View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Carolina Digital Repository

# A Primer on Receiver Operating Characteristic Analysis and Diagnostic Efficiency Statistics for Pediatric Psychology: We Are Ready to ROC

Eric A. Youngstrom,[1,2] PhD
[1]*Department of Psychology and* [2]*Department of Psychiatry, University of North Carolina at Chapel Hill*

All correspondence concerning this article should be addressed to Eric A. Youngstrom, PhD, Department of Psychology, University of North Carolina at Chapel Hill, Davie Hall CB 3270, Chapel Hill, NC 27599-3270, USA. E-mail: eay@unc.edu

Received March 6, 2013; revisions received May 27, 2013; accepted July 9, 2013

**Objective** To offer a practical demonstration of receiver operating characteristic (ROC) analyses, diagnostic efficiency statistics, and their application to clinical decision making using a popular parent checklist to assess for potential mood disorder. **Method** Secondary analyses of data from 589 families seeking outpatient mental health services, completing the Child Behavior Checklist and semi-structured diagnostic interviews. **Results** Internalizing Problems raw scores discriminated mood disorders significantly better than did age- and gender-normed *T* scores, or an Affective Problems score. Internalizing scores <8 had a diagnostic likelihood ratio <0.3, and scores >30 had a diagnostic likelihood ratio of 7.4. **Conclusions** This study illustrates a series of steps in defining a clinical problem, operationalizing it, selecting a valid study design, and using ROC analyses to generate statistics that support clinical decisions. The ROC framework offers important advantages for clinical interpretation. Appendices include sample scripts using SPSS and R to check assumptions and conduct ROC analyses.

**Key words** diagnostic efficiency; evidence-based medicine; receiver operating characteristic analysis; sensitivity and specificity.

A 10-year-old girl comes to our medical clinic for a psychological evaluation assessing factors that might contribute to problems adhering to her diabetes management regimen. Her parents complete the standard paperwork, including an Achenbach Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001), which our clinic uses as a brief, broad measure to identify if there are any emotional or behavioral concerns that might complicate treatment or warrant intervention in their own right. She earns a raw score of 27 on Internalizing problems, and a 7 on a Diagnostic and Statistical Manual of Mental Disorders (DSM)-oriented Affective Problems scale, and a *T*-score of 76, based on comparing her score with other girls in her age range in the standardization sample. The CBCL is widely used in clinical settings and in research, and it has accumulated evidence of validity for diagnoses of depression (Warnick, Bracken, & Kasl, 2008). But what does the score mean in the context of this individual patient? We know that youths with depression tend to score higher, on average, on these scales, but can we translate her score into an estimate of the probability that this girl has depression? What should we do next . . . more assessment? Refer for treatment for depression?

Signal detection theory (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000) and Bayesian methods (Bayes & Price, 1763; Kruschke, 2011) provide a statistical and conceptual framework for taking the research data and translating them into direct answers to these practical clinical questions. This primer illustrates the application of receiver operating characteristic (ROC) analysis and related diagnostic efficiency statistics to a research data set, using two popular statistical programs, SPSS and R, to run the

*Journal of Pediatric Psychology* 39(2) pp. 204–221, 2014
doi:10.1093/jpepsy/jst062
Advance Access publication August 21, 2013
*Journal of Pediatric Psychology* vol. 39 no. 2 © The Author 2013. Published by Oxford University Press on behalf of the Society of Pediatric Psychology.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

analyses. The primer compares and contrasts traditional ways of assessing criterion validity versus ROC, and illustrates methods for checking assumptions, running the main analyses, and generating figures. Graphical methods play a central role in the ROC approach to evaluating tests. The primer provides guidance about making informed choices of cut scores, and then packaging the findings in a way that promotes clinical decision making. Table I lays out a larger context of where ROC and related methods fit in a fully developed program that moves from basic assessment research to clinical decisions with an individual patient. There have been recent advances in guidelines and recommendations for STAndardized Reporting of tests of Diagnostic assessments (the STARD Guidelines; Bossuyt et al., 2003) and tools to help critically evaluate reporting of results (Whiting et al., 2011); there are excellent treatments of how to apply Bayesian methods to clinical decision making within an evidence-based medicine (EBM) framework (Straus, Glasziou, Richardson, & Haynes, 2011). These constitute important foreground and background for the role of ROC, guiding decisions about which assessment methods are contenders for clinical use, and how to implement them in practice. ROC and related methods are the engine for statistically appraising a test's performance at classifying cases into groups correctly—such as those with versus without mood disorder. These methods also provide a statistical process for comparing different tests and deciding whether one is superior for making these classification decisions. Although these methods are often presented in the context of diagnostic

Table I. *Steps in Designing, Conducting, Reporting, and Interpreting Receiver Operating Characteristic Analyses to Support Clinical Decision Making*

| Step | Research study | Clinical application to patient |
|---|---|---|
| 1. Define the clinical topic and criterion variable (e.g., diagnosis) | Operational definition of dependent variable (''reference standard''), usually defined as dichotomous, yes/no or present/absent (Bossuyt et al., 2003) | Definition of clinical decision that test result will help evaluate (Straus et al., 2011) |
| 2. Select the predictor (i.e., ''index test'') | Select ''index test'' (Bossuyt et al., 2003). If multiple candidates available, pick based on effect size from group comparisons with strong designs | Critically review published studies to focus on designs that are likely to yield unbiased and clinically generalizable estimates (Whiting et al., 2011) |
| 3. Select an appropriate sample and research design | Make sure that the criterion diagnosis was made blind to the predictor test result. Have study inclusion and exclusion criteria, clinical, and demographic characteristics duplicate the intended clinical usage as much as possible | Check the methods of the study generating the ROC results for strong, unbiased designs (Whiting et al., provide a checklist). Decide whether sample and patient characteristics are a good match |
| 4. Determine the criterion validity of the predictor | **Conventional methods:** *t*-test and Cohen's *d* to compare groups, or point-biserial correlations or phi coefficient<br>**Better methods:** ROC analysis followed up with diagnostic efficiency statistics[a]<br>**Best methods:** ROC followed by multilevel diagnostic likelihood ratios[a] | Focus on effect sizes and validity of design. If article or manual only reports *d*, can convert to estimates of AUC and sensitivity–specificity. Prefer designs that use unbiased and clinically generalizable definitions, even if effect size looks smaller than biased designs |
| 5. Compare performance versus other samples or tests | Use tests of independent AUCs to compare published results,[a] or more powerful tests of dependent curves if multiple measures available in the same sample (DeLong et al., 1988; Hanley & McNeil, 1983; Robin et al., 2011)[a] | Critically appraise different tools and decide what would be most appropriate for patient (Straus et al., 2011) |
| 6. Optimize cut-score thresholds for decisions | Evaluate costs and benefits and alter choice of cut score depending on clinical setting, goal, and utilities (Kraemer, 1992; Swets et al., 2000) | Discuss risks, benefits, and patient preferences, and adjust wait-test and test-treat thresholds to decide next action |
| 7. Evaluate clinical applicability | Look at test positive rate (''level'') (Straus et al., 2011), positive and negative predictive powers under plausible scenarios; present natural frequencies (Gigerenzer & Hoffrage, 1995)[a] | Use probability nomogram or applet to combine test result DLR with other information from risk factors, independent tests; conduct ''sensitivity analyses'' to illustrate range of probabilities; discuss next action with patient (Straus et al., 2011) |
| 8. Make the results and test easy to use | Report findings according to STARD recommendations (Bossuyt et al., 2003). Provide DLRs in article; provide nomogram or link to applet; compare results with other tests so reader can make informed decision[a] | Have ''portfolio'' with nomogram, decision support information available for commonly used tests and presenting problems (Youngstrom, 2013) |

[a]Denotes statistical analysis detailed in this primer.

decisions, they could be used with any dichotomous variable, such as predicting treatment responder/nonresponder status, or probability of dropping out of treatment. Beyond the scope of this primer, these methods can extend to scenarios with multiple categories (Robin et al., 2011) or continuous dependent variables (Kruschke, 2011).

This primer concentrates on the case where there is a clinically important dependent variable with two categories, such as mood disorder status, and our goal is to appraise test scores as predictors of that status and describe their diagnostic efficiency in clinical practice. In fundamental ways, this reverses the traditional research design: Instead of sorting a large group of cases into two groups, those with and those without depression, and then using a *t*-test or a nonparametric analog to evaluate whether the group distributions are significantly different, ROC flips the variables so that the categories are the dependent variable and the test score is the predictor. The presentation here relies on a minimum of statistical formulae, and the Appendix presents syntax to duplicate these analyses in SPSS and R. The data used here are available as well, so that interested readers can duplicate the analyses and then ''reverse engineer'' them to apply the methods to new data.

ROC is a more natural model of how clinicians need to work. We obtain test results for an individual person, and then we need to make high-stakes decisions about the person's chances of having a diagnosis or particular outcome. The raw data we need for these methods are readily available. Any data set that generated a *t*-test or a $\chi^2$ could be reanalyzed using ROC. As we will see, ROC methods generate the building blocks to link group data to individual probabilities of diagnosis, and from there to clinical decisions about the next action.

This primer will use data from a project designed to evaluate several behavior checklists as potential aids in the evaluation of mood disorder (Youngstrom et al., 2005; NIH R01 MH066647) to illustrate the steps in designing, analyzing, and applying ROC analyses. The project enrolled a consecutive case series at an outpatient clinic, had caregivers complete the CBCL, and had highly trained interviewers complete semi-structured diagnostic interviews using the Schedule for Affective Disorders and Schizophrenia for Child and Adolescents (KSADS; Kaufman et al., 1997). A consensus review process finalized diagnoses, synthesizing clinical and interview findings, but staying blind to the results of the CBCL and other checklists to prevent criterion contamination. Table II presents key demographic and clinical characteristics; additional details about method and procedure are in the article by Youngstrom et al. (2005). Rather than following the conventional sections of a primary research report, the primer follows the steps delineated in Table I and provides more information about design and analytic choices than typically would be included in a research report. Additional technical details are embedded as comments in the example syntax in the Appendices.

## Steps in Applying ROC Analyses to Data and to Individual Cases

The next sections follow the sequence outlined in Table I, discussing issues in data analysis and application to clinical

Table II. *Descriptive Statistics for Clinical and Demographic Variables, and Bivariate Tests of Association With Mood Disorder Status*

| Variable | Any mood ($n = 241$) | No mood ($n = 348$) | Test statistic | *p* | Effect size |
|---|---|---|---|---|---|
| Age in years | | | | | |
| *M* | 11.70 | 9.93 | $t$ (587 $df$) = 6.49 | <.0005 | $d = .62$ |
| *SD* | 3.47 | 3.08 | Levene's $F = 3.08$ | .080 | |
| Female | $n = 118$ (49%) | $n = 120$ (35%) | $\chi^2$ (1 $df$) = 12.40 | <.0005 | phi = .15 |
| Race (African American %) | $n = 208$ (86%) | $n = 316$ (91%) | $\chi^2$ (3 $df$) = 4.55 | .208 | – |
| Comorbid diagnoses (count) | | | | | |
| *M* | 3.29 | 2.17 | $t$ (425.1 $df$) = 10.25 | <.0005 | $d = .86$ |
| *SD* | 1.43 | 1.09 | Levene's $F = 23.96$ | <.0005 | |
| Internalizing raw total | | | | | |
| *M* | 20.26 | 12.08 | $t$ (426.7 $df$) = 10.31 | <.0005 | $d = .91$ |
| *SD* | 10.38 | 7.97 | Levene's $F = 19.85$ | <.0005 | |
| Internalizing *t*-score | | | | | |
| *M* | 67.77 | 59.98 | $t$ (541.9 $df$) = 9.71 | <.0005 | $d = .80$ |
| *SD* | 9.24 | 10.03 | Levene's $F = 4.88$ | .028 | |
| Affective disorders raw | | | | | |
| *M* | 4.29 | 3.01 | $t$ (485.7 $df$) = 5.77 | <.0005 | $d = .49$ |
| *SD* | 2.75 | 2.51 | Levene's $F = 4.96$ | .026 | |

cases. The presentation weaves these together, because the direct connection of analysis to clinical decision making is a strength of ROC. Keeping the clinical goal in mind also clarifies many considerations about research design and analyses.

### *Step 1. Define the Clinical Topic and the Criterion Variable*

The first step in using ROC methods is to select a clinical problem and operationally define it. In our example, the clinical issue is evaluating whether someone has ''depression.'' The operational definition should specify what constitutes ''depression''—does it subsume dysthymic disorder? Depression not otherwise specified? Also crucial is deciding the research design and construction of the ''reference standard.'' There are now guidelines about reporting the results of studies evaluating diagnostic efficiency (Bossuyt et al., 2003) and checklists for evaluating diagnostic validity and identifying possible sources of bias (Whiting et al., 2011).

The choice of whether to use a ''broad'' or ''narrow'' definition of depression deserves some thought. Focusing on a more narrow definition will change the results of the ROC analysis, and it also determines how the results should be used in practice. Focusing only on major depressive disorders may make it easier to detect the target cases, because the target cases will have a more severe and clear presentation; but it also could make it harder to classify the other cases correctly, because cases with dysthymic disorder might also score high on a measure of internalizing problems, but be classified as ''not major depression'' by the narrow definition of the reference standard (Zhou, Obuchowski, & McClish, 2002).

Likewise, when a clinician applies the results of the ROC analyses, it is vital to keep in mind the operational definition of the diagnosis or outcome. Our example will predict ''any mood disorder,'' guided by the logic that our goal is to identify cases for further evaluation. Inasmuch as we would also want to detect dysthymic disorder or other mood disorders and adjust our treatment planning similarly, it makes sense to use a broad definition. Our operational definition of depression included diagnoses of mood disorder (bipolar disorder with depression, unipolar depression, dysthymic disorder, depression not otherwise specified) based on a semi-structured diagnostic interview of both the youth and parent by highly trained and closely supervised raters who then reviewed findings with a licensed psychologist to produce a consensus diagnosis (Youngstrom et al., 2005). Both the reliability and validity of the diagnoses were high based on the methods.

### *Step 2. Select the Predictor*

The next step is to select the assessment instrument to evaluate as a potential predictor of the diagnosis. In the medical decision-making literature, the predictor is often called the ''index test'' (Bossuyt et al., 2003), and the specific cut score or interpretation algorithm is sometimes called the ''referent'' (Kraemer, 1992). Here we use *predictor*, recognizing that it could be predicting either a concurrent diagnostic status or a future outcome. All of the usual criteria in selecting a research measure apply: It should have adequate reliability, good construct validity, and so forth. Reliability is important for the precision of classification.

Criterion validity is the crucial element for a candidate for ROC analysis, though—the potential predictor needs to be statistically associated with the reference standard diagnosis, or there is no point in studying it further. When designing a new study or selecting a test as a clinician, criterion validity helps triage the instruments. Whether articles report group-based statistics such as correlation coefficients, *t*-tests, or $\chi^2$, the result needs to be statistically significant for the instrument to be a contender for individual classification.

The CBCL is a logical candidate for a predictor because it is well-validated, widely used, and has demonstrated criterion validity for anxiety and mood disorders. A quick PubMed search finds several articles that have already applied ROC to the CBCL (Ferdinand, 2008), and a recent meta-analysis reviewing performance of the CBCL for predicting several diagnoses (Warnick et al., 2008). However, the studies were neither consistent in which scale they used as the predictor, nor in the operational definition of the target diagnoses. We focus on the Internalizing Problems score, because it has high reliability and has demonstrated criterion validity with regard to mood disorders (Achenbach & Rescorla, 2001). We also will test whether the DSM-oriented Affective Problems score performs significantly better, given that experts selected its item content to be more specific to mood disorder (Achenbach & Rescorla, 2001; Lengua, Sadowski, Friedrich, & Fisher, 2001).

Another key consideration is the amount of shared method variance between the predictor and criterion. Shared method variance will exaggerate the apparent association. If the predictor involves someone reading a questionnaire, and the reference standard is someone else reading another questionnaire aloud to the participant as a structured interview, the source methods are similar, and the correlation between the ''predictor'' and ''criterion'' will be extremely high (cf. Steer, Cavalieri, Leonard, & Beck, 1999). Such a design would overestimate how helpful the predictor would be in a clinical setting. At the other

end of the validity continuum would be a reference standard that incorporates information from multiple sources, such as structured or semi-structured interviews with the parent and the youth, along with direct observation of mental status, integration of developmental and treatment history, and perhaps even neurocognitive or biological assay results. Synthesizing information from multiple sources will avoid spuriously inflating the association due to shared method variance between the predictor and criterion (Campbell & Fiske, 1959). The concept of external validity or generalizability is vital: More valid designs use predictors and criteria that best model what would be useful in clinical practice.

### *Step 3. Select an Appropriate Sample and Research Design*

Not all samples will be well-suited for ROC analyses. If the criterion diagnosis was made based on the predictor, then there is ''criterion contamination,'' and the results will literally be too good to be true. Blinding, or recusing the predictor from the construction of the criterion diagnosis, is essential to generate valid estimates of the accuracy of the prediction when clinicians will use the predictor by itself (Bossuyt et al., 2003). The present study was designed to evaluate the diagnostic efficiency of several tests, so the CBCL was gathered by a separate research assistant, and the criterion diagnoses were blind to CBCL results (Youngstrom et al., 2005).

Sample composition also is a major consideration. Ideally, the circumstances of data collection will closely mimic how clinicians might use the test in practice. Inclusion and exclusion criteria for the study sample should approximate the clinical context for intended use. Consecutive case series designs or random sampling would provide a strong degree of validity (Straus et al., 2011).Many research designs that would be valid for other purposes could produce dangerously misleading results if repurposed for an ROC analysis. A common example would be designs that combine a clinical diagnostic group with healthy controls (Barrera & Garrison-Jones, 1988), or samples that blend distilled groups that initially were screened with a variety of exclusion criteria that increase the internal validity of the design for its original purpose, such as an efficacy clinical trial, but reduce the generalizability. Changing the composition of the comparison group will directly influence the diagnostic specificity of the predictor (compare Tillman & Geller, 2005, with Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). Clinicians usually are not confronted with decisions about whether the individual has depression versus no mental health issue at all; but rather they are trying to decide whether depression is a concern out of the full spectrum of typical diagnoses that might present at a clinic.

The sample here was a consecutive case series at an urban community mental health center, with the only inclusion requirement being an ability to complete the interview and measures in English, and the only exclusion being a diagnosis of cognitive disability or pervasive developmental disorder. Table II reports the demographics, basic clinical features, and CBCL descriptive statistics.

### *Step 4. Determine the Criterion Validity of the Predictor*

An ROC analysis is one way of testing the criterion validity of a predictor. Despite several advantages, it is not yet a common way of reporting results in the pediatric psychology literature. Articles and technical manuals are much more likely to present statistical significance, correlation coefficients, or effect sizes such as *d* (Cohen, 1988). For example, the CBCL manual reports a point-biserial correlation of .45 between the Internalizing Problems score and clinical diagnoses of mood disorder based on 134 youths seen at the clinic in Rochester, Vermont (Achenbach & Rescorla, 2001, p. 130). Effect sizes are fungible, and meta-analysis capitalizes on the fact that it is possible to convert one effect size into the other (Lipsey & Wilson, 2001). Cohen's *d* and the area under the curve (AUC) from an ROC analysis both quantify the amount of separation between the distribution of score for the two groups of interest, those with and without the criterion diagnosis.

Table II reports the results of a *t*-test and Levene's *F*-test of the homogeneity of variance for the CBCL scales. Table III reports the correlations among variables. The presence of mood disorder was positively correlated with age and female gender, consistent with risk of depression increasing in adolescence and in females (Cyranowski, Frank, Young, & Shear, 2000). Mood diagnosis also showed medium and large correlations with the CBCL scales compared with Cohen's (1988) rules of thumb. Age and gender show a small but significant correlation with the raw Internalizing Problems score, but not the *T*-score, reflecting how the age and sex norms adjust for the tendency of female adolescents to have somewhat higher raw score on average (Achenbach & Rescorla, 2001). Overall, the findings indicate good criterion validity for the CBCL scales. However, the results do not provide guidance about how to interpret an individual case's scores.

#### Examining Criterion Validity via ROC

ROC analyses use the same variables as *t*-test or point-biserial correlations, but using the index test as the input, and the diagnostic category as the criterion. ROC evaluates

Table III. *Correlations Among Variables (*N = *589)*

| Variable | Female | Age in years | Internalizing raw score | Internalizing *T* score | Affective disorder raw |
|---|---|---|---|---|---|
| Any mood diagnosis | 0.15***[a] | 0.26***[,b] | 0.41***[,b] | 0.37***[,b] | 0.24***[,b] |
| Female | | 0.19***[,b] | 0.10*[,b] | 0.05[b] | 0.06[b] |
| Age in years | | | 0.09* | 0.06 | −0.13** |
| Internalizing raw | | | | 0.94*** | 0.76*** |
| Internalizing *T* | | | | | 0.73*** |

[a]Phi coefficient.

[b]Point-biserial correlation; all others are Pearson *r* correlations.

*$p < .05$, **$p < .005$, ***$p < .0005$, two-tailed.

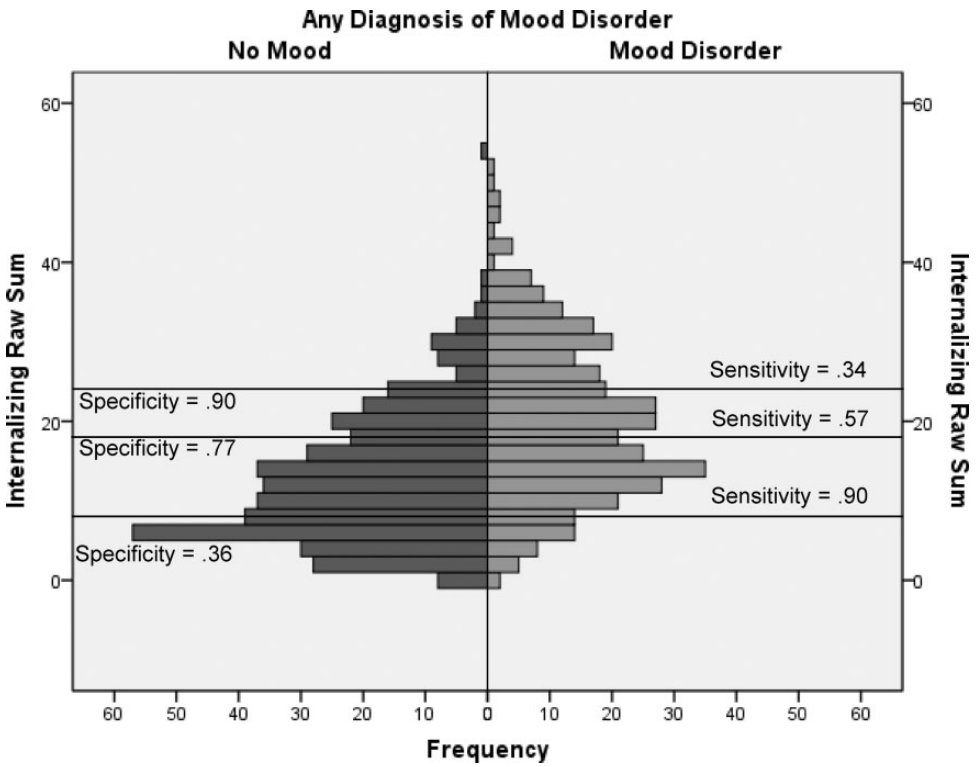


**Figure 1.** Population pyramid of raw Internalizing Problems score distributions for those with a diagnosis of any mood disorder versus no mood disorder, *N* = 589. *Note.* Generated in SPSS. Superimposed lines indicate three proposed cut scores: raw Internalizing score of 8+ (90% sensitivity), 18+ (maximum kappa = 0.34, based on 41% prevalence), and 24+ (90% specificity).

the trade-off between diagnostic specificity versus sensitivity. *Specificity* refers to the accuracy of the test for the cases that do not have the target condition. Its complement is the ''false alarm'' rate, or how often cases that do not have the diagnosis would incorrectly score positive on the index test—specificity plus false alarm rate always sum to 1.0 (see glossary in Figure 5 for derivation and summary of terms). *Sensitivity* describes accuracy among those who do have the diagnosis. It is always possible to achieve perfect sensitivity by diagnosing all cases with the condition, but this would also have a 100% false alarm rate and specificity of 0%. Conversely, perfect specificity is always attainable by never diagnosing any cases; of course, this strategy also yields a sensitivity of 0%, as none of the cases with the disorder would be diagnosed, either. Neither of these strategies is useful in most clinical applications (cf. Kraemer, 1992; Pulleyblank, Chuma, Gilbody, & Thompson, 2013; Youngstrom, 2013). Ideally, there would be a cut score or threshold on the predictor that would separate those with the diagnosis from those without it. Moving the cut score higher on Internalizing Problems, where high scores denote more pathology, improves the specificity of a test and reduces the false alarm rate, but at the price of potentially reducing sensitivity to cases that have the diagnosis.

Figure 1 is a ''population pyramid'' or ''back to back histogram'' comparing the distribution of raw Internalizing

Problems scores for cases without mood disorder versus those with any mood disorder based on the KSADS reference standard. If the cut score were set at a 0 or higher, all of the cases with mood disorder would exceed the threshold, yielding 100% sensitivity; but all the cases without mood disorder also exceed that threshold, resulting in 100% false alarms. The score distribution was higher in the ''Mood'' group, consistent with the results of the *t*-test and the point-biserial correlation. Raising the cut score to a 1, so that scores of 0 are considered ''negative'' test results, but scores of 1 or higher are ''positive'' test results, would correctly classify seven of the cases without mood disorder, reducing the false alarm rate to 98%—still unimpressive. However, even this small adjustment in the cut score misclassifies one of the cases with a mood diagnosis, reducing the sensitivity to 99.6%. If there is any overlap in the two distributions, then it is impossible to find a cut score that could separate the two groups with 100% accuracy, delivering perfect sensitivity and specificity at the same time. Moving to the other extreme, the cut score would need to be 53 or higher to eliminate all false alarms, at which point the sensitivity would have dropped to 0.

An ROC curve plots the sensitivity of the test as a function of the false alarm rate (or sometimes the specificity, producing a mirror image with the same AUC). Most software packages present false alarms on the x-axis and sensitivity on the y-axis. Figure 2 presents the ROC plot for the three index tests plotted simultaneously. The top right corner has the coordinates (false alarm = 100%, sensitivity = 100%). It corresponds to setting the cut score at 0, with a 100% test positive rate. The empirical ROC curve then raises the cut score one point at a time, plotting the combination of the false alarm rate and sensitivity, until the highest scores observed in the data are plotted. The ROC curve visually summarizes the trade between decrement in sensitivity and improving specificity (false alarm reduction) as the cut score becomes more stringent. A perfectly discriminating test would reach the top left corner, including 100% sensitivity and 0% false alarms on the curve. The diagonal line represents chance performance. Visually, the closer the ROC curve comes to the top left corner, and the further it is from the random ROC line on the diagonal, the better job it does discriminating the target condition. The raw Internalizing score appears to be doing the best of the three index tests based on the position of its curve.

ROC can quantify the accuracy of the test by estimating the AUC. The AUC can be estimated using a variety of parametric distributional assumptions, or it can be estimated nonparametrically (Zhou et al., 2002). It also is possible to test the AUC against the null hypothesis of chance performance, or that the AUC in the population is 0.50.

With nonparametric estimation, ROC actually requires fewer distributional assumptions than would *t*-test, analysis of variance, or correlation. ROC does not assume a normal distribution for the index test, so skewness and

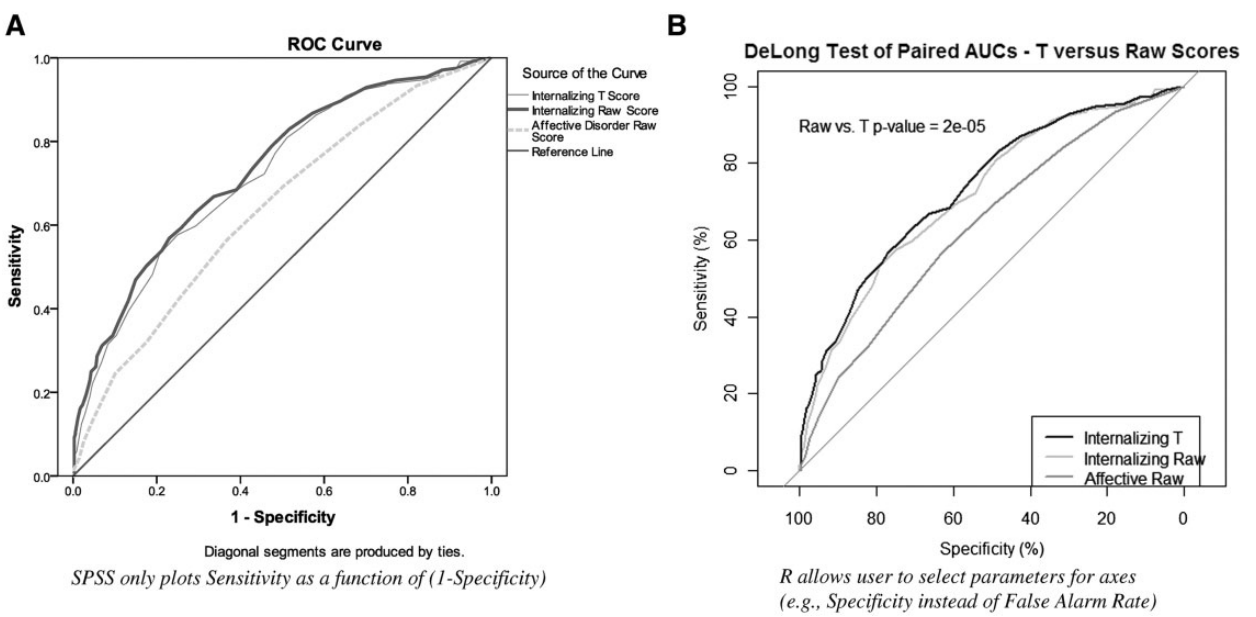


**Figure 2.** Receiver Operating Characteristic (ROC) curves for index tests from the CBCL predicting mood disorder diagnoses (41% base rate; $N = 589$). (**A**) SPSS ROC procedure – plotting three index tests (**B**) pROC Package in R – DeLong test of difference between Internalizing raw and *T* scores.

kurtosis evident in Figure 1 are not intrinsically problematic. Unlike *t*-test or analysis of variance, nonparametric ROC does not assume homogeneity of variance, either (cf. Table II). However, there are situations where distributions can create problems. If the score distribution within a diagnostic group is bimodal, or if there are regions where the score frequencies do not progress monotonically, then the derived estimates will not behave monotonically, either. Similarly, if the group that has a lower median score on the index test also has higher extreme scores, either due to outliers or overdispersion, then estimates of test accuracy in the extreme score range will not be accurate. These are examples of what are termed ''degenerate'' distributions (Zhou et al., 2002), and both are evident in Figure 1: The highest observed Internalizing score comes from a case without mood disorder, and there are ''notches'' in both histograms where moderately high scores are slightly less common than the slightly more elevated scores. A variety of smoothing operations, or bootstrapping, could address degeneracy. In practice, Kraemer (1992) offers a rule of thumb of not reporting sensitivity or specificity unless there are at least 10 cases at each marginal position in a $2 \times 2$ table of the data—in other words, only report diagnostic efficiency statistics when there are at least 10 cases that have the diagnosis, 10 that do not, 10 that test positive, and 10 that test negative. Functionally, this means ignoring the extremely low and high cut score thresholds, and concentrating on the score ranges where the data will be most informative. A visual plot such as a population pyramid will often be the most efficient way of detecting these potential problems for the ROC analysis.

Table IV presents the AUC statistics for all three predictors. All were statistically significant, and the 95% confidence intervals do not include the null hypothesis of 0.50. The AUC quantifies the degree of nonoverlap in the mood and nonmood groups of scores. Conceptually, the AUC can be interpreted as the probability that a randomly selected case with mood disorder would have a higher score on the index test than a randomly selected case without mood disorder. Swets and others have offered benchmarks for gauging AUCs, suggesting that values $\geq 0.9$ are ''excellent,'' $\geq 0.80$ ''good,'' $\geq 0.70$ ''fair,'' and $< 0.70$ ''poor.'' These are probably appropriate for engineering and some biomedical applications, but in the context of mental health diagnoses, they are less representative. The AUC is constrained by the reliability and validity of the reference standard: If the criterion diagnosis is imperfect, then it is impossible for the AUC to reach 1.00 (Kraemer, 1992; Pepe, 2003). In practice, many of the best-performing behavior checklists and inventories currently available deliver AUC estimates in the 0.7–0.8 range under clinically realistic conditions and with valid reference standard diagnoses. When questionnaires produce AUCs greater than 0.90, it is more likely to indicate design flaws rather than exceptional discriminative validity (Youngstrom et al., 2006).

### *Step 5. Compare Performance Versus Other Samples or Tests*

How do the three index tests compare with each other in terms of discriminating mood disorders? The AUC estimate is highest for the raw Internalizing score, and the confidence intervals for it do not overlap with the confidence interval for the Affective Problems score, indicating that they are significantly different. Because all three predictors were evaluated in the same sample, much more statistically powerful methods can test whether their performance differs significantly. SPSS does not include any of these methods as of version 20, but it is possible to use the method proposed by Hanley and McNeil (1983) using six pieces of information from the sample: The two AUC values and their standard errors, plus the correlation between the two predictors in the subgroup without the diagnosis and the subgroup with the diagnosis. The appended SPSS syntax uses the ''split file'' routine as a simple way of generating the two correlations. The Internalizing raw and *T*-scores correlated $r = .937$ in the cases without mood, and $r = .952$ in the cases with mood. Plugging these numbers plus the AUC and standard errors into the formula from Hanley and McNeil yields $z = 1.60$, $p = .1098$, suggesting that the two are not significantly different (the section titled ''Step 5A'' in the Excel spreadsheet implements the necessary calculations if the reader wants to use the method). In contrast, both identify mood

Table IV. *Area Under the Curve From Receiver Operating Characteristic Analyses*

| | | | | 95% Confidence interval | |
|---|---|---|---|---|---|
| Index test | Area under curve | Standard error | *p* value | Lower | Upper |
| Internalizing *t*-score | 0.720 | 0.021 | <.0005 | 0.678 | 0.761 |
| Internalizing raw score | 0.735 | 0.021 | <.0005 | 0.694 | 0.776 |
| Affective disorders raw score | 0.638 | 0.023 | <.0005 | 0.593 | 0.683 |

disorder significantly better than the Affective Problems scale, $z = 5.81$, $p < .00005$ for raw Internalizing and $z = 4.57$, $p < .00005$ for *T*-scores. These results indicate that the Affective Problems scale is significantly less accurate than either of the other scales at discriminating mood disorders; an ironic finding, given that it was designed to more closely conform to the DSM diagnostic criteria.

The pROC package in R (Robin et al., 2011) includes several statistical tests of the difference between two paired ROC curves estimated in the same sample. The Hanley and McNeil approach is one option, but pROC combines it with bootstrapping to provide more accurate estimates of the standard errors, defaulting to 2000 replications sampled with replacement. pROC also includes the DeLong test for paired ROC curves (DeLong, DeLong, & Clarke-Pearson, 1988), which also has more power and precision than the methods currently available in SPSS. Based on the DeLong test, the difference in performance between the raw and *T*-score also achieves statistical significance, $p = .00002$ (shown in Figure 2, Panel B). As shown here, small differences in AUC can be statistically significant if the predictors are strongly correlated; it is important to use appropriate statistical tests rather than just inspecting confidence intervals.

It also is possible to compare diagnostic performance between different samples. Ferdinand (2008) reported that the CBCL Affective Problems scale earned an AUC of 0.83 with a standard error of $\pm 0.03$, predicting semi-structured interview diagnoses of major depression and dysthymia. Hanley and McNeil also provided a formula for testing the difference of AUC values derived from independent samples, and the two AUC coefficients and their standard errors are sufficient statistics. Comparing Ferdinand's results with those in Table IV, the Affective Problems score performed significantly less well in the present data, $z = 5.05$, $p < .00005$ (available in ''Step 5B'' in the Excel spreadsheet).

The CBCL manual reported a point-biserial correlation instead of an AUC. Hasselbad and Hedges (1995) provided formulae for converting *r*, *d*, sensitivity and specificity, or descriptive data parameters into each other (see Supplementary Excel file, ''Supporting tools for converting other published results into AUC estimates''). The correlation of 0.45 for Internalizing and diagnosis translates to a *d* of 1.01 and an AUC of 0.762. Comparing these values with the AUC from the present sample generates a $z = 1.00$, $p = .3192$, indicating that the differences between the estimates in our data and Achenbach's clinic are not statistically significant. The online Supplementary Excel spreadsheet also implements the Hanley and McNeil (1983) test. If several different published estimates were available, then all of the effect sizes could be converted into the same metric and then tested for homogeneity. Both the Hanley and McNeil test and the meta-analytic test of homogeneity address the generalizability of the results across samples. If these indicate significant differences, then a next step would be to identify variables moderating diagnostic accuracy. Clinicians confronted with significantly different estimates should focus on the estimates generated by the more valid design (Whiting et al., 2011) and where the participants look most similar demographically and clinically to the patient being evaluated (Straus et al., 2011).

### *Step 6. Optimize Cut Score Thresholds for Decisions*

The next step is to select a cut score and evaluate the diagnostic efficiency statistics. The choice of optimal threshold depends on three sets of factors: (1) the intended use of the test, (2) the base rate of the disorder in the clinical setting, and (3) the relative costs and benefits attached to correct classification and errors. If the goal is to use an index test as a screener, then high sensitivity is more important than specificity, because the goal is to avoid missing cases that truly have the target diagnosis; and conversely, applications using the index test as diagnostic confirmation would put more of a premium on specificity (Kraemer, 1992). The base rate directly affects the overall accuracy of classifications, as well as the positive and negative predictive powers of the test, whereas sensitivity and specificity are algebraically unrelated to base rate (Pepe, 2003) (see Glossary as well). *Positive predictive power* describes the percentage of cases testing positive that actually have the diagnosis, and *negative predictive power* is the accuracy rate of negative test results. These are clinically intuitive and helpful rates, but they change as a function of the rate of the diagnosis (as will become obvious in the following examples). Cohen's *kappa* is a measure of accuracy that adjusts for both the base rate and the percentage of cases testing positive; in fact, kappa is the special case of a more general family of methods for calibrating test performance, where kappa weights the costs of false-positive and false-negative results equally (Kraemer, 1992). There are more advanced approaches that can integrate the costs and benefits attached to the assessment when selecting optimized decision thresholds (Kraemer, 1992; Swets et al., 2000).

Without all of the cost and benefit utilities available, there are three pragmatic approaches: (1) pick a desired sensitivity, and evaluate the rest of the test performance around that, or conversely start with an desired specificity and work from there (Pepe, 2003); (2) select a threshold

Table V. *Different Optimal Threshold and Multilevel Diagnostic Likelihood Ratios for Internalizing Raw Scores*

| | | | | | | | Prevalence of 41% | | Prevalence of 10% | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cut Score | Sensitivity | Specificity | Kappa | Level | DLR+ | DLR− | PPV | NPV | PPV | NPV |
| 90% Sensitivity: 8+ | **0.896** | 0.362 | 0.229 | 0.744 | 1.405 | 0.287 | 0.493 | 0.834 | 0.135 | 0.969 |
| Max. Kappa: 18+ | 0.568 | 0.770 | **0.344** | 0.368 | 2.473 | 0.560 | 0.631 | 0.720 | 0.216 | 0.941 |
| 90% Specificity: 24+ | 0.336 | **0.905** | 0.253 | 0.194 | 3.544 | 0.733 | 0.711 | 0.663 | 0.283 | 0.925 |
| *Multilevel DLRs (based on sample quintiles)* | | | | | | | | | | |
| 0–6 | – | – | – | ~20% | 0.24 | – | 0.140 | – | 0.026 | – |
| 7–11 | – | – | – | ~20% | 0.62 | – | 0.301 | – | 0.065 | – |
| 12–16 | – | – | – | ~20% | 0.90 | – | 0.385 | – | 0.091 | – |
| 17–23 | – | – | – | ~20% | 1.57 | – | 0.521 | – | 0.149 | – |
| 24+ | – | – | – | ~20% | 3.54 | – | 0.711 | – | 0.283 | – |
| *Multilevel DLRs (based on more informative thresholds)* | | | | | | | | | | |
| 0–7 | – | – | – | 26% | **0.29** | – | 0.166 | – | 0.031 | – |
| 8–23 | – | – | – | 55% | **1.03** | – | 0.417 | – | 0.103 | – |
| 24–30 | – | – | – | 11% | **2.31** | – | 0.615 | – | 0.204 | – |
| 31+ | – | – | – | 8% | **7.40** | – | 0.837 | – | 0.451 | – |

*Note*. Boldface denotes the parameter specified a priori to select the test cut threshold.

based on the maximum kappa, recognizing that the kappa estimate itself is tied to the base rate, and will not generalize to settings with different rates (Kraemer, 1992); or (3) divide the index test score into multiple ranges, and estimate the diagnostic efficiency separately for each range (Straus et al., 2011). Because the raw Internalizing score performed significantly better than the Affective Problems scale and the *T*-score, we use it to illustrate different ways of evaluating the cut scores.

SPSS lists the sensitivity and false alarm rate for all observed scores by requesting that it print the coordinates of the curve (/PRINT COORDINATES subcommand). These can be copied and pasted into Excel and then transformed using the calibrations Kraemer provides (see Supplementary Excel spreadsheet section labeled ''Step 6''). The pROC ''coordinates'' function also generates all of the sensitivity, specificity, and positive and negative predictive powers, but not the kappa. Table V reports the cut scores that provide ~0.90% sensitivity (scores of 8+), maximize kappa (in a sample with a base rate of 41%, scores of 18+), and provide ~90% specificity (scores of 24+). Table V also has positive and negative predictive power estimates, based on a 41% rate of mood disorder.

### *Step 7. Evaluate Clinical Applicability*

Counter-intuitively, negative results on a highly sensitive test are more decisive than positive results. If the threshold is set low to improve sensitivity, and a case scores even lower, then it is unlikely that they have the diagnosis in question. EBM refers to this as the ''SnNOut'' heuristic—on a Sensitive test, a Negative result rules the diagnosis Out. Conversely, on highly Specific tests, Positive scores are more helpful at ruling a diagnosis In, the SpPIn heuristic (Straus et al., 2011).

Rather than relying on the SnNOut and SpPIn heuristics, though, EBM advocates using Bayes' theorem to synthesize the prior probability of diagnosis, often estimated as the base rate, with the information from the test result, to generate a revised, posterior probability estimate. Bayes' theorem has been well known and discussed for centuries, but it has not had great uptake in clinical decision making because the formula is usually presented as combinations of probabilities (McFall & Treat, 1999). Cognitive psychologists have advocated presenting results as ''natural frequencies'' instead of probabilities (Gigerenzer & Hoffrage, 1995). Figure 3 presents the results for the threshold that maximized kappa in the sample, a raw cut score of 18+ on Internalizing. The lower half of the figure illustrates how the base rate directly changes the positive and negative predictive powers in a new setting. Cognitive psychologists suggest using this format to present test results to patients as well as in research to facilitate understanding.

Other alternatives include using online calculators (simple examples are included in ''Step 7'' in the Excel spreadsheet) or a probability nomogram (Figure 4) to combine prior probabilities with test results. To use the probability nomogram, diagnostic likelihood ratios (DLRs) are calculated. These are the proportion of cases with the diagnosis scoring in a given range divided by the proportion of the cases without the diagnosis scoring in the same range (Straus et al., 2011). In the simple case where there is one cut score, the DLR for a positive test result

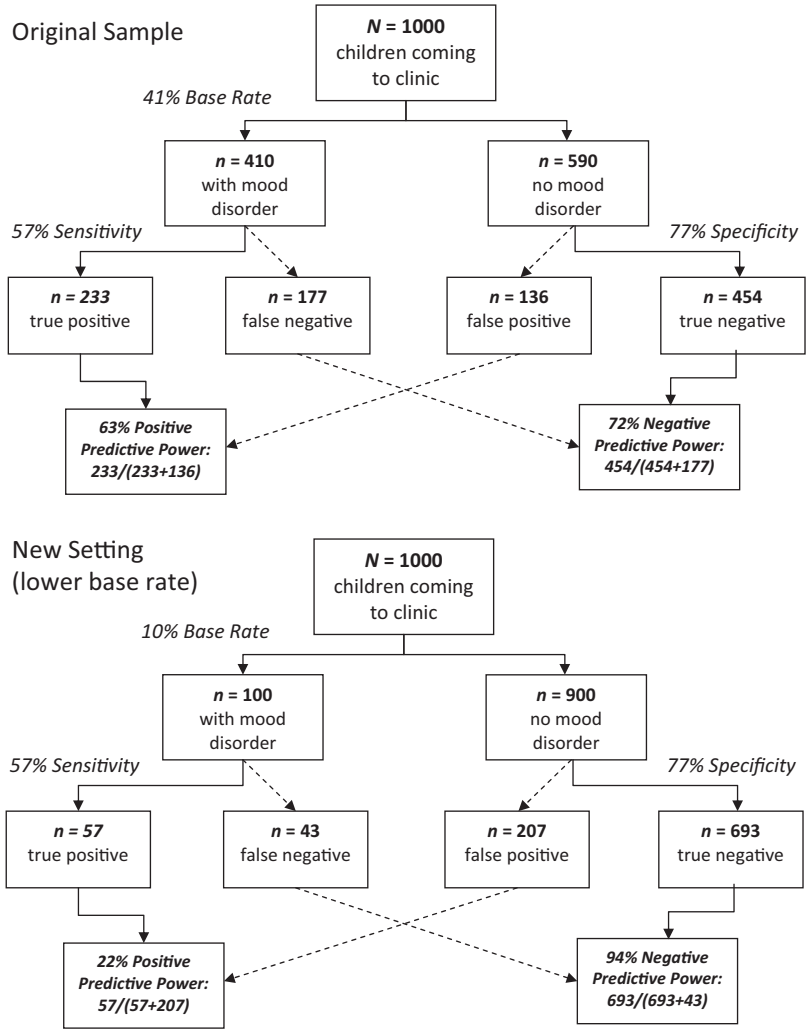


**Figure 3.** Natural frequencies illustrating performance of test and effects of base rate.

is the sensitivity divided by the false alarm rate. However, if adopting the DLR framework, then it often preserves more information to divide the index test into multiple levels of scores, such as ''low risk,'' ''indeterminate,'' and ''high risk.'' The DLRs then can be estimated for each range (Straus et al., 2011). The SPSS syntax appended illustrates doing this by dividing the Internalizing score into quintiles, and also by developing an alternate scoring defining more extreme low and high score ranges to increase the information value. Estimating the DLRs is straightforward using the CROSSTABS procedure in SPSS (see appended syntax).

To apply the DLRs to an individual case using the probability nomogram, one would begin by finding the prior probability on the left hand line, and plotting the DLR corresponding with the test result on the middle line (Jenkins, Youngstrom, Washburn, & Youngstrom, 2011). Connecting the dots and extending across the third line provides the posterior probability. If the reader starts with the sample base rate on the left hand line, and connects it with the DLR from Table V on the middle line, then the estimate on the third line should correspond to the predictive value reported in Table V. Using the probability nomogram results in large improvements in accuracy compared with intuitive, impressionistic interpretation of the same information (Jenkins et al., 2011).

Because the DLR is derived from the sensitivity and specificity, it is independent of the base rate, and it is more likely to generalize outside of the sample where it was developed (Pepe, 2003). Having established that the results in our sample appear consistent with other published reports increases our confidence in the generalizability of these thresholds and the estimates of diagnostic efficiency. The DLR approach addresses the problem of changing base rates (which can be a major issue otherwise—see the natural frequencies in the bottom half of Figure 3, or the

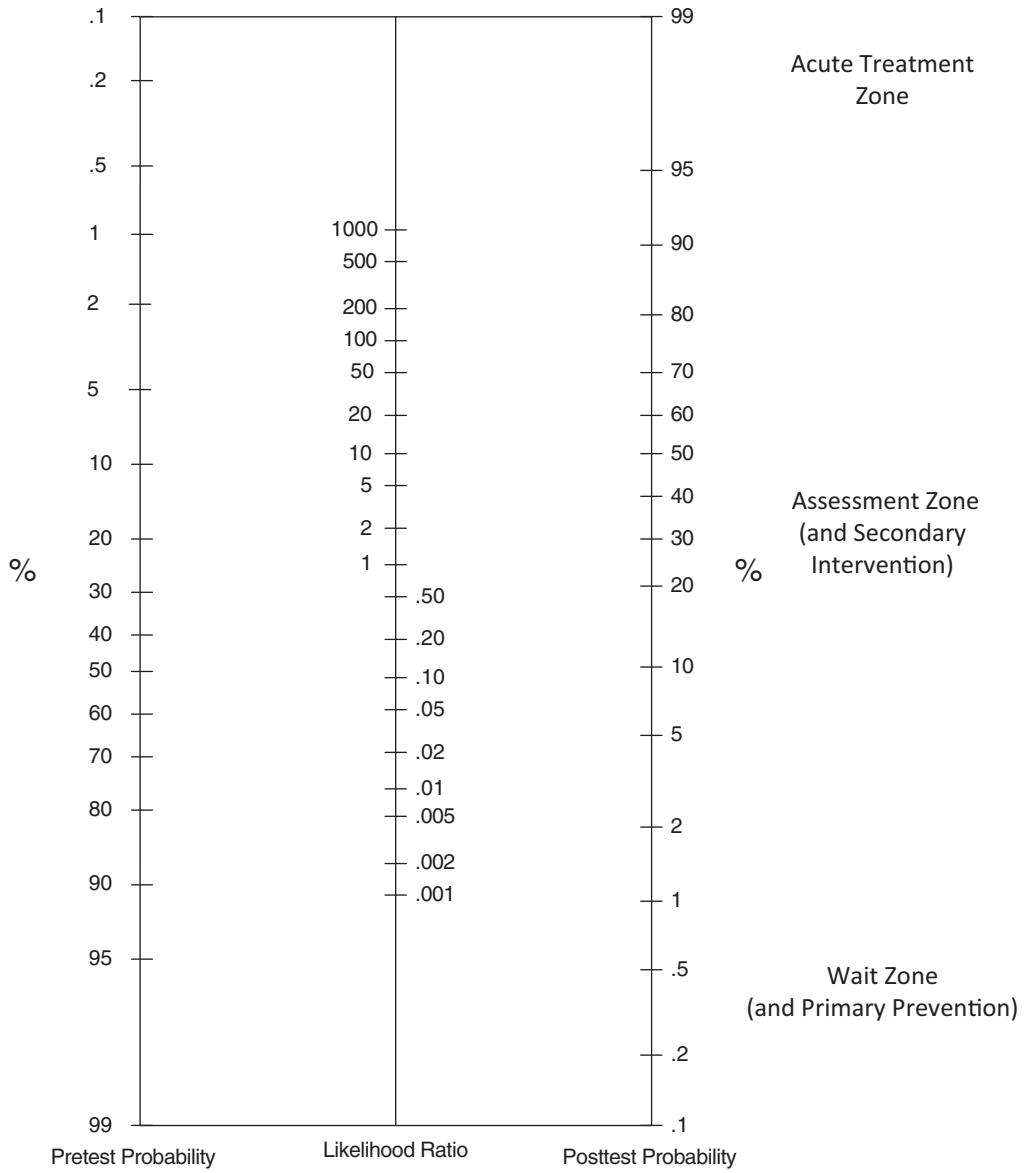


**Figure 4.** Probability nomogram for combining probability with diagnostic likelihood ratios. *Note.* Straus et al. (2011) provide the rationale and examples of using the nomogram. Jenkins et al. (2011) illustrate using it with a case of possible pediatric bipolar mood disorder.

estimated predictive powers in the last column of Table V, both of which are based on a 10% prevalence of mood disorder that more likely approximates the base rate in nonmental health settings).

Rather than simply focusing on statistical significance, ROC focuses attention on the effect size and the impact on individual clinical decisions. The test positive rate will determine the costs associated with follow-up assessment. In Table V, using the threshold attached to a 90% sensitivity results in 74% of the original sample testing positive. Screening using Internalizing scores of 8+ would require follow-up with almost three-quarters of the families!

EBM has moved toward using multilevel DLRs and then comparing the posterior probability with two major decision thresholds, the Wait-Test and the Test-Treat threshold (Straus et al., 2011). If the posterior probability falls below the Wait-Test threshold, then the diagnosis is considered ruled out; if it exceeds the Test-Treat threshold, then it is ''ruled in'' and the next clinical action is to develop a treatment plan. If the probability falls in between, then the next action would be to select additional assessments that could revise the probability. This threshold approach facilitates discussion with patients about their values and preferences, which can be used to adjust the thresholds (cf. Pulleyblank et al., 2013). The framework

can also incorporate prevention and targeted intervention as well as acute treatment (Youngstrom, 2013).

### *Step 8. Make the Results and Test Easy to Use*

Research reports can follow the STARD reporting guidelines to ensure that clinically relevant information about the design, analyses, and results is presented clearly and thoroughly (Bossuyt et al., 2003). After comparing several index tests, it will be possible to make clear recommendations about which perform significantly better. Presenting the DLRs will make it easier for clinicians to use Bayesian methods to integrate test results with other risk factors, generating posterior probabilities (Straus et al., 2011). Including a copy of the probability nomogram (Figure 4) or a Web link to an online calculator makes it even more feasible for clinicians to use the information in real time. Reporting the mean, *SD*, and *n* for both the group with and without the target diagnosis facilitates weighting the sample results appropriately in future meta-analyses (Hasselbad & Hedges, 1995). In clinical settings, practitioners can track the local base rates of diagnoses and common presenting problems, select assessment tools that have demonstrated discriminative validity, and have the DLRs available along with means of integrating different pieces of information in real time, such as probability nomograms or software applets (Youngstrom, 2013).

### *Summary of Results—Evaluating CBCL Against Mood Disorders Criterion*

The goals of the analyses in the demonstration project were to evaluate the diagnostic efficiency of the CBCL for detecting diagnoses of depression, to compare results with other published findings, to compare tests with each other in the same sample, and to develop DLRs to facilitate interpretation of test results for individual cases. Results found that that the CBCL scales offered statistically significant discrimination between cases with mood disorder versus all other outpatient cases. However, the Internalizing score provided significantly great discriminative validity based on either the Hanley and McNeil or DeLong procedures for comparison. The CBCL provided better discrimination at low score ranges, as indexed by DLRs. High scores increased the odds of a mood disorder being present, but the CBCL scores also showed high rates of false-positive results due to other conditions, such as anxiety disorders, also yielding high scores. The CBCL results still produce clinically meaningful changes in the probability of mood disorder in clinical settings with low to moderate rates of mood disorders, effectively ruling mood disorders out in most

| | | **Condition** (based on "Reference Standard") Positive | Negative | |
|---|---|---|---|---|
| **Test Outcome** | Positive | True Positive (TP) | False Positive (FP) *Type I Error* | **Positive Predictive Value***: Accuracy of positive test result TP / Σ (TP, FP) |
| | Negative | False Negative (FN) *Type II Error* | True Negative (TN) | **Negative Predictive Value***: Accuracy of negative test result TN / Σ (TN, FN) |
| | | **Sensitivity**: Accuracy of test among those that have the condition TP / Σ (TP, FN) | **Specificity**: Accuracy of test among those that do not have the condition TN / Σ (TN, FP) | |

**Base rate of condition***: Prevalence of the condition in the sample
Σ(Condition Positive) / Σ(Total *N*) = Σ(TP, FN)/ Σ(TP, FN, FP, TN)

**"Level" of Test*** (or "Test Positive Rate"): Percentage of cases scoring positive on the test
Σ(Test Positive) / Σ(Total *N*) = Σ(TP, FP)/ Σ(TP, FN, FP, TN)

**Percentage Correct*** (or "Efficiency" of Test): Raw percentage of cases classified correctly
Σ(True Positive, True Negative) / Σ(Total *N*) = Σ(TP, TN)/ Σ(TP, FN, FP,TN)

**False Alarm Rate**: Rate of false positives among those that do not have condition; 1 – Specificity
Σ(False Positive) / Σ(False Positive, True Negative) = Σ(FP)/ Σ( FP,TN)

* This parameter is algebraically linked to the base rate of the condition

**Figure 5.** Glossary of diagnostic efficiency terms.

cases, and identifying a subset of cases warranting further evaluation.

## General Summary

ROC analysis has become popular in machine learning, engineering, and EBM, as well as being advocated for use in clinical and pediatric psychology (McFall & Treat, 1999; Swets et al., 2000). The raw data it uses are readily available. ROC methods reorganize the variables to focus on the information value and classification of individual cases. The results can be combined via Bayes' theorem with other information about the patient or clinical setting to develop statistical prediction rules (Swets et al., 2000) or posterior probability estimates that guide the next clinical action (Straus et al., 2011). There are a variety of advanced topics that go beyond the scope of this primer, including scenarios where there are more than two categories, or with continuous dependent variables. Another important area of work is determining optimal sequences when multiple tests are available (Kraemer, 1992). Logistic regression analyses provide a way of testing whether combinations of tests show significant incremental validity, as well as making it possible to test whether variables statistically moderate the diagnostic efficiency of predictors (Hosmer & Lemeshow, 2000). ROC and associated techniques, such as estimating DLRs, are straightforward to implement with recent versions of SPSS, although estimating kappa coefficients and predictive values requires computations outside of SPSS. The free pROC package (Robin et al., 2011) for R is currently the most fully developed and documented procedure for estimating ROC curves, confidence intervals, and performing bootstrapped tests of paired and unpaired ROC curves. Meta-analytic methods also make it straightforward to compare results from one sample with benchmarks reported in technical manuals and articles, even if they did not use ROC methods. Experts have talked about the potential value of ROC and Bayesian methods for improving clinical decision making for decades (McFall & Treat, 1999; Meehl, 1954). The techniques are now available in all major commercial statistical software packages. As we have seen, the data for ROC are readily available, and EBM has developed models and supports for using ROC and DLRs in real time. We are poised for these methods to start delivering on their promise, and hopefully, this primer and the appended resources will facilitate more applications in pediatric psychology.

## Supplementary Data

Supplementary data can be found at: http://www.jpepsy.oxfordjournals.org

## Funding

Data collection supported by the National Institute of Mental Health, R01 MH066647, ''Improving the Assessment of Juvenile Bipolar Disorder,'' PI: R. Youngstrom.

*Conflicts of interest*: E.Y. has consulted with Lundbeck Pharmaceuticals about assessment tools for use in clinical trials.

## Appendix

### Appendix A: SPSS Syntax

```
title 'ROC Primer'.
   * Syntax written by Eric Youngstrom, Ph.D., March 6, 2013.
   * Syntax will run on '605. ROC Primer.sav'.
   * Data consist of 589 cases presenting to community mental health center as part of NIH R01MH066647, PI: E. Youngstrom.
   * Build Table II. Descriptives of clinical, demographic characteristics.
   frequencies /var cgender, crace, agechild, anymood.
   descriptives /var agechild anymood tint intn_r.1 affd_r.1 /statistics default skew kurtosis.
   *Tests of bivariate association between mood diagnosis and clinical, demographic variables.
   t-test /var agechild comorbid intn_r.1 tint affd_r.1 /groups anymood (1 0).
   crosstabs cgender crace by anymood / stat chisq /cell count row col asresid.
   * Building Table III. Correlations for variables.
   correlations /var anymood cgender agechild intn_r.1 tint affd_r.1 /stat desc /missing listwise.
   *Building Figure 1. Population pyramid splitting Internalizing Raw Score by AnyMood diagnosis.
   * Note the evidence of "degeneracy" - including "notches" in the distributions where the frequencies are not monotonic,
   and also that the No Mood group has the highest scores.
   XGRAPH CHART=[HISTOBAR] BY intn_r.1 [s] BY anymood[c] /COORDINATE SPLIT=YES.
```

```
* Building Table IV- AUC & SE estimates, coordinates of the ROC curve (sensitivity & false alarm rate for each cut score), and Figure 2.
roc tint intn_r.1 affd_r.1 by anymood (1) /print se coordinates /plot curve (ref).
* Set up Hanley & McNeil (1983) test of paired ROC AUC values estimated from the same sample.
* Estimate correlations between index test variables in the subgroups with and without Mood Diagnoses.
sort cases by anymood.
split file by anymood.
correlations /variables tint intn_r.1 affd_r.1 .
split file off.
* Find quintile thresholds for multilevel likelihood ratios.
frequencies /variables intn_r.1 / ntiles (5).
* Divide Internalizing Raw Score into quintiles.
recode intn_r.1 (0 thru 6.999 = 1) (7 thru 11.999 = 2) (12 thru 16.999 = 3) (17 thru 23.999 = 4) (24 thru hi = 5) into intgroup5.
frequencies /variables intgroup5.
* Estimate Diagnostic Likelihood Ratios (DLRs) (reported in Table V).
crosstabs intgroup5 by anymood /cell count col.
* Calculate DLR by dividing column percentage for Mood Disorder group by column percentage for No Mood Disorder group.
* Estimate alternate thresholds for DLRs, lumping scores with DLR values close to 1.0 into a large "indeterminate" range
and creating more extreme high score segment, following Kraemer's rule of thumb to keep about 10 cases at each marginal position.
recode intn_r.1 (0 thru 7.999 = 1) (8 thru 23.999 = 2) (24 thru 30.999 = 3) (31 thru hi =4) into intgroup5alt.
* Estimate alternate (DLRs) (reported in Table V, bottom panel).
crosstabs intgroup5alt by anymood / cell count col exp.
*Easter Egg: Logistic regression analyses.
* This tests several additional research questions:
(a) do T scores or Affective Problems provide any predictive increment after controlling for Raw Internalizing?
(b) do youth gender (female = 1) or age in years provide any increment after controlling for Internalizing?
(c) does gender moderate the association between Internalizing and diagnosis? (FEMxInt multiplies CGender dummy code x Internalizing).
logistic regress anymood /enter intn_r.1 /enter tint affd_r.1 /enter cgender agechild /remove tint affd_r.1 / enter femxint.
```

## Appendix B: R Syntax

```
# Sample R syntax for J Pediatric Psychology article
# Written by Eric Youngstrom, Ph.D., March 24, 2013
# Uses same data set as SPSS example; imports data as SPSS file
# Data file is '605. ROC Primer.sav'
# Data consist of 589 cases presenting to community mental health center
# as part of NIH R01 MH066647, PI: E. Youngstrom
# This is a list of packages that need to be installed to generate analyses
# and output; the file path and other details will vary depending upon
# your local R installation
install.packages("corrgram")
library("corrgram", lib.loc="C:/Users/eyoungst/Documents/R/win-library/2.13")
install.packages("Hmisc")
library("Hmisc", lib.loc="C:/Program Files/R/R-2.13.0/library")
# Import SPSS data file
roc605datav2<-spss.get("c:/EAY WIP/Numbered projects/605. ROC paper for JPP/605. ROC paper for JPP data.sav", use.value.labels=TRUE)
```

```
   # Get basic descriptives and check that
file imported correctly
   summary (roc605datav2)
   # Attach file to simplify calls for
variables
   attach(roc605datav2)
   # Histogram; note that R is case sensi-
tive (whereas SPSS syntax is not)
   hist(tint)
   # Build Table II. Descriptives of clinical
& demographic characteristics
   summary (roc605datav2)
   # Need to get SDs for continuous
variables
   dv<-(roc605datav2$[3:6])
   sd(dv)
   t.test(intn.r.1~anymood)
   # Building Table III- Correlations among
variables
   rcorr(as.matrix(dv))
   # This is an example of a scatterplot
matrix as a way of checking
   # distributions and for outliers
   corrgram(as.matrix(dv),
lower.panel=panel.pts,
upper.panel=panel.ellipse)
   chisq.test(cgender, anymood)
   # Building Figure 1- "Population Pyramid"
(aka "Back to Back Histogram")
   # Splitting Internalizing Raw Score by
AnyMood diagnosis
   poppyramid<-histbackback(spli-
t(intn.r.1,anymood), ylab="Raw Score",
   main = 'Population Pyramid of Raw
Internalizing')
   #Just adding color to the figure
   barplot(-poppyramid$left,
col="gray",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   barplot(poppyramid$right, col="red",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   #  Alternate Figure 1- "Population
Pyramid" (aka "Back to Back Histogram")
   # Splitting Internalizing T Score by
AnyMood diagnosis
   poppyramid<-histbackback(split(tint,
anymood), ylab="Raw Score",
   main = 'Population Pyramid of
Internalizing T Score')
   # Just adding color to the figure
   barplot(-poppyramid$left,
col="gray",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   barplot(poppyramid$right, col="red",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   # Second Alternate Figure 1-
"Population Pyramid" (aka "Back to Back
Histogram")
   # Splitting Affective DSM-Oriented
Score by AnyMood diagnosis
   poppyramid<-histbackback(split(aff-
d.r.1,anymood), ylab="Raw Score",
   main = 'Population Pyramid of
Affective Disorders DSM Score')
   #Just adding color to the figure
   barplot(-poppyramid$left,
col="gray",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   barplot(poppyramid$right, col="red",
   horiz=TRUE,    space=0,    add=TRUE,
axes=FALSE)
   # Building Table IV- AUC and SE estimates,
coordinates of the ROC curve,
   # and building Figure 2- Plot of ROC
curves
   # Note that this may take a while to run,
   # because it is drawing 2000
bootstrapped replicates
   # For nonparametric estimation, which
is the default in SPSS and pROC in R,
   # the significance test for comparing
the observed AUC to the null hypothesis is
identical
   # to the Mann-Whitney U test (Zhou et
al., 2002), which is the nonparametric
analog to t-test.
   # pROC also could do Hanley & McNeil
(1983) test, but defaults to DeLong,
   # which has more statistical power
   rocobj1 <- plot.roc(anymood, tint,
main="DeLong Test of Paired AUCs - T
versus Raw Scores",
   percent=TRUE, col="gray")
   rocobj2    <-    lines.roc(anymood,
intn.r.1, percent=TRUE, col="red")
   rocobj3    <-    lines.roc(anymood,
affd.r.1, percent=TRUE, col="yellow")
```

```
   testobj <- roc.test(rocobj1, rocobj2)
   text(40, 80, labels=paste("Raw vs. T p-value =", format.pval(testobj$p.value)),
   adj=c(0, .5)
   legend("bottomright", legend=c("Internalizing T", "Internalizing Raw", "Affective Raw"),
   col=c("gray", "red", "yellow"), lwd=2)
   # Find Quintile thresholds to examine multilevel diagnostic likelihood ratios
   # Create categories
   intgroup5 <-cut(intn.r.1, breaks=c(0,7,12,17,24,max(intn.r.1))
   table(intgroup5,anymood)
   # Create alternate categories, lumping scores with DLRs close to 1 into
   # a large "indeterminate" band, and creating more extreme high score
   # segment, following Kraemer's rule of thumb to keep about 10 cases at
   # each marginal position
   intgroup5alt <-cut(intn.r.1, breaks=c(0,8,24,31,max(intn.r.1))
   table(intgroup5alt,anymood)
```

## References


Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont.

Barrera, M., & Garrison-Jones, C. V. (1988). Properties of the Beck Depression Inventory as a screening instrument for adolescent depression. *Journal of Abnormal Child Psychology, 16*, 263–273.

Bayes, T., & Price, R. (1763). An Essay Towards Solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., ... de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41–44.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cyranowski, J. M., Frank, E., Young, E., & Shear, K. (2000). Adolescent onset of the gender difference in lifetime rates of major depression. *Archives of General Psychiatry, 57*, 21–27.

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.

Ferdinand, R. F. (2008). Validity of the CBCL/YSR DSM-IV scales Anxiety Problems and Affective Problems. *Journal of Anxiety Disorders, 22*, 126–134.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843.

Hasselbad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167–178. doi:10.1037/0033-2909.117.1.167

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

Jenkins, M. M., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42*, 121–129. doi:10.1037/a0022506

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., ... Ryan, N. (1997). Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*, 980–988.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines*. Newbury Park, CA: Sage.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Academic Press.

Lengua, L. J., Sadowski, C. A., Friedrich, W. N., & Fisher, J. (2001). Rationally and empirically derived dimensions of children's symptomatology: Expert ratings and confirmatory factor analyses of the CBCL. *Journal of Consulting and Clinical Psychology, 69*, 683–698.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage Publications.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessment with signal detection theory. *Annual Review of Psychology, 50*, 215–241. doi:10.1146/annurev.psych.50.1.215

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Wiley.

Pulleyblank, R., Chuma, J., Gilbody, S. M., & Thompson, C. (2013). Decision curve analysis for assessing the usefulness of tests for making decisions to treat: An application to tests for prodromal psychosis. *Psychological Assessment*. doi:10.1037/a0032394

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Muller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77. doi:10.1186/1471-2105-12-77

Steer, R. A., Cavalieri, T. A., Leonard, D. M., & Beck, A. T. (1999). Use of the Beck Depression Inventory for primary care to screen for major depression disorders. *General Hospital Psychiatry, 21*, 106–111.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). *Evidence-based medicine: How to practice and teach EBM* (4th ed.). New York, NY: Churchill Livingstone.

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

Tillman, R., & Geller, B. (2005). A brief screening tool for a prepubertal and early adolescent bipolar disorder phenotype. *American Journal of Psychiatry, 162*, 1214–1216.

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the child behavior checklist and strengths and difficulties questionnaire: A systematic review. *Child and Adolescent Mental Health, 13*, 140–147. doi:10.1111/j.1475-3588.2007.00461.x

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*, 529–536. doi:10.1059/0003-4819

Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining evidence-based medicine innovations with psychology's historical strengths to enhance utility. *Journal of Clinical Child & Adolescent Psychology, 42*, 139–159. doi:10.1080/15374416.2012.736358

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., ... Findling, R. L. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7*, 507–517.

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Comparing the effects of sampling designs on the diagnostic accuracy of eight promising screening algorithms for pediatric bipolar disorder. *Biological Psychiatry, 60*, 1013–1019. doi:10.1016/j.biopsych.2006.06.023

Zhou, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York, NY: Wiley.