converted to False Positives. Accordingly, Specificity drops from .892 to .649. In general, as we make our threshold for predicting diagnosis more liberal (i.e., as we increase the Selection Ratio), Sensitivity will increase, and Specificity will decrease.

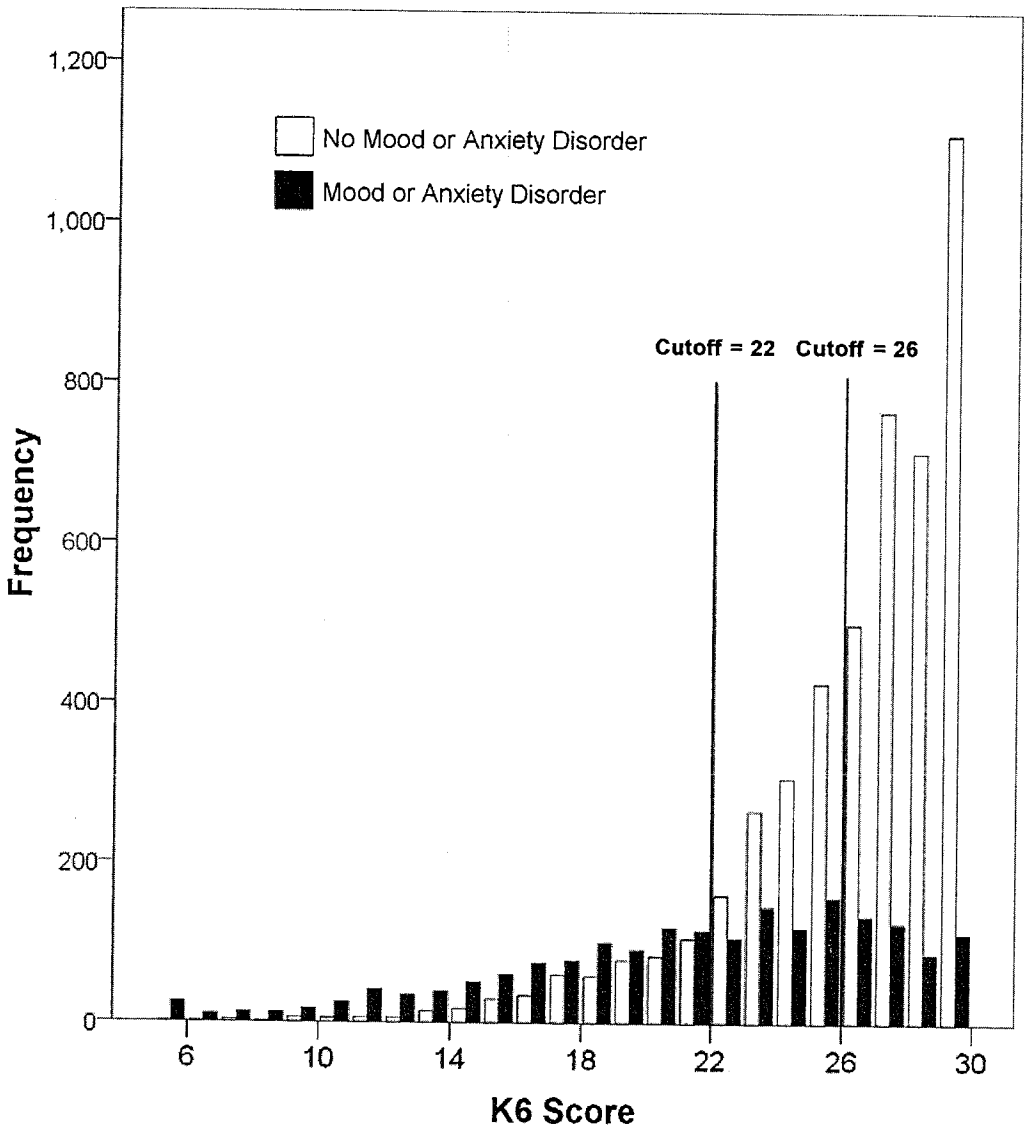

FIGURE 37.2. Histograms of K6 scores for respondents who did and did not receive a mood or anxiety disorder diagnosis. Lower scores indicate greater psychological distress.

Figure 37.3 shows this inverse relationship across a wide range of cutoff values. If we set a conservative threshold that demands strong evidence of problems (in the case of the K6, this means a low score) before predicting disorder, then Specificity will be high and Sensitivity will be low. As we make our threshold more liberal, requiring less evidence of problems before we predict disorder, Specificity will decrease and Sensitivity will increase. It should be clear that general statements like "the Sensitivity of this scale when predicting depression is .8" are not very informative, given that Sensitivity depends on the cutoff that we choose.

Sensitivity and Specificity are conditional on outcomes (the interview-based classifications of present and absent expressed as the rows in Figure 37.1). Two additional indexes of test performance are conditional on our predictions (the columns of Figure 37.1). Positive Predictive Power is the probability that someone we predict to have a disorder actually has a disorder. It is computed as Valid Positive Percent/Selection Ratio, which in the current example equals .643. Negative Predictive Power is the probability that someone we predict not to have a disorder in fact does not have a disorder. It is computed as Valid Negative Percent/(100 – Selection

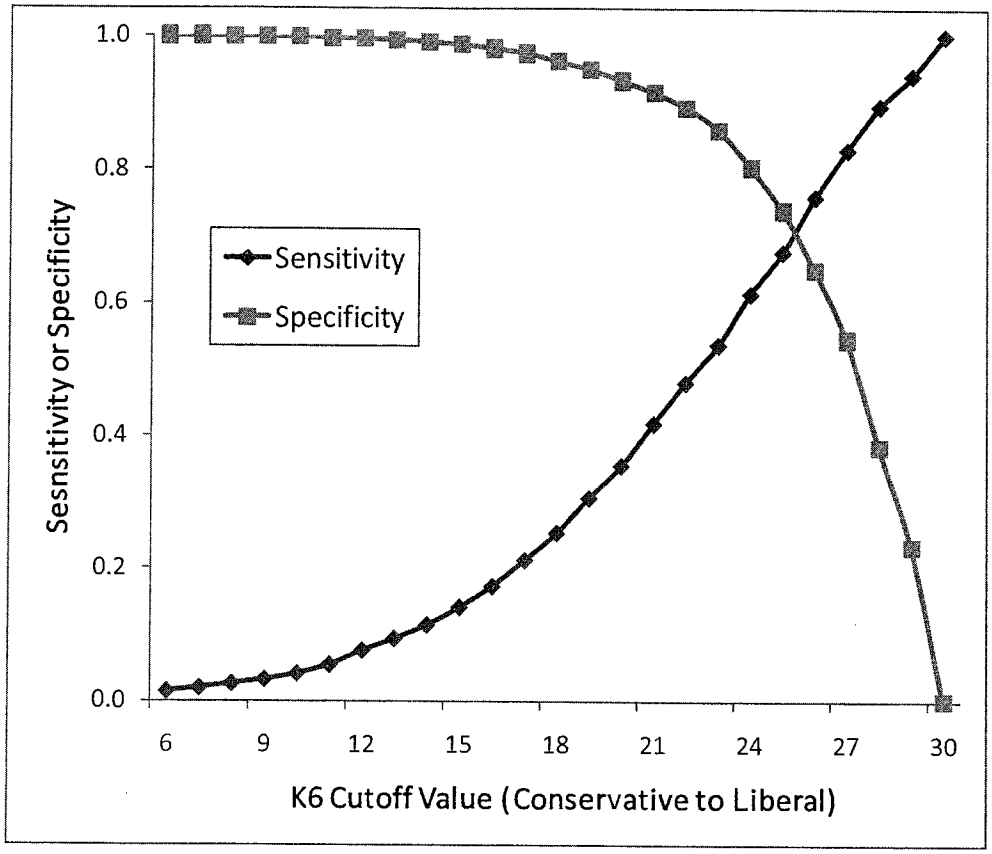

FIGURE 37.3. Sensitivity and Specificity as a function of K6 cutoff values.

Ratio), which in the current example equals .811. Like Sensitivity and Specificity, Predictive Power is influenced by our cutoff for predicting disorder. Making a cutoff more liberal usually will decrease Positive Predictive Power because as we move toward the not-disordered side of the distribution (the right side of Figure 37.2), we will usually pick up more not-disordered individuals relative to those with disorders. This will increase the percentage of False Positives more than the percentage of Valid Positives. The same change to a more liberal threshold will usually result in an increase in Negative Predictive Power because as we move toward the not-disordered end of the distribution, we will lose a higher percentage of False Negative individuals than Valid Negative individuals. Figure 37.4 shows the empirical relationship between the selected cutoff and both Positive and Negative Predictive Power for the current sample.

Positive and Negative Predictive Power are also strongly influenced by the BR or prevalence of a phenomenon. At a given threshold, increasing prevalence implies relatively more Valid Positive than False Positive cases, and relatively fewer Valid Negative than False Positive cases. Thus, as prevalence increases, Positive Predictive Power will increase and Negative Predictive Power will decrease. Figure 37.5 shows the expected change in Positive and Negative Predictive Power with increasing prevalence in the current sample.

The discussion thus far shows why Sensitivity, Specificity, Positive Predictive Power, and Negative Predictive Power provide better information about test performance than an overall measure of percent correct. They focus our attention on specific goals (e.g., finding people who need help versus not squandering resources on people who do not need help) rather than on a general goal of overall accuracy. But because these indexes refer to test performance at only one of many possible criteria or thresholds, and because they refer to test performance at only one observed BR, their generalizability to

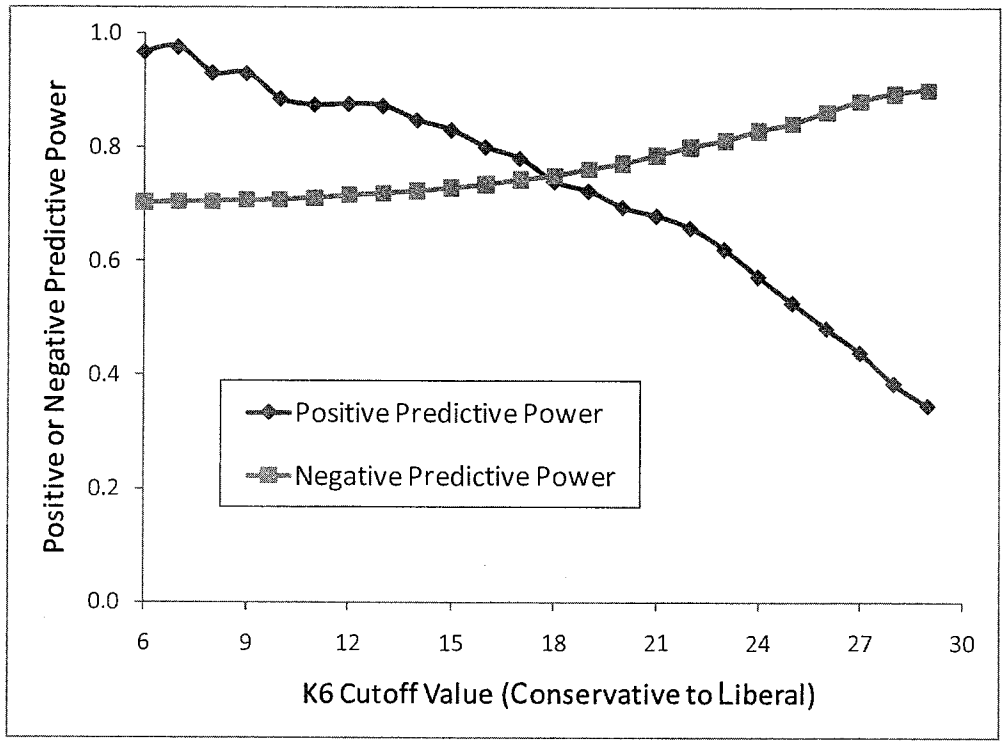

FIGURE 37.4. Positive and Negative Predictive Power as a function of K6 cutoff values.

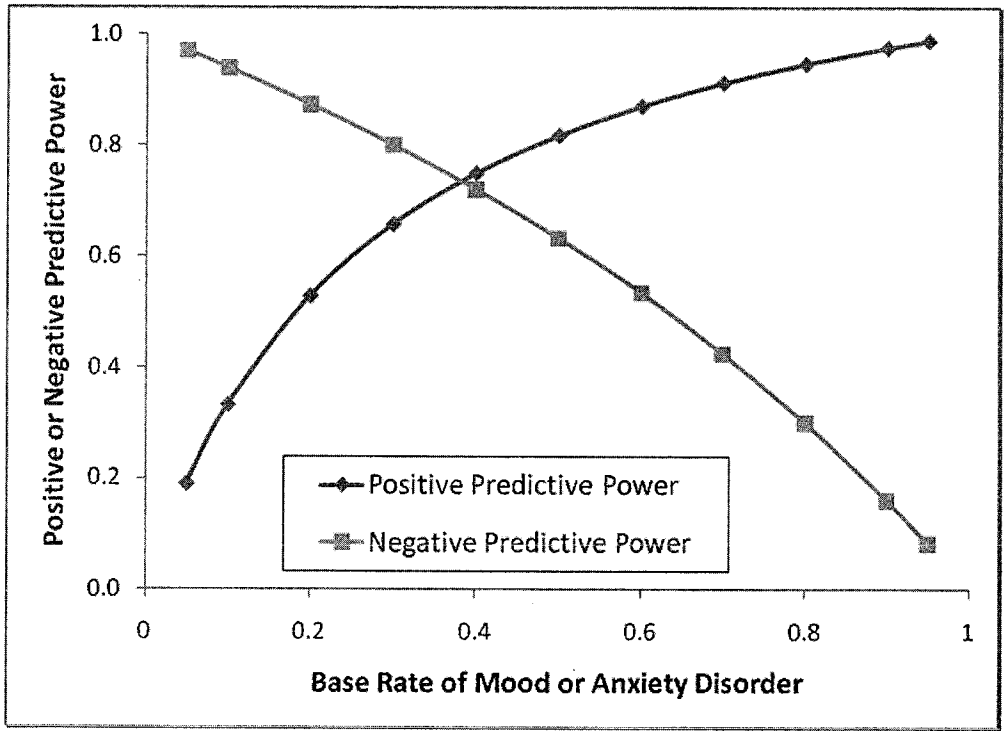

FIGURE 37.5. Positive and Negative Predictive Power as a function of the base rate (or prevalence) of a mood or anxiety disorder, assuming a cutoff of 22 on the K6.

new samples and applications is questionable. If the relative importance of avoiding False Negative versus False Positive mistakes differs in a new setting, thereby implying a different cutoff score, or if the prevalence differs in the new setting, then it will be difficult to predict how a scale will function in the new setting on the basis of reports of Sensitivity, Specificity, Positive Predictive Power, and Negative Predictive Power in previous studies. It would be far more useful to have a measure of test performance that better generalizes across samples and thresholds. Next, we provide an overview of just such an index, the area under the ROC curve. Subsequently, we describe how decision-theory methods can be used to select a threshold or cutoff value that optimizes practical utility in a context characterized by a particular BR and set of values.

QUANTIFYING DISCRIMINATORY POWER: APPLICATION OF ROC ANALYSIS

ROC analysis, an analytic approach based on signal detection theory, yields a quantitative index of how well an assessment strategy detects or predicts a signal of interest or discriminates two signals of interest (e.g., the presence of a disorder, the occurrence of violence, response to treatment, recidivism, and so on). Originally, engineers developed ROC analysis to quantify how well a human receiver detected electronic signals in the presence of noise, and ROC analysis acquired its name from its application to radar-detection problems during World War II (Pierce, 1980). Unlike the indexes of test performance reviewed thus far, the ROC-based index is independent of the BRs or prevalence of a phenomenon, the selected cutoff score, and the values or utilities placed on the four potential decision-making outcomes. Subsequent decision-theory approaches then can be used to optimize cutoff selection for the assessment strategy with maximal discriminatory power in a particular context, which necessarily will be influenced by phenomenon BRs and user-specified values. The sequential employment of ROC analysis and decision-theory approaches provides psychologists with a powerful pair of tools for the selection and application of valid and practically useful assessment and prediction methods (e.g., Swets, 1996; Swets et al., 2000).

We illustrate the use of ROC methods by continuing our analysis of the K6 screen for psychological distress and illness. We now will use the language of signal detection theory to refer to the four cells in Figure 37.1 (e.g., Hits, Misses, False Alarms, and Correct Rejections, rather than Valid Positives, False Negatives, False Positives, and Valid Negatives). Figure 37.6 presents the ROC curve for the K6 scale as a predictor of the interview-based diagnostic outcome discussed thus far. The axes of the ROC plot are the hit and false alarm rates (i.e., the proportions of Valid Positives and False Positives, respectively), and each point on the ROC curve corresponds to a pair of hit and false alarm rates that results from the use of a specific cutoff value. The hit rate can be computed as Hits/(Hits + Misses), and the false alarm rate corresponds to False Alarms/(False Alarms + Correct Rejections). In more traditional language, the ROC curve is a plot of Sensitivity against 1 – Specificity at all possible cutoff values. A few pairs of false alarm and hit rates are indicated by their associated K6 cutoff values. For example, counting K6 scores less than or equal to 22 as positive cases (because lower scores indicate more distress on the K6) produces a false alarm rate of .107 and a hit rate of .480, and a cutoff score of 28 produces false alarm and hit rates of .616 and .895, respectively. The cutoff value of 28 corresponds here to a liberal criterion or cutoff, which results in a substantial hit rate but also a high false alarm rate. In contrast, the markedly conservative cutoff value of 19 results in a very low false alarm rate (.050) but also an unimpressive hit rate (.307). Thus, the cutoff changes from maximally conservative to maximally liberal as one moves along an 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). Because lower K6 scores indicate greater pathology, lower scores index more conservative cutoffs. On other measures in which higher scores indicate greater pathology, however, higher scores correspond to more conservative cutoffs because fewer positive cases are identified by high cutoff scores.

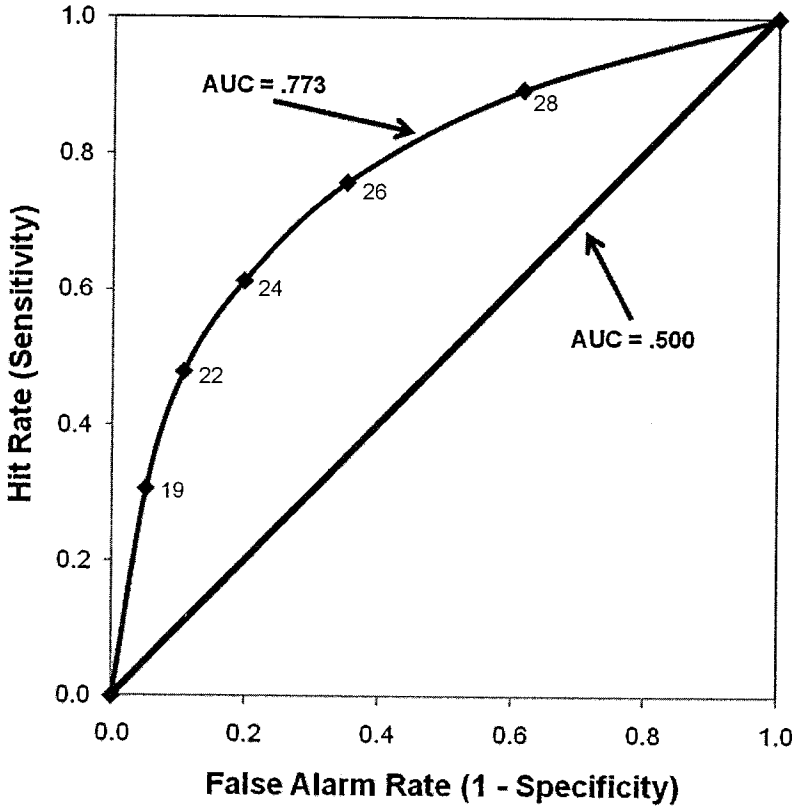

FIGURE 37.6. Receiver operating characteristic curve for K6 scale, with five labeled cutoff values ranging from conservative (19) to liberal (28). AUC = area under the curve.

The area under the ROC curve (AUC) quantifies the discriminative power of an assessment or prediction method independently of the cutoff value, unlike traditional accuracy indexes. 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). An ROC curve that lies on the main diagonal (see Figure 37.6) 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 performance of the diagnostic system increases, the distance of the observed ROC curve from the chance line increases.

The AUC for the K6 as a predictor of mood or anxiety disorder diagnoses in the current illustrative data set is .773, with a standard error of .007 and a 95% confidence interval estimate ranging from .763 to .783. The AUC value 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 correctly by the assessment method (Green & Swets, 1966; Hanley & McNeil, 1982). In the current context, this value indicates that the K6 score will be lower 77.3% of the time for a randomly selected individual with a mood or anxiety disorder than for a randomly selected individual without a mood or anxiety disorder. A z test demonstrates that the observed AUC value is significantly greater than the chance value of .500 ($z = 41.949$, $p < .0001$), indicating that K6 scores are a significant signal of the presence of a mood or anxiety disorder.

Figure 37.7 juxtaposes the ROC curve for the K6 with that for the K10, a 10-item screen that includes four additional items (Kessler et al., 2002, 2003).

The AUC value for the K10 is .782 (*SE* = .006, 95% confidence interval [CI] = .772–.792), and it is significantly greater than the chance value of .50, $z = 44.373$, $p < .0001$. The difference between the AUC values for the K10 and K6 is .008 (*SE* = .002, 95% CI = .005–.012) and is significantly greater than 0, $z = 5.004$, $p < .001$. This set of results indicates that the K10 shows a statistically but not practically significant advantage over the K6 for detection of the presence of a mood or anxiety disorder. Inspection of Figure 37.7 suggests that the K10 may have a small advantage over the K6 when more liberal cutoff values are employed (e.g., 24 and greater). Thus, when the practical goal in a particular context is to use a screening device to distinguish those with very few symptoms from those with more than very few symptoms, the K10 may be slightly preferable to the K6, in spite of the inclusion of four addition items. When the goal instead is to distinguish those reporting significant symptoms from the remainder of respondents, the K6 and K10 perform very similarly.

Formal methods are available to compare two ROC curves either (a) at a single hit or false alarm rate point on the curve (McNeil & Hanley, 1984) or (b) across a range of user-specified false alarm rate values (e.g., Y. He & Escobar, 2008; McClish, 1989). For example, we could use these methods to evaluate whether the Sensitivity of the K10 is significantly greater than the Sensitivity of the K6 at a false alarm rate of .500. Or, we could evaluate whether the discriminatory power of the K10 is significantly greater than that of the K6 for false alarm rate values ranging from .500 to .800.

In their initial report on the psychometric properties of the K6 and K10, Kessler et al. (2002) reported AUC values of .876 and .879, respectively. The notably higher values presumably reflect in part their use of a broader gold standard index: the 12-month diagnosis of any anxiety disorder, any mood disorder, or any nonaffective psychosis as well as a Global Assessment of Functioning score between 0 and 70. Similar to the present findings,

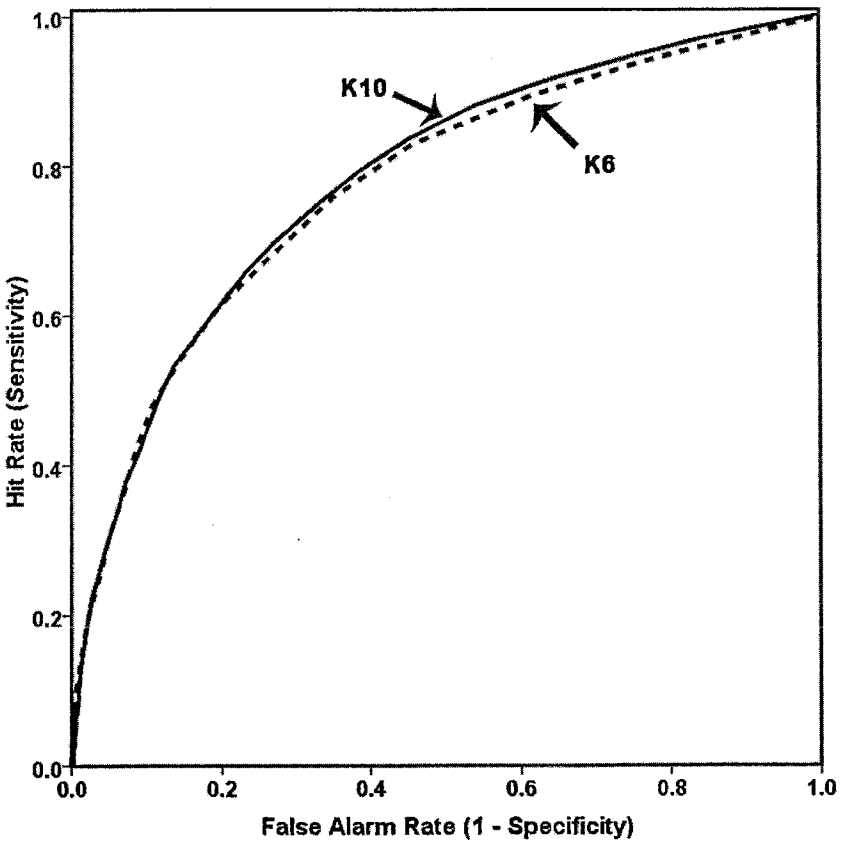

FIGURE 37.7. Receiver operating characteristic curves for K6 and K10 scales.

however, administration of the K10 did not enhance substantially the information acquired from administering the K6. Given the negligible increase in discriminatory power associated with use of the longer K10 in both Kessler et al.'s initial report and the current illustrative analyses, we hereafter use the K6 in all analyses.

Of course, the AUC value for a scale will depend on the outcome that it is predicting. For illustrative purposes, Figure 37.8 presents K6 ROC curves for four of the six anxiety disorder diagnoses that currently are available in the NCS-R: generalized anxiety disorder, panic disorder, specific phobia, and social phobia. Table 37.1 lists the AUC values, standard errors, 95% CIs, and z statistics and evaluates whether discriminatory power is significant for each diagnosis. The discriminatory power of the K6 is significantly greater than chance for all four diagnoses. Not surprisingly, the performance of this six-item screening measure varies across diagnoses; it is significantly worse for detection of specific phobia than for the other three diagnoses, all $p < .001$, and it is significantly better for detection of generalized anxiety disorder than for panic disorder, social phobia, and specific phobia, all $p < .05$.

ROC methods initially were parametric and assumed to be appropriate only when the underlying distributions were normal and showed 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 (e.g., Hanley & McNeil, 1982). Both parametric and nonparametric methods allow the user to compare AUC values with chance performance values and to compare

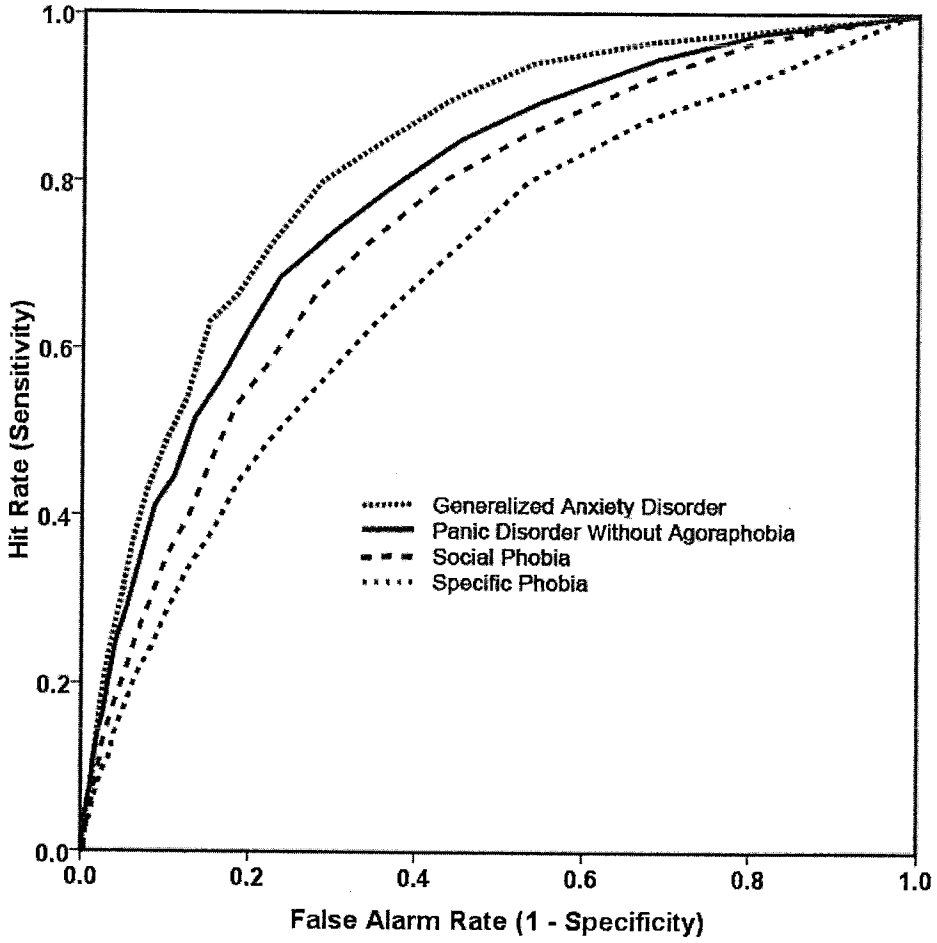

FIGURE 37.8. Receiver operating characteristic curves for K6-based detection of four anxiety disorders.

TABLE 37.1

Receiver Operating Characteristic Analysis Results for K6 Screening of Four Anxiety Disorders

Disorder	AUC	*SE*	*z*	*p*	95% CI
Generalized anxiety disorder	.825	.0102	31.899	.0001	.815–.834
Panic disorder	.788	.0141	20.430	.0001	.778–.797
Social phobia	.751	.00985	25.454	.0001	.740–.761
Specific phobia	.686	.0100	18.608	.0001	.675–.698

Note. AUC = area under the curve; CI = confidence interval.

two independent or dependent AUC values (e.g., DeLong, DeLong, & Clarke-Pearson, 1988; Hanley & McNeil, 1982, 1983; Metz, Wang, & Kronman, 1984).

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 person (or other stimulus) into one of two categories but also indicate their confidence level for the accuracy of their classification, typically on a 5- or 7-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 (see Macmillan & Creelman, 1991). When ROC analysis is used to quantify the performance of assessment and prediction strategies, the first strategy most commonly is employed.

ROC approaches typically have been applied only to dichotomous classification decisions, although diagnosticians often are called on to make classifications into more than two discrete categories. Fortunately, Scurfield (1996) generalized signal detection theory (SDT) analysis to account for unidimensional classifications into three or more categories, and recently there has been a flurry of developments on this front (e.g., X. He, Gallas, & Frey, 2010; Li & Zhou, 2009). DeCarlo (1998) also has shown how conventional SDT models are special cases of the generalized linear model, suggesting a wide variety of potential extensions, including incorporation of predictors and covariates into ROC analyses.

A variety of software programs are available for ROC analysis. In a review, Stephan, Wesseling, Schink, and Jung (2003) recommended the use of Analyse-It Software (available for purchase at http://www.analyse-it.com), AccuROC (which no longer is available), or MedCalc (available for purchase at www.medcalc.be). Both Analyse-It and MedCalc have user-friendly interfaces and nice graphics, and they report results for both single AUC values and for the difference in two AUC values as well as associated standard errors, confidence intervals, and statistical tests. SPSS, in contrast, currently does not provide an evaluation of the difference between two AUC values. SAS, which was not included in the Stephan et al. (2003) review, provides a macro as well as several statements within the logistic procedure relevant to ROC analysis. Gönen (2007) detailed how other features of SAS can be employed to conduct far more extensive ROC analyses. STATA also provides a number of parametric and nonparametric ROC analysis options, including comparison of two AUC values. All analyses for this chapter were conducted using MedCalc version 11.2.1.0; parametric and nonparametric findings were nearly identical.

SELECTING A CUTOFF: APPLICATION OF DECISION AND INFORMATION THEORY

Although ROC analysis provides an index of discriminatory power that is independent of cutoff values, BRs, and the values or utilities placed on the four decision-making outcomes, it does not provide

the optimal cutoff value or illustrate how the ideal cutoff value varies as a function of the hit and false alarm rates, BRs, and values (Hsiao, Bartko, & Potter, 1989; Mossman & Somoza, 1989; Murphy et al., 1987; Somoza et al., 1994; Swets et al., 2000). Having first used ROC methods to identify the assessment or prediction strategy with the greatest discriminatory power, users next must select an optimal cutoff value, which necessarily involves specification of a function to be maximized. Thus, there is no true and unique optimal cutoff value, and the usefulness of a diagnostic test can vary widely across the contexts in which it is employed as a function of cutoff selection. In the next section, we provide an overview of two common approaches to selecting optimal cutoff values that incorporate hit and false alarm rates, BRs, and utilities in their criterion function.

Decision Theory Approach to Cutoff Specification

Meehl and Rosen (1955) and Somoza and 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; Swets, 1992). This approach allows the user to place a differential value on (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 (i.e., variation in the perceived benefits and costs associated with the four outcomes). As summarized in the following equation, the overall utility of a specific cutoff value is a function of the hit rates and false alarm rates (*HR* and *FAR*) that result from a given cutoff value, a BR estimate (expressed as a proportion), and the values or utilities placed on each of the four decision-making outcomes (*UH* = utility for hits, *UM* = utility for misses, *UFA* = utility for false alarms, and *UCR* = utility for correct rejections):

$$U_{overall} = (BR)(HR)(UH) + (BR)(1 - HR)(UM) + (1 - BR)(FAR)(UFA) + (1 - BR)(1 - FAR)(UCR). \quad (1)$$

Each term in $U_{overall}$ is the product of the probability of a particular outcome (e.g., the probability of a Hit is *BR*HR*) and the utility of that outcome (e.g., *UH*). Thus, $U_{overall}$ is a utilities-weighted sum of the probabilities of the four decision-making outcomes. Utilities typically range between 0 and 1, where a value of 0 represents the least desired outcome and a value of 1 indicates the most desired outcome. Typically, therefore, hits and correct rejections are assigned utilities ≥ .5, whereas misses and false alarms are assigned utilities ≤ .5. Suppose, for example, that we wanted to instantiate the common decision goal of maximizing percent correct in the previous example. We would assign the maximal value of 1 to correct detection of individuals with an anxiety or mood disorder (i.e., *UH* = 1) and to correct rejection of individuals without an anxiety or mood disorder (i.e., *UCR* = 1). The minimal value of 0 would be assigned to failure to detect individuals with an anxiety or mood disorder (i.e., *UM* = 0) and to failure to reject individuals without an anxiety or mood disorder (i.e., *UFA* = 0). We then would compute $U_{overall}$ for all possible cutoff values (e.g., the 25 potential K6 cutoff scores) and a range of prevalence rates. These steps readily can be instantiated in Excel.

Figure 37.9 depicts how the overall utility of various K6 cutoff values changes as a function of phenomenon BRs in a specific decision context, assuming that the decision goal is to maximize percent correct. As the BR of either a mood or anxiety disorder increases from .100 to .900, the optimal cutoff value increases markedly from 13 to 30. More generally, whenever the decision goal is to maximize percent correct, the ideal cutoff value necessarily becomes more conservative and results in fewer positive classifications as the BR of a phenomenon decreases, so that false alarms do not become too frequent.

The potentially marked influence of changes in BRs on the utility of cut scores commonly is ignored in both research and applied contexts. Cutoff scores determined to be *optimal* during scale development may be reified and used without modification across contexts in which BRs vary widely. For example, an optimal cutscore might be determined in an initial study in which the BRs for the phenomenon of interest are higher (perhaps because of oversampling persons with disorders) than in the context in which the resulting measure and cutscore commonly are applied. As a result, the cutscore that was

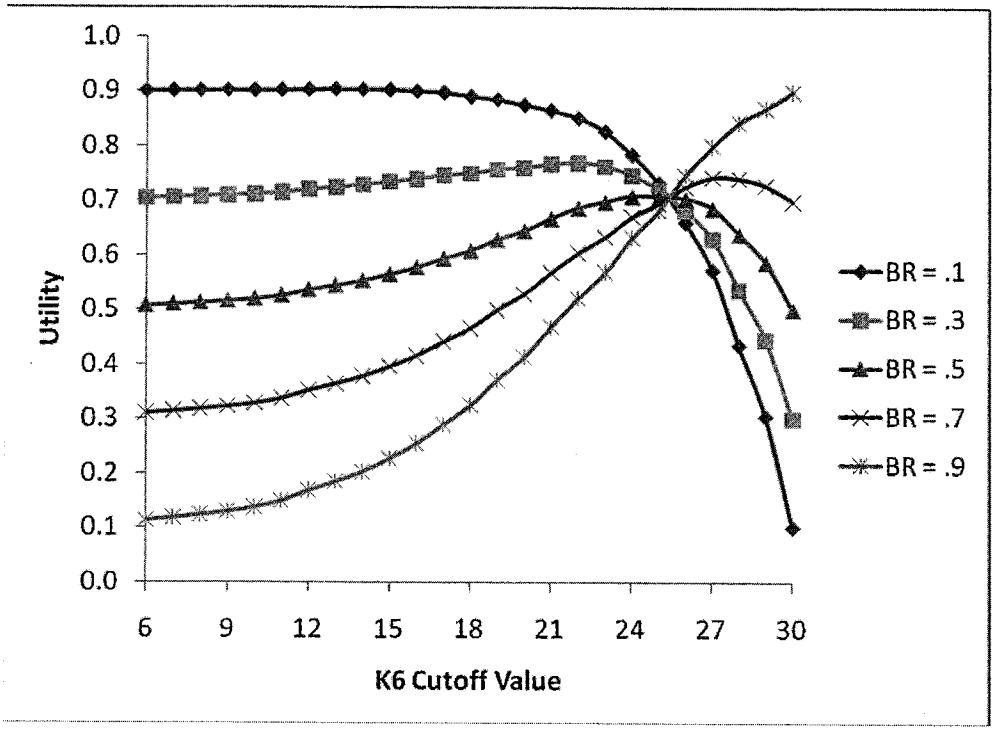

FIGURE 37.9. Utility of K6 cutoff values as a function of the base rate (BR) of mood or anxiety disorders while maximizing percent correct.

optimal in the higher BR context is too liberal in the lower BR context, resulting in a notable increase in the relative frequency of false alarms. Alternatively, the ideal cutoff score for a measure that emerges from work with a large community sample might be unacceptably conservative when applied within a clinical context.

Exhibit 37.2 illustrates the potential impact of ignoring the influence of disorder prevalence on the practical utility of the K6. Exhibit 37.2A presents the classification results for the 6,656 individuals in the current sample, given two assumptions. First, the BR of a mood or anxiety disorder is assumed to be .285, the observed value for the current sample. Second, the cutoff score is 22, the optimal value assuming that the proportion correct is being maximized for the given prevalence. This cutoff is associated with a hit rate of .480 and a false alarm rate of .107. Overall proportion correct is .775 [i.e., (911 + 4248)/6656)]. Exhibit 37.2B then presents the results if the same cutoff value of 22 is employed in a context in which the BR of mood or anxiety disorders is .500, rather than .285. This might occur in a clinical context, for example. Overall proportion correct drops to .686. Finally, Exhibit 37.2C presents the results if the ideal cutoff value is used for a context in which the BR of mood or anxiety disorders is .500. This cutoff value is higher (24), as would be expected when the proportion of positive cases increases, and it is associated with a hit rate of .613 and a false alarm rate of .197. Notably, overall proportion correct now increases to .708, and a significantly greater proportion of positive cases is detected. Thus, it is critical for researchers both to provide and to make use of BR-specific guidance on the cutoff values that optimally balance correct and incorrect decisions.

A decision goal of maximizing percent or proportion correct frequently is selected to sidestep the need to specify values or utilities for each of the four decision-making outcomes. This default approach makes equally strong implicit assumptions, however. In the present context, choosing to maximize percent correct is predicated on the assumption that correctly identifying those *without* mood or anxiety disorders is just as important as correctly identifying those *with* mood or anxiety disorders, although some might argue the latter is more valuable.

Exhibit 37.2
Influence of Base Rates on Proportion Correct Classification of Mood or Anxiety Disorders on the Basis of K6 Scores, Using a Cutoff of 22

		K6-based classification		
		Disorder present	**Disorder absent**	**Total**
A: Base rate = .285, cutoff = 22 (*HR* = .480, *FAR* = .107), proportion correct = .775				
Interview-based classification ("Truth")	Disorder present	911	988	1,899
	Disorder absent	509	4,248	4,757
	Total	1,420	5,236	6,656
B: Base rate = .500, cutoff = 22 (*HR* = .480, *FAR* = .107), proportion correct = .686				
Interview-based classification ("Truth")	Disorder present	1,597	1,731	3,328
	Disorder absent	356	2,972	3,328
	Total	1,953	4,703	6,656
C: Base rate = .500, cutoff = 24 (*HR* = .613, *FAR* = .197), proportion correct = .708				
Interview-based classification ("Truth")	Disorder present	2,040	1,288	3,328
	Disorder absent	654	2,674	3,328
	Total	2,694	3,962	6,656

Note. *HR* = hit rate; *FAR* = false alarm rate.

Analogously, this approach stipulates that erroneously classifying a person as having a mood or anxiety disorder is just as problematic as failing to identify a person with a mood or anxiety disorder, although some might perceive the latter to be more serious. Consideration of plausible value specifications is facilitated by inspection of the following utility ratio (Somoza & Mossman, 1991):

$$\text{Utility Ratio} = (UCR - UFA) / (UH - UM). \quad (2)$$

Maximizing percent correct essentially specifies a utility ratio of 1.0, whereby the difference between the values placed on correct versus incorrect decisions about negative cases in the numerator is the same as the difference between the values placed on correct versus incorrect decisions about positive cases in the denominator. Alternative value specifications could capture the greater perceived importance of decisions about positive cases than negative cases, however. For example, we might stipulate that $UH = 1$, $UCR = .75$, $UM = 0$, and $UFA = .25$, resulting in a utility ratio of .5 (i.e., we care twice as much about decisions regarding positive cases than negative cases). Alternatively, pronounced concerns about the negative consequences or side effects of case identification or treatment might lead one to place greater value on decisions about negative cases ($UH = .75$, $UM = .25$, $UCR = 1$, $UFA = 0$), thereby specifying a utility ratio of 2.0 (i.e., we care twice as much about decisions regarding negative than positive cases). Thus, value configurations that weight decisions about positive and negative cases equally correspond to a utility ratio of 1.0, configurations that weight decisions about positive cases far more than negative cases produce utility ratios less than 1.0, and configurations that weight decisions about negative cases far more than positive cases produce utility ratios greater than 1.0.

Figure 37.10 illustrates how the overall utility of various K6 cutoff values changes as a function of the utility ratio, or the relative value placed on decisions about positive versus negative cases. The BR is held fixed at .285 for all computations, as this is the probability of a mood or anxiety disorder in the current data set. As greater importance is placed on decisions about positive cases (i.e., as the utility

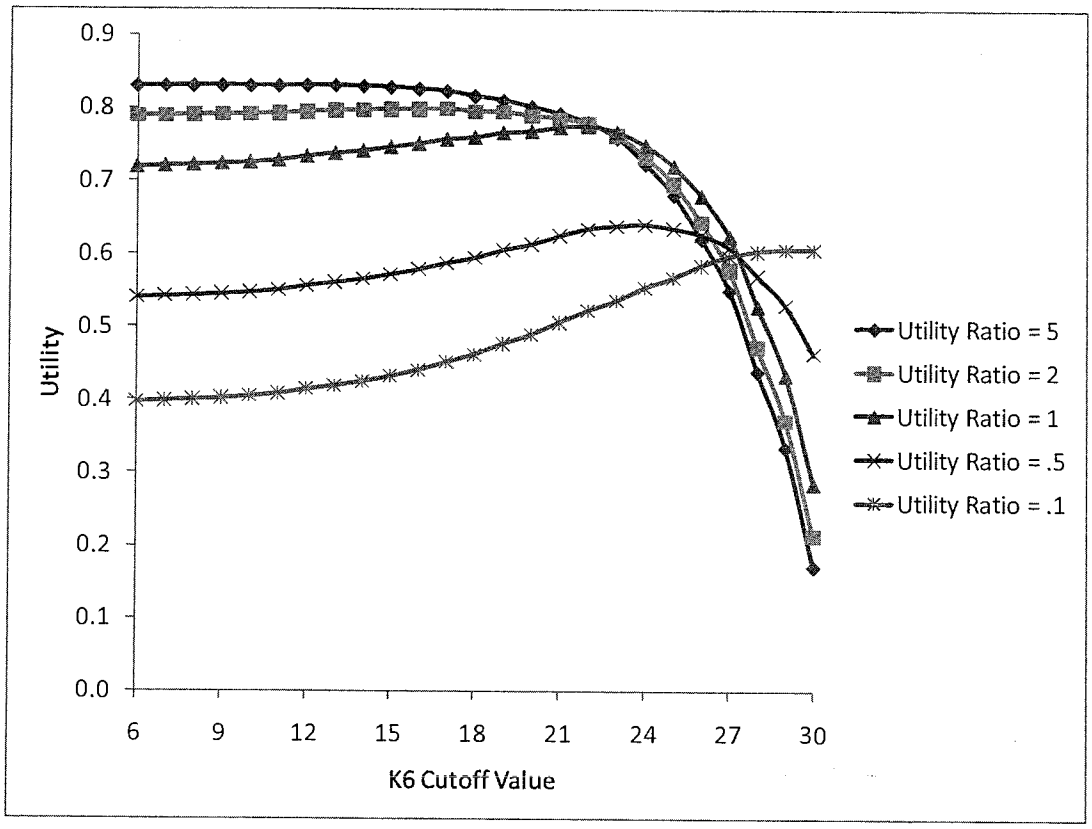

FIGURE 37.10. Utility of K6 cutoff values as a function of the relative value placed on positive versus negative cases, assuming the base rate of mood or anxiety disorders is .285. See text for more information.

ratio decreases), the most useful threshold increases in value from 13 for a utility ratio of 5.0 (decisions about negative cases more important) to 30 for a utility ratio of 0.1 (decisions about positive cases more important). When decisions about positive and negative cases are valued equally (i.e., the utility ratio = 1.0), the optimal cutoff value is 22. Not surprisingly, as the value placed on positive cases increases, the cutoff becomes more liberal.

Exhibit 37.3 illustrates how ignoring implicit assumptions about the equal importance placed on decisions about positive and negative cases when maximizing proportion correct may result in the selection of unnecessarily conservative cutoff scores. Exhibit 37.3A presents the classification results for the 6,656 individuals in the current sample, given two assumptions. First, the prevalence of a mood or anxiety disorder is assumed to be .285, the observed value for the current sample. Second, the cutoff score is 22, which is the optimal value assuming that proportion correct is being maximized for the given prevalence rate (i.e., *UH* = 1, *UCR* = 1, *UM* = 0, and *UFA* = 0; the utility ratio = 1.0). Using a cutoff of 22 produces a hit rate of .480 and a false alarm rate of .107. Under these conditions, the proportion correct for positive cases (Sensitivity) is .480, and the proportion correct for negative cases (Specificity) is .893. Exhibit 37.3B then presents the results if the cutoff score is 24, which is ideal in a context in which accurate decisions about those with a mood or anxiety disorder are construed as twice as important as accurate decisions about those without a mood or anxiety disorder (i.e., *UH* = 1, *UCR* = .75, *UM* = 0, and *UFA* = .25; the utility ratio = .5). The cutoff score of 24 is associated with a hit rate of .613 and a false alarm rate of .197. Under these conditions, the proportion correct for positive cases increases to .613 (an increase of 13 percentage points), whereas the proportion correct for negative cases declines to .803 (a decrease of 9 percentage points). This example highlights how implicitly placing equal importance on decisions about positive and negative cases when one in actuality places far greater importance on positive than negative

Exhibit 37.3
Influence of Relative Value Placed on Positive and Negative Cases on Proportion Correct Classification of Presence or Absence of Mood/Anxiety Disorders on the Basis of K6 Scores

		K6-based classification		
		Disorder present	Disorder absent	Total
A: Cutoff = 22 (*HR* = .480, *FAR* = .107), proportion correct (positive case) = .480, proportion correct (negative case) = .893				
Interview-based classification ("Truth")	Disorder present	911	988	1,899
	Disorder absent	509	4,248	4,757
	Total	1,420	5,236	6,656
B: Cutoff = 24 (*HR* = .613, *FAR* = .197), proportion correct (positive case) = .613, proportion correct (negative case) = .803				
Interview-based classification ("Truth")	Disorder present	1,164	735	1,899
	Disorder absent	935	3,822	4,757
	Total	2,099	4,557	6,656

Note. *HR* = hit rate; *FAR* = false alarm rate.

decisions may result in the use of unnecessarily conservative cutoff scores, resulting in decreased accuracy for positive cases.

More generally, the practical utility of assessment and prediction devices could be enhanced greatly if researchers routinely provided optimal cutoff scores for a range of BRs and utility ratios during measure development. Although the utility approach has been criticized because it requires the user to quantify both BRs and utility ratios,[1] it is important to recognize that proceeding instead by ignoring BR effects and maximizing percent correct also makes stringent assumptions that can exert marked influences on overall accuracy. In other words, no absolute optimal cutoff value exists in the absence of prevalence information and assumptions about the meaning of optimal.

Information Theory Approach to Cutoff Specification

To finesse the use of subjective utilities, Metz, Goodenough, and Rossmann (1973) proposed that an information theory (Shannon & Weaver, 1949) analysis of the ROC curve provides a natural optimization function (information gain, or I_{gain}) for the selection of an optimal threshold (see also Mossman & Somoza, 1989; Somoza, Soutullo-Esperon, & Mossman, 1989; Somoza, Steer, Beck, & Clark, 1994):

$$I_{gain} = (BR)(HR)(log2(HR/G)) + (BR)(1 - HR)(log2[(1 - HR)/(1 - G)]) + (1 - BR)(FAR)(log2(FAR/G)) + (1 - BR)(1 - FAR)(log2[(1 - FAR)/(1 - G)]), \quad (3)$$

where G = Selection Ratio (expressed as a proportion).

According to Metz et al.'s (1973) approach, information gain refers to the reduction of uncertainty about the true classification of a person that results from administering the diagnostic measure. For our example, information gain refers to the difference between the uncertainties about the mood or anxiety disorder status of an individual before and after knowing the individual's K6 score.

Inspection of the criterion functions specified by decision theorists ($U_{overall}$) and information

[1]In some settings, there may be data available on the relative cost (time, expense, productivity) of false positives and false negatives that can be used to facilitate utility specification. When such data are not available, expert ratings may provide subjective utilities that are useful as a starting point.

theorists (I_{gain}) reveals that both incorporate the false alarm rate, the hit rate, and the BR (expressed as a proportion). I_{gain} maximizes information gain, however, whereas $U_{overall}$ maximizes utility. Interestingly, $U_{overall}$ is a general case of I_{gain}, because I_{gain} provides an alternative specification of the utilities of the four outcomes (Metz et al., 1973; Somoza & Mossman, 1992a, 1992b). Thus, Metz et al.'s (1973) approach to criterion selection sidesteps the necessity of explicitly specifying the outcome utilities. Variability in prevalence continues to exert an influence on cutoff selection in the information theory approach, however.

Figure 37.11 depicts the influence of BRs on information gain for K6 cutoff values. Two effects are visible in the figure. First, information gain from administering the K6 is maximal for a BR of .5 and declines markedly when BRs are extremely low or high. This reflects the far greater a priori uncertainty about a case in which both positive and negative outcomes are equally likely. In contrast, markedly unequal BRs for positive and negative outcomes provide extensive a priori information about the most likely outcome (i.e., one could simply predict the more prevalent category for each case and be correct the overwhelming majority of the time). Second, the optimal cutoff becomes more liberal as BRs increase. For the K6, ideal cutoffs range from 22, when the prevalence of a mood or anxiety disorder is 5%, to 26, when the prevalence is 95%. Ignoring the influence of BRs on information gain has similarly deleterious effects to those illustrated in Exhibit 37.2 for overall utility.

Comparison of Two Approaches and Recommendations

Table 37.2 contrasts the optimal K6 cutoff scores for varying BRs for four optimization functions: I_{gain}, $U_{overall}$ assuming decisions about positive cases are twice as important as decisions about negative cases, $U_{overall}$ assuming decisions about positive and negative cases are equivalent in value, and $U_{overall}$ assuming decisions about negative cases are twice as important as decisions about positive cases. These values range from 7 to 30, making it

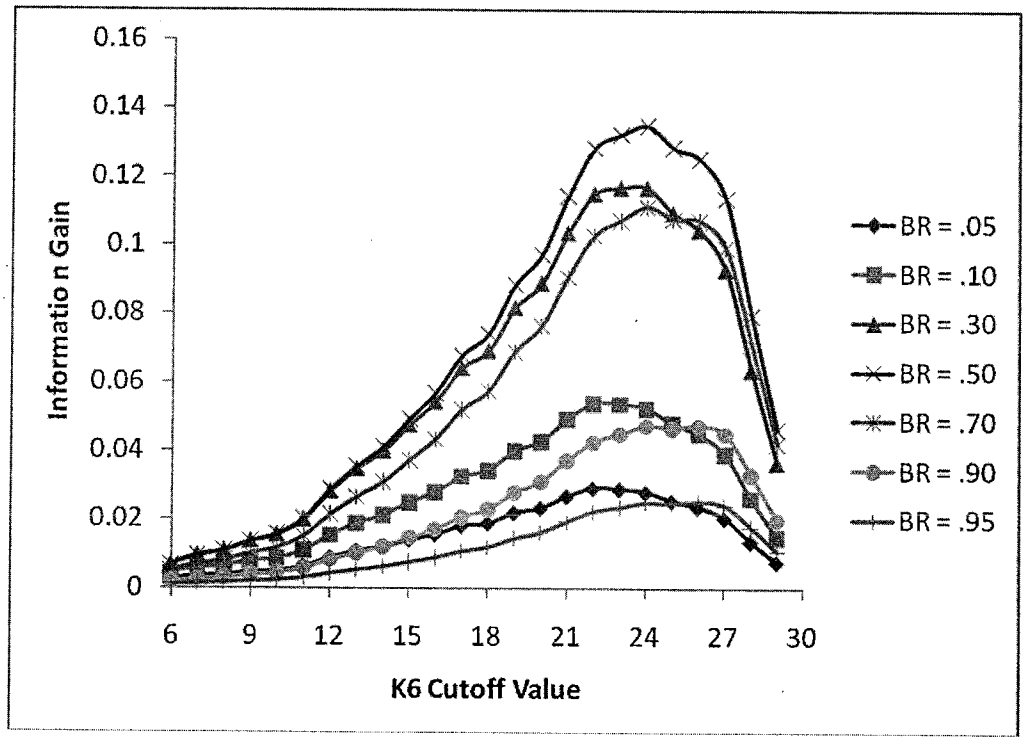

FIGURE 37.11. Information gain associated with K6 cutoff values as a function of the base rate (BR) of a mood or anxiety disorder.

TABLE 37.2

Optimal K6 Cutoff Values as a Function of Base Rates and Optimization Function

	Base rate						
Optimization function	**.05**	**.10**	**.30**	**.50**	**.70**	**.90**	**.95**
Maximizing I_{gain}	22	22	23	24	24	26	26
Maximizing $U_{overall}$: Decisions about positive cases twice as important as decisions about negative cases	13	17	24	27	30	30	30
Maximizing $U_{overall}$: Decisions about positive and negative cases equal in importance	7	13	22	24	27	30	30
Maximizing $U_{overall}$: Decisions about negative cases twice as important as decisions about positive cases	7	9	17	22	26	30	30

Note. I_{gain} = information gain; $U_{overall}$ = overall utility.

evident that both users and developers of assessment and prediction devices will benefit from attending to three factors when selecting a cutoff score that maximizes practical utility: (a) the BRs of the phenomenon in the context in which the device is employed, (b) whether maximizing utility or information is preferred, and (c) the relative importance of decisions about positive versus negative cases of the phenomenon if maximizing utility is preferred. It is not critical to the profitable use of this information that either exact BRs be known or utility ratios be specified precisely. If the user is wholly unable to specify even an approximate utility ratio, then the cutoff value that maximizes information for the approximate prevalence rate should be selected. We suspect that most users will be in a position to articulate a clear preference between the three options provided in the table, however. In this case, cutoffs predicated on utility maximization are recommended. More generally, we urge those developing assessment and prediction devices to provide a similar table of ideal cutoff values rather than a single cutoff value that may not be robust to variations in BRs and utility ratios.

CONCLUSION

Contemporary evaluation of the performance of assessment and prediction strategies in psychological science entails the completion of a two-step strategy that distinguishes the discriminative and decisional aspects of psychological measurement. First, ROC methods can be used to quantify the power of our measures to discriminate between two mutually exclusive states of interest. Indexes drawn from signal detection theory, such as AUC, assess performance independently of the selected cutoff value, the BRs of the phenomenon of interest, and the values placed on the four decision-making outcomes, unlike traditional indexes such as Sensitivity, Specificity, Positive Predictive Power, and Negative Predictive Power. Second, decision-theory methods can be employed to select a cutoff value that maximizes either the practical utility of or the information gained by test administration in a particular decision-making context, as defined by both BRs and the relative values or utilities placed on the different outcomes. This approach highlights the context specificity of optimal cutoffs or thresholds, prompting a recommendation that researchers who develop new measurement strategies routinely report optimal cutoff values for a range of potential BRs and four potential decision-making goals. Notably, precise specification of local BRs or utilities is not critical to the profitable use of this information, which will obviate the need to make stringent assumptions about the generalizability of BRs and utilities across decision-making contexts and will enhance the practical applicability of our measurement strategies.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309–319. doi:10.1037/1040-3590.7.3.309

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*, 186–205.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*, 837–845. doi:10.2307/2531595

Gönen, M. (2007). *Analyzing receiver operating characteristic curves with SAS*. Cary, NC: SAS Institute Inc.

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

Hanley, J. A. (1988). The robustness of the "binormal" assumptions in fitting ROC curves. *Medical Decision Making, 8*, 197–203. doi:10.1177/0272989X 8800800308

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

Hanley, J. A., & McNeil, B. J. (1983). Method for comparing the area under two ROC curves derived from the same cases. *Radiology, 148*, 839–843.

He, X., Gallas, B. D., & Frey, E. C. (2010). Three-class ROC analysis-toward a general decision theoretic solution. *IEEE Transactions on Medical Imaging, 29*, 206–215. doi:10.1109/TMI.2009.2034516

He, Y., & Escobar, M. (2008). Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with application to genomic studies. *Statistics in Medicine, 27*, 5291–5308. doi:10.1002/sim.3335

Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). Diagnosing diagnoses. *Archives of General Psychiatry, 46*, 664–667.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., . . . Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in nonspecific psychological distress. *Psychological Medicine, 32*, 959–976. doi:10.1017/S0033291702006074

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., . . . Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry, 60*, 184–189. doi:10.1001/archpsyc.60.2.184

Kessler, R. C., Berglund, P., Chiu, W-T., Demler, O., Heeringa, S., Hiripi, E., . . . Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research, 13*, 69–92. doi:10.1002/mpr.167

Kessler, R. C., & Merikangas, K. R. (2004). The National Comorbidity Survey Replication (NCSR): Background and aims. *International Journal of Methods in Psychiatric Research, 13*, 60–68. doi:10.1002/mpr.166

Kessler, R. C., & Üstün, T. B. (2004). The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research, 13*, 93–121. doi:10.1002/mpr.168

Li, J. L., & Zhou, X. H. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference, 139*, 4133–4142. doi:10.1016/j.jspi.2009.05.043

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge, England: Cambridge University Press.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making, 9*, 190–195.

McNeil, B. J., & Hanley, J. A. (1984). Statistical approaches to the analysis of receiving operating characteristic (ROC) curves. *Medical Decision Making, 4*, 137–150. doi:10.1177/0272989X 8400400203

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194–216. doi:10.1037/h0048070

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*, 283–298. doi:10.1016/S0001-2998(78)80014-2

Metz, C. E., Goodenough, D. J., & Rossmann, K. (1973). Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology, 109*, 297–303.

Metz, C. E., Wang, P. L., & Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In F. Deconinck (Eds.), *Information processing in medical imaging* (pp. 432–445). The Hague, the Netherlands: Nijhoff.

Mossman, D., & Somoza, E. (1989). Maximizing diagnostic information from the dexamethasone suppression test: An approach to criterion selection using receiver operating characteristic analysis. *Archives of General Psychiatry, 46*, 653–660.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic tests. *Archives of General Psychiatry, 44*, 550–555.

Pierce, J. R. (1980). *An introduction to information theory: Symbols, signals, and noise* (2nd ed.). New York, NY: Dover.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology, 40*, 253–296.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*, 396–408. doi:10.1037/1040-3590.17.4.396

Somoza, E., & Mossman, D. (1991). "Biological markers" and psychiatric diagnosis: Risk-benefit balancing using ROC analysis. *Biological Psychiatry, 29*, 811–826. doi:10.1016/0006-3223(91)90200-6

Somoza, E., & Mossman, D. (1992a). Comparing and optimizing diagnostic tests: An information-theoretical approach. *Medical Decision Making, 12*, 179–188.

Somoza, E., & Mossman, D. (1992b). Comparing diagnostic tests using information theory: The INFO-ROC technique. *Journal of Neuropsychiatry and Clinical Neurosciences, 4*, 214–219.

Somoza, E., Soutullo-Esperon, L., & Mossman, D. (1989). Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory. *International Journal of Bio-Medical Computing, 24*, 153–189. doi:10.1016/0020-7101(89)90029-9

Somoza, E., Steer, R. A., Beck, A. T., & Clark, D. A. (1994). Differentiating major depression and panic disorders by self-report and clinical rating scales: ROC analysis and information theory. *Behaviour Research and Therapy, 32*, 771–782. doi:10.1016/0005-7967(94)90035-3

Stephan, C., Wesseling, S., Schink, T., & Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clinical Chemistry, 49*, 433–439. doi:10.1373/49.3.433

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532. doi:10.1037/0003-066X.47.4.522

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychological diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1–26. doi:10.1111/1529-1006.001

APA Handbooks in Psychology

APA Handbook of Research Methods in Psychology

VOLUME 1

Foundations, Planning, Measures, and Psychometrics

Harris Cooper, *Editor-in-Chief*
Paul M. Camic, Debra L. Long, A. T. Panter,
David Rindskopf, and Kenneth J. Sher, *Associate Editors*

American Psychological Association • Washington, DC

Published by
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
www.apa.org

To order
APA Order Department
P.O. Box 92984
Washington, DC 20090-2984
Tel: (800) 374-2721; Direct: (202) 336-5510
Fax: (202) 336-5502; TDD/TTY: (202) 336-6123
Online: www.apa.org/pubs/books/
E-mail: order@apa.org

In the U.K., Europe, Africa, and the Middle East, copies may be ordered from
American Psychological Association
3 Henrietta Street
Covent Garden, London
WC2E 8LU England

American Psychological Association Staff
Gary R. VandenBos, PhD, *Publisher*
Julia Frank-McNeil, *Senior Director, APA Books*
Theodore J. Baroody, *Director, Reference, APA Books*
Kristen Knight, *Project Editor, APA Books*

Typeset in Berkeley by Cenveo Publisher Services, Columbia, MD

Printer: Maple-Vail Book Manufacturing Group, York, PA
Cover Designer: Naylor Design, Washington, DC

Library of Congress Cataloging-in-Publication Data

APA handbook of research methods in psychology / Harris Cooper, editor-in-chief.
v. cm.
Includes bibliographical references and index.
Contents: v. 1. Foundations, planning, measures, and psychometrics—v. 2. Research designs : quantitative, qualitative, neuropsychological, and biological—v. 3. Data analysis and research publication.
ISBN-13: 978-1-4338-1003-9
ISBN-10: 1-4338-1003-4
1. Psychology—Research—Methodology—Handbooks, manuals, etc. 2. Psychology—Research—Handbooks, manuals, etc. I. Cooper, Harris M. II. American Psychological Association. III. Title: Handbook of research methods in psychology.
BF76.5.A73 2012
150.72'1—dc23

2011045200

British Library Cataloguing-in-Publication Data
A CIP record is available from the British Library.

Printed in the United States of America
First Edition

DOI: 10.1037/13619-000