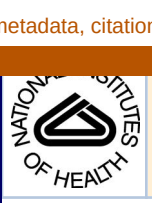

Published in final edited form as:

Assessment. 2015 April ; 22(2): 198–207. doi:10.1177/1073191114540748.

Adapting the Posterior Probability of Diagnosis (PPOD) Index to Enhance Evidence-Based Screening: An Application to ADHD in Primary Care

Oliver Lindhiem[1], **Lan Yu**[1], **Damion J. Grasso**[2], **David J. Kolko**[1], and **Eric A. Youngstrom**[3]

[1]University of Pittsburgh, School of Medicine, Department of Psychiatry

[2]University of Connecticut, Department of Psychiatry

[3]University of North Carolina at Chapel Hill, Department of Psychology and Psychiatry

Abstract

This study adapts the Posterior Probability of Diagnosis (PPOD) Index for use with screening data. The original PPOD Index, designed for use in the context of comprehensive diagnostic assessments, is overconfident when applied to screening data. To correct for this overconfidence, we describe a simple method for adjusting the PPOD Index to improve its calibration when used for screening. Specifically, we compare the adjusted PPOD Index to the original index and Naïve Bayes probability estimates on two dimensions of accuracy, discrimination and calibration, using a clinical sample of children and adolescents (N = 321) whose caregivers completed the Vanderbilt Assessment Scale to screen for Attention-Deficit/Hyperactivity Disorder (ADHD) and who subsequently completed a comprehensive diagnostic assessment. Results indicated that the adjusted PPOD Index, original PPOD Index, and Naïve Bayes probability estimates are comparable using traditional measures of accuracy (sensitivity, specificity, AUC) but the adjusted PPOD Index showed superior calibration. We discuss the importance of calibration for screening and diagnostic support tools when applied to individual patients.

Keywords

evidence-based assessment; screening; item response theory; calibration; ADHD

Mental and behavioral health disorders, unlike many medical conditions, are diagnosed on the basis of co-occurring symptoms (usually self-report or caregiver-report) rather than more objective diagnostic tests such as blood tests or MRIs. For this reason, developing evidence-based screening and assessment strategies is particularly important for the fields of psychology and psychiatry. Confidence in a diagnosis is crucial for deciding whether to initiate treatment, pursue additional testing, or rule-out a diagnosis. Proponents of evidence-based medicine (EBM) offer principles for incorporating sound Bayesian reasoning into diagnostic assessments and screening (e.g., Straus, Glasziou, Richardson, & Hayes, 2011) as well as practical recommendations for applying these guidelines to clinical psychology (e.g.,

Corresponding Author: Oliver Lindhiem, Ph.D., University of Pittsburgh, School of Medicine, Department of Psychiatry, 3811 O'Hara St., Pittsburgh, PA 15213, Office: 537 Bellefield Towers, Phone: 412-246-5909, Fax: 412-246-5341, lindhiemoj@upmc.edu.

Youngstrom, 2012). This study represents one example of how EBM principles can be applied to the routine screening of mental and behavioral health disorders catalogued in the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013). Specifically, we describe a simple method for adjusting the Posterior Probability of Diagnosis (PPOD) Index (Lindhiem, Kolko, & Yu, 2013) to enhance its accuracy and clinical utility as a screening tool (versus a diagnostic tool).

The Posterior Probability of Diagnosis (PPOD) Index

In an earlier paper, we introduced the PPOD Index (Lindhiem, Kolko, & Yu, 2013) which was developed as a Bayesian diagnostic-support tool for quantifying the degree of confidence associated with a diagnosis and facilitating a means to communicate this information to patients. Figure 1 shows a graphical depiction of the conceptual difference between traditional symptom counts (*a la* DSM) and the PPOD Index. Traditional diagnoses are based on symptom counts with a cutoff at which patients abruptly go from not having a diagnosis to having a diagnosis. In contrast, the PPOD Index is a continuous measure that quantifies the likelihood that a patient meets or exceeds a latent diagnostic threshold. Based on a latent trait model, the PPOD Index is calculated using item response theory (IRT) and Bayesian methods. Latent trait scores (θ) are first estimated using IRT software. Then the PPOD Index is calculated from the posterior distribution of θ for an individual patient's pattern of symptoms. This is done by numerically integrating the posterior distribution of θ above a diagnostic threshold. In its current form, the PPOD Index can be applied to DSM diagnoses without hierarchical rules or "skip outs" such as Oppositional Defiant Disorder and Conduct Disorder. This method has two advantages over traditional diagnostic approaches. First, the PPOD Index is based on a patient's individual *pattern* of symptoms and risk factors. Second, this method quantifies the degree of confidence associated with a diagnosis in probabilistic terms (0%–100%). Although the PPOD Index does not eliminate the need to ultimately make categorical clinical decisions, it allows a clinician to quantify the confidence associated with each diagnosis and communicate this critical information to patients and their families. From a patient-centered perspective, this information assists in shared decision making and treatment planning (Straus, Tetroe, & Graham, 2011).

Discrimination and Calibration

In order for a diagnostic/screening support tool to be clinically useful, it must be accurate not only at the group level but also when applied to individual cases. Discrimination and calibration are two aspects of predictive model accuracy, and their relative importance depends on the intended use of the model. Discrimination refers to how well a model can predict a category or outcome such as a disease, and is typically measured using metrics including sensitivity, specificity, and area-under-the- curve (AUC; e.g., Kraemer, 1992). Calibration must be defined carefully, as the term can have different meanings in different contexts. The term calibration is often used in a general sense to mean how well any statistical model fits actual data, and is generally evaluated using goodness-of-fit statistics. In the context of probabilistic models for predicting binary outcomes, however, the term calibration has a much more precise meaning. In this context, calibration refers to the consistency between predicted probabilities and the proportion of empirical observations

(e.g., Jiang, Osl, Kim, & Ohno-Machado, 2012; Redelmeier, Bloch, & Hickam, 1991; Spiegelhalter, 1986). For example, if the posterior probability of a disease is estimated at .85 does the patient truly have an 85% chance of having the disease? If the model is well calibrated, for every 100 patients with a .85 posterior probability of the disease, 85 would actually have the disease. Calibration in this sense is evaluated using Brier scores and related indices (e.g. Brier, 1950; Ferro, 2007; Spiegelhalter, 1986) or the Hosmer-Lemeshow (HL) goodness-of-fit statistic. Throughout this manuscript, we use the term calibration with this narrower definition. In this sense, predictive models can have good discrimination (in terms of sensitivity, specificity, and AUC) but still be poorly calibrated. If the purpose of a model is simply to minimize Type 1 errors (false positives) and Type 2 errors (false negatives) at the group level, then discrimination is of primary importance. If, however, the model is intended to be used as a tool to aid individualized decision-making, then calibration is of equal importance. Many models used for estimating the posterior probability of diseases, such as Naïve Bayes models, have adequate discrimination but are poorly calibrated (Jiang et al., 2012).

Diagnostics Versus Screening

The PPOD Index, as described in our original paper (Lindhiem et al., 2013), was developed as a tool for clinicians to quantify confidence associated with a final diagnosis. In order to use the PPOD Index in the context of screening (versus a final diagnosis), however, an adjustment is necessary to account for the nature of the screening data. Intuitively, we should expect a higher degree of confidence in a diagnosis based on a thorough diagnostic interview conducted by a trained clinician in conjunction with input from parents, teachers, and clinical observations (i.e., multiple sources of information and methods of acquiring information). In contrast, the veracity of screening data, whether based on self-report or parent-report, is always questionable. Any probabilistic prediction about the likelihood of a diagnosis that is based on screening data should be made cautiously. In other words, PPOD Index values should be more conservative when applied to a screening tool than when applied to gold-standard diagnostic data from a structured clinical interview.

We expect, therefore, that the originally proposed PPOD Index is over-confident (too many values close to 0.0 or 1.0) when applied to screening data. In order to apply the PPOD Index to screening data, it is necessary to adjust for the veracity or “believability” of the data. This “extra step” is needed to improve the accuracy of the PPOD Index (in terms of calibration) in much the same way that shrinkage estimators improve the predictive accuracy of regression models. In other words, the original PPOD Index, while appropriate for diagnostic data, “overfits” screening data. In order to be applied to screening data, an adjustment is necessary to correct for this “overfitting” by making the index values more conservative.

Current Study

In this study, we extend the PPOD Index by exploring its application as a screening tool (versus a diagnostic tool). Specifically, we describe a method to adjust the PPOD Index to improve its accuracy when used as a screening tool. Specifically, we compare the accuracy

of the original PPOD Index with the adjusted PPOD Index and Naïve Bayes probability estimates in terms of discrimination and calibration. We illustrate the method for the diagnosis of ADHD, Inattentive Type in a clinical sample of children and adolescents referred for treatment due to disruptive behavior problems. This study uses the same sample as our previous paper on the PPOD Index (Lindhiem et al., 2013), but different variables.

Method

Participants

Participants in this study were parent-child dyads (N = 321) consisting of a clinical sample of boys (n = 207; 65%) and girls (n = 114; 35%) who were referred for services due to disruptive behavior. Children ranged in age from 5 to 12 (M = 8.00; SD = 1.97). Eight (2.5%) children were reported as Hispanic, 67 (21%) Black/African American, 259 (81%) White, and not reported 3 (0.9%) children. None were reported as American Indian/Alaskan Native, Asian, or Native Hawaiian/Pacific Islander. Parent relationship to child was reported as biological mother (n = 291; 91%), biological father (n = 16; 5%), adopted mother (n = 8; 2.5%), adopted father (n = 1; 0.3%), or grandmother (n = 4; 1.3%). Two hundred three (64%) parents were married/remarried and living with their spouse, 70 (22%) were single and never married, 30 (9%) were divorced, 14 (4.4%) were separated from their spouse, and 1 (0.3%) was a widow/widower. Parent education levels were reported as follows: 1 (0.3%) junior high (9^{th} grade); 6 (1.9%) with some high school (10^{th} or 11^{th} grade), 63 (20%) with a high school degree or GED, 62 (19%) with some college (at least 1yr), 52 (16%) with an Associate Degree (2 years), 94 (30%) with 4-year college degree, and 41 (13%) with Graduate/Professional Training. Most parents were employed either full-time (n = 176; 55%) or part-time (n = 48; 15%). Median household income was in the $50,000 – $74,999 range. The number of adults in the home ranged from one to five (M = 1.93; SD = 0.63) and the number of children in the home ranged from zero to six (M = 1.60; SD = 1.13).

Measures

Screen for ADHD Symptoms—Symptoms of ADHD were assessed using the Vanderbilt Assessment Scale-Parent Version (VAS-Parent; Wolraich, Hannah, Baumgaertel, & Feurer, 1998). Items 1 through 9 of the VAS-Parent assess the 9 symptoms of ADHD, Predominantly Inattentive Type. Each item is rated on 4-point Likert scale (0 = Never; 1 = Sometimes; 2 = Often; 3 = Very Often). Due to sample size considerations, each item was binarized and re-coded as a “symptom” (1 = Often or Very Often) or “not a symptom” (0 = Never or Sometimes). Although polytomous IRT models have been developed to handle likert responses, they require larger (minimum N = 500) sample sizes (Reise & Yu, 1990). Using this dichotomous variable, Cronbach’s α was high in the current sample (.88). Additional psychometric properties of the VAS-Parent are described in detail in the literature (Wolraich, Lambert, Doffing, Bickman, Simmons, & Worley, 2003).

Diagnostic Status—Final consensus ADHD diagnoses were based on an abbreviated version of the K-SADS (Kaufman, Birmaher, Brent, & Rao, 1997). The K-SADS is a diagnostic interview for DSM-based diagnostic categories with well-established reliability and validity. Diagnostic interviews were conducted separately with both the parent and

child. Final diagnoses were determined during weekly team meetings with input from the clinician who conducted the interviews and the medical director (a child psychiatrist).

Recruitment Procedure

Participants were recruited from primary-care offices in the Pittsburgh area. Families who met study criteria were then scheduled for an intake assessment that included a diagnostic interview. Each family met with one of four Masters-level clinicians with additional training in clinical assessments and diagnostics. Parents completed the VAS-Parent during the intake assessment. The total minutes spent with families during the assessment ranged from 105 to 335 minutes (M = 155.48; SD = 29.15).

Data Analyses

Confirmatory factor analysis—A two-factor confirmatory mode using the maximum likelihood estimation with robust standard errors (MLR) estimator was fitted to the VAS-Parent ADHD symptoms using Mplus Version 6.1 (Muthen & Muthen, 2010) to check the assumption, explicit in the *DSM*, that ADHD has two distinct subtypes. The indicators (items) were treated as categorical variables. We emphasized the magnitude of factor loadings in the CFA and the fit and information values reflected in IRT models. Item fit was checked using item level diagnostic statistics and the summed score χ^2. A p-value larger than 0.05 was used as a cut-off for good item fit.

Two-parameter logistic (2PL) IRT model—We used IRTPRO (Cai, du Toit, & Thissen, 2011) to estimate latent trait scores (θ) and standard errors for each patient using a two-parameter logistic (2PL) IRT model for dichotomous items. Scoring was based on the expected a posteriori (EAP) estimation method (Bock & Mislevy, 1982) and assuming a standard normal prior distribution. We also estimated threshold parameters (βs) and discrimination parameters (αs) for each of the ADHD symptoms.

Posterior Probability of Diagnosis (PPOD) Index—The PPOD Index for each patient was estimated by numerically integrating the posterior distribution of θ above the diagnostic threshold, as described in our earlier paper (Lindhiem et al., 2013), using a MATLAB (MathWorks, 2011) program specifically created for this purpose. The diagnostic threshold is defined as the θ level associated with the DSM criteria for a given disorder. For ADHD, Inattentive Type, the diagnostic threshold is therefore the lowest θ out of all symptom patterns of six or more symptoms, which was estimated at θ = 0.12. The PPOD Index was therefore defined as the following posterior probability: $p(\theta \geq 0.12 \mid \text{response pattern})$. Because IRTPRO uses 60 quadrature points ranging from −3.0 to 2.9 in increments of 0.1, we selected the two thresholds on either side of 0.12 (Lower Bound = 0.2 and Upper Bound = 0.1) which were averaged as the final PPOD Index estimate. The posterior distribution of θ for each response pattern can be represented using the following form of Bayes Theorem for discrete values of θ,

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{\sum_{\theta}p(D|\theta)p(\theta)},$$

where *D* is a response pattern (in this case a symptom profile) and $p(\theta)$ is the probability mass at θ. Figure 2 shows an example of the posterior distribution of θ for one symptom pattern (out of 512 possible patterns).

The Adjusted PPOD Index—We adjusted the PPOD Index by reapplying Bayes' theorem to estimate the posterior probability of a final consensus diagnosis given a particular PPOD Index value from the screen. The equation for the adjusted PPOD Index can be represented by the following equation,

$$p(\mathrm{Dx}|\mathrm{PPOD})=\frac{p(\mathrm{PPOD}|\mathrm{Dx})p(\mathrm{Dx})}{p(\mathrm{PPOD})},$$

where p(Dx | PPOD) is the adjusted PPOD Index, "Dx" is the final consensus diagnosis, and "PPOD" is a particular PPOD Index value from the screen. It should be noted that p(Dx) is the base rate of the diagnosis in the dataset, which can either be calculated directly from the dataset or estimated using historical data from the clininc or setting in which the Ajusted PPOD Index will be used. Because several PPOD Index values were sparse or not represented in the data set, we then applied a repeated k-nearest neighbor smoothing algorithm to reduce noise and allow for estimates of missing values. Figure 3 graphically summarizes the resulting values of the adjusted PPOD Indexes alongside the values of the original PPOD Index values for all the cases in the dataset.

Sensitivity to misspecification—As with any model, its performance will be affected by misspecification of model parameters. The adjusted PPOD Index is directly proportional to the base rates, so the degree of bias depends on the latent trait level. For low latent trait levels (most cases), the bias will be trivial regardless of the degree of misspecification. For high latent trait levels (few cases), the clinical significance of the bias would depend of the degree of misspecification. But even in this case, typical levels of misspecification will result in little bias. For example, the base rate in this study (.49) has a standard error of .03. The 95% CI for the base rate is .43–.55. Even with significant misspecification (actual base rate of .43 or .55), the adjusted PPOD Index values would be biased by an average of 5% and never more than 10%.

Primary data analyses—Data analyses were conducted using SPSS version 21 and STATA 12.0. We compared the accuracy of the original PPOD Index and adjusted PPOD Index in terms of discrimination and calibration. Discrimination was evaluated in term of sensitivity, specificity, positive predictive value, negative predictive value, and ROC analyses. Calibration was evaluated using Brier scores, Spiegelhalter's *z* statistic, and the Hosmer-Lemeshow (HL) goodness-of-fit statistic. The HL statistic is based on a chi-square distribution, with high chi-square values and low *p*-values indicating poor calibration (Hosmer & Lemeshow, 2004).

Results

Factor loadings from the confirmatory factor analysis ranged from 0.68 to 0.87 for factor 1 (Inattentive) and from 0.68 to 0.85 for factor 2 (Hyperactive/Impulsive). The correlation

between the two factors was 0.56. The item fit χ^2 was insignificant for all items, indicating good item fit within each factor. The item parameter estimates for the two-parameter (2PL) model are summarized in Table 1. We see for example that "does not seem to listen when spoken to directly" has the lowest threshold parameter ($\beta = -0.55$). In other words, a child would only need to exceed the ADHD, Inattentive Type trait level of $\theta = -.55$ before there is a 50% chance that his or her parent would endorse this item as "often" or "very often". The item "loses things necessary for tasks or activities (toys, assignments, pencils, or books)" had the highest threshold parameter ($\beta = .22$). A child would need to exceed the ADHD, Inattentive Type trait level of $\theta = .22$ before his or her parent would have a 50% chance of endorsing this item as "often" or "very often." All nine symptoms of ADHD, Inattentive Type had good discrimination parameters (all α parameters above 1.0) and ranged from 1.58 ("does not seem to listen when spoken to directly") to 3.72 ("has difficulty keeping attention to what needs to be done").

Symptom Counts, Diagnostic Categories, and PPOD Indices

Table 2 summarizes the values for the original PPOD Index and the adjusted PPOD Index. Values for the original PPOD Index ranged from 0.00 to > .99. Many of the original PPOD Index values were close to 0.0 or 1.0, indicating high confidence. Values close to 0.0 indicate high confidence of no diagnosis, whereas values close to 1.0 indicate high confidence of a diagnosis. In comparison, the truncated range (.08 to .91) of the Adjusted PPOD Index indicates more conservative estimates. The difference between the two ranges can readily be seen in Figure 3.

Discrimination and Calibration

Table 3 summarizes the accuracy of each PPOD index and probability estimates from the Naïve Bayes algorithm in terms of both discrimination and calibration. In terms of discrimination, all three indices performed comparably as measured by area under the ROC curve (AUC), sensitivity, specificity, positive predictive value (PPV), and Negative Predictive Value (NPV). However, the original PPOD Index (Spiegelhalter's *z*-statistic = 10.30, $p < 0.001$; Hosmer-Lemeshow $\chi^2(10) = 191.43$, $p < .001$) and Naïve Bayes probability estimates (Spiegelhalter's *z*-statistic = 17.86, $p < 0.001$; Hosmer-Lemeshow $\chi^2(10) = 473.63$, $p < .001$) were poorly calibrated. (For both Spiegelhalter's *z* and Hosmer-Lemeshow χ^2 significant *p*-values indicate poor calibration.) In contrast, the adjusted PPOD Index evidenced good calibration, Spiegelhalter's *z*-statistic = −0.89, p = .81, and Hosmer-Lemeshow $\chi^2(10) = 4.91$, p = .90. Figure 4 depicts calibration plots for the original PPOD Index and the adjusted PPOD Index. Perfect calibration is represented by the diagonal lines. The shape (backward "S") of the calibration plot for the original PPOD Index is characteristic of an "over-confident" model, with many predictions above .95 and below .05. In contrast, the data points on the calibration plot for the adjusted PPOD fall closer to the diagonal.

Discussion

The PPOD Index was developed to answer the question, "What is the likelihood (0–100%) that an individual patient meets or exceeds the diagnostic threshold for a particular disorder

given his or her pattern of symptoms?" (Lindhiem et al., 2013). In this study, we extend the PPOD Index by adapting it for use as a screening aid. Specifically, we proposed a simple method to make the PPOD Index more conservative when applied to screening data—which typically come from settings where the target condition will be rare, thereby improving its calibration. The adjustment takes into account the base rate of the new criterion (in this case a final consensus diagnosis) and results in enhanced calibration. Accurate calibration is vital when a clinical tool will be used to guide clinical decision-making at an individual level. We applied the PPOD Index to screening data from the parent-report form of the VAS and demonstrated a method to enhance accuracy in terms of calibration. Our results suggest that the original PPOD Index was poorly calibrated (over-confident) when applied to screening data, but is easily adjusted by re-applying Bayes' theorem to predict the probability of the new criterion.

Clinical Implications

Implementation of the PPOD Index in pediatric and other primary care settings has the potential to minimize inaccuracies in diagnoses that stem from a reliance on data from unstructured diagnostic interviews or screening instruments, which yield a lower degree of confidence in diagnosis as compared to gold-standard diagnostic evaluations informed by multiple sources. Unstructured interviews remain the most common method of making diagnoses in practice, despite extensively documented shortcomings in terms of accuracy and vulnerability to biases (Garb, 1998; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). As discussed earlier, rating scales can have adequate discriminability, but poor calibration in the sense that they are inconsistent regarding how well their predicted probabilities map onto individual patients' observed outcomes. The adjusted PPOD Index is designed to provide information that will help clinicians to interpret screening results and, in turn, make more informed clinical decisions. Specifically, the adjusted PPOD Index informs the clinician how likely his or her patient is to have a particular disorder based on the patient's pattern of responses on a particular screening instrument. Whereas the original PPOD Index does a similar task for actual symptom patterns, we modified the adjusted PPOD Index to yield more appropriately conservative estimates.

The information provided by the adjusted PPOD Index may inform a clinician's decision regarding whether or not to seek a second opinion, perform additional assessment of a particular diagnosis, or to obtain information from additional sources. On a larger scale, data derived from the adjusted PPOD Index may allow healthcare settings to flag sub-populations of patients for targeted assessment and treatment. For example, in situations in which rating scale data is obtained by patients at intake and recorded in their medical records, the adjusted PPOD Index, which could be programmed into electronic medical record algorithms, could indicate to clinicians which patients need additional attention. The entire process of identifying high-risk patients could occur in the background, prior to clinician's involvement. This notion is attractive, especially given the many competing demands that exist in healthcare settings and the scarcity of resources, including staff time and sufficient insurance reimbursement, to address them (Gardner, Kelleher, Pajer, & Campo, 2003; Knapp & Foy, 2012; Wren, Bridge, & Birmaher, 2004). With technological advances happening rapidly and electronic medical records becoming more prevalent and

sophisticated, the feasibility of implementing the adjusted PPOD Index within healthcare settings is quite promising.

A second benefit of supplementing rating scale data with the adjusted PPOD Index is the ability to inform consumers of mental health services about how confident a clinician is regarding her or her child's diagnostic status. Providing confidence information in a way that is intuitive to consumers empowers them to make informed decisions regarding treatment and whether to pursue a second opinion. Further, it communicates to patients information that reflects the seriousness or certainty of the condition, which may influence their adherence to treatment recommendations. With growing recognition that creating informed consumers of mental health is critical for establishing widespread practice of evidence-based assessment and treatment (e.g., Bielavitz & Pollack, 2011; Nakamura, Chorpita, Hirsch, Daleiden, Slavin,, Amundson, & Vorsino, 2011), there appears to be high demand for a clinical decision-support tool such as the PPOD Index.

We chose to illustrate the adjusted PPOD Index for enhancing ADHD screening given the significant shortcomings of current practices for assessing and diagnosing pediatric ADHD emphasized in the recent literature (Dalsgaard, Nielsen, & Simonsen, 2013; Zelnik, Bennett-Back, Miari, Goez, & Fattal-Valevski, 2012). Self-administered ADHD measures can achieve high sensitivity, but with mediocre specificity, which is partially attributable to the fact that ADHD shares symptom characteristics with a number of other psychiatric disorders, making differential diagnosis rather complex (Klein, Pine, & Klein, 1998; Youngstrom, Arnold, & Frazier, 2010; Zelnik et al., 2012). Consequently, the American Academy of Child and Adolescent Psychiatry (AACAP) has established best practice parameters that recommend a comprehensive and rigorous evaluation that involves assessing ADHD symptoms, frequently comorbid disorders, and disorders that share common symptom characteristics (e.g., stress- or trauma-related disorders, learning disorders, mood disorders) using multiple sources (e.g., school, caregiver, child) and employing multiple methods (i.e., in-depth interviews, self-administered assessments, observation; Parker & Corkum, 2013; Pliszka, 2007). Not surprisingly, "real world" barriers impede clinicians' ability to adopt and implement such an approach, especially in pediatric primary care, where the bulk of ADHD is diagnosed and treated (Wolraich, Bard, Stein, Rushton, & O'Connor, 2010). One survey of practicing pediatricians found that only a quarter reported incorporating all of the recommended components on a routine basis (Wolraich et al, 2010). Many clinicians rely heavily on self- or parent-report rating scales to determine whether a child meets criteria for an ADHD diagnosis (Robinson, 2005). Reliance on rating scales as the primary justification for diagnosis is troubling given their tendency to generate a substantial number of false positives (Parker & Corkum, 2013). A literature review of 13 studies examining psychometric properties of ADHD rating scales reported specificities as low as 44% (Snyder, Hall, Cornwell, & Quintana, 2006). This suggests that a number of children who present with characteristics of ADHD symptoms but who would not meet criteria for ADHD provided a comprehensive diagnostic evaluation are erroneously diagnosed with and treated for ADHD.

Indeed, ADHD is the most commonly diagnosed neuropsychiatric disorder in children and adolescents (Biederman & Faraone, 2005). A large, nationally represented study examining

change in prescription rates in the U.S. over the years 2002 to 2010 reported a 46% increase in ADHD medication prescriptions (Chai et al., 2012). Another study involving rigorous diagnostic assessments of children ages 5–13 from South Carolina and Oklahoma found that about one out of every 20 children was prescribed ADHD medication despite not meeting diagnostic criteria for ADHD (Wolraich et al., 2012). This is noteworthy given the documented side effects and potential health risks associated with children's long-term use of psychostimulants (Evans, Morrill, & Parente, 2010).

In addition, suboptimal specificity makes diagnosis vulnerable to a number of confounding factors and biases. These include known health disparities associated with increased likelihood of an ADHD diagnosis in Whites relative to racial/ethnic minorities, high versus low socioeconomic status, and children with versus without private health insurance (Morgan, Staff, Hillemeier, Farkas, & Maczuga, 2013; Kowatch et al., 2013). Also, children with caregivers who are stronger advocates and who have greater knowledge of ADHD and special education policies have greater odds of being assessed for and diagnosed with ADHD (Bussing, Gary, Mills, & Garvan, 2003).

Limitations and Future Directions

In order to apply the adjusted PPOD Index, one must have historical data, including base rates, specific to one's sample or clinic. In the current study, we also dichotomized our symptom data and ran a 2PL IRT model due to the sample size. This likely resulted in loss of information. A larger sample (500+) would allow for estimation using a polytomous IRT model. Finally, most DSM diagnoses are not based on pure symptom counts as with ADHD or Oppositional Defiant Disorder. Diagnostic criteria for other disorders, such as Posttraumatic Stress Disorder or Autism Spectrum Disorder, include additional clustering rules. The PPOD Index would need to be modified to accommodate these additional criteria. It will be also useful to conduct simulation studies to examine the comparative performance of the original PPOD Index and the adjusted PPOD Index under various conditions.

Conclusions

Although the PPOD Index does not eliminate the need to ultimately make a categorical decision (e.g., additional testing, referral to a specialist), it allows a clinician to quantify the likelihood of a diagnosis. This level of confidence is clinically useful information. For example, a provider might encourage all patients with a 25% or higher probability of a given disorder to return for a follow-up appointment within a specified timeframe. The adjusted PPOD Index has the potential to influence clinical decisions made on the basis of rating scales by quantifying the degree of confidence in the predicted probability. Although outcomes from the original PPOD Index and adjusted PPOD Index performed comparably at predicting a final consensus diagnosis in terms of AUC, the adjusted PPOD Index had superior calibration. The original PPOD Index was over-confident (too many values close to 0.0 or 1.0) whereas the adjusted PPOD Index made predictions that were consistent with the observed proportion of final diagnoses. Whether one should use the original PPOD Index or the adjusted PPOD Index depends on the nature of question and the data that the index is being applied to. The original PPOD Index is appropriate for applications to diagnostic data whereas the adjusted PPOD Index should be used for applications to screening data.

Acknowledgements

This study was supported by grants from the National Institute of Mental Health (NIMH) to the first author (MH 093508) and forth author (MH 063272). We acknowledge the contributions of the research and clinical staff of the Service for Kids in Primary-Care (*SKIP*) program, Janelle Higa, and Charles Bennett.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Fourth Edition. Washington, DC: American Psychiatric Association; 2000. Text Revision.

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Fifth Edition. Washington, DC: American Psychiatric Association; 2013.

Biederman J, Faraone SV. Attention-deficit hyperactivity disorder. The Lancet. 2005; 366:237–248.

Bielavitz S, Pollack DA. Effective mental health consumer education: A preliminary exploration. The Journal of Behavioral Health Services & Research. 2011; 38(1):105–113. [PubMed: 20358303]

Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement. 1982; 6:431–444.

Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950; 78(1):1–3.

Bussing R, Gary FA, Mills TL, Garvan CW. Parental explanatory models of ADHD: Gender and cultural variations. Social Psychiatry and Psychiatric Epidemiology. 2003; 38(10):563–575. [PubMed: 14564385]

Cai, L.; du Toit, SHC.; Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling. Chicago, IL: Scientific Software International; 2011.

Chai G, Governale L, McMahon AW, Trinidad JP, Staffa J, Murphy D. Trends of outpatient prescription drug utilization in US children, 2002–2010. Pediatrics. 2012; 130:23–31. [PubMed: 22711728]

Dalsgaard S, Nielsen HS, Simonsen M. Five-Fold increase in national prevalence rates of attention-deficit/hyperactivity disorder medications for children and adolescents with autism spectrum disorder, attention-deficit/hyperactivity disorder, and other psychiatric disorders: A Danish register-based study. Journal of Child and Adolescent Psychopharmacology. 2013

Evans WN, Morrill MS, Parente ST. Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children. Journal of Health Economics. 2010; 29:657–673. [PubMed: 20739076]

Ferro C. Comparing probabilistic forecasting systems with the Brier score. Weather and Forecasting. 2007; 22:1076–1088.

Garb, HN. Studying the clinician: Judgment research and psychological assessment. Washington, DC: American Psychological Association; 1998.

Gardner W, Kelleher KJ, Pajer KA, Campo JV. Primary care clinicians' use of standardized tools to assess child psychosocial problems. Ambulatory Pediatrics. 2003; 3(4):191–195. [PubMed: 12882596]

Hosmer, DW.; Lemeshow, S. Applied logistic regression. 2nd ed.. Hoboken, NJ: John Wiley & Sons; 2004.

Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association. 2012; 19:263–274. [PubMed: 21984587]

Kaufman J, Birmaher B, Brent D, Rao U. 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. 1997; 36(7):980–988. [PubMed: 9204677]

Klein RG, Pine DS, Klein DF. Resolved: Mania is mistaken for ADHD in prepubertal children. Journal of the American Academy of Child & Adolescent Psychiatry. 1998; 37(10):1993–1096.

Knapp PK, Foy JM. Integrating mental health care into pediatric primary care settings. Journal of the American Academy of Child & Adolescent Psychiatry. 2012; 51(10):982–984. [PubMed: 23021473]

Kowatch RA, Youngstrom EA, Horwitz S, Demeter C, Fristad MA, Birmaher B, Findling RL. Prescription of psychiatric medications and polypharmacy in the LAMS cohort. Psychiatric Services. 2013; 64:1026–1034. [PubMed: 23852186]

Kraemer HC. Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. Psychoneuroendocrinology. 1992; 17:527–536. [PubMed: 1287676]

Lindhiem O, Kolko DJ, Yu L. Quantifying diagnostic uncertainty using item response theory: The posterior probability of diagnosis index. Psychological Assessment. 2013; 25:456–466. [PubMed: 23356682]

Lowell DI, Carter AS, Godoy L, Paulicin B, Briggs-Gowan MJ. A randomized controlled trial of Child FIRST: A comprehensive home-based intervention translating research into early childhood practice. Child Development. 2011; 82:193–208. [PubMed: 21291437]

MathWorks, Inc.. MATLAB. Natick, MA: 2011.

Morgan PL, Staff J, Hillemeier MM, Farkas G, Maczuga S. Racial and ethnic disparities in ADHD diagnosis from kindergarten to eighth grade. Pediatrics. 2013; 132:85–93. [PubMed: 23796743]

Muthen, L.; Muthen, B. Mplus 6.1. Los Angeles, CA: 2010.

Nakamura BJ, Chorpita BF, Hirsch M, Daleiden E, Slavin L, Amundson MJ, Vorsino WM. Large-scale implementation of evidence-based treatments for children 10 years later: Hawaii's evidence-based services initiative in children's mental health. Clinical Psychology: Science and Practice. 2011; 18(1):24–35.

Parker A, Corkum P. ADHD Diagnosis: As simple as administering a questionnaire or a complex diagnostic process? Journal of Attention Disorders. 2013

Pelham WE, Foster EM, Robb JA. The economic impact of attention-deficit/hyperactivity disorder in children and adolescents. Ambulatory Pediatrics. 2007; 7:121–131. [PubMed: 17261491]

Pliszka S. Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry. 2007; 46:894–921. [PubMed: 17581453]

Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: How to compare Brier scores. Journal of Clinical Epidemiology. 1991; 44(11):1141–1146. [PubMed: 1941009]

Reise SP, Yu J. Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement. 1990; 27(2):133–144.

Rettew DC, Lynch AD, Achenbach TM, Dumenci L, Ivanova MY. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. International Journal of Methods in Psychiatric Research. 2009; 18:169–184. [PubMed: 19701924]

Robinson LM. Promoting multidisciplinary relationships: a pragmatic framework for helping service providers to work collaboratively. Canadian Journal of Community Mental Health (Revue canadienne de santé mentale communautaire). 2005; 24(1):115–127. [PubMed: 16568625]

Snyder SM, Hall JR, Cornwell SL, Quintana H. Review of clinical validation of ADHD behavior rating scales. Psychological Reports. 2006; 99(2):363–378. [PubMed: 17153805]

Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine. 1986; 5:421–433. [PubMed: 3786996]

StataCorp.. Stata Statistical Software: Release 12. College Station, TX: 2011.

Straus, SE.; Glasziou, P.; Richardson, WS.; Haynes, RB. Evidence-based medicine: How to practice and teach it. 4th ed.. Philadelphia: Churchill Livingstone; 2011.

Straus SE, Tetroe JM, Graham ID. Knowledge translation is the use of knowledge in health care decision making. Journal of Clinical Epidemiology. 2011; 64:6–10. [PubMed: 19926445]

Wolraich ML, Bard DE, Stein MT, Rushton JL, O'Connor KG. Pediatricians' attitudes and practices on ADHD before and after the development of ADHD pediatric practice guidelines. Journal of Attention Disorders. 2010; 13:563–572. [PubMed: 19706877]

Wolraich ML, Hannah JN, Baumgaertel A, Feurer ID. Examination of DSM-IV critieria for attention deficit /hyperactivity disorder in a county-wide sample. Journal of Developmental & Behavioral Pediatrics. 1998; 19:162–168. [PubMed: 9648041]

Wolraich ML, Lambert W, Doffing MA, Bickman L, Simmons T, Worley K. Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. Journal of Pediatric Psychology. 2003; 28:559–568. [PubMed: 14602846]

Wolraich ML, McKeown RE, Visser SN, Bard D, Cuffe S, Neas B, Danielson M. The prevalence of ADHD: Its diagnosis and treatment in four school districts across two states. Journal of Attention Disorders. 2012

Wren FJ, Bridge JA, Birmaher B. Screening for childhood anxiety symptoms in primary care: Integrating child and parent reports. Journal of the American Academy of Child and Adolescent Psychiatry. 2004; 43(11):1364–1371. [PubMed: 15502595]

Youngstrom EA. Future directions in psychological assessment: Combining evidence based medicine innovations with psychology's historical strengths to enhance utility. Journal of Clinical Child & Adolescent Psychology. 2012; 42:139–159. [PubMed: 23153181]

Youngstrom EA, Arnold LE, Frazier TW. Bipolar and ADHD comorbidity: Both artifact and outgrowth of shared mechanisms. Clinical Psychology: Science and Practice. 2010; 17:350–359. [PubMed: 21278822]

Zelnik N, Bennett-Back O, Miari W, Goez HR, Fattal-Valevski A. Is the Test of Variables of Attention reliable for the diagnosis of attention-deficit hyperactivity disorder (ADHD)? Journal of Child Neurology. 2012; 27:703–707. [PubMed: 22378668]

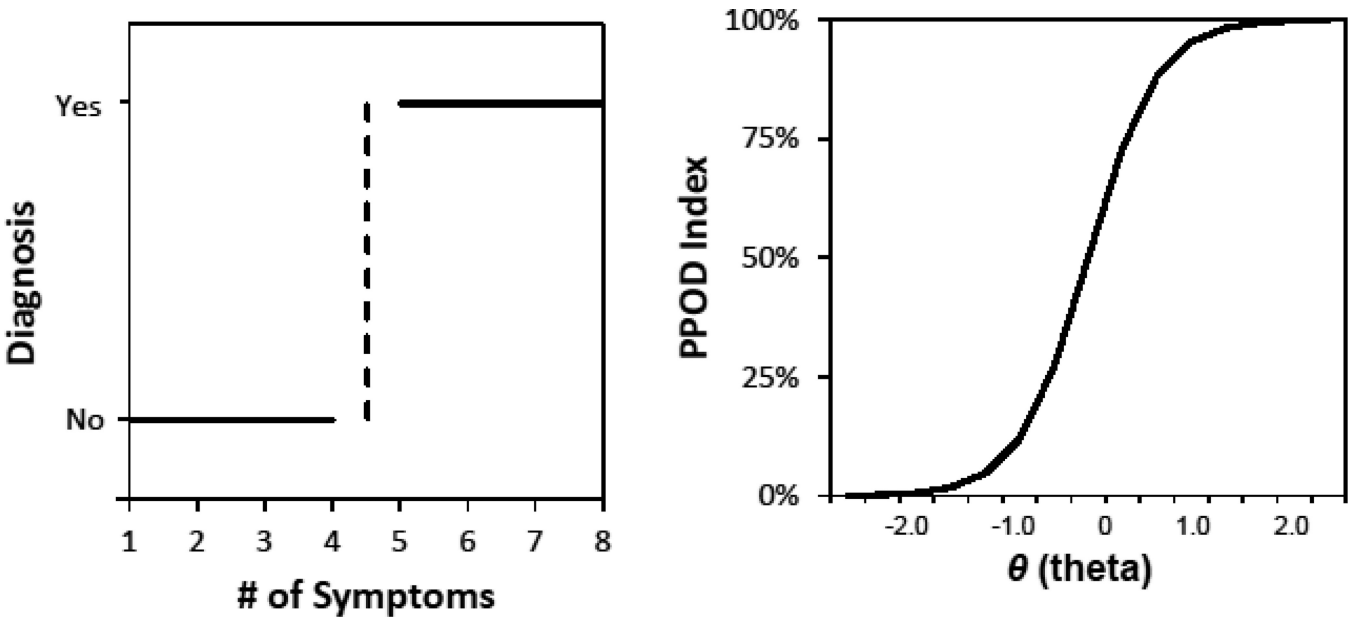

Figure 1.
A graphical depiction of the conceptual difference between traditional symptom counts and the PPOD Index.

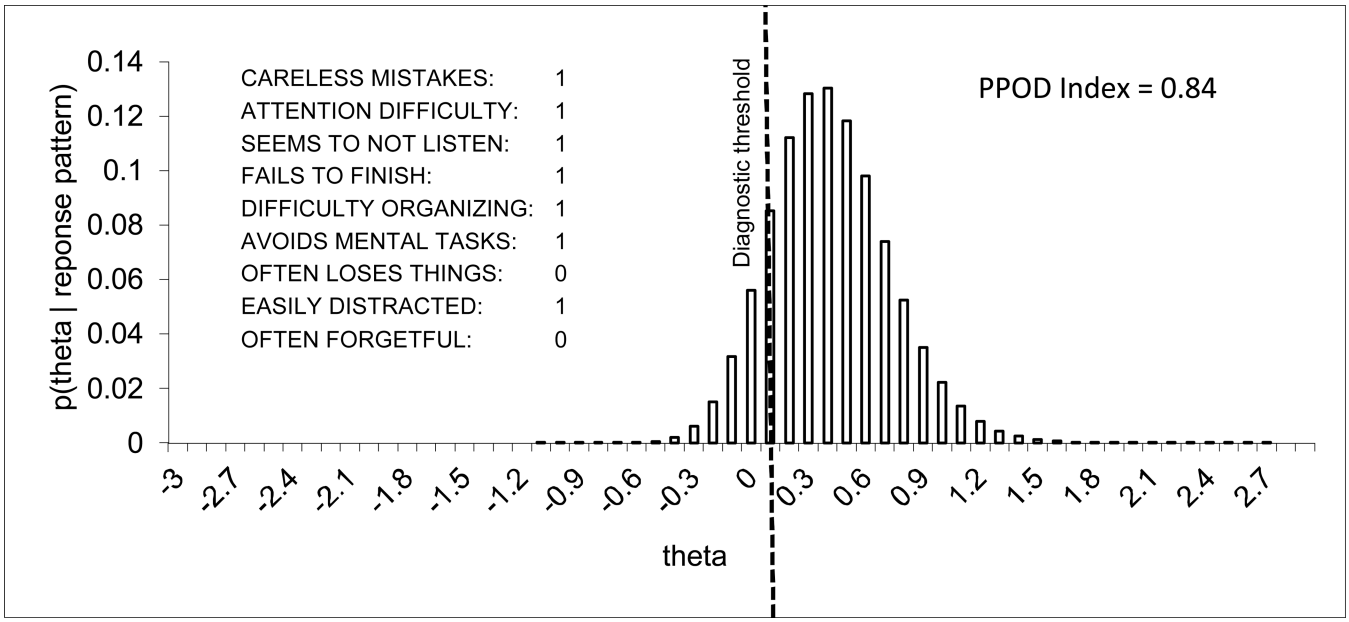

Figure 2.

The PPOD Index is estimated by numerically integrating the posterior distribution of θ above the diagnostic threshold for individual symptom patterns.

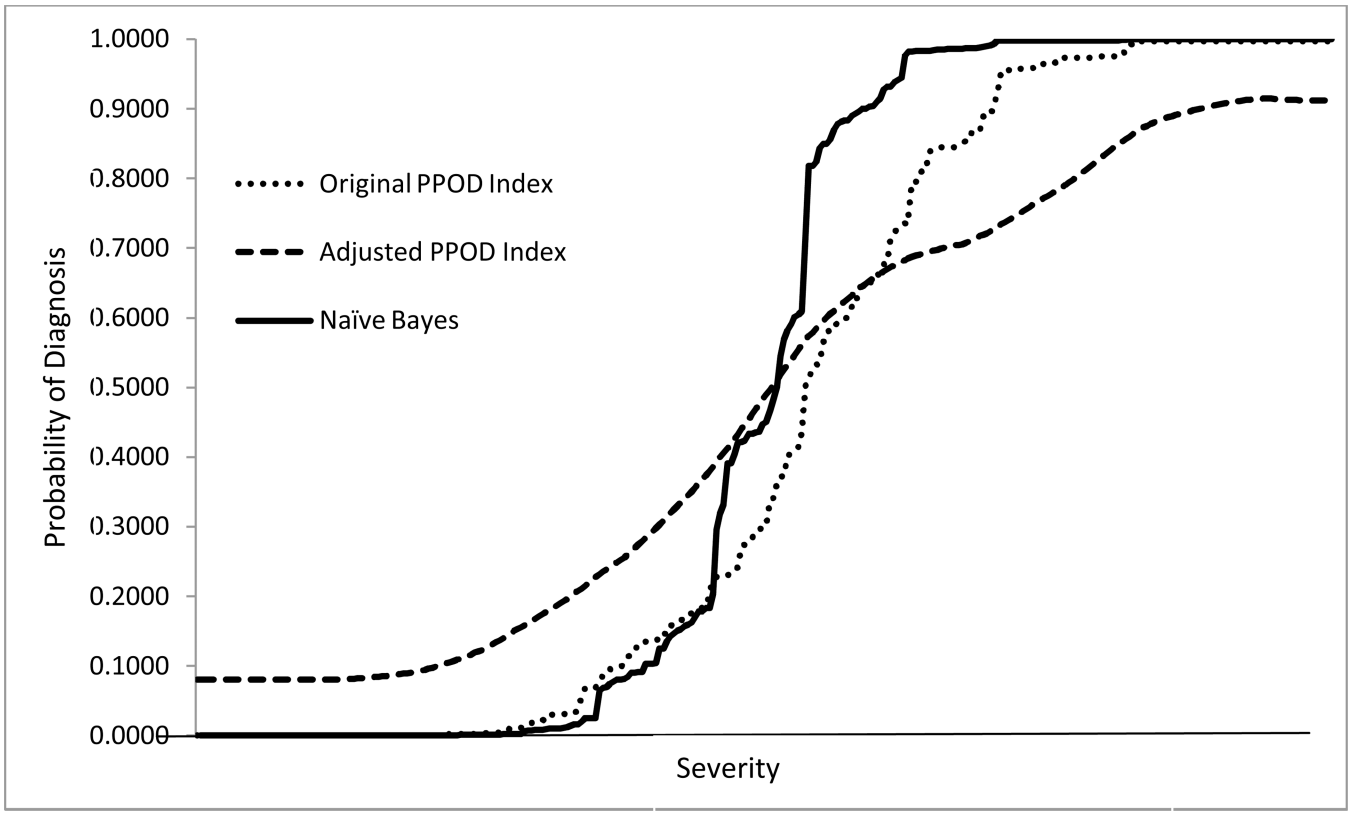

Figure 3.
Curves showing the original PPOD Index values, adjusted PPOD Index values, and Naïve Bayes probability estimates for all cases (N = 321) in order or severity.

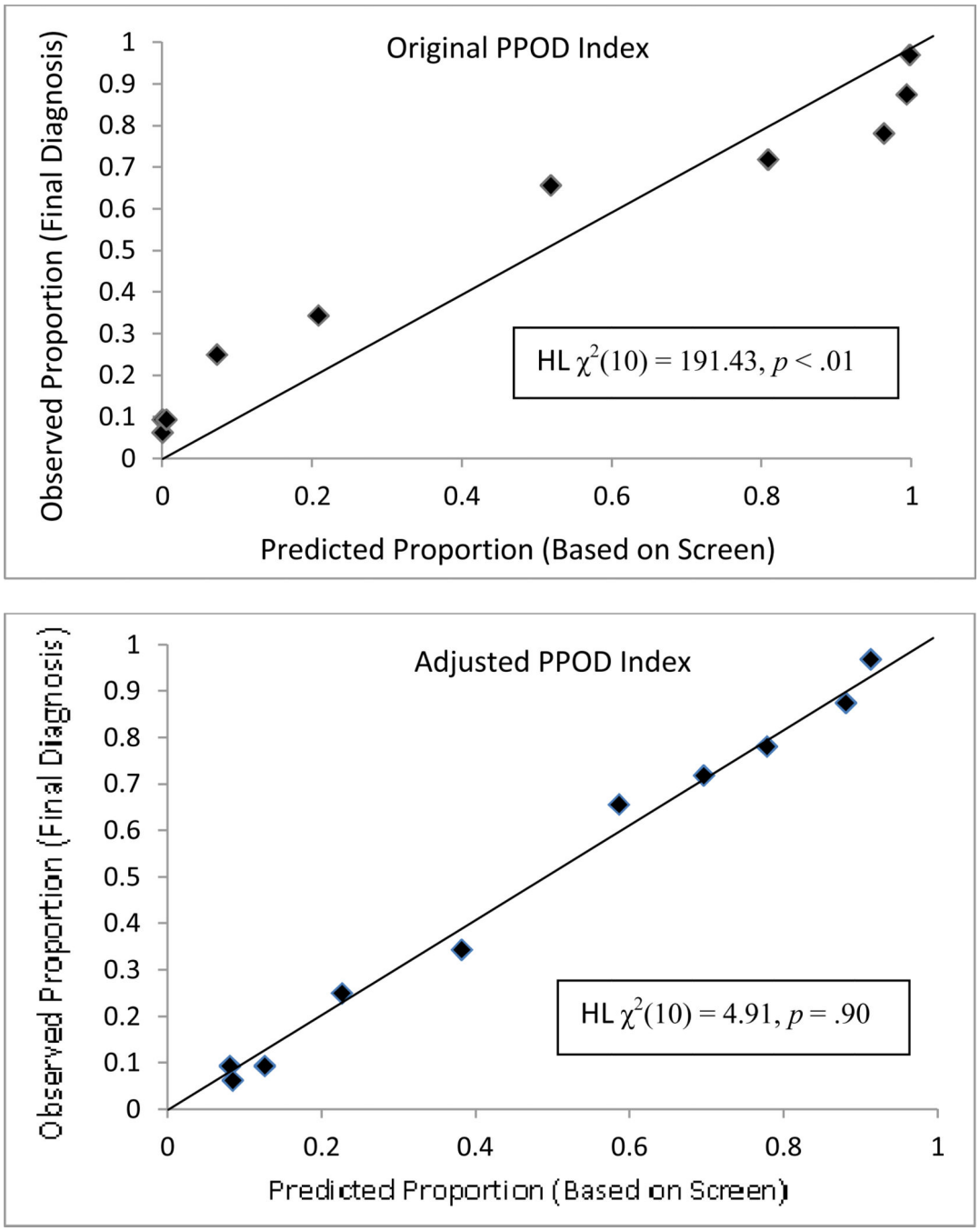

Figure 4.
Calibration plots for the original PPOD Index and adjusted PPOD Index. Perfect calibration is represented by the diagonal lines.

Table 1

Item Parameters for the Symptoms of ADHD, Predominantly Inattentive Type, Sorted by Ascending Order of β (2PL Model)

	Item Parameters		Standard Errors	
Item	**α**	**β**	**σ_α**	**σ_β**
Does not seem to listen when spoken to directly	1.58	−0.55	0.23	0.14
Is easily distracted by noises or other stimuli	2.98	−0.51	0.53	0.13
Has difficulty keeping attention to what needs to be done	3.72	−0.41	0.63	0.11
Does not pay attention to details or makes careless mistakes with, for example, homework	3.01	−0.16	0.45	0.10
Does not follow through when given directions and fails to finish activities (not due to refusal or failure to understand)	2.95	−0.06	0.45	0.10
Avoids, dislikes, or does not want to start tasks that require ongoing mental effort	2.19	0.07	0.32	0.11
Has difficulty organizing tasks and activities	2.69	0.09	0.41	0.10
Is forgetful in daily activities	2.92	0.19	0.53	0.09
Loses things necessary for tasks or activities (toys, assignments, pencils, or books)	2.32	0.22	0.37	0.10

Table 2

Categorical Diagnoses, Symptoms Counts, PPOD Indices

Symptom Counts	Categorical DSM Diagnosis	Original PPOD Index Range	Adjusted PPOD Index Range
9	YES	.99	.86 – .91
8	YES	.96 – .99	.73 – .86
7	YES	.72 – .92	.68 – .73
6	YES	.50 – .81	.57 – .69
5	NO	.16 – .51	.33 – .57
4	NO	.08 – .23	.23 – .42
3	NO	.01 – .07	.15 – .23
2	NO	.00 – .01	.10 – .16
1	NO	.00	.08 – .10
0	NO	.00	.08

Table 3

Discrimination and Calibration of the Original PPOD Index, Adjusted PPOD Index, and Naïve Bayes Probability Estimates

	Original PPOD Index	Adjusted PPOD Index	Naïve Bayes
Area Under the Curve (AUC)	.878 (p <.001)	.880 (p <.001)	.865 (p <.001)
Sensitivity	.795	.821	.819
Specificity	.842	.806	.823
Positive Predictive Value (PPV)	.827	.853	.814
Negative Predictive Value (NPV)	.813	.778	.828
Brier Score	.149	.137	.158
Spiegelhalter's z	10.30 (p <.001)	−0.89 (p =.81)	17.86 (p <.001)
Hosmer-Lemeshow $\chi^2(10)$	191.43 (p <.001)	4.91 (p =.90)	473.63 (p <.001)