Portability of a Screener for Pediatric Bipolar Disorder to a Diverse Setting

Andrew J. Freeman

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements of the degree of Master of Arts in the Department of Psychology.

Chapel Hill
2010

Approved by:

Eric Youngstrom, Ph.D.

Andrea Hussong, Ph.D.

David Thissen, Ph.D.

Abstract

Andrew Freeman: Portability of a Screener for Pediatric Bipolar Disorder to a Diverse Setting

(Under the direction of Eric Youngstrom)

The purpose of the study is to examine differential item functioning when moving from an Academic setting to a community setting, differential item functioning when extracting ten items in a community setting, and comparative diagnostic efficiency of the extracted items to the embedded items. Differential item functioning using Samejima's Graded Response Model indicated that across samples, total observed scores were similar across levels of mania. ROC indicated that the ten extracted items discriminated well. Sum scores less than 18 substantially decreased the probability of bipolar disorder, while sum scores greater than 18 substantially increased the probability of bipolar disorder in youth. Findings suggest that the extracted items perform similarly to the embedded items in the community setting.

Table of Contents

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

List of Abbreviations

DIF - Differential Item Functioning

EA - Embedded Academic

EC - Embedded Community

GBI - General Behavior Inventory

GRM - Graded Response Model

IRT - Item Response Theory

KSADS - Schedule for Schizophrenia and Affective Disorders - Child version

PBD - Pediatric Bipolar Disorder

P-GBI - Parent-reported General Behavior Inventory

PGBI-10M - Parent-reported 10 item General Behavior Inventory

ROC - Receiver-Operating Characteristic

Q-ROC - Quality Receiver-Operating Characteristic

I. Introduction

Clinic visits associated with pediatric bipolar disorder (PBD) have increased forty-fold in the last decade to almost 1% of outpatient visits (Moreno, et al., 2007). General population prevalence estimates suggest that up to 1.9% of youth are affected with bipolar disorder (Van Meter, Moreira, & Youngstrom, 2009). These figures suggest that clinicians may be identifying less than one-half of youth with PBD in their office (Moreno, et al., 2007). The discrepancy and controversy (e.g., Biederman, Klein, Pine, & Klein, 1998) surrounding PBD is most likely due to chronic underdiagnosis of PBD (Blader & Carlson, 2007; Youngstrom, Youngstrom, & Starr, 2005) and possibly over-diagnosis as seen in lack of agreement between clinical and research diagnoses (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009).

PBD is characterized by periods of time where youth experience elevated mood, increased energy, irritability, grandiosity, and decreased need for sleep (Geller, et al., 1998). Episodes in youth are often long and symptom severity fluctuates (Birmaher, et al., 2006; Wozniak, Biederman, Kiely, & Ablon, 1995). Thus, shifting moods make identifying the index mood episode difficult. Additionally, symptoms of PBD overlap with multiple other disorders such as ADHD (Bowring & Kovacs, 1992; Kim & Miklowitz, 2002). Therefore, accurate diagnosis is made by correctly identifying symptoms' frequency, intensity, number, and duration (Quinn & Fristad, 2004). Retrospective studies of adults with bipolar disorder have shown that a pediatric onset is associated with worse clinical outcomes such as: Increased substance misuse (Wittchen, et al., 2007), more frequent cycling (Leverich, et al., 2007; Perlis, et al., 2004), and increased rates of suicidal acts (Angst, Stassen, Clayton, & Angst, 2002; Goldstein, et al., 2005). Current hypotheses suggest that poor clinical outcomes are due to the repeated "kindling" of mood over time (Post, Susan, & Weiss, 1992). Therefore, early identification and early treatment are thought to be best for a severe mental illness for which a correct diagnosis is often obtained at least one year after the first visit to a doctor for emotional or behavior problems and regularly as long as 11 years (Hirschfeld, 2001; Lish, Dime-Meenan, Whybrow, Price, & Hirschfeld, 1994; Stang, et al., 2006).

Youngstrom et al. (2004) have shown that both broad and narrow band measures are adequate identifiers of PBD. Whereas broadband measures such as the Achenbach Child Behavior Checklist are widely used clinically (Belter & Piotrowski, 2001; Clemence & Handler, 2001), the CBCL does not measure the core symptoms of mania (Achenbach & Rescorla, 2001). Multiple narrowband measures of mania have been developed for the past three decades (e.g., Depue, 1981); however, the Beck Depression Inventory (Beck, Steer, Ball, & Ranieri, 1996) is the only narrowband mood measure that appears on lists of commonly used instruments (Clemence & Handler, 2001). The narrowband measures of bipolar disorder are often developed for selecting college students at risk (e.g., Depue, 1981) or outpatient psychiatric units associated with hospitals (e.g., Hirschfeld, et al., 2000; Pavuluri, Henry, Devineni, Carbray, & Birmaher, 2006).

The validity of measures examining emotions could change when moving from a rarefied university or academic medical center setting to community mental health centers due to changes in socio-economic status, ethnicity of participants, acuity of presenting problem, types of present problems, and differences in patterns of comorbidity (Kowatch, Youngstrom, Danielyan, & Findling, 2005; Neighbors, Jackson, Campbell, & Williams, 1989; Youngstrom & Green, 2003). Messick (1989) states that validity is relative to purpose and setting. As a result, the utility and functioning of a measure must be repeatedly demonstrated. For example, the Mood Disorder Questionnaire has a well-documented deficit in its ability to identify bipolar disorder in non-academic settings (Hirschfeld, et al., 2003; Miller, Klugman, Berv, Rosenquist, & Ghaemi, 2004). The lack of discrimination outside of highly specialized settings could be due to Berkson's bias - distortion of statistical and meaningful conclusions due to pre-existing bias in sampling technique (Berkson, 1946). Traditionally, Berkson's bias results from comparing hospital samples to healthy community controls, because the exposure to risk factors is different even when controlling for gender, age, ethnicity, and SES. In measurement, this is reflected in the target population changing. For example, an academic medical center sample consisting of youth with PBD, attention deficit hyperactivity disorder without mood disorder, and youth who are healthy will be unable to identify items that will have robust generalizability for screening for PBD because the probability of other conditions such as conduct disorder has not been controlled (Geller, et al., 2002). Measures developed or tested in highly selected samples might not generalize out to community mental health because the target population is changed. This possibility was demonstrated across multiple measures of bipolar symptoms, which all had better psychometric performance under more "distilled" academic medical center type sampling conditions (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006). For a measure to be used in widespread screening of a diagnosis, the measure should be robust across diverse samples (Kraemer, 1992; Strauss, Richardson, Glasziou, Haynes, & Strauss, 2005).

Screening measures present entirely different challenges to measure writers and researchers than do more typical narrowband measures meant to cover a specific disorder or construct. Typically, the items on a given measure are viewed as a sampling of all potential items that pertain to the construct (DeVellis, 2003). For example, a general measure of mania should contain items that provide balanced coverage across all symptoms and associated features. Additionally, the varying levels of each symptom should be queried either through individual item's response options or alternate items. However, accurate diagnosis of mania could rely not on measuring the entire construct of mania, but on using items that discriminate mania from other conditions the best (Guyatt & Rennie, 1993; Strauss, et al., 2005). Items on diagnostic measures should purposefully be selected for having high discrimination amongst conditions.

Youngstrom et al. (2008) developed a brief screener for PBD that selected the 10 most discriminating items of the General Behavior Inventory (GBI). Originally, the GBI was meant to provide symptomatic coverage of subthreshold mania and identify adults at risk of bipolar disorder in college samples (Depue, 1981). However as a screening measure, the GBI has four significant shortcomings for both PBD and a community mental health population. First, the GBI is a self-report measure. Youth with PBD generally have poor insight into their symptoms (Dell'Osso, et al., 2002; Youngstrom, Findling, & Calabrese, 2003). Second, the GBI is written at a university reading level, with complex sentences and descriptions of the phenomena. Best practices guidelines for writing items are to use simple sentences that measure single constructs (DeVellis, 2003;

Gronlund, 1988). Third, the items often are a conglomerate of multiple symptoms, or they juxtapose multiple aspects of functioning and behavior. Fourth, the GBI is a long measure - encompassing 73 items, 2,896 words, and 7 pages in 11 point font. The 10 item parent report GBI (PGBI-10M) removed the issue of compromised insight and decreased the burden, while maintaining adequate diagnostic efficiency.

The PGBI-10M is not technically a "short-form" because it does not systematically cover depression as does the original GBI (Smith, McCarthy, & Anderson, 2000). "Scale carving" is a more apt descriptor of the PGBI-10M. Scale carving is traditionally used when the goal is to reduce burden on participants, while continuing to measure the construct of interest. The PGBI-10M represents the ten most discriminating items carved from the parent report GBI. Scale carving is relatively rarely studied; however, it could have a number of deleterious effects. First, scale carving can change the observed score mean and standard deviation (Desai & Braitman, 2005). For diagnostic measures of PBD, decision-making is typically based upon the sum of the raw scores (e.g., Youngstrom, et al., 2004). Therefore, changes in mean or standard deviation would indicate that alternate cut scores should be used. Second, responses to items can cahnge depending on the sequence of the items (Hamilton & Shuminsky, 1990; Knowles, 1988; Steinberg, 1994). Items originally near the end of the measure could be highly discriminating due to position, but once extracted, those items' ability to discriminate could lessen. Therefore, the extraction of the ten best discriminating items from the context of sixty-three other mood related items might result in a change to the discriminatory ability of each of the ten individual items on the GPBI-10M and the scale as a whole.

The items on the parent report GBI provides a context of querying about mood and change in mood. The removal of this context could result in a context effect. Context effects are traditionally defined as the interaction between the content of the prior item with current item (Schuman, Presser, & Ludwig, 1981). Tourangeau, Rips, & Rasinski (2000) expanded the definition of a context effect to include not just prior items, but also survey mode, interviewer characteristics, changes in survey form, and other external factors such as temperature. The process of identifying a response is multi-staged (Tourangeau & Rasinski, 1988). First, items are read and interpreted. Second, relevant memories are recalled. Third, these memories are applied to the item. A context effect is any external factor that could cause error or misinterpretation of information to occur in the evaluation of an item. During each of these steps, the lack of the prior 63 items - the "context" - could cause responses to change in the PGBI-10M as compared to the original parent report GBI.

Standards of Evidence Based Medicine imply that a good diagnostic measure should be able to accurately and efficiently discriminate between cases and non-cases, regardless of length, construct coverage, or changes in item functioning (Guyatt & Rennie, 1993; Kraemer, 1992; Strauss, et al., 2005). Discrimination between cases and non-cases is based upon both a measure's sensitivity and specificity. *Sensitivity* is the proportion of those that have the criterion diagnosis that are correctly identified. *Specificity* is the proportion of cases without the criterion diagnosis correctly identified. Sensitivity divided by (1-Specificity) yields the diagnostic likelihood ratio associated with a positive test result (Deeks & Altman, 2004). As the likelihood ratio increases, the probability of having the target condition increases. Conversely, the negative diagnostic likelihood ratio aids in preventing over-diagnosis. The negative diagnostic likelihood ratio is (1 - Sensitivity) divided by specificity. The parent reported GBI has shown adequate diagnostic discrimination in a racially and socio-economically diverse setting (Youngstrom, Meyers, et al., 2005). For the PGBI-10M to be useful diagnostically, the ten independent items must continue to discriminate between cases and non-cases. The PGBI-10M needs to increase the probability of PBD being the correct diagnosis (i.e., high scores should be associated with a high likelihood ratio, resulting in a higher probability of PBD when all other conditions are equal).

In the present study, the PGBI-10M will be examined for transportability into new settings. *Specific aims include:*

1) Examine differential item functioning and differential test functioning of the ten items on the parent reported GBI between two socio-economic and racially distinct samples.

2) Examine the differential item functioning and differential test functioning of the extracted ten items in the form of the PGBI-10M compared to the embedded ten items in the form of the full parent report GBI.

3) Examine the diagnostic efficiency of the PGBI-10M when administered separately compared to the 10-items embedded within the parent-reported GBI.

II. Method

Participants

Participants were 2252 youths presenting at either an urban academic medical center (n = 813) or an urban community mental health center (n = 1439) in the Midwest divided into two subsamples. Inclusion criteria for the current study at both sites were: 1) Youths between the ages of 5 years and 18 years, 2) Both caregiver and youth provided written consent and assent, 3) both caregiver and youth presented for the assessment, and 4) both caregiver and youth were conversant in English. The total sample was split into four groups: Embedded Academic (EA), Embedded Community (EC), Extracted, and Both.

The EA group consisted of 813 youths and their caregivers from an academic medical center. The EC group consisted of 481 youths from the community mental health center. The primary caregivers of the EA and EC youth completed the full parent-reported GBI. The Extracted group consisted of 799 youths from the community mental center, whose parents completed the PGBI-10M only as standalone measure during general intake to the clinic. The Both group consisted of 159 youths from the community mental health center, whose parents completed both the PGBI-10M at general intake and then later completed full parent-reported GBI during an expanded research protocol. Table 1 displays the demographic information for EA, EC, and Both groups.

Recruitment. The academic medical center site had multiple pharmacotherapy trials open for bipolar spectrum disorders, unipolar depression, schizophrenia, attention-deficit/hyperactivity disorder, unipolar depression, and post-traumatic stress disorder (as described in Findling, et al., 2001). Youths were referred by community mental health workers or responded to advertisements. Youths and caregivers willing to participate in treatment protocols were included if their initial symptoms appeared to match the enrollment criteria for open trials. Additionally, the sample was enriched for mood disorders with offspring of parents with bipolar disorder who were receiving treatment at an affiliated adult mood disorders clinic.

The community mental health center site consisted of youths and caregivers presenting at a Midwestern urban community mental health center for treatment. Using a consecutive case series design at intake, all youth and caregiver pairs were asked to participate in an assessment research study. When research capacity was full, a random subsample of all intakes was offered the chance to participate. All youth - regardless of initial presentation - between the ages of 5 years and 18 years were eligible to participate in the current study.

Measures

Schedule for Affective Disorders and Schizophrenia for Children(KSADS). The KSADS is a semi-structured interview that queries symptoms from common Axis I disorders from both the parent and child. The KSADS-PL-Plus is an amalgamation of the mood modules from the Washington University KSADS (Geller, et al., 2001) and the KSADS Present & Lifetime version (Kaufman, et al., 1997). The Washington University KSADS includes additional symptoms and associated features of depression and mania beyond those included in the KSADS Present & Lifetime version.

Parent Report General Behavior Inventory. The parent report GBI is a modified GBI in that all questions now query the parent about the mood and behavior of his/her offspring (Youngstrom, Findling, Danielson, & Calabrese, 2001). The parent report GBI consists of 73 items measuring depressive, hypomanic, and mixed symptoms of mood disorder. Participants answer “Never or Hardly Ever” to “Very Often or Almost Constantly” on a four point Likert scale about their offspring. An example of a depressive symptom item is: “Have there been times of three days or more, when your child was not physically ill, that your child were so tired or worn out that it was very difficult or even impossible for your child to do normal every day activities?” An example of a hypomanic item is: “Has your child experienced periods of several days or more when, although your child was feeling unusually happy and intensely energetic (clearly more than your child’s usual self), your child also was physically restless, unable to sit still, and had to keep moving or jumping from one activity to another?” An example of a mixed item is: “Has your child’s mood or energy shifted rapidly back and forth from happy to sad or high to low?” The parent report GBI consists of two scales, like the original GBI, Depression (Cronbach's α = .96) and Hypomanic/Biphasic (Chronbach's α = .92, as reported in Youngstrom, et al., 2004).

10-item General Behavior Inventory. The PGBI-10M was developed from the parent report GBI using item response theory to determine the 10 best discriminating items (Youngstrom, et al., 2008). The PGBI-10M consists of items from only the hypomanic/biphasic scale of the parent-report GBI. Participants answer "Never or Hardly Ever" to "Very Often or Almost Constantly" on a four-point Likert scale about their offspring (Cronbach's α = .92).

Procedure

The protocol for Embedded Academic, Embedded Community, and Both groups were similar. Caregivers provided written consent for the youth to participate in the study. Youth provided written assent to participate in the study. The same research assistant interviewed both caregiver and youth sequentially with the KSADS. Caregivers completed the parent reported GBI that was included in an additional battery of assessment measures.

Recruitment for the Embedded Community and Both groups occurred during a general clinical intake. During this time, the PGBI-10M was administered. The Both group consists of individuals who completed both the PGBI-10M, agreed to participate in the assessment study, and presented for the assessment study. The Extracted Group received the PGBI-10M as part of standard clinical care, and de-identified data were coded for comparison to the other versions.

Diagnosis. Research assistants were highly trained. Symptom level ratings were compared with a reliable rater for new raters for at least 5 interviews rating along and then 5 interviews leading. A new rater passed a session if he/she achieved an overall $\kappa>=.85$ at the item level of the entire interview and a $\kappa=1.0$ at the diagnostic level. All cases were reviewed using the Longitudinal Evaluation of All Available Data (LEAD) procedure (Spitzer, 1983). After completing the interview process, the research assistant met with a licensed clinical psychologist to review the case. During the LEAD meeting, the research assistant presented the KSADS symptoms and diagnoses, family history, and any information available from intake (e.g., intake diagnoses, chart review of diagnoses, prior treatment history, and school history). Both the licensed clinical psychologist and the research assistant were blind to the parent-report GBI and the PGBI-10M.

III. Results

Introduction to Item Response Theory

Item response theory is a collection of models that allows for the evaluation of an item's functioning on an underlying trait by quantifying its properties. The *discrimination parameter* represents the relationship between the item response and the latent trait. The *difficulty parameter* represents how much of the latent trait a person needs before choosing a given response. For items with only two categories, the discrimination and difficulty parameters define the item characteristic curve. Samejima's graded response model (GRM) (1969) is a generalized two parameter logistic model. GRM estimates a boundary response function that combines the discrimination parameter and the difficulty parameters. The boundary response functions for each threshold are calculated by: $P_{ik}^{*}(\theta_s) = \frac{e^{a_i(\theta_s - b_{ik})}}{(1+e^{a_i(\theta_s - b_{ik})})}$. $P_{ik}^{*}(\theta_s)$ represents the probability that an examinee with an ability level (θ) will respond to item *i* at or above category *k*. The boundary response equations allow for the estimation of α and *b* parameters. The α parameter, often the *discrimination parameter*, is represented by the item characteristic curve. The *b* parameter, often referred to as the *difficulty parameter*, determines the horizontal inflection point of the item characteristic curve. In GRM, the *difficulty parameter* represents the threshold, or level of latent trait, at which a higher response will be chosen at least 50% of the time. The combination of the α parameter and *b* parameters results in the boundary response function. Although item response theory's origins are in large scale educational testing, the principles and meanings of the parameters are applicable to psychopathology (Thissen & Steinberg, 1988). The *discrimination parameter* is a measure of how related an item is to the trait, similar to a factor loading. The *difficulty parameter* is a measure of the level of illness on a dimension that a person needs to endorse the item. In the current paper, the parameters of GRM indicate both how related a specific symptom is to mania and how much mania a youth must have before his/her caregiver will endorse a more frequent response.

Differential item functioning (DIF) occurs when two groups of people with the same ability level do not have the same probability of choosing identical responses (Lord, 1980). Likelihood ratio tests of model fit are currently considered best practices because they allow for direct tests of both the discrimination and difficulty parameters (Camilli & Shepard, 1994; Thissen, Steinberg, & Wainer, 1993). Likelihood ratio tests of DIF occur by the following steps: 1) Constraining all parameters and all items to be equal; 2) Constrain all parameters and all items equal except for *item i*; 3) Compare the -2(loglikelihoods) for the two models by subtraction, if they are not significantly different then no DIF is assumed; 4) If there is significant DIF, then constrain all items and parameters to be equal except for the difficulty parameters of *item i*. During step 4, two different -2(loglikelihoods) are produced that allow for the determination of whether the *disrcimination parameter* or the *difficulty parameters* are indicating DIF (Thissen, 2001). Due to the number of multiple comparisons being made in the DIF process, the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) as implemented in Excel (Thissen, Steinberg, & Kuang, 2002) was used to control the false discovery rate. Items showing DIF were graphed and visually inspected to evaluate the extent of DIF (Steinberg & Thissen, 2006).

Evaluation of Item Response Theory Assumptions

The GRM assumes unidimensionality and local independence. Unidimensionality is the presence of a single underlying latent trait. Local independence posits that the relationship among any set of items is due only to the single underlying latent trait. Satisfying the unidimensionality assumption usually satisfies the local independence assumption.

Using Comprehensive Exploratory Factor Analysis (Browne, Cudeck, Tateneni, & Mels, 2004), an ordinary least squares exploratory factor analyses of the polychoric correlation matrix were used to test the assumption of unidimensionality for the ten items in the EA, EC, and Extracted groups. Ordinary least squares exploratory factor analysis allows for the factoring of ordered categories that typically violate the assumption of multivariate normality associated with maximum likelihood estimation procedures (Wirth & Edwards, 2007). Table 2 displays the unrotated first order factor loadings. All items loaded strongly on a single underlying latent trait. A unidimensional structure was supported by graphing a scree plot (Cattell, 1966), minimum average partials (Velicer, 1976), and Glorfeld's extension of Horn's parallel analysis (Glorfeld, 1995). Each of these methods indicated a single factor solution.

A confirmatory factor analysis with one latent variable for each of the three samples was fit using Mplus 5.21 (Muthen & Muthen, 1998-2007) to examine whether local dependence occurred. A CFI greater than .95 and RMSEA less than .10 are considered indicators of acceptable fit (Hu & Bentler, 1998). The single factor model displayed poor fit in all three groups (all CFIs < .95 & RMSEAs > .10). Therefore, modification indices were examined in each of the three samples. In all three samples, the residual correlation between items 3 and 7 was the largest. When these two items were allowed to correlate, the model fit was deemed acceptable (all CFIs < .95 and RMSEAs < .10). The correlation between items 3 and 7 indicate local dependence. Local dependence is when two or more items correlate with each other beyond what is explained by the underlying factor. Local dependence can occur in the underlying dimension or at a surface level (Thissen, Bender, Chen, Hayashi, & Wiesen, 1992). Surface local dependence occurs when tests contain similar or redundant items causing overestimation of the quantity of information provided by the test which is often seen as upwardly biased discrimination parameter estimates. Examining the text of the items suggests that surface local dependence is occurring. Additionally, Item 8 also queries about mood shifts; however, it does not have significant modification indices suggesting that there is not a second latent factor. In addition a seperate item level confirmatory factor analyses indicated poor model fit when these three items were part of a seperate factor or a bifactor was added (CFIs < .95, RMSEA > .10). As DIF can be assumed to be the examination of whether a second factor alters item response, each set of DIF analyses were ran three times: including all ten items, excluding only item 3, and excluding item 7. Results were not substantially or substantively different across the three analyses. Discrimination parameters for the two items were slightly lower when the other item was excluded. Results presented are from analysis including all ten items.

Aim 1: Differential Item Functioning of the 10 embedded items between an Academic Medical Center and Community Mental Health Center.

For the portability analyses, EA was the reference group and EC was the focal group. Table 3 displays the item parameters and g^2 goodness of fit index for the 10 items. Items 5, 9, and 10 displayed no evidence of DIF after controlling for the false discovery rate, *p*s > .05. Items 3 and 6 had significantly lower discrimination parameters in the EC group than in EA group. The discrimination DIF on rapid mood and energy shifts (Item 3) and the elated mood or energy with sleep disturbance (Item 6) suggested that these might be significantly worse markers of mania in the EC group than the EA group. However, item 2 (which examines only elated mood) discriminated significantly better in the EC group than the EA group. Elated mood had a stronger relationship in the EC group than the EA group with mania. However, on examination of the items' ICCs in Figure 1, the practical effect size of the difference was relatively small. On the difficulty parameters, Items 1, 4, 7, and 8 showed significant differences, *p*s <.05. Items 7 and 8 produced significantly lower scores in the EC group than the EA group because they required a small to moderate amount less of mania to endorse a higher response. Item 1 was significantly easier for the EC group than the EA group. Item 4 was significantly more difficult for EC group than EA group at the extreme scores. Figure 2 indicates that even though most of the items functioned differently between the two settings, as a scale the 10 items were functioning similarly across settings because small differences in opposite directions cancelled each other out. In both samples, the 10 items produced nearly identical observed scores for individuals with the same amount of mania after controlling for group mean differences.

Aim 2: Differential Item Functioning of the 10 items embedded and extracted at a Community Mental Health Center

For the context effect analyses, EC was the reference group and Extracted was the focal group. Table 4 displays the item parameters and g^2 goodness of fit index for the 10 items. After controlling for the false discovery rate, only item 7 showed significant DIF. Figure 3 displays that at all levels Item 7 produced lower scores for the Extracted items compared to EC items. Higher responses for querying about "mood and energy always at the extremes" (Item 7) required less mania when it was extracted. Figure 4 indicates that even though Item 7 showed some DIF, the 10 items together were functioning similarly. Therefore, context effects did not appear to affect responses on the ten items.

Aim 3: Diagnostic Efficiency of the extracted 10 items

Receiver Operating Characteristic (ROC) curves examined diagnostic efficiency by comparing the sensitivity and false alarm rate (1-specificity) for each score (Altman & Bland, 1994). An area under the curve (AUROC) of .50 would indicate chance performance. The PGBI-10M was compared to the 10 items embedded in the parent report-GBI. Figure 5 displays the ROC curves for the PGBI-10M compared to the 10 items on the parent reported GBI. The PGBI-10 significantly predicted PBD, AUROC = .79, 95% C.I. = .69- .90. The 10 items embedded in the parent report GBI significantly predict PBD, AUROC = .80, 95% CI = .71 - .89. The two AUROC curves were compared using compared using $z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}}$, which controls for the related covariance between the two dependent measures (Hanley & McNeil, 1983). The two curves were not significantly different, $z = .06$, $p = .95$.

Clinically, the ability to determine the probability of a positive or negative diagnosis for individual cases is important. The Quality Reciever Operating Characeristic (Q-ROC) curve indicates which scores on a test provide the most optimally specific score, the most optimally sensitive score, and the most optimally balanced score (Kraemer, 1992). The Q-ROC is a plot of calibrated sensitivity on the y-axis and calibrated specificity on the x-axis. Calibrated sensitivity is: $\frac{Sensitivity\ -Level\,(q)}{1-Level\,(q)}$. Calibrated specificity is: $\frac{Specificity\ -(1-Level\,(q))}{Level\,(q)}$. Sensitivity is the proportion of individuals with PBD that are correctly identified as having PBD. Specificity is the proportion of individuals without PBD correctly identified as not having PBD. Level (q) is the proportion of individuals with a positive test regardless of actual diagnosis. Q-ROC rewards correct identification of true positives and true negatives, while penalizing for false positives and false negatives.

Figure 6 displays the Q-ROC for the PGBI-10M. Visually, the optimally sensitive cut score would have the highest elevation on the y-axis. The Q-ROC curve indicated that the optimal cut score for maximizing sensitivity is 5. For a score of 5 or higher, the positive diagnostic likelihood ratio was 1.3, while the negative diagnostic likelihood ratio was less than 0.1. Thus, a score of 4 or lower would substantially decrease the probability of a PBD diagnosis, but a score of 5 or higher would not substantially increase the probability of a PBD diagnosis. The most optimally specific score will be located farthest to the right. The Q-ROC curve indicated that the optimal cut score for maximizing specificity is 17. A score of 18 or higher, the positive diagnostic likelihood ratio was 3.4, while the negative diagnostic likelihood ratio was .5. The score that optimally balances both sensitivity and specificity will be in the top right. Q-ROC does not clearly indicate a single, optimally balanced cut score. Therefore, the scores were broken into six subgroups: Very Low (0.0 - .9), Low (1.0 - 4.9), Low to Neutral (5.0 - 9.9), Neutral (10.0 - 14.9), High (15.0 - 17.9), Very High (18.0 - 30.0) (same cut scores as Youngstrom, et al., 2008). The diagnostic likelihood ratios were: .1 (Very Low), .1 (Low), .1 (Low to Neutral), .1 (Neutral), .5 (High), and 2.9 (Very High). Scores less than 18 substantially decrease the probability of a PBD diagnosis, while scores greater than 18 increased the probability of PBD.

IV. Discussion

The first specific aim of this project was to examine the portability of the ten best discriminating items between an academic medical center and community mental health center. The data indicated that the 10 items as a test function similarly across samples and context. When moving from a sample where individuals have a higher income, are primarily Caucasian, and probands often have been selected for mood disorder to a sample with lower income, primarily African-American, and lower rates of mood disorde, the 10 items continue to function similarly as a test. At the item level, querying rapid mood/energy shifts and elated mood with sleep disturbance was mildly less discriminating in the community mental health sample. Questioning elated mood was slightly more discriminating in the community mental health sample. Parents were more likely to endorse "mood and energy at the extremes" in the community mental health sample than at the academic medical center, while they were less likely to endorse "elated mood with hyperactivity and high energy" at the community mental health center. Visual examination of the effects suggested that differences on both the discrimination and difficulty parameters were small. Additionally, the item level differences balanced themselves across the scale. After controlling for mean differences, the total observed score represented equivalent levels of mania between the two samples even though individual items showed differences across the two samples. This result corresponds to findings that suggest the parent-reported GBI functions similarly in a diverse sample as it does in a more selected sample (Youngstrom, Meyers, et al., 2005). Therefore, the 10 best discriminating items of the parent report GBI show similar functioning across samples.

The second specific aim was to examine whether context effects occur from extracting the ten items from the full parent-reported GBI. The findings indicated that context did not have a strong effect on parental responses to the ten items. Nine of the ten items showed no differences in their relationship to mania or to the amount of mania required to endorse any particular response when they were administered by themselves or within the context of the full length parent report GBI. The one exception was the item querying extreme mood and energy. On the extracted, free-standing ten items on the PGBI-10M, parents were slightly more likely to endorse higher response categories at similar levels of mania. These results appear consistent with the suggestion by Steinberg (2001) that precise items are less likely to be affected by context. Item response is most likely due to respondents pooling prior memories, evaluating the consistency of those memories, and evaluating the similarities amongst the memories (Tourangeau, et al., 2000). Vague items are more likely to pull for memories that are not consistent or similar. The precision of the GBI items is most likely negating the role of context.

The third specific aim was to examine the diagnostic efficiency of the PGBI-10M. The results indicate that the PGBI-10M predicted PBD similar to the 10 items embedded within the parent report GBI. On one hand, low scores (≤18) on the PGBI-10 substantially decrease the probability of diagnosis (DLRs < .5); on the other hand, high scores (≥18) indicate a moderate increase in the probability of a PBD diagnosis (DLR = 2.9). The diagnostic likelihood ratios are substantially weaker than those found on longer parent report questionnaires of PBD (e.g., Youngstrom, et al., 2004). Compared to the diagnostic likelihood ratios from the Embedded Academic, the PGBI-10M continues to rule out a diagnosis of PBD very well when compared to the original diagnostic likelihood ratios from the Embedded Academic (Youngstrom, et al., 2008). However, the PGBI-10M does a less well job of increasing the probability of a PBD diagnosis for high scorers than the original Embedded Academic positive diagnostic likelihood ratios suggest. Therefore, clinicians should use scores of 18 or less to help rule out a diagnosis of PBD, whereas higher scores should prompt a systematic screening of mania symptoms (Youngstrom, Freeman, & Jenkins, 2009).

The primary strength of this study is the large, multi-site, diverse sample of youth with reports of mania symptoms. In testing the utility of the PGBI-10M, the test performed well across sites, suggesting that it is portable and resistant to context effects. Additionally, the current study reflects one of the first attempts to study item level functioning in youth with PBD.

The diverse sample is also its primary weakness. Due to the differences in both socio-economic status and race between the academic medical center and the community mental health center, the item level differences cannot be attributed with certainty to any single factor. Although the effect sizes are small, the sample differences prevent a conclusion about whether certain items (e.g., elated mood) are better predictors for Caucasians or African-Americans or for lower versus higher socio-economic status. However, item response theory allows for group differences in mean scores, because it focuses the analyses on the relationship between the item and the latent trait (Thissen, Steinberg, & Gerrard, 1986). Clinically, the most significant limitation is the relatively small sample size used to compute the diagnostic likelihood ratios. The small sample size results in relatively unstable estimates of prediction. This concern is mitigated somewhat by the similarity between the DLR estimates generated here as compared to estimates from independently published samples (Youngstrom, et al., 2008).

Future studies should examine whether the item level differences are due to differences in race and ethnicity or due to differences in socio-economic status. Knowing these differences and whether they have large effect sizes could aid clinicians in determining lines of questioning and the weight to place on different symptoms dependent upon demographic information. Additionally, examining the predictive validity of the PGBI-10M in larger samples will aid in the accuracy and precision of clinically useful cut scores.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 3: receiver operating characteristic plots. *British Medical Journal, 309*(6948), 188.

Angst, F., Stassen, H. H., Clayton, P. J., & Angst, J. (2002). Mortality of patients with mood disorders: follow-up over 34-38 years. *Journal of affective disorders, 68*(2-3), 167-181.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *J Pers Assess, 67*(3), 588-597.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57*(6), 717-726.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Berkson, J. (1946). Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin, 2*(3), 47-53.

Biederman, J., Klein, R. G., Pine, D. S., & Klein, D. F. (1998). Resolved: mania is mistaken for ADHD in prepubertal children. *J Am Acad Child Adolesc Psychiatry, 37*(10), 1091-1096; discussion 1096-1099.

Birmaher, B., Axelson, D., Strober, M., Gill, M. K., Valeri, S., Chiappetta, L., et al. (2006). Clinical course of children and adolescents with bipolar spectrum disorders. *Arch Gen Psychiatry, 63*(2), 175-183. doi: 63/2/175 [pii]
10.1001/archpsyc.63.2.175

Blader, J. C., & Carlson, G. A. (2007). Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996-2004. *Biol Psychiatry, 62*(2), 107-114. doi: S0006-3223(06)01446-6 [pii]
10.1016/j.biopsych.2006.11.006

Bowring, M. A., & Kovacs, M. (1992). Difficulties in diagnosing manic disorders among children and adolescents. *J Am Acad Child Adolesc Psychiatry, 31*(4), 611-614.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2004). CEFA: Comprehensive Exploratory Factor Analsysis, Version 2.00 [Computer Software and Manual]. Columbus. Retrieved from http://quantrm2.psy.ohio-state.edu/browne

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research, 1*(2), 245 - 276.

Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment, 76*(1), 18-47.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *British Medical Journal, 329*(7458), 168-169.

Dell'Osso, L., Pini, S., Cassano, G. B., Mastrocinque, C., Seckinger, R. A., Saettoni, M., et al. (2002). Insight into illness in patients with mania, mixed mania, bipolar depression and major depression with psychotic features. *Bipolar Disord, 4*(5), 315-322.

Depue, R. A. (1981). A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: A conceptual framework and five validation studies. *Journal of Abnormal Psychology, 90*(5), 381-437.

Desai, S., & Braitman, K. A. (2005). The Effects of Scale Carving on Instruments Assessing Violence. *Journal of Family Violence, 20*(2), 101-107.

DeVellis, R. F. (2003). *Scale development : theory and applications*: Thousand Oaks, Calif. : SAGE Publications, c2003.

Findling, R. L., Gracious, B. L., McNamara, N. K., Youngstrom, E. A., Demeter, C. A., Branicky, L. A., et al. (2001). Rapid, continuous cycling and psychiatric co-morbidity in pediatric bipolar I disorder. *Bipolar Disorders, 3*(4), 202-210.

Geller, B., Craney, J. L., Bolhofner, K., Nickelsburg, M. J., Williams, M., & Zimerman, B. (2002). Two-year prospective follow-up of children with a prepubertal and early adolescent bipolar disorder phenotype. *The American journal of psychiatry, 159*(6), 927-933.

Geller, B., Williams, M., Zimerman, B., Frazier, J., Beringer, L., & Warner, K. L. (1998). Prepubertal and early adolescent bipolarity differentiate from ADHD by manic symptoms, grandiose delusions, ultra-rapid or ultradian cycling. *J Affect Disord, 51*(2), 81-91. doi: S0165-0327(98)00175-X [pii]

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., DelBello, M. P., et al. (2001). Reliability of the Washington University in St. Louis Kiddie Schedule for Affective Disorders and Schizophrenia (WASH-U-KSADS) mania and rapid cycling sections. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(4), 450-455.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*(3), 377-393.

Goldstein, T. R., Birmaher, B., Axelson, D., Ryan, N. D., Strober, M. A., Gill, M. K., et al. (2005). History of suicide attempts in pediatric bipolar disorder: Factors associated with increased risk. *Bipolar Disorders, 7*(6), 525-535.

Gronlund, N. E. (1988). *How to construct achievement tests (4th ed.)*. Englewood Cliffs, NJ, US: Prentice-Hall, Inc.

Guyatt, G. H., & Rennie, D. (1993). Users' guides to the medical literature. *Journal of the American Medical Association, 270*(17), 2096-2097.

Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology, 59*(6), 1301-1307. doi: 10.1037/0022-3514.59.6.1301

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*(3), 839-843.

Hirschfeld, R. M. (2001). Bipolar spectrum disorder: improving its recognition and diagnosis. *The Journal of clinical psychiatry, 62 Suppl 14*, 5-9.

Hirschfeld, R. M., Holzer, C., Calabrese, J. R., Weissman, M., Reed, M., Davies, M., et al. (2003). Validity of the mood disorder questionnaire: a general population study. *The American journal of psychiatry, 160*(1), 178-180.

Hirschfeld, R. M., Williams, J. B., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E., Jr., et al. (2000). Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire. *The American journal of psychiatry, 157*(11), 1873-1875.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453. doi: 10.1037/1082-989x.3.4.424

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., et al. (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 and Adolescent Psychiatry, 36*(7), 980-988.

Kim, E. Y., & Miklowitz, D. J. (2002). Childhood mania, attention deficit hyperactivity disorder and conduct disorder: a critical review of diagnostic dilemmas. *Bipolar Disord, 4*(4), 215-225.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*(2), 312-320. doi: 10.1037/0022-3514.55.2.312

Kowatch, R. A., Youngstrom, E. A., Danielyan, A., & Findling, R. L. (2005). Review and meta-analysis of the phenomenology and clinical characteristics of mania in children and adolescents. *Bipolar Disorders, 7*(6), 483-496.

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

Leverich, G. S., Post, R. M., Keck, P. E., Jr., Altshuler, L. L., Frye, M. A., Kupka, R. W., et al. (2007). The poor prognosis of childhood-onset bipolar disorder. *The Journal of pediatrics, 150*(5), 485-490.

Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A., & Hirschfeld, R. M. (1994). The National Depressive and Manic-depressive Association (DMDA) survey of bipolar members. *Journal of affective disorders, 31*(4), 281-294.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*: Hillsdale, N.J. : L. Erlbaum Associates, 1980.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.).* (pp. 13-103). New York, NY England: Macmillan Publishing Co, Inc
American Council on Education.

Miller, C. J., Klugman, J., Berv, D. A., Rosenquist, K. J., & Ghaemi, S. N. (2004). Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. *Journal of affective disorders, 81*(2), 167-171.

Moreno, C., Laje, G., Blanco, C., Jiang, H., Schmidt, A. B., & Olfson, M. (2007). National trends in the outpatient diagnosis and treatment of bipolar disorder in youth. *Archives of general psychiatry, 64*(9), 1032-1039.

Muthen, L. K., & Muthen, B. O. (1998-2007). Mplus User's Guide. (Version 5.2). Los Angeles: Muthen & Muthen.

Neighbors, H., Jackson, J., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. *Community Mental Health Journal, 25*(4), 301-311.

Pavuluri, M. N., Henry, D. B., Devineni, B., Carbray, J. A., & Birmaher, B. (2006). Child mania rating scale: development, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*(5), 550-560.

Perlis, R. H., Miyahara, S., Marangell, L. B., Wisniewski, S. R., Ostacher, M., DelBello, M. P., et al. (2004). Long-term implications of early onset in bipolar disorder: data from the first 1000 participants in the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biol Psychiatry, 55*(9), 875-881.

Post, R. M., Susan, R., & Weiss, B. (1992). Sensitization, kindling, and carbamazepine: an update on their implications for the course of affective illness. *Pharmacopsychiatry, 25*(1), 41-43.

Quinn, C. A., & Fristad, M. A. (2004). Defining and identifying early onset bipolar spectrum disorder. *Current psychiatry reports, 6*(2), 101-107.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *Int J Methods Psychiatr Res, 18*(3), 169-184. doi: 10.1002/mpr.289

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4), 100-100.

Schuman, H., Presser, S., & Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly, 45*(2), 216-223. doi: 10.1086/268652

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102-111.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*(5), 399-411.

Stang, P. E., Frank, C., Kalsekar, A., Yood, M. U., Wells, K., & Burch, S. (2006). The clinical history and costs associated with delayed diagnosis of bipolar disorder. *MedGenMed, 8*(2), 18. doi: 528163 [pii]

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66*(2), 341-349.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*(2), 332-342.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*(4), 402-415. doi: 10.1037/1082-989x.11.4.402

Strauss, S. E., Richardson, W. S., Glasziou, P., Haynes, R. B., & Strauss, S. E. (2005). *Evidence Based Medicine (3rd Edition)* (3rd ed.). London: Churchill Livingstone.

Thissen, D. (2001). IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. *Unpublished ms*.

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report*. Research Memorandum, 92-2. University of North Carolina at Chapel Hill. Chapel Hill.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*(3), 385-395. doi: 10.1037/0033-2909.104.3.385

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118-128.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementaion of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67 - 113). Hillsdale, NJ: Erlbaum.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314. doi: 10.1037/0033-2909.103.3.299

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York, NY US: Cambridge University Press.

Van Meter, A., Moreira, A. L., & Youngstrom, E. A. (2009). *Meta-analysis of Epidemiological Studies of Pediatric Bipolar Disorder*. Paper presented at the

American Academy of Child and Adolescent Psychiatry Annual Convention, Honolulu, HI.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. doi: 10.1007/bf02293557

Wirth, R. J., & Edwards, M. C. (2007). Item Factor Analysis: Current Approaches and Future Directions. *Psychological Methods, 12*(1), 58-79.

Wittchen, H. U., Frohlich, C., Behrendt, S., Gunther, A., Rehm, J., Zimmermann, P., et al. (2007). Cannabis use and cannabis use disorders and their relationship to mental disorders: a 10-year prospective-longitudinal community study in adolescents. *Drug and alcohol dependence, 88 Suppl 1*, S60-70.

Wozniak, J., Biederman, J., Kiely, K., & Ablon, J. S. (1995). Mania-like symptoms suggestive of childhood-onset bipolar disorder in clinically referred children. *Journal of the American Academy of Child & Adolescent Psychiatry, 34*(7), 867-876.

Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Who Are the Comorbid Adolescents? Agreement Between Psychiatric Diagnosis, Youth, Parent, and Teacher Report. *Journal of Abnormal Child Psychology, 31*(3), 231-245.

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C. A., Bedoya, D. D., et al. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(7), 847-858.

Youngstrom, E. A., Findling, R. L., Danielson, C. K., & Calabrese, J. R. (2001). Discriminative validity of parent report of hypomanic and depressive symptoms on the General Behavior Inventory. *Psychological Assessment, 13*(2), 267-276.

Youngstrom, E. A., Frazier, T. W., Demeter, C. A., Calabrese, J. R., & Findling, R. L. (2008). Developing a 10-Item Mania Scale From the Parent General Behavior Inventory for Children and Adolescents. *The Journal of clinical psychiatry*, e1-e9.

Youngstrom, E. A., Freeman, A. J., & Jenkins, M. M. (2009). The assessment of children and adolescents with bipolar disorder. *Child and adolescent psychiatric clinics of North America, 18*(2), 353-390, viii-ix.

Youngstrom, E. A., & Green, K. W. (2003). Reliability generalization of self-report of emotions when using the Differential Emotions Scale. *Educational and Psychological Measurement, 63*(2), 279-295.

Youngstrom, E. A., Meyers, O., Demeter, C. A., Youngstrom, J., Morello, L., Piiparinen, R., et al. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7*(6), 507-517.

Youngstrom, E. A., Meyers, O., 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*(9), 1013-1019.

Youngstrom, E. A., Youngstrom, J. K., & Starr, M. (2005). Bipolar diagnoses in community mental health: Achenbach child behavior checklist profiles and patterns of comorbidity. *Biological Psychiatry, 58*(7), 569-575.

Table 1. Demographic characteristics of the Embedded Academic, Embedded Community, and Both groups.

	Embedded Academic (n=)	Embedded Community (n=)	Both (n=)
Gender	60.8% Male	58.4% Male	65.4% Male
	39.2% Female	41.6% Female	34.6% Female
Ethinicity	13.4% African-American	83.3% African-American	91.2% African-American
	78.8% Caucasian	9.2% Caucasian	3.8% Caucasian
Age	11.45 (3.3)	10.8 (3.4)	10.0 (3.4)
Comorbidity	2.1 (1.3)	2.7 (1.4)	2.6 (1.2)
Primary Diagnosis	Bipolar 1: 23.2%	Bipolar 1: 3.4%	Bipolar 1: 1.3%
	Other Bipolar: 20.3%	Other Bipolar: 11.4%	Other Bipolar: 10.1%
	Unipolar: 22.5%	Unipolar: 31.0%	Unipolar: 21.4%
	Behavior: 23.0%	Behavior: 45.1%	Behavior: 57.2%
	Other: 11.1%	Other: 9.1%	Other: 9.4%

Table 2. Factor loadings for the Embedded Academic, Embedded Community, and Extracted 10 items.

Item	Embedded Academic	Embedded Community	Extracted Community
1	.82	.68	.73
2	.80	.65	.76
3	.74	.76	.79
4	.87	.76	.84
5	.79	.73	.84
6	.78	.76	.78
7	.74	.72	.75
8	.86	.77	.81
9	.86	.76	.81
10	.67	.66	.71

Note: The first three eigenvalues for the Embedded Academic are: 6.67, .78, and .52. The first three eigenvalues for the Embedded Community are: 5.73, .89, and .59. The first three eigenvalues for the Extracted Community are: 6.50, .81, and .52.

Table 3. Discrimination and difficulty parameter estimates from Differential Item Functioning Results comparing Embedded Academic to Embedded Community.

Item	Content	Group	α	b_1	b_2	b_3	α DIF (p)	*b* DIF (p)
No DIF								
5	Happy+Energy+Rage	EA	1.93	.17	1.11	1.79	.0 (1.00)	8.6 (.04)
		EC	1.92	.32	1.20	2.10		
9	Happy+Energy+Anger	EA	2.12	-.20	.99	1.77	2.5 (.11)	2.8 (.42)
		EC	2.49	-.11	.89	1.74		
10	Racing Thoughts	EA	1.64	.48	1.78	2.69	3.80 (.05)	7.4 (.06)
		EC	1.28	.33	1.68	2.85		
More discriminating in the Academic Sample than ACI Sample								
3	Rapid Mood/Energy Shift	EA	2.17	-.53	.57	1.40	7.5 (.01)*	7.5 (.06)
		EC	1.63	-.75	.39	1.42		
6	Happiness/Energy + Sleep Disturbance	EA	2.17	.19	1.21	1.85	3.9 (.05)*	11.3 (.01)
		EC	1.74	.07	1.01	1.88		
Less discriminating in the Academic Sample than ACI Sample								
2	Happy	EA	1.58	.65	1.75	2.76	4.2 (.04)*	7.9 (.05)
		EC	2.02	.69	1.85	2.72		
More difficult at Academic Sample than ACI Sample								
7	Mood+Energy at Extremes	EA	1.87	-.48	.72	1.55	1.6 (.21)	39.2 (<.01)*
		EC	1.64	-.80	.21	1.24		
More difficult at ACI Sample than Academic Sample								
1	Happy+Energy+Hyperactivity	EA	1.62	-.20	.69	1.36	5.0 (.03)*	27.9 (<.01)
		EC	2.06	-.02	.81	1.73		
Item more difficult at ACI Sample at average levels, but more difficult at higher levels for Academic Sample								
8	Mood Switching across days	EA	2.25	.03	1.26	1.94	2.3 (.13)	15.8 (<.01)*
		EC	2.65	.27	1.11	1.79		
Item more difficult at average and extremely high levels at ACI Sample, but more difficult at high levels at Academic Sample								
4	Happy+Energy	EA	2.13	.30	1.37	2.03	3.2 (.07)	11.5 (.01)*
		EC	2.61	.44	1.28	2.17		

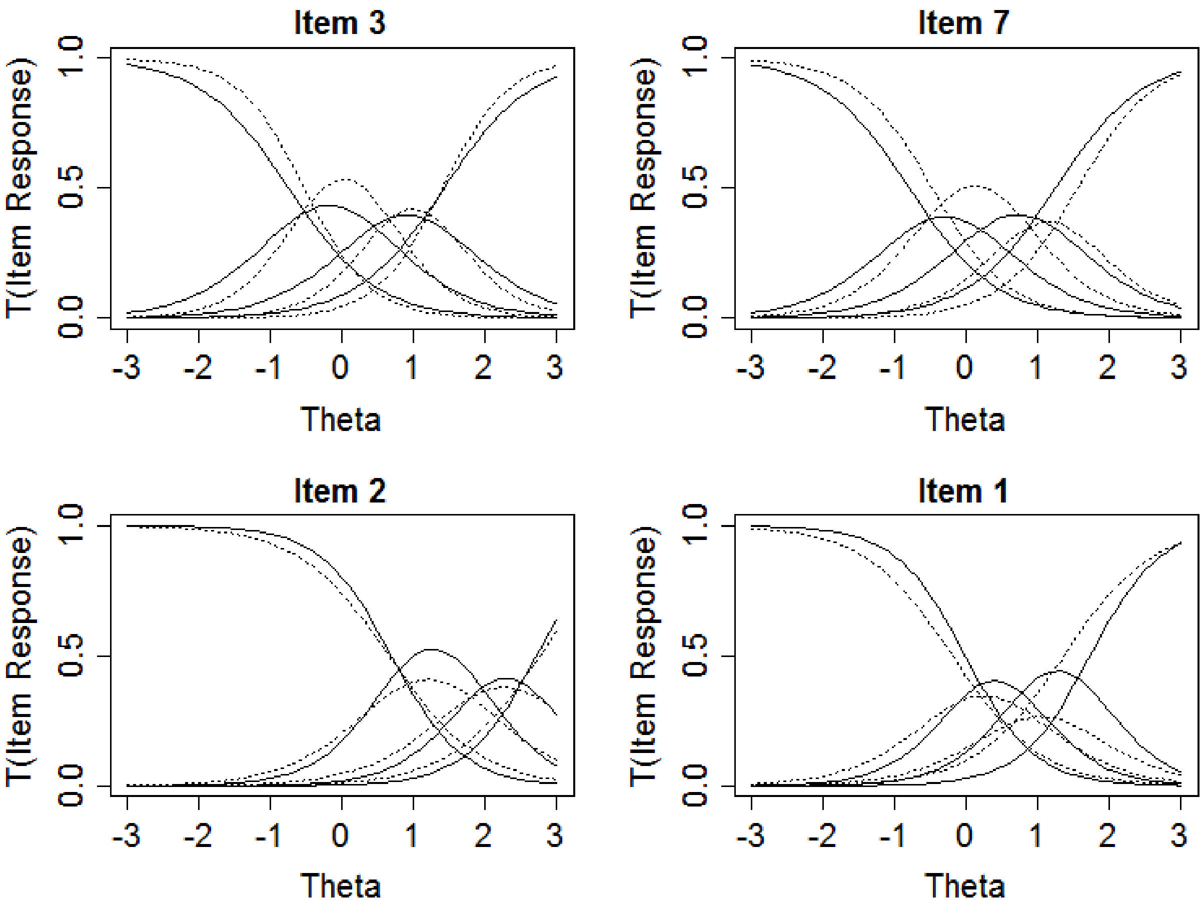
Note: *Indicates significantly different after Benjamini-Hochberg Correction.

Table 4. Differential Item Functioning Results comparing Embedded Community to the Extracted Community.

Item	Content	Group	α	b_1	b_2	b_3	α DIF (p)	*b* DIF (p)
No DIF								
1	Happy+Energy+Hyperactivity	EC	1.58	-.20	-.78	1.78	.0 (1.00)	13.3 (<.01)
		Extracted	1.59	-.20	.71	1.38		
2	Happy	EC	1.80	.44	1.54	2.54	1.8 (.18)	4.9 (.18)
		Extracted	1.53	.66	1.8	2.84		
3	Rapid Mood/Energy Shift	EC	2.09	-.49	.71	1.56	.1 (.75)	5 (.17)
		Extracted	2.15	-.56	.55	1.40		
4	Happy+Energy	EC	2.42	.35	1.23	2.21	2 (.16)	9.2 (.03)
		Extracted	2.07	.30	1.4	2.07		
5	Happy+Energy+Rage	EC	2.45	.00	.95	1.72	5.2 (.02)	9.4 (.02)
		Extracted	1.91	.18	1.14	1.82		
6	Happiness/Energy + Sleep Disturbance	EC	1.89	.23	1.16	2.04	1.1 (.29)	4.2 (.24)
		Extracted	2.12	.18	1.22	1.88		
8	Mood Switching across Days	EC	2.21	-.02	1.11	2.03	.0 (1.00)	5.5 (.13)
		Extracted	2.24	.03	1.27	1.95		
9	Happy+Energy+Anger	EC	2.30	-.27	.85	1.78	.5 (.48)	4 (.26)
		Extracted	2.14	-.20	.99	1.77		
10	Racing Thoughts	EC	1.57	.26	1.45	2.45	.0 (1.00)	12.3 (<.01)
		Extracted	1.58	.49	1.82	2.76		
More Mania to endorse higher responses if Item is Embedded								
7	Mood+Energy at Extremes	EC	1.86	-.29	1.04	2.06	.0 (1.00)	29 (<.01)*
		Extracted	1.88	-.52	.68	1.51		

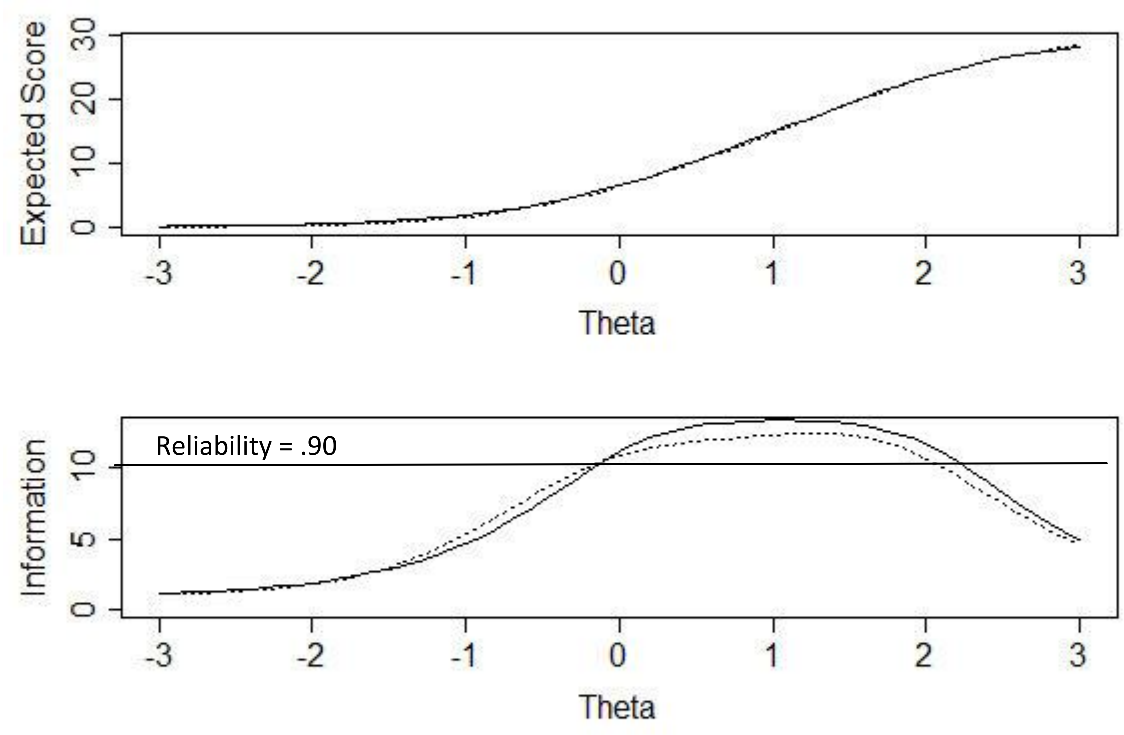
Note: *Indicates significantly different after Benjamini-Hochberg Correction.

Figure 1. Boundary Response Functions for selected items showing DIF between the Embedded Academic and Embedded Community.



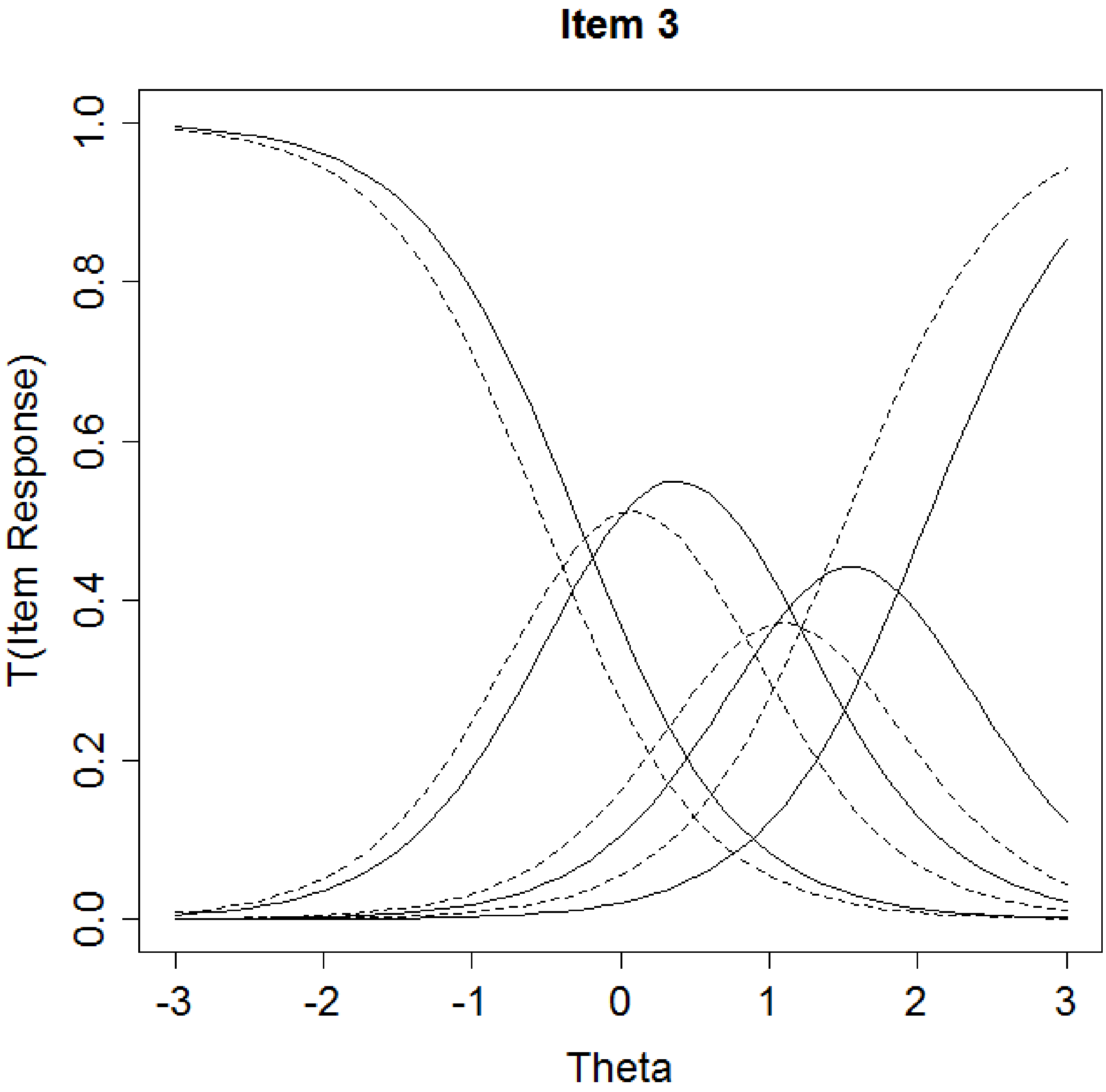
Note: Solid line is Embedded Community. Dotted line is Embedded Academic. Item 3 is more discriminating in EA than EC. Item 7 is more difficult in EA than EC. Item 2 is less discriminating in EA than EC. Item 1 is less difficult in EA than EC.

Figure 2. Test Characteristic and Test Information Curves comparing the ten items of the Embedded Academic to the same ten items in the Embedded Community.



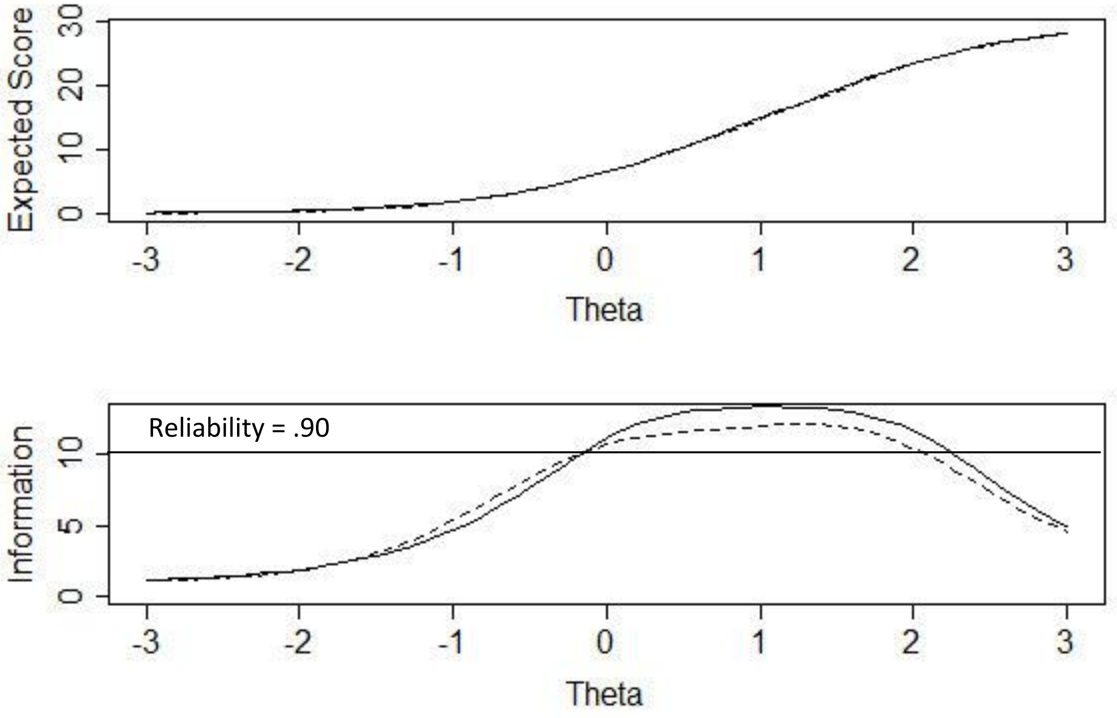
Note: Dotted line is Embedded Academic. Solid line is Embedded Community.

Figure 3. Boundary Response Function for Item 7 showing lower difficulty for the Embedded Community compared to the Extracted.



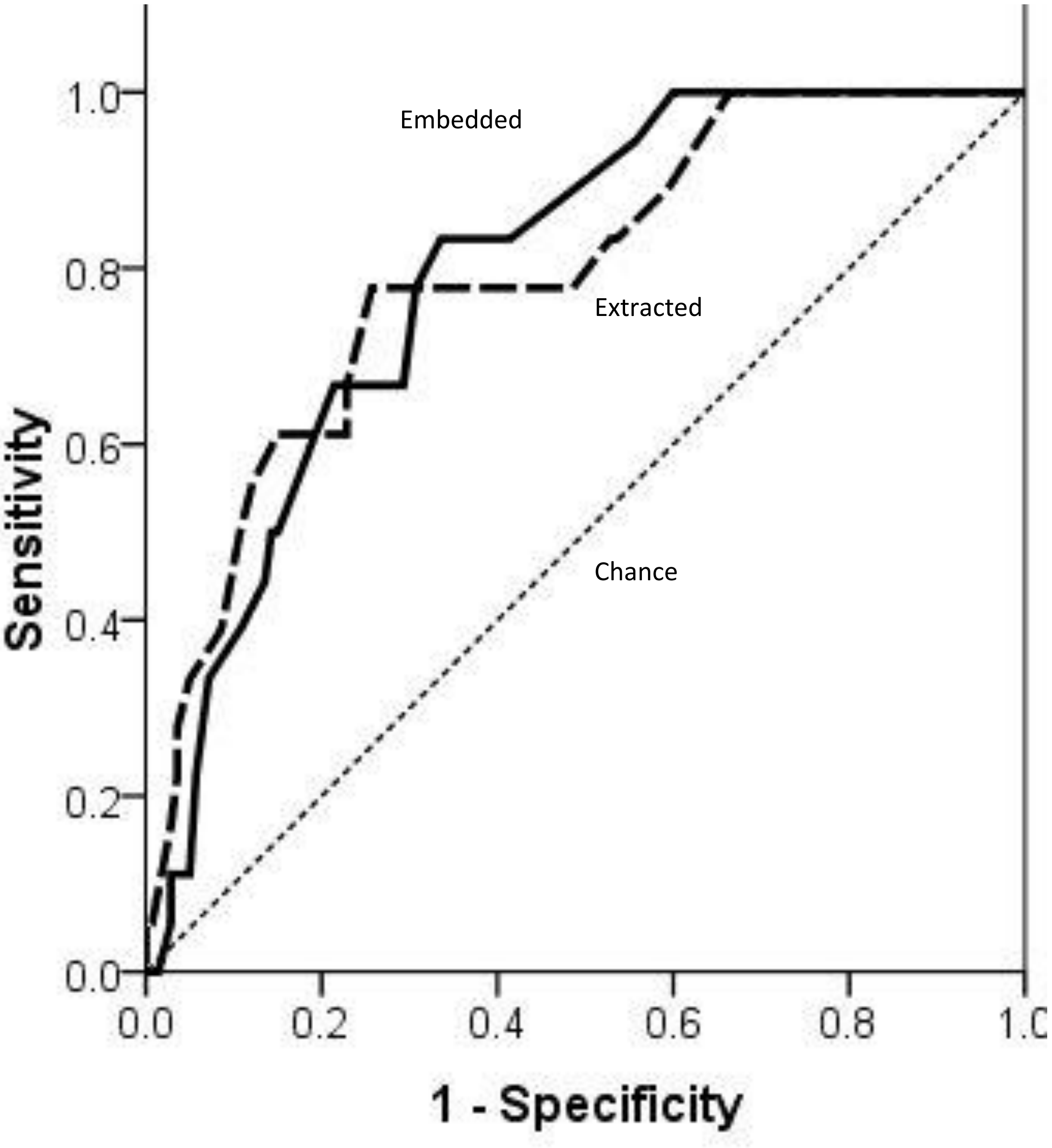
Note: Solid line is Embedded Community. Dashed line is Extracted.

Figure 4. Test Characteristic and Test Information Curves comparing the ten items of the Embedded Community to the same ten items in the Extracted.



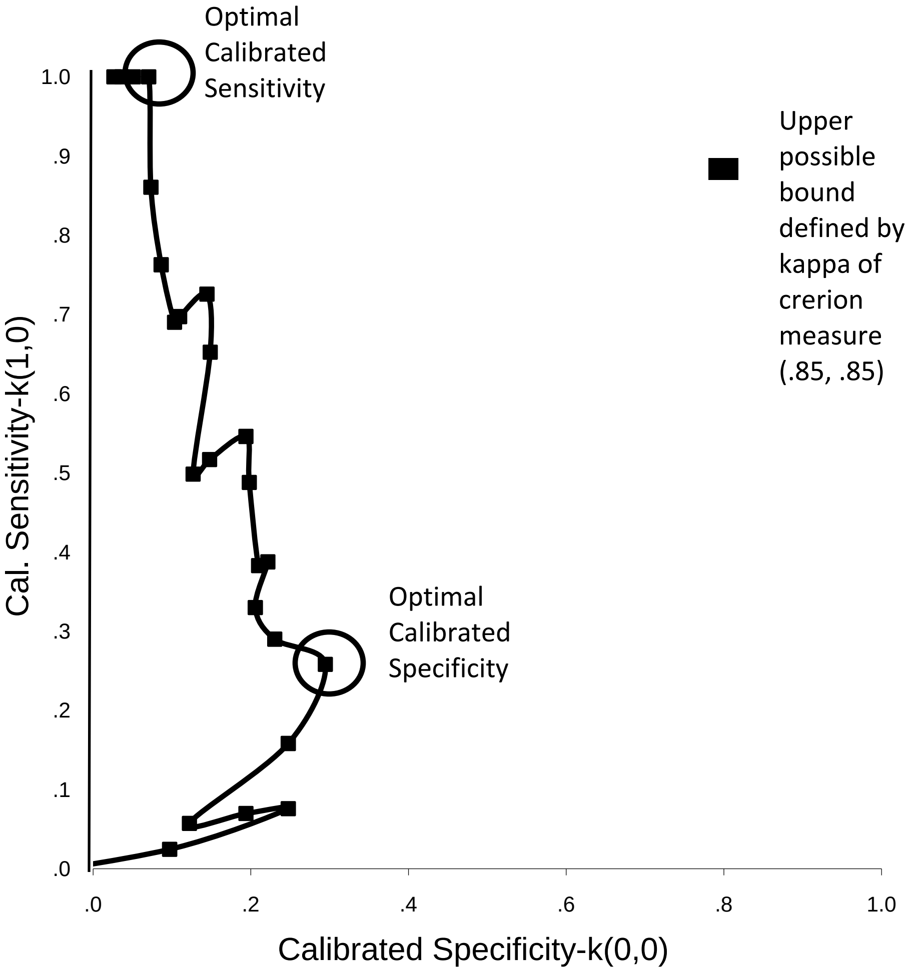
Note: Solid line is Embedded Community. Dashed line is Extracted.

Figure 5. Receiver Operating Characteristics comparing PGBI-10M to the 10 items embedded within the parent reported GBI at discrimination bipolar disorder from all other diagnoses.



Note: Chance is equal to an Area under the curve of .50.

Figure 6. Quality Receiver Operating Characteristic Curve of the PGBI-10M


Optimal
Calibrated
Sensitivity
Upper
possible
bound
defined by
kappa of
crerion
measure
(.85, .85)
Optimal
Calibrated
Specificity
Cal. Sensitivity-k(1,0)
Calibrated Specificity-k(0,0)