UNDERSTANDING THE ACCURACY OF TESTS WITH CUTTING SCORES: 
THE SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE MODEL

ALAN G. GLAROS
University of Missouri-Kansas City

REX B. KLINE
Concordia University
Montreal, Quebec, Canada

While researchers usually are concerned about psychometric properties of psychological tests estimated using large samples, most clinical decision-makers must evaluate the accuracy of test results for individuals. This is particularly true as regards tests that have cutting scores to determine, for example, whether to assign a particular diagnosis or accept an applicant into a training program. This paper reviews a conceptual model that may foster improved understanding of test outcomes for individuals. The terms “sensitivity,” “specificity,” and “predictive value” are defined, and the relations of positive and negative predictive values to population base rates are emphasized. Examples from the psychological literature are presented to illustrate the utility of these concepts in clinical decision-making with psychological tests. Implications for test users, test developers, and instructors are discussed.

Many psychological tests used in clinical, educational, and occupational settings have cutting scores (e.g., T scores of 70 for MMPI scales) intended to assist examiners in making decisions (e.g., diagnose as depressed, predict success or failure in a training program, hire or reject). While there is a voluminous literature devoted to such tests (e.g., Cronbach & Gleser, 1965; Meehl & Rosen, 1955; Rimm, 1963), much of it is very technical and directed primarily at psychometricians. In addition, much of this literature is concerned with the effects of different parameters (e.g., population base rates, selection ratios, or test validity) on the overall accuracy of test outcomes for large samples of people. Unfortunately, much of this work is only partially applicable to actual clinical assessment. That is, most users of psychological measures deal with test results for individual people, and there has been little elaboration in the psychometric literature of models to assist clinicians in the evaluation of the accuracy of test outcomes in this context.

The purpose of this paper is to present a conceptual framework that can be applied to tests with cutting scores to determine the degree of confidence that can be placed in different test outcomes for individuals. This framework—the sensitivity, specificity, and predictive value model—also can be applied to the evaluation of the strengths and liabilities of specific tests relative to alternative measures. The sensitivity, specificity, and predictive value model originally was developed in medicine for the evaluation of laboratory screening procedures. The feature of this model most valuable to users of psychological tests is that it incorporates important parameters identified in the psychometric literature (such as population base rates) and allows determination of the confidence that may be placed in different test outcomes.

We thank Joel Ager, Alan Bass, Howard Beckman, Robert Burack, and David Lachar for reviewing earlier drafts of this paper and for their suggestions.

Reprint requests should be addressed to Alan G. Glaros, Ph.D., Department of Behavioral Science, University of Missouri-Kansas City, 650 East 25th Street, Kansas City, Missouri 64108.
The sensitivity, specificity, and predictive value model will be outlined, and its role in identification of the confidence that test users may place in results of tests with cutting scores will be emphasized. Several examples from the psychological literature of the application of the sensitivity, specificity, and predictive value model will be reviewed to demonstrate its practical utility. Finally, limitations on the application of this model to psychological assessment will be considered.

SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE

*Sensitivity* is defined as the capacity of an assessment instrument or battery to yield a positive result for a person with the diagnostic condition or attribute of interest. To use a medical analogy, sensitivity is the proportion of "diseased" individuals who obtain scores above the cutting point of a screening test. That is,

\[
\text{Sensitivity} = \frac{\text{Diseased persons with positive test results}}{\text{All diseased persons}} \tag{1}
\]

Similarly, *specificity* reflects the capacity of an assessment instrument to yield a negative result for a person without a diagnostic condition or attribute. That is, the proportion of "nondiseased" persons who obtain normal-range scores on the screening test equals specificity:

\[
\text{Specificity} = \frac{\text{Nondiseased persons with negative test results}}{\text{All nondiseased persons}} \tag{2}
\]

Both sensitivity and specificity were defined first by Yerushalmy (1947) and have been an important part of the medical literature since that time (Griner, Mayewski, Mushlin, & Greenlan, 1981; Vecchio, 1966).

In practice, sensitivity and specificity are easy to calculate. Table 1 presents hypothetical data on the diagnosis of schizophrenia using an equally hypothetical "QRS" test. Using the margin totals, it is easily determined that 120 individuals are classified properly as schizophrenic. Of these, a total of 90 obtained clinical-range scores on the "QRS" test. Using formula (1), the sensitivity of the QRS test can be computed quickly as 90/120 = .75. In calculating the specificity of the "QRS" test, it can be noted that 400 of the 500 nonschizophrenics were classified properly by the test. Using formula (2), the specificity of the "QRS" is equal to 400/500 = .80.

Table 1
**Hypothetical Results of Test on a Population of Schizophrenics and Nonschizophrenics**

<table>
<thead>
<tr>
<th>Test result</th>
<th>True status</th>
<th>Schizophrenic</th>
<th>Nonschizophrenic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenic</td>
<td>90</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Nonschizophrenic</td>
<td>30</td>
<td>400</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity and specificity are empirical observations from reliably categorized diagnostic or attribute groups. Also note that sensitivity and specificity are entirely de-
Sensitivity, Specificity, and Predictive Value

termined by the established cutting score of the test. Obviously, as cutting scores are altered, sensitivity and specificity values of tests will change as well. Our concern in this presentation is not how to establish optimal cutting scores, but with determination of the predictive accuracy of tests with published cutting scores. (See Cronbach & Gleser, 1965; Cureton, 1957; Dawes, 1962; Meehl & Rosen, 1955, for discussions about establishing optimal cutting scores.)

Because sensitivity and specificity are estimated using different samples of people, these indices can vary independently of one another. Both sensitivity and specificity can be high and approach the perfect case of a test that is 100% sensitive and 100% specific, or both can be low. Another property is that both indices are independent of sample size and population base rates. A test that is 75% sensitive for schizophrenia should identify correctly 75 of 100, 600 of 800, or 30 of 40 schizophrenics. Similarly, a test that is 80% specific should classify correctly 160 of 200, 40 of 50, or 800 of 1000 nonschizophrenics. Sensitivity and specificity are also alternative ways to represent the external validity of a test. In general, as the ability of the test to discriminate diagnostic groups of interest increases, so do its sensitivity and specificity.

Sensitivity and specificity are useful indices of test accuracy when measures are applied to groups of persons with known characteristics (e.g., successful training program graduates or not; schizophrenic or not). This situation is not, however, representative of actual clinical assessment. That is, tests typically are used in the evaluation of persons whose status as regards some attribute or condition is unknown. The assessor does not obtain both positive and negative results from a test given to one person on a single occasion. The task for the assessor is, instead, to determine the "correctness" of positive or negative test results for individuals. The question is, how much confidence can an assessor have in positive or negative test findings?

An example of the relation of test sensitivity and specificity to the confidence that decision-makers may place in specific test outcomes follows. Suppose that the base rate of severe depression in a particular setting is 10%. A screening scale—the "XYZ" test—will be used in an attempt to identify depressed persons. Results from previous research with this measure indicate that its sensitivity is equal to 60% and its specificity is equal

| Table 2 |
| Calculation of Positive and Negative Predictive Value |

A. Outcomes of predictions of "XYZ" test

<table>
<thead>
<tr>
<th>True status</th>
<th>Depressed</th>
<th>Not depressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>Not depressed</td>
<td>4</td>
<td>72</td>
</tr>
</tbody>
</table>

B. General calculation of predictive value

<table>
<thead>
<tr>
<th>Test result</th>
<th>Attribute</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A</td>
<td>B + A</td>
</tr>
<tr>
<td>Negative</td>
<td>C</td>
<td>D + C</td>
</tr>
<tr>
<td></td>
<td>A + C</td>
<td>B + D</td>
</tr>
</tbody>
</table>
to 80%. Presented in the upper portion of Table 2 (section A) are expected outcomes if the “XYZ” test were used with 100 people. Classification according to the “XYZ” scale would result in correct diagnosis of 78 individuals, while 22 cases would be mislabeled. Considering the prevalence of depression in this setting and these sensitivity and specificity values, what is the degree of confidence that assessors could place “XYZ” test outcomes of persons whose true status as regards depression is unknown?

In answering this question, knowledge of the predictive values of test results will prove useful (Griner et al., 1981). The positive predictive value of a test result is the likelihood that a person with a positive test finding actually has the predicted attribute or “disease.” In personnel selection applications, positive predictive value also has been called the “success ratio,” which indicates the proportion of applicants accepted on the basis of their test scores who actually succeed in their subsequent job performance (Cascio, 1982).\(^1\)

In contrast, the negative predictive value of a test result is the likelihood that a person with a negative test sign does not have the “disease.” In schematic form,

\[
\text{Positive predictive value} = \frac{\text{Diseased persons with a positive test sign}}{\text{All persons with a positive test sign}} \quad (3)
\]

\[
\text{Negative predictive value} = \frac{\text{Nondiseased persons with a negative test sign}}{\text{All persons with a negative test sign}} \quad (4)
\]

In computing the positive predictive value for the “XYZ” scale, it can be noted that classification according to this measure identified 24 individuals as depressed. Of these 24, 6 actually were depressed. The positive predictive value of the “XYZ” test is equal to 6/24 = .25. Similarly, of the 76 persons identified as not depressed by the “XYZ” test, 72 were classified properly; the negative predictive value is equal to 72/76 = .95. A positive result from the “XYZ” test has only a 25% probability of being correct, while a negative test result has a 95% likelihood of being correct. Thus, negative test outcomes on the “XYZ” scale warrant more confidence than positive findings.

The general case for computing sensitivity, specificity, and positive and negative predictive values for any test is easily described. Using the notation in the lower portion of Table 2 (section B), these values can be quickly computed as follows: Sensitivity = A/(A + C); specificity = D/(B + D); positive predictive value = A/(A + B); and negative predictive value = D/(C + D). (The base rate of the “disease” = [A + C]/[A + B + C + D].)

\(^1\)This computation for the positive predictive value is considerably less cumbersome than an equivalent expression of Bayes Theorem:

\[
\text{Positive predictive value} = \frac{\text{p(D+/T+) = p(D+) \times p(T+/D+)} \times \begin{bmatrix} \text{p(D+) \times p(T+/D+) + p(D-) \times p(T-/D-)} \end{bmatrix}}{\begin{bmatrix} \text{p(D+) \times p(T+/D+) + p(D-) \times p(T-/D-)} \end{bmatrix}}
\]

where \(p(D+)\) = base rate of the attribute of interest, \(p(D-) = [1 - \text{base rate}], p(T+/D+) = \text{sensitivity}, \) and \(p(T-/D-) = [1 - \text{specificity}].\)

Equivalently, negative predictive value also can be expressed via Bayes Theorem:

\[
\text{Negative predictive value} = \frac{\text{p(D-/T-) = p(D-) \times p(T-/D-)} \times \begin{bmatrix} \text{p(D-) \times p(T-/D-) + p(D+) \times p(T+/D+)} \end{bmatrix}}{\begin{bmatrix} \text{p(D-) \times p(T-/D-) + p(D+) \times p(T+/D+)} \end{bmatrix}}
\]

where \(p(T-/D-)\) = specificity, and \(p(T+/D+) = [1 - \text{specificity}].\)
The predictive values of tests with cutting scores vary as a function of their sensitivity and specificity values and the population base rate of the disorder of interest. In general, higher sensitivity and specificity values are associated with higher predictive values. How does population base rate affect predictive values?

As an example of the effect of base rates upon predictive values, suppose that the “XYZ” test above (sensitivity = 60%, specificity = 80%) will be used in a setting in which the base rate of severe depression is 30% rather than 10%. In a sample of 100 persons, 30 would be expected to be severely depressed. Because the “XYZ” scale is 60% sensitive, a total of 18 of these 30 depressed people would obtain a clinical-range test score, and 12 would obtain false-negative results. Of the 70 nondepressed individuals, 56 would obtain normal “XYZ” scores, and 14 would attain false-positive results (because the test is 80% specific).

Using the notation in the lower portion of Table 2, A = 18, B = 14, C = 12, and D = 56. Thus, the positive predictive value of the “XYZ” scale when the base rate of depression is 30% equals $A/(A + B) = 18/(18 + 14) = .56$, and the negative predictive value is $D/(C + D) = 56/(12 + 56) = .82$. Thus, a positive result from the “XYZ” test in a setting in which the base rate of depression is 30% has a 56% chance of being correct, while a negative test result has a 82% chance of being correct. Compared to the 10% base rate condition, the predictive value of a positive “XYZ” result is much higher (56% vs. 25%), while the predictive value of a negative “XYZ” result is somewhat lower (82% vs. 95%) at a higher base rate condition.

Assuming given values for the sensitivity and specificity of a test, when the base rate of the disorder of interest is low, the predictive value of a negative test result will be greater than the predictive value of a positive test result. That is, when the disorder is relatively rare, a positive test finding is typically not useful in confirming its presence (Cronbach, 1984; Lanyon & Goodstein, 1982; Meehl & Rosen, 1955; Rosen, 1954). Alternatively, as the base rate of the disorder approaches 100%, the positive predictive value of a test typically will be much greater than its negative predictive value. Thus, when the base rate of some condition is high and a test is used for selection purposes, a positive test finding generally should warrant more confidence than a negative test finding.

Other examples of the relation of sensitivity, specificity, and population base to the predictive values of tests follow. Suppose the base rate of some disorder in a population is 2%. A screening test for this disorder has a sensitivity of 95% and a specificity of 90%. What are the positive and negative predictive values of this test under this base rate condition?

For this problem, we can assume arbitrarily that the population has a size of 10,000; selection of any other large number will have no effect on calculations of positive and negative predictive values. Because the prevalence of the disorder is 2%, we can assume that 200 people have the disorder and 9800 do not. Of the 200 “diseased” people, the test will identify correctly 190 (95%). Similarly, of the 9800 non-afflicted people, a total of 8,820 (90%) will obtain normal-range scores on the screening test. Using the general notation in Table 2, A = 190, B = 980, C = 10, and D = 8,820. The positive predictive value of the test equals $190/(190 + 980) = .16$, and the negative predictive value is $8,820/(10 + 8,820) = .99$. That is, of all individuals who obtain clinical-range scores on the screening test, approximately 16% will be expected actually to have the predicted condition. A total of 99% of all individuals with normal-range test scores would be expected to be unafflicted.

If the base rate of the disorder is not 2% in another clinical setting, test sensitivity and specificity still can be used to compute predictive values. Suppose the same screening test will be used with a population in which the prevalence of the disorder is 60%. Because test sensitivity and specificity are invariant of population base rate, these values
are still, respectively, 95% and 90%. In a population of 10,000 people, a total of 6,000 would have the disorder. Of these 6,000, a total of 5,700 (95%) would be expected to score above the cutting point of the screening test. Of the 4,000 people without the disorder, 3,600 (90%) would obtain normal-range test scores. Using the notation of Table 2, \( A = 5,700, B = 400, C = 300, \) and \( D = 3,600. \) The positive predictive value of the test equals \( 5,700/(5,700 + 400) = .93, \) and the negative predictive value equals \( 3,600/(300 + 3,600) = .92. \) At a higher base rate condition, the positive predictive value of the screening test is substantially higher, while the negative predictive value is marginally lower.

Figure 1 presents the positive and negative predictive values of a test that has a sensitivity value of 90% and a specificity value of 90% as a function of different base rate conditions. As the base rate increases, positive predictive value increases, while negative predictive value decreases.

![Figure 1](image)

**Fig. 1.** Predictive values for a test that is 90% sensitive and 90% specific under different base rate conditions.

The sensitivity, specificity, and predictive value model also can be applied to the comparison of alternative assessment procedures. For example, Wiggins (1973) suggested that the classification accuracy of tests with cutting scores be compared against random assessment procedures. Random assessment means that diagnoses of “diseased” and “not diseased” would be assigned to people by chance (e.g., randomly diagnose 1-2% of all individuals seen in an outpatient setting as schizophrenic).

Suppose that one wished to compare the predictive accuracy of the “XYZ” depression screening scale described above (which is 60% sensitive and 80% specific) against random diagnosis in a setting in which the base rate of depression is 10%. For this example, random assessment would mean that a diagnosis of depression is assigned randomly to 10% of all cases, regardless of presenting information. Thus, the tactic of
random diagnosis would have a sensitivity equal to 10% (1 of 10 depressed persons would be identified correctly), and a specificity equal to 90% (1 in 10 nondepressed persons would be identified incorrectly).

In an earlier example, we showed that if the “XYZ” scale were administered to 100 people, 6 of 10 depressed and 72 of 90 nondepressed people would be identified correctly. A total of 22 (4 depressed and 18 nondepressed) cases would be identified incorrectly by the “XYZ” test. In contrast, if random assessment were used with 100 people, 1 of 10 depressed and 81 of 90 nondepressed people would be identified correctly; a total of 18 cases would be identified incorrectly. Thus, fewer overall prediction errors are made with random assessment than with the “XYZ” scale.

Should assessors abandon use of the “XYZ” scale because they can place more confidence in outcomes yielded by random diagnosis of depression? In order to answer this question, awareness of the positive and negative predictive values for both procedures is necessary. The positive and negative predictive values of the “XYZ” test under a 10% base rate were calculated earlier and equal 25% and 95%, respectively. In computing the positive predictive value for random assessment, it can be noted that of the 10 persons with a positive test, only 1 actually is depressed. Using formula (3), the positive predictive value of random prediction is equal to 1/10 = .10. Of the 90 individuals who did not receive a diagnosis of depression, 81 were classified properly. Using formula (4), the negative predictive value is equal to 81/90 = .90. Thus, a positive result using random diagnosis has a 10% likelihood of being correct, while a negative result from random prediction has a 90% chance of being correct.

Results of these calculations for the predictive values of the “XYZ” scale and random diagnosis suggest an apparent anomaly. That is, while use of the “XYZ” scale results in more total classification errors than random assignment of diagnoses, the test nevertheless provides more accurate predictions on the basis of both classes of outcomes (prediction of depression and nondepression). Therefore, test users must not be concerned solely with the overall error rate, but also with the accuracy of specific test results.

The information outlined in Figure 1 also can be used to evaluate the incremental utility of this test compared to random diagnostic procedures. For example, at a base rate of 5%, a test that is 90% sensitive and 90% specific will have a positive predictive value of .33. In contrast, the positive predictive value of a random diagnostic procedure (where 5% of all people are given a diagnosis that designates the disorder) will be 5%. Use of the screening test would result in an increase of the chance that someone with a positive test results really had the disorder of 28% over that obtained with random diagnosis.

APPLICATIONS OF THE SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE MODEL TO THE PSYCHOLOGICAL LITERATURE

We have presented a straightforward conceptual framework that may assist users of psychological tests to understand better strengths and limitations of assessment procedures that involve cutting scores. While this framework has been used extensively in medicine, the predictive accuracy of psychological screening tests has not often been described in this way. This framework could be applied easily to psychological scales given that certain information is available. All that is needed to compute test sensitivity and specificity are results from empirical investigations that have used a test with samples that exhibit the attribute of interest and those that do not. Once the sensitivity and specificity for a particular test are established, the expected positive and negative predictive values under several base rate conditions can be determined. Examples of these procedures will be discussed.
MacAndrew (1965) administered the MMPI to alcoholic and non-alcoholic men and constructed a 49-item alcohol abuse screening scale based on items that differentiated these groups. Using a total sample of 300 subjects of each group, MacAndrew established a raw score cut-off of 24 for his screening scale. Application of this cutting-score in MacAndrew's sample resulted in accurate classification of 82% of all subjects. Of the alcoholic men, a total of 248 were classified correctly, while a total of 242 psychiatric outpatients were classified accurately. Thus, in this sample the MacAndrew scale had a sensitivity of 83% (248/300 = .83) and a specificity of 81% (242/300 = .81). The positive predictive value of the MacAndrew scale for this sample equals 248/(248 + 58) = 81%, while the negative predictive value equals 242/(52 + 242) = 82%.

While these predictive values for MacAndrew's alcoholism scale are respectable, they are relevant only for testing situations in which the base rate of alcoholism is 50% (Miller, 1976). As demonstrated earlier, the predictive value of a test will vary as a function of the base rate. To understand the predictive utility of MacAndrew's scale in situations in which the alcoholism base rate differs from 50%, the predictive values must be adjusted accordingly. For example, assuming that MacAndrew's scale is generally 83% sensitive and 81% specific, what would its predictive values be when used in a setting in which the base rate of alcoholism is only 10%?

The predictive values can be calculated as follows. If the base rate of alcoholism is 10%, then within a sample of 1,000 individuals (any large sample size can be used), 100 would be expected to be alcoholic. Of these 100, 83 will be identified correctly by MacAndrew's scale, and 17 will obtain false negative results. A total of 81% of the remaining 900 (729) cases will be classified accurately and 171 cases will have false positive results. The positive predictive value of MacAndrew's screening scale under a 10% base rate equals 83/(83 + 171) = 33%, while the negative predictive value equals 729/(17 + 729) = 98%. Both of these values are quite different from those derived assuming a base rate of 50%.

Sensitivity, specificity, and predictive value also can be applied to less structured, less "objective" assessment situations. For example, Pfeffer et al. (1981) examined the usefulness of a 10-item mental status interview in the diagnosis of senile dementia. Of 70 patients assigned a diagnosis of dementia, a total of 25 obtained scores in the clinical range (sensitivity = 36%); all 125 nondemented control subjects had scores in the normal range (specificity = 100%). Assuming that the prevalence of dementia in a population over 65 years of age is 15% (Mortimer & Schuman, 1981), the negative predictive value of Pfeffer et al.'s (1981) interview is 90%. On the other hand, the positive predictive value is 100%. Thus, this mental status interview is somewhat more effective in identifying the presence of dementia than in determining that an elderly person is not demented.

The concepts of test sensitivity, specificity, and predictive value can be applied to more complex selection situations than use of a single test scale. For example, Voelker, Lachar, and Gdowski (1983) investigated the usefulness of the Personality Inventory for Children (PIC) in predicting whether hyperactive children responded positively to stimulant medication (methylphenidate). The PIC is a parent-informant, true-false inventory with 16 scales that reflect different aspects of child behavioral, emotional, and cognitive status (Wirt, Lachar, Klinedinst, & Seat, 1984).

Of the 46 children in the entire sample, a total of 27 (59%) were rated as being "good responders" to methylphenidate, while 19 (41%) showed little or no improvement and were rated as being "poor responders." A discriminant function analysis was conducted to identify the weighted linear combination of PIC scales that best differentiated the "good" and "poor" medication responders. When the resulting discriminant function was used to reclassify the children in these two groups, 78% of the "good" and 68% of the "poor" medication responders were classified correctly. Thus, the sensitivity of this PIC discriminant function in this sample was 78%, and the specificity was 68%.
The proportion of children in this study who were "good" medication responders (59%) is consistent with other estimates of the proportion of hyperactive children who improve subsequent to administration of stimulant medication (Freedman, 1971), so a 59% base rate will be assumed for this example. Given that 21 of 26 of the "good" medication responders and 13 of 19 of the "poor" medication responders were classified correctly by the PIC, the positive predictive value of this discriminant function equals $21/(21 + 6) = .78$, while the negative predictive value is $13/(6 + 13) = .68$. In short, the PIC discriminant function identified by Voelker et al. (1983) may be useful in predicting the response to stimulant medication of children diagnosed as hyperactive.

Kazdin, Colbus, and Rodgers (1986) recently investigated whether different diagnostic groups of children could be differentiated with several objective interview and behavioral rating scales. A total of 85 depressed and 85 nondepressed inpatients (diagnosed using Research Diagnostic Criteria; RDC; Spitzer, Endicott, & Robins, 1978) were included in the sample (base rate of depression = 50%). Kazdin et al. studied several measures, but only their findings with regard to the Children's Depression Inventory (CDI; Kovacs, 1981) are described in this example. The CDI is a child-informant inventory that assesses emotional, behavioral, and cognitive symptoms of depression (higher scores indicate higher levels of distress). Kazdin et al. also administered a rephrased version of the CDI (to indicate that the child's depression was being rated) to the children's mothers. Children were classified as depressed when they obtained scores above the cut-off on either the child (raw score >11) or parent (raw score >17) versions of the CDI.

A total of 91% (77 of 85) children assigned RDC diagnoses of depression were classified as depressed by use of the above rule (sensitivity = 91%), while 72% (61 of 85) of the children assigned other diagnoses were classified as depressed (specificity = 27%). Using the notation presented in Table 2, $A = 77$, $B = 61$, $C = 8$, $D = 24$, the positive predictive value of this CDI classification procedure equals $77/(77 + 61) = .56$, and the negative predictive value equals $24/(8 + 24) = .75$. Thus, when the base rate of depression is 50%, a total of 56% of positive CDI classifications would be expected to be accurate, while a higher proportion (75%) of the negative results would be correct.

Assuming a sensitivity of 91% and a specificity of 27%, the positive and negative predictive values of this CDI classification rule under any other base rate condition can be calculated. For example, suppose that the base rate of depression at a mental health facility for children is estimated to be 5% (based upon the frequency with which this diagnosis was assigned to all children seen over the last 2 years). In a sample of 1,000 children (again, any large number can be used), 50 would be expected to be clinically depressed; 46 (91%) would be identified correctly by the above CDI procedure, and 4 would not. Of the remaining 950 nondepressed children, 257 (27%) would obtain correct CDI results, and 693 would not. The positive predictive value would be $46/(46 + 693) = .06$, and the negative predictive value equals $257/(4 + 257) = .98$. Under a base rate condition of 5%, this CDI classification procedure would be of little use in accurately detecting the presence of depression, but would be more effective in ruling out its presence.

**LIMITATIONS OF THE SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE FRAMEWORK WITH PSYCHOLOGICAL TESTS**

While the model reviewed in this paper has proven to be very useful in medicine, there are two important limitations with regard to application of this conceptual framework with psychological tests. The first problem involves estimation of population base rates. As discussed, calculation of positive and negative predictive values requires an estimate of population base rate. While estimation of the base rates of medical or psychological disorders can be difficult (Griner et al., 1981; Meehl & Rosen, 1955),
lack of base rate information in some settings need not lead to the automatic dismissal of the framework outlined here. As suggested by Meehl and Rosen (1955), "our ignorance of base rates is nothing more subtle than our failure to compute them" (p. 213). Sources of information such as case records or tabulations of the frequencies of certain diagnoses may provide initial base rate estimates.

The use of imprecise (but not grossly inaccurate) estimates of population base rate may not actually have a large impact upon calculations of expected predictive values, especially for tests with high sensitivity and specificity (Griner et al., 1981). For example, suppose that a test is 90% sensitive and 90% specific. If the prevalence of some disorder is mistakenly estimated as 20% when it is actually 40%, the expected positive predictive value estimate using the erroneous base rate will be in error by being 17% too low (positive predictive value based on 20% base rate = 69%; based on 40% base rate = 86%), while the expected negative predictive value will be in error by being 4% too high (negative predictive value based on 20% base rate = 97%; based on 40% base rate = 93%). With higher sensitivity and specificity values, the effect of an incorrect base rate estimate of the same magnitude upon calculated positive and negative predictive values becomes smaller, while lower sensitivity and specificity test values increase the errors associated with imprecise base rate estimates.

Another limitation concerns estimation of the sensitivity and specificity indices based upon diagnostic classifications that are not reliable or valid. For example, Kline (1988) reviewed the results of several recent investigations of the convergence of psychiatric diagnoses assigned to children and scores from related behavior rating scales completed by parents. Because most of the objective checklists used in these studies have cutting scores, many researchers have calculated sensitivity and specificity indices based upon the diagnostic groups included in their investigations. Interpretation of these results is confounded, however, by the fact that several researchers have not estimated the reliability of the diagnostic classifications that they have studied. In view of the fact that the evidence for the validity and interrater reliability of several diagnostic categories for children (and adults) is weak, the generalizability and accuracy of sensitivity and specificity estimates derived using unsubstantiated diagnostic distinctions are dubious. When psychiatric diagnostic groups (e.g., depressed vs. nondepressed) are used to estimate the sensitivity and specificity of psychological tests, evidence with regard to the reliability of each diagnostic category should be presented. For instance, Kazdin et al. (described above) reported that the kappa value (which indicates chance-corrected interrater agreement) for RDC depression diagnoses assigned in their study was a satisfactory 75%.

CONCLUSION

Our presentation of sensitivity, specificity, and predictive value has implications for test users, test developers, and instruction in assessment. For test users, the utility of the framework presented here is, we hope, fairly obvious. First, test users can quantitatively estimate the accuracy of different test outcomes. Secondly, test users can compare the accuracy of two or more alternative tests. To accomplish these aims, test users will need to concern themselves more with sensitivity/specificity and base rate parameters, rather than with the overall proportion of correct predictions made by a test.

For developers of tests with cutting scores, our suggestions are both simple and direct: (1) calculate and publish sensitivity and specificity estimates; and (2) prepare and publish a graph of positive and negative predictive values under varying population base rates (e.g., Figure 1) for new assessment instruments. The preparation of such a graph would ease the calculational burden on the test user, who simply could consult the graph to determine the positive or negative predictive value of an instrument for a given base rate.
Finally, our framework has implications for instruction and training. Instructors of assessment courses need to stress the importance of population base rates and their effect upon the degree of confidence that should be placed in particular test outcomes. This emphasis on base rates and on sensitivity/specificity values could lead trainees to more rapid diagnostic sophistication.

REFERENCES
