Running Head: SHORT FORMS FOR SELF-REPORT GBI

Evaluating and Validating Short Forms of the General Behavior Inventory Mania and Depression Scales for Self-Report of Adolescent and Young Adult Mood Symptoms

Abstract

Objective: To evaluate short forms of free self-report mania and depression scales, evaluating their reliability, content coverage, criterion validity, and diagnostic accuracy. **Method:** Youths age 11 to 18 years seeking outpatient mental health services at either an Academic medical clinic (*N*=427) or urban Community mental health center (*N*=313), completed the General Behavior Inventory (GBI) and other rating scales. Youths and caregivers completed semi-structured interviews to establish diagnoses and mood symptom severity, with GBI scores masked during diagnosis. Ten- and seven-item short forms, psychometric projections, and observed performance were tested first in the Academic sample and then externally cross-validated in the Community sample. **Results:** All short forms maintained high reliability (all alphas >.80 across both samples), high correlations with the full length scales (*r* .85 to .96), excellent convergent and discriminant validity with mood, behavior, and demographic criteria, and diagnostic accuracy undiminished compared to using the full length scales. Ten-item scales showed advantages in terms of coverage; the 7 Up showed slightly weaker performance. **Conclusions:** Present analyses evaluated and externally cross-validated short forms that maintain high reliability and content coverage, and show strong criterion validity and diagnostic accuracy – even when used in an independent sample with very different demographics and referral patterns. The short forms appear useful in clinical applications including initial evaluation, as well as in research settings where they offer an inexpensive quantitative score. Short forms are available in more than two dozen languages. Future work should further evaluate sensitivity to treatment effects and cultural invariance.

Keywords: Depression, mania, short form, screening, self-report, children and adolescents, sensitivity and specificity

Evaluating and Validating Short Forms of the General Behavior Inventory Mania and Depression Scales for Self-Report of Adolescent and Young Adult Mood Symptoms

There is tremendous need for brief, reliable, widely validated measures of mood symptoms. Mood disorders are associated with high personal, social, and economic burden as well as high risk of suicide. They frequently go years or decades undiagnosed (Drancourt et al., 2013; Hirschfeld, Lewis, & Vornik, 2003; Marchand, Wirth, & Simon, 2006), and inter-rater reliability for diagnosing depression (Regier et al., 2012) and bipolar disorders (Jensen-Doss, Youngstrom, Youngstrom, Feeny, & Findling, 2014; Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009) is poor when not using semi-structured interviews, rating scales, or both. Clinical trials for mood disorders often use interviews as the basis for measuring symptom severity and outcomes (Cerimele, Goldberg, Miller, Gabrielson, & Fortney, 2019). These interviews have limitations in terms of requiring training to ensure consistent calibration and prevent drift across raters or sites (Mackin, Targum, Kalali, Rom, & Young, 2006), and they do not transport easily to clinical practice, where they may not be reimbursed by payers and may not fit into psychotherapy or medication check sessions. A validated, brief self-report rating of hypomanic and depressive symptoms could play key roles in screening, enrollment in trials, calibration across sites, quantifying symptom severity, and tracking progress and outcomes in larger systems of care (Guo et al., 2015; Youngstrom et al., 2013).

The General Behavior Inventory (GBI; Depue et al., 1981) offers a promising way of addressing these gaps. The GBI was originally developed in the late 1970s, and deliberately included a wide range of behaviors and associated features beyond the diagnostic symptoms of depression or mania, which was prescient given changes in subsequent DSM and ICD criteria for mood episodes. Depue and colleagues put the GBI through a remarkably extensive program of validation, including five studies published in the original monograph. GBI scores correlate meaningfully with family history of mood disorder (Klein, Depue, & Slater, 1985), mood symptoms and diagnoses in offspring of parents with bipolar (Klein, 1984), high levels of cortisol and dopamine metabolites (Depue, Luciana, Arbisi, Collins, & Leon, 1994; Depue, Kleiman, Davis, Hutchinson, & Krauss, 1985), concurrent prediction of diagnostic status in teens (Danielson, Youngstrom, Findling, & Calabrese, 2003) and adults (Pendergast et al., 2014)-- including prospective prediction of diagnostic evolution (Alloy, Abramson, Urosevic, Nusslock, & Jager-Hyman, 2010; Findling et al., 2013), tracking with fluctuations in mood states (Lovejoy & Steuerwald, 1995; Mallon, Klein, Bornstein, & Slater, 1986), and sensitivity to treatment response (Findling et al., 2007; Youngstrom et al., 2013). The GBI produced some of the largest effect sizes in recent meta-analyses of diagnostic accuracy for identification of bipolar spectrum disorders in youths (Youngstrom, Genzlinger, Egerton, & Van Meter, 2015) and adults (Youngstrom, Egerton, et al., 2018). The GBI is also free to use clinically and in research, eliminating a cost barrier to access (Beidas et al., 2015).

However, there are some pragmatic issues that have undermined the utility of the GBI. One of the biggest hurdles is length. The canonical version has 73 items scored on a four-point scale from 0 (*Never or hardly ever*) to 3 (*Very often; almost constantly*), along with 3 validity items that do not factor in the scale scoring. The number of items does not tell the full story, either. The items are complex, pairing different mood and energy levels to describe the characteristic present. These “mood swings”/mixed (“biphasic,” in Depue’s terms) items are challenging from a traditional test development perspective, as they are “double-barreled” and show large cross-loadings in factor analyses. At the same time, the complicated items are also those that best distinguish between diagnoses, so they have been retained as candidate items for short forms (Youngstrom et al., 2018). From the respondent’s perspective, the items have an 11th grade reading level, because of the length and compound structure. If it were possible to carve short forms from the GBI, they might substantially reduce the time and cognitive complexity for respondents. If they retained adequate reliability and validity, they could play a key role in filling the gap in adequate assessment of mood symptoms from intake to progress and outcome tracking.

There have been multiple forays into developing GBI short forms for research (e.g., Lewinsohn, Klein, & Seeley, 2000) and potential clinical use. Two that are particularly promising are the 7 Up-7 Down (Youngstrom, Murray, Johnson, & Findling, 2013) and 10 item versions of mania and depression scales (Youngstrom et al., 2018). The 7 item scale was built with a primary focus on factor structure, whereas the 10 item versions picked items that maximally discriminated between diagnostic groups while also loading on a single factor). Both of these have shown high internal consistency reliability and criterion validity, as well as univocal factor structure. Yet both also have key unanswered questions at present.

The 10-item forms (mania and two alternate 10-item depression forms) were built and validated using parent report about a youth's (age 5 to 18 years) mood and behavior (Youngstrom, Frazier, Findling, & Calabrese, 2008; Youngstrom et al., 2018). Parent report appears to be one of the most valid sources of information about youth hypomanic symptoms and behavior (see Youngstrom et al., 2015; cf. Youngstrom, Joseph, & Greene, 2008), but parent report also only correlates moderately with youth report on the GBI (Youngstrom, Meyers, Youngstrom, Calabrese, & Findling, 2006) – similar or a little better than typical cross informant agreement in general (De Los Reyes et al., 2015). More concerning is the evidence that parents and youths do not notice or endorse the same symptoms at similar levels of the latent trait (Freeman, Youngstrom, Freeman, Youngstrom, & Findling, 2011), and different factors may influence the credibility of parent versus youth report (Youngstrom et al., 2011). Thus, items that are highly discriminating when completed by the parent may be much less accurate used as a self-report.

In contrast, the 7 Up-7 Down was built using self-report data. The items selected wound up being quite different than the parent-derived forms, with only three mania items overlapping between 7 Up and 10M, and four between 7 Down and Depression 10 Form B (only one with Form A). 7 Up-7 Down internal consistency and criterion correlations again were excellent, but the research to date has not had criterion diagnoses available to perform receiver operating characteristic (ROC) analyses (Kraemer, 1992). Thus it is unknown how well the 7 Up-7 Down might perform as a screener or diagnostic aid; nor are diagnostic likelihood ratios available to support interpretation at the individual case level (Straus, Glasziou, Richardson, & Haynes, 2018). Content analysis of the items on the 7 Up-7 Down also indicates that the scales cover fewer facets of mood symptoms than the 10-item versions (see Figure 1).

Our goal was to test and externally cross-validate short forms of adolescent self-report. We report psychometric information for five GBI youth self-report short scales: two measuring hypomanic/manic symptoms (the 7 Up and the 10M), and three measuring depressive symptoms (the 7 Down and the 10D form A and form B). We compare the reliability to projections using the formulae recommended by Smith et al. (Smith, McCarthy, & Anderson, 2000), and follow the same blueprint for examining factor, criterion, and discriminative validity as done with the parent-report versions of the items (Youngstrom et al., 2018). We also use two independent samples (Youngstrom, Halverson, Youngstrom, Lindhiem, & Findling, 2018) that differ in terms of demographic and clinical characteristics to evaluate generalizability of the reliability, criterion and discriminative validity of these scales.

Method

Participants

The academic sample (*N*=427) consisted of families seeking outpatient services at an academic medical center. The community sample (*N*=313) consisted of families presenting to a large urban community MH center. All families sought outpatient mental health services for youth between the ages of 11 and 18 years and were conversant in English; 160 youth (38%) from the academic sample received a BP diagnosis, as did *n*=41 (13%) in the community sample. Families received modest compensation for the full day interview, supported by grants from the Stanley Medical Research Institute (redacted) and NIH R01MH(redacted for peer review).

Table 1 presents both samples' descriptive statistics and effects effect sizes for differences. The Academic sample had higher rates of mood disorders because of recruitment for offspring studies and clinical trials. The Community sample had more externalizing disorders and attention-deficit/hyperactivity disorder (ADHD) as well as much lower average SES.

Diagnoses

All diagnoses used Longitudinal Expert evaluating All Data (LEAD; Spitzer, 1983) consensus meetings to integrate findings from semi-structured interviews with family psychiatric history and prior treatment history (but not rating scales – to avoid criterion contamination). Highly-trained interviewers used versions of the Kiddie Schedule for Affective Disorders and Schizophrenia – Present and Lifetime version (KSADS-PL; Kaufman et al., 1997), augmented in 2001 with the mood items from the Washington University KSADS (Geller et al., 2001). Interviewers met sequentially with youths and caregivers, using clinical judgment and additional interviewing to reconcile disagreements. Inter-rater reliability was high (item level kappa >.85 across both sites).

Measures

General Behavior Inventory (GBI; Depue et al., 1981). The two main scales are the depressive, containing 46 items, and the hypomanic/biphasic scale, with 28 items--the recommended scoring includes one item on both. The hypomanic/biphasic scale contains both mixed items—asking about mood swings, and juxtaposing manic and depressive aspects—as well as more univocal hypomanic items.

Mood severity ratings. Interviewers also did the Young Mania Rating Scale (YMRS; Young, Biggs, Ziegler, & Meyer, 1978) and Children's Depression Rating Scale-Revised (CDRS-R; Poznanski, Freeman, & Mokros, 1985) as measures of manic and depressive symptom severity, again based on caregiver and youth response and direct observation during the interview. Symptoms needed to be attributable to mood disorders, and not some other psychiatric condition, to count in the score (cf. Yee et al., 2015).

Achenbach System of Empirically Based Assessment. Youths completed the Youth Self Report (YSR) and primary caregivers did the Child Behavior Checklist (CBCL) about their youth (Achenbach & Rescorla, 2001). Both have 118 problem items rated from 0 (*Not True (as far as you know)*) to 2 (*Very True or Often True*). The ASEBA provides multiple empirically derived scales. Present analyses use the Externalizing scale as the convergent criterion for hypomanic/manic and Internalizing scales for depressive scales (following Youngstrom et al., 2018).

Statistical Analyses

We followed the steps used in prior work with the parent report version of the GBI (Youngstrom et al., 2018). We again used the Academic sample to project performance of the 7- and 10-item short forms when tested in the Community sample. Descriptive statistics and effect sizes compared samples, which had markedly different demographic characteristics and referral patterns. Reliability, factor analyses and a series of correlations between short forms and various severity rating tested content and criterion validity. ROC analyses evaluated discriminative validity, with DeLong's test (DeLong, DeLong, & Clarke-Pearson, 1988) checking whether there was significant shrinkage in the diagnostic accuracy for the short form. We computed multilevel diagnostic likelihood ratios (DiLRs) to provide practical guidelines for using these new short scales as part of the initial evaluation process. DiLRs are the ratio of the percentage of cases with the target condition versus the percentage of comparison cases falling in a given score range (Straus et al., 2018).

As in the 2018 paper, we used the formulae from Smith et al. (2000) to project probable Cronbach's alpha, content coverage, attenuation of criterion correlations, and savings in length for the community sample, using estimates from the academic sample in the formulae. We used both rational content mapping (see Figure 1) and factor analysis to investigate the content coverage. Samejima's graded response item response theory (IRT) model quantified reliability, information (an extension of reliability, looking at the precision of score estimates across the range of the underlying trait), and item characteristics. We used the same set of criterion variables as in the 2018 paper, swapping the positions of youth report on the YSR for caregiver report on the CBCL to maintain mono-method comparisons, and used Steiger's test of dependent correlations to compare the criterion correlations for the short and full-length scales. We used the same a priori ranking of predicted criterion correlation sizes as with the 2018 paper, with the expectation that youth-youth correlations would be boosted by shared method variance compared to parent-youth correlations (Podsakoff, MacKenzie, & Podsakoff, 2012). In contrast, in the parent-rated GBI paper, it was parent-parent correlations that were highest due to the boost from shared method variance. Analyses used SPSS version 25, except for the IRT analyses, which used the R package *mirt* version 1.3.1. Please see supplemental materials for a detailed rational for the rankings of the criterion correlations (https://en.wikiversity.org/wiki/Evidence_based_assessment/Instruments/General_Behavior_Inventory#Rationale_for_the_Ranking_of_Expected_Criterion_Correlations), which also has links to archived syntax and sufficient data to reproduce the analyes.

Procedure

The university, hospital, and community mental health center IRBs reviewed and approved all protocols. After guardian consent and teen assent, an interviewer sequentially met with them, using clinical judgement to resolve any differences in opinion. Youths completed GBI while their caregiver was doing the KSADS interview. The diagnoses were made without access to the GBI, ASEBA, or other scales. We followed the STARD guidelines (Bossuyt et al., 2003) for study design and reporting of results.

Results

Missing Data Analyses

The main source of missing data was the ASEBA, which was added several years into one of the projects (so missing on the first set of cases); it was 91% complete for youth report, and 88% for the caregiver report. The completion rate for other variables ranged from 96% and up. Propensity scores (Guo & Fraser, 2010) used clinical and demographic variables to model probability of observation, and sensitivity analyses included the propensity scores to evaluate potential response bias.

Demographics and Descriptive Statistics

As reported previously, the Academic sample had higher SES and more families self-identifying as White; it also had higher rates of mood disorder due to targeted recruitment for active research protocols, and thus also higher average scores on mood rating scales. The Community sample showed higher rates of disruptive behavior disorders (ADHD, oppositional defiant disorder/ODD, conduct disorder/CD), as well as post-traumatic stress disorder (PTSD); thus ASEBA Externalizing scores also were significantly higher. The effect sizes for differences ranged from small to very large. In consequence, the Community sample offered a stringent test of whether the short forms would continue to work in a setting that differed on demographic and clinical characteristics (rather than using a random split of the Academic sample).

Factor Structure and Content Coverage of Short Forms

The 73 items of the full length GBI form two scales, Depressed and Hypomanic/Biphasic (Depue et al., 1981). Prior work (Danielson et al., 2003; Youngstrom, Findling, Danielson, &

Calabrese, 2001) grouped the 73 items into 20 parcels of 3 or 4 items each: eight parcels with Hypomanic/Biphasic items; and twelve parcels with Depression items. Figure 1 shows the coverage in terms of sampling from different parcels. The 10-M included items from 6 of 8 hypomanic/biphasic parcels; the 7 Up draws from only three. The 10Da draws from nine parcels, the 10Db from 7 parcels, and the 7 Down from three. The GBI only contains one item with suicidal ideation (#73); the 10Da and 10Db omit it to maintain parallel content; however, the 7 Down includes it.

Exploratory factor analyses evaluated the dimensionality of the short forms using scree plots, Velicer's minimum average partials (MAP) and parallel analysis using the 95th percentile of 1000 simulations). Models included 30 items (10M, 10Da, 10Db) together to see if they retained two factors, and similar analyses pooled the 14 items from the 7 Up-7 Down. Separate analyses tested the unidimensionality of the 10 and 7 item sets as a precursor to the IRT analyses. For the 30 item analyses, all heuristics indicated two factor solutions (except parallel analysis retaining only one in the community sample); two factors always was the indicated solution for the 7 Up-7 Down. When the 10 or 7 items were analyzed separately, one factor was always the best fitting solution. The items loaded as expected, with a PROMAX rotation providing simple structure with very highly correlated factors ($r > .7$). Items showed adequate to strong loadings on the hypothesized factor (smallest loading in the single factor solutions was .41; median loading = .65), and modest cross-loadings in the two factor models. The most ambiguous item (#53 in the full set) showed loadings of .34 and .36 in the community sample; it is one of the "mixed" or "biphasic" items in Depue's parlance. Supplemental tables include the eigenvalues and the factor loadings for all twelve analyses in both samples. Additionally, we used the community sample to fit confirmatory factor analyses testing the unidimensionality of the short forms. Fit was acceptable, with CFI values of .94 to .97, and RMSEAs between .05 to .08. These are reassuring, particularly taking an a priori model, fitting it to a new independent sample, with no model adjustments; and yet more reassuring given the medium to large differences in demographic and clinical characteristics. A copy of the syntax and data are archived to make full details available here: DOI 10.17605/OSF.IO/QAGYD.

Reliability and Precision of the Short Forms

We used the formula from Smith et al. to project the internal consistency alpha for the short forms. The observed alpha equaled or exceeded the projected value in all ten instances (five short forms, evaluated separately in the Academic and Community samples); see Table 2. Table 2 also included standard error of measurement (*SEM*), a standard error of the difference score (*SEd*) for two administrations of the same form, and critical values for 90% and 95% confidence that scores showed "reliable change" (Jacobson et al., 1991). IRT analyses showed that all of the 10 item short forms had reliability >.80 between theta levels one standard deviation below average to three standard deviations above the average trait level for each mood score. The 7 Up showed a narrower range with good reliability, from -.5 to 2.5, whereas the 7 Down fell in between. The supplemental materials include a table of the item theta and discrimination parameters for the graded response model, as well as calibration when all the short form items are included in the same analyses.

Retained Content Coverage of the Short Forms

Smith et al. (2001) also provided a formula for projecting the correlation between the short form and full-length version, which estimates the degree of content coverage, expressed as shared variance. Projected correlations ranged from *r*=.75 (7 Up) to .86 (either 10Da or 10Db). The observed correlations were all higher than projected, ranging from *r*=.85 (7 Up) to .94 (all three 10 item forms) in the Community sample, based on an embedded administration format. The 10 item versions deliberately selected items to provide good content coverage, pulling from different item parcels. The two parcels omitted from the 10M item pool were grandiosity (cf. Van Meter, Burke, Kowatch, Findling, & Youngstrom, 2016) and mood extremes; the concept of changes in mood and energy is baked into the content of items throughout the GBI (even on other parcels). Of the twelve depression parcels, only cognitive disturbance was not included on either 10 item short form. The 7 Up and 7 Down each drew from three parcels, contributing to their lower correlation with the full length scales (Table 2), although these still remained exceptionally high (*r*=.85 to .93).

Criterion Validity of the Short Forms: Convergent & Discriminant Correlations

Table 3 presents correlations for a range of criteria, sorted in descending order of expected criterion correlation based on convergent versus discriminant traits (e.g., mood scales expected to show higher correlations than Externalizing scales), shared method variance (e.g., two youth report scales expected to show higher correlation than youth with parent or interview ratings), as well as attenuation due to dichotomization (e.g., mood severity scores measured continuously should show higher correlations than categorical diagnoses based on the same interview). Steiger's test of dependent correlations compared the 10M and 7 Up correlations to the 28-item Hypomanic/Biphasic correlations. The criteria and ranking as the same as used in the prior study with the parent-reported items (Youngstrom et al., 2018), except that the YSR scales are added and expected to be highly correlated.

The convergent correlations were strong, and the discriminant correlations substantially lower, following the general pattern we hypothesized: The pattern of observed coefficients correlated with the predicted ranks .76 to .88 in the Academic and .71 to .76 in the community samples. In the Academic sample, the 10M produced three criterion correlations that were slightly but significantly smaller than the 28 item original, and one that was larger, with the largest different being a *q* of .06, a small effect size for the difference (Cohen, 1988). There were no significant differences between the 10 and 28 item correlations in the Community sample. In contrast, 10 of the correlations in the Academic and 9 in the Community sample were significantly different comparing the 7 and 28 item versions, with the 7 item scale always producing smaller correlations.

The largest difference was *r*=.58 vs .41, Cohen's *q*=.23, were Cohen suggested .1 as small and .3 as a medium-sized difference. The median *q* comparing 7 Up and 28-item correlations was .05, well in the small range.

For the depression scales, the convergent correlations were all highly significant, and very large within youth informant (.61 to .74), around .40 with the interview rated depression, .30 to .40 with diagnoses of mood disorder, and .15 to .23 with parent report (see Table 4). The pattern of coefficients correlated .82 to .84 in the Academic and .71 to .78 with expected ranks in the Community sample. The short forms showed significant differences from the criterion correlations for the 46-item scale in 2 to 7 cases, but there was no consistent direction (sometimes the short forms produced higher, sometime lower correlations) and only one discrepancy would have survived a Bonferroni correction. It was for the 7 Down in the Community sample, which correlated .61 with YSR Internalizing, versus a .69 for the 46-item with Internalizing, *q*=.14, again a small effect.

Overall, the results strongly support the convergent and discriminative validity of the short forms, with only mild differences from the correlations using the full-length version – with the exception of the 7 Up showing small but systematic shrinkage in the validity coefficients. The replication in the Community sample provides a conservative test, given the large differences in demographic and clinical factors.

Discriminative Validity of the Short Forms

Receiver operating characteristic (ROC) analyses evaluated the discriminative validity of the mania scales predicted any bipolar diagnosis, and the depression scales predicting the presence of any mood disorder. We compared the 10M and 7 Up to the performance of the 28-item scale and also the Externalizing scale on the Achenbach. Although Externalizing does not include several symptoms specific to mania, prior studies have established that it is sensitive to bipolar disorder

(Mick, Biederman, Pandina, & Faraone, 2003; Youngstrom et al., 2015). The inter-rater reliability of the KSADS consensus diagnoses set an upper range of AUC ~.925 instead of the theoretical limit of 1.00 (Kraemer, 1992). The 28, 10, and 7 item versions all delivered AUCs from .62 to .66 across both samples (Table 5), never significantly different from each other in head-to-head or cross-sample comparisons. These correspond to "fair" accuracy, with Cohen's *d* values of .43 to .58 (medium-sized). These results aligned with the meta-analytic estimate of performance for a youth-report scale using a clinically meaningful, not a distilled or healthy control comparison group (95% CI: .49 to .73; Youngstrom, Genzlinger, et al., 2015). Unlike prior work with parent-report versions, the accuracy of the GBI scales did not differ from the Achenbach Externalizing score (cf. Youngstrom et al., 2015; Youngstrom et al., 2018).

Similarly, all of the depression forms had AUC values ranging from .63 to .69, corresponding to moderate diagnostic accuracy given the clinically challenging comparison group. These translate to medium-sized *d* values. Again, there were no significant differences moving to the Community sample, indicating good generalizability. Nor were there any significant differences in accuracy between measures within sample (based on the correlated-measure DeLong tests). The Internalizing scale showed similar accuracy to the GBI scales in both samples.

Comparison to Parent Report. We previously evaluated the accuracy of parent report on the GBI short forms (Youngstrom et al., 2018), but roughly half of that sample was age 5 to 10 years – younger than the present sample. We therefore re-ran ROC analyses for the cases that had both the youth and parent report on the same scales. The AUCs were all +.10 to +.20 higher for when parents completed the same scales, and the parent performance was always significantly superior based on DeLong tests.

Diagnostic Likelihood Ratios

Diagnostic likelihood ratios (DiLRs) facilitate the interpretation of scores with individual cases. A DiLR compares what percentage of people who have a condition get scores in the reference range, versus the corresponding percentage of people who do not have the condition yet also score in the same range. Evidence-Based Medicine recommends using DiLRs because they make it easy to use Bayes' Theorem to update probabilities of diagnoses for individual cases: the updated odds are the product of the odds of a diagnosis multiplied by the DiLR (Straus et al., 2018). Free websites and apps will do the algebra.

A graphical alternative would be to use the probability nomogram (Figure 3), avoiding the need to do any computation. Examining the nomogram shows that DiLR values close to 1 do not change the probability. DiLRs smaller than 1 decrease the updated probability, and values greater than 1 increase the predicted probability. DiLRs of 2+ (or <.5) can be useful, especially when combined with other findings about the case. DiLRs > 5 (or <.2) are even more helpful, and greater than 10 (or <.1) are often clinically decisive, moving a 50% prior probability to >90% or <10%.

We followed the same steps as in prior work to created multi-level DiLRs, keeping more information from low and very high risk scores: (1) we used the Academic sample to define score quintiles, and then split the top quintile to have an extremely high decile to handle positive skew; (b) we then compared the fractions of cases with versus without the target diagnoses falling in each segment, (c) adjusting boundaries to avoid degenerate score distributions (e.g., where the DiLRs do not progress monotonically across segments due to sampling error). Table 6 has the DiLRs for the 10M and 7 Up to predict the probability of a bipolar spectrum disorder (bipolar I, II, cyclothymic disorder, or Other Specified Bipolar and Related Disorder), as well as the DiLRs for the 10Da, 10Db, and 7 Down depression scales predicting any mood disorder (including major depression, dysthymic/persistent depressive disorder, and Other Specified Depressive Disorder as well as bipolar disorders). For all five scales, low scores cut the odds of the mood diagnoses roughly in half. High scores roughly double the odds. For the depression scales, using the top decile to define a "very high" risk range more than quintuples the odds of a mood disorder diagnosis.

Sensitivity Analyses to Effects of Propensity Scores and Outliers

As noted above, rates of missing data were low, and differences between completers and missing cases tended to be small. Sensitivity analyses used propensity scores (estimated separately for the Academic and Community samples). In nine of ten analyses, the short form still predicted the target diagnosis significantly, with regression weights staying similar size or even increasing after covarying propensity scores. The exception was the 7 Up in the Community sample, which dropped to a trend when adjusting for propensity scores.

Lastly, regression analyses checked for multivariate outliers, with diagnoses (including comorbidities), age and sex as predictors of the short form scores. Cook's distance did not identify any influential outliers in either sample; Mahalanobis' distance identified four cases with an unusual constellation of predictor scores in the Community sample, and one case had a high Studentized deleted residual. Omitting these cases did not change any substantive results.

Discussion

Our goal was to rigorously evaluate 10-item short forms assessing manic and depressive symptoms, using the item sets derived based on previous work with parent-report on the General Behavior Inventory (Youngstrom et al., 2008; Youngstrom et al., 2018). The present paper's contributions are to replicate and extend prior work with parent-report on the GBI short forms to use with self-report, looking at reliability, content coverage, factor structure, criterion validity across a suite of 20 criterion variables that cover a range of convergent and discriminant criteria, and using three different source methods (youth report, parent report, and clinical interview), as well as testing the diagnostic accuracy for discriminating cases with mood disorders from all other comers to outpatient clinics. Extending the validation work to include self-report is vital in clinical contexts because in many settings, adolescents and young adults may self-refer for services. In those situations, self-report is the easiest data source for initial assessments. It also provides key information about client perceptions of severity for goal setting and measuring treatment response. Similar to prior work with parent report on the same scales (Youngstrom et al., 2018), we followed recommended practices for selecting items, projecting the changes in internal-consistency and content coverage, checking that we retained the factor structure, and then systematically testing for changes in criterion validity (Smith et al., 2000). We also capitalized on having two demographically and clinically different samples as a way of testing the generalizability of the short form performance (e.g., "geographic transportability;" Konig et al., 2007).

A secondary goal was to also evaluate the 7 Up-7 Down (Youngstrom et al., 2013), which was developed using partially overlapping data sets, starting with self-report as the source of information. Although the 7 Up-7 Down was built and validated across multiple international samples, none included the research-grade diagnoses necessary to do an unbiased ROC analysis and to estimate diagnostic likelihood ratios. The present paper fills this gap with regard to the diagnostic accuracy of a measure that has gained a fair degree of popularity for use with teens and adults. Rather than publish these analyses as a separate brief report, we took the opportunity to compare the 7-item and 10-item versions head-to-head, letting us offer guidance about whether one is clearly preferable. The core analyses thus looked at five candidate short forms: the 7 Up and a 10 item mania scale, benchmarked against the 28 item full length Hypomanic/Biphasic scale; and the 7 Down and two parallel 10 item Depression forms, A & B, compared to the 46 item scale.

All five scales performed as well or better than expected in terms of internal consistency, preserving the Depression-Hypomania distinction when used in tandem, and acting univocally (single factor) when analyzed separately. The content mapping showed that the 10 item scales sampled broadly from the mood symptoms and facets. In contrast, the 7 Up and 7 Down concentrated on three facets each. The narrowness contributed to lower correlations with the full-length scales, reflecting less breadth of coverage. Still, the part-whole correlations would be considered excellent and well above the projected values (Smith et al., 2000). Part of the robustness stems from the complexity of the GBI items: Most ask about changes in mood and energy, often juxtaposing symptoms. Because the items are inherently multifaceted, short forms that lean heavily on three parcels still possess a fair degree of breadth. The IRT analyses showed that reliability of estimates was high across the range of trait scores likely to be encountered in clinical settings, though they would not provide as accurate estimates at the low end of nonclinical ranges.

The short forms showed statistically valid and moderately helpful diagnostic accuracy, even when used in a community mental health clinic with medium- to large differences in demography and clinical concerns. The accuracy and effect sizes match projections based on prior meta-analyses (Youngstrom et al., 2015). The present analyses "pre-shrunk" estimates to minimize the degree of further attenuation likely to be encountered in new samples and settings (Youngstrom, Salcedo, Frazier, & Perez Algorta, 2019). These include using an independent replication sample, as well as making sure that the samples included high rates of comorbidity and challenging differential diagnoses--both of which can worsen the diagnostic specificity of the test. Even under these conditions, the scales continued to provide clinically useful information about diagnostic probabilities, as well as fairly precise estimates of mood symptom severity to use in treatment goal-setting and measuring outcomes.

Which Scale to Use?

The present paper compares five short forms versus two full length versions, as well as a widely used incumbent, the Achenbach System of Empirically Based Assessment (Achenbach & Rescorla, 2001), that offers broad band scales of related constructs: Externalizing versus hypomania/mania, and Internalizing versus depression symptoms. When completed by youths, all of the scales performed similarly, so all could claim validation (in both samples). Ties go to the short forms in most clinical and research applications, though. If one could obtain essentially identical results with a scale that is 64% to 85% shorter, the savings in time and participant burden are decisive. So the full length versions are the least appealing, and not recommended except for research applications where comprehensive item coverage would be justified.

The similar performance between the short forms and the Achenbach scales suggests that if a clinician already is using the YSR, the short forms would be redundant and not worth adding clinically for an initial evaluation. They might still have a role for measuring treatment response (Youngstrom et al., 2013), as the Achenbach is much longer (and would incur a financial cost with each administration). If the clinician or agency was not already using the Achenbach, then the GBI short forms offer a free alternative, but narrowly focused on mood disorders. These could be combined with other "best of the free" scales (Beidas et al., 2015) for other issues to create a more comprehensive initial evaluation. Using the "law of the vital few" Pareto Principle (Youngstrom, 2020; Youngstrom & Van Meter, 2016), it will often be feasible to combine six to eight brief scales as a core battery covering the bulk of the common clinical questions. The Assessment Center toolkit (https://www.hgaps.org/assessment-center.html) built by the Society for Clinical Child and Adolescent Psychology and Helping Give Away Psychological Science (https://hgaps.org) is a free prototype of this.

Which short form to use? For depression, any of the three short forms would be a good choice. The 10Da and 10Db have broader content coverage, and the parallel forms could be a virtue in outcome evaluation, where one could be used at baseline and the other for progress checks, or alternate them when the same client is filling them out repeatedly. The 10-item scales also cover the depression-specific "low positive affect" and anhedonia dimension from the tripartite model of depression and anxiety (Clark & Watson, 1991; Gaylord-Harden, Elmore, Campbell, & Wethington, 2011). Users could use Form A for recruitment purposes and Form B to measure depression severity at baseline (or vice versa). Clinicians also could switch between Forms A and B to measures change during treatment, lessening boredom and practice effects. Computer adaptive testing could of course use the full item set to construct other dynamic short forms. Of note, only the 7 Down includes a suicidal ideation item, and none of them includes an item assessing suicidal behavior or non-suicidal self-injury. If these are important to assess clinically, other scales will be needed (Millner & Nock, 2020). If asking about suicidal ideation changes the risk management or liability concerns, then the 10-items forms would be preferable to the 7 Down.

For manic symptoms, the 10-item version offers small but replicated advantages over the 7 Up that include higher reliability, more extensive content coverage, higher validity coefficients, and better discriminative validity. The differences are not huge, but nor is the cost of using a free scale that is three items longer. The evidence suggests that the 10-item version usually should be preferred.

Another consideration is choice of informant – should we have the youth or a parent or familiar caregiver fill out the scales? Although this paper concentrated on the youth self-report versions, prior work investigated parent-completed 10 item versions, and found their performance excellent (Youngstrom, Van Meter, et al., 2018). In the meta-analysis looking at mania scales, caregiver report showed significantly greater discriminative validity than youth or teacher report (Youngstrom et al., 2015), and caregiver report outperforms self-report in direct comparisons in every sample published to date (e.g., Youngstrom, Gracious, Danielson, Findling, & Calabrese, 2003; Youngstrom et al., 2005). The difference in discriminative validity of parent versus youth report for depression has not been frequently compared, but the parent advantage is likely to be smaller, if any. Both the participant-observer distinction and the modest typical parent-youth agreement about youth mood and behavior (De Los Reyes et al., 2015) provide theoretical and empirical context for expecting differences, and indeed, the item content of scales developed based on different informants showed very little overlap in item content. Given the evidence that parent-report may be more valid, one clinical option is to gather both, or to augment youth report with parent short forms, particularly when the assessment question includes potential bipolar disorder. On the other hand, in settings where parents are not easily accessible, such as foster care, forensic, and some hospital settings, the self-report options are valid and clinically useful (Youngstrom, Morton, & Murray, 2020). It is possible that the validity of self-report may increase with age or with repeated mood episodes, due to improve insight (when not hypomanic) or meta-cognitive ability: The accuracy of top-tier self-report in adults (AUC~.76; Youngstrom, Egerton, et al., 2018) is closer to the effect size for parent report in youths (AUC~.77) than youth self-report (AUC~.66) (Youngstrom et al., 2015).

Limitations and Future Directions

It would be ideal to have more information about the psychometrics of the 10 item depression scales when they are used in a standalone, extracted format. Technical psychometric information has been published for the 10M in an extracted versus embedded format (Freeman et al., 2012), finding negligible differences; and other studies have used the 7 Up-7 Down in an extracted format (albeit not with direct comparison to full length versions or with criterion diagnoses). The depression items have many of the same technical features as the hypomania items. The body of evidence suggests that the psychometrics are likely remain good. In addition, 10D Form A and 10M were used in an extracted format in multiple clinical trials, showing sensitivity to treatment effects (Findling et al., 2012; Youngstrom et al., 2013); whereas 10D Form B and the 7 Up and 7 Down have not yet been studied as outcome measures to our knowledge.

Future studies also should check for measurement invariance across different languages and societal groups. Prior work with other instruments has found that small to moderate cultural differences exist in the endorsement rates of specific items and scales (He, Burstein, Schmitz, &

Merikangas, 2013; Ivanova, Achenbach, et al., 2007; Rescorla et al., 2007a; 2007b; Warnick, Bracken, & Kasl, 2008), though the factor structure has tended to generalize across societies (Ivanova, Dobrean, et al., 2007). The GBI 10-item short forms have been translated into more than two dozen languages as part of measuring secondary outcomes for European studies, so there are high-quality, professional translations, but no published tests of differential item functioning or other aspects of performance as yet.

A major next step would be to examine multivariate models for combining information from different scales (e.g., depression and mania), different informants (e.g., youth and parent), and other risk factors and clinical findings (e.g., family history of mood disorder, age of onset) to examine incremental validity and to develop decision support algorithms and optimal sequences. We intend to do that as a next step with the existing data, and we would welcome efforts to replicate and extend in other data and with additional predictors.

Clinical Implications

Any of the short forms are much shorter, highly reliable in the ranges likely to be encountered in clinical applications, and deliver similar levels of diagnostic accuracy to the full-length version. Perhaps most importantly, all are free. The psychometric properties generalized from an Academic sample to a Community sample that had major demographic and referral pattern differences. The diagnostic likelihood ratios make it easy to combine GBI scores with other assessment findings to estimate personalized probabilities of mood disorders. The standard errors and critical change scores help with clinical applications to evaluate treatment response and change in mood symptom severity. Translations to other several languages are already available, including English, Spanish, Portuguese, and Chinese, as well as a score of other European languages) for the 10-item Mania and Depression Form A, opening up opportunities for detailed exploration of cross-cultural invariance and accelerating dissemination and implementation (https://trello.com/b/dYUKlNRP/translated-measures-dashboard).

References

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

Beidas, R. S., Stewart, R. E., Walsh, L., Lucas, S., Downey, M. M., Jackson, K., . . . Mandell, D. S. (2015). Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive & Behavioral Practice, 22*, 5-19. doi:10.1016/j.cbpra.2014.02.002

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal*, *326,* 41-44. doi:10.1136/bmj.326.7379.41

Cerimele, J. M., Goldberg, S. B., Miller, C. J., Gabrielson, S. W., & Fortney, J. C. (2019). Systematic review of symptom assessment measures for use in measurement-based care of bipolar disorders. *Psychiatric Services, 70*, 396-408. doi:10.1176/appi.ps.201800383

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*, 316-336.

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

Danielson, C. K., Youngstrom, E. A., Findling, R. L., & Calabrese, J. R. (2003). Discriminative validity of the General Behavior Inventory using youth report. *Journal of Abnormal Child Psychology, 31*, 29-39.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141*, 858-900. doi:10.1037/a0038498

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

Depue, R. A., Slater, J. F., Wolfstetter-Kausch, H., Klein, D. N., Goplerud, E., & Farr, D. 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*, 381-437. doi:10.1037/0021-843X.90.5.381

Drancourt, N., Etain, B., Lajnef, M., Henry, C., Raust, A., Cochet, B., . . . Bellivier, F. (2013). Duration of untreated bipolar disorder: missed opportunities on the long road to optimal treatment. *Acta Psychiatrica Scandinavica, 127*, 136-144. doi:10.1111/j.1600-0447.2012.01917.x

Findling, R. L., Youngstrom, E. A., Zhao, J., Marcus, R., Andersson, C., McQuade, R., & Mankoski, R. (2012). Respondent and item level patterns of response of aripiprazole in the acute treatment of pediatric bipolar I disorder. *Journal of Affective Disorders, 143*, 231-235. doi:10.1016/j.jad.2012.04.033

Freeman, A. J., Youngstrom, E. A., Frazier, T. W., Youngstrom, J. K., Demeter, C., & Findling, R. L. (2012). Portability of a screener for pediatric bipolar disorder to a diverse setting. *Psychological Assessment, 24*, 341-351. doi:10.1037/a0025617

Freeman, A. J., Youngstrom, E. A., Freeman, M. J., Youngstrom, J. K., & Findling, R. L. (2011). Is caregiver-adolescent disagreement due to differences in thresholds for reporting manic symptoms? *Journal of Child and Adolescent Psychopharmacology, 21*, 425-432. doi:10.1089/cap.2011.0033

Gaylord-Harden, N. K., Elmore, C. A., Campbell, C. L., & Wethington, A. (2011). An examination of the tripartite model of depressive and anxiety symptoms in African American youth: stressors and coping strategies as common and specific correlates. *Journal of Clinical Child* *and Adolescent Psychology, 40*, 360-374. doi:10.1080/15374416.2011.563467

Geller, B., Zimerman, B., Williams, M., Bolhofner, K., Craney, J. L., DelBello, M. P., & Soutullo, C. (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 & Adolescent Psychiatry, 40*, 450-455.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Los Angeles, CA: Sage.

Guo, T., Xiang, Y. T., Xiao, L., Hu, C. Q., Chiu, H. F., Ungvari, G. S., . . . Wang, G. (2015). Measurement-Based Care Versus Standard Care for Major Depression: A Randomized Controlled Trial With Blind Raters. *American Journal of Psychiatry, 172*, 1004-1013. doi:10.1176/appi.ajp.2015.14050652

He, J. P., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): the factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology, 41*, 583-595. doi:10.1007/s10802-012-9696-6

Hirschfeld, R. M., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry, 64*, 161-174.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bilenberg, N., . . . Verhulst, F. C. (2007). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology, 75*, 729-738. doi:10.1037/0022-006X.75.5.729

Ivanova, M. Y., Dobrean, A., Dopfner, M., Erol, N., Fombonne, E., Fonseca, A. C., . . . Chen, W. J. (2007). Testing the 8-syndrome structure of the child behavior checklist in 30 societies.

Journal of Clinical Child and Adolescent Psychology, 36, 405-417. doi:10.1080/15374410701444363

Jensen-Doss, A., Youngstrom, E. A., Youngstrom, J. K., Feeny, N. C., & Findling, R. L. (2014). Predictors and moderators of agreement between clinical and research diagnoses for children and adolescents. *Journal of Consulting & Clinical Psychology, 82*, 1151-1162. doi:10.1037/a0036657

Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., . . . Ryan, N. (1997). Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version (K-SADS-PL): Initial reliability and validity data. *Journal of the American Academy of Child & Adolescent Psychiatry, 36*, 980-988. doi:10.1097/00004583-199707000-00021

Klein, D. N. (1984). *Cyclothymia in the adolescent offspring of bipolar depressives: Validating a behavioral risk index against the genetic high risk paradigm.* (44), Univ. Microfilms International, US, (10-B)

Konig, I. R., Malley, J. D., Weimar, C., Diener, H. C., Ziegler, A., & German Stroke Study, C. (2007). Practical experiences on the necessity of external validation. *Statistics in Medicine, 26*, 5499-5511. doi:10.1002/sim.3069

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

Lewinsohn, P. M., Klein, D. N., & Seeley, J. (2000). Bipolar disorder during adolescence and young adulthood in a community sample. *Bipolar Disorders, 2*, 281-293. doi:10.1034/j.1399-5618.2000.20309.x

Lovejoy, M. C., & Steuerwald, B. L. (1995). Subsyndromal unipolar and bipolar disorders: comparisons on positive and negative affect. *Journal of Abnormal Psychology, 104*, 381-384.

Mackin, P., Targum, S. D., Kalali, A., Rom, D., & Young, A. H. (2006). Culture and assessment of manic symptoms. *British Journal of Psychiatry, 189*, 379-380.

Mallon, J. C., Klein, D. N., Bornstein, R. F., & Slater, J. F. (1986). Discriminant validity of the General Behavior Inventory: an outpatient study. *Journal of Personality Assessment, 50*, 568-577.

Marchand, W. R., Wirth, L., & Simon, C. (2006). Delayed diagnosis of pediatric bipolar disorder in a community mental health setting. *Journal of Psychiatric Practice, 12*, 128-133.

Mick, E., Biederman, J., Pandina, G., & Faraone, S. V. (2003). A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. *Biological Psychiatry, 53*, 1021-1027. doi:10.1016/S0006-3223(03)00234-8 |

Millner, A. J., & Nock, M. K. (2020). Self-injurious thoughts and behaviors. In E. A. Youngstrom, M. J. Prinstein, E. J. Mash, & R. Barkley (Eds.), *Assessment of Disorders in Childhood and Adolescence* (5th ed.). New York, NY: Guilford Press.

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

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63,* 539-569.

Poznanski, E. O., Freeman, L. N., & Mokros, H. B. (1985). Children's Depression Rating Scale - Revised. *Psychopharmacology Bulletin, 21,* 979-989.

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2012). DSM-5 Field Trials in the United States and Canada, Part II: Test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry, 170*, 59–70. doi:10.1176/appi.ajp.2012.12070999

Rescorla, L., Achenbach, T. M., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., . . . Verhulst, F. (2007a). Epidemiological Comparisons of Problems and Positive Qualities Reported by Adolescents in 24 Countries. *Journal of Consulting and Clinical Psychology, 75,* 351-358. doi:10.1037/0022-006X.75.2.351

Rescorla, L. A., Achenbach, T. M., Ginzburg, S., Ivanova, M., Dumenci, L., Almqvist, F., . . . Verhulst, F. (2007b). Consistency of teacher-reported problems for students in 21 countries. *School Psychology Review, 36,* 91-110.

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. *International Journal of Methods in Psychiatric Research, 18,* 169-184. doi:10.1002/mpr.289

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

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

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2018). *Evidence-based medicine: How to practice and teach EBM* (5th ed.). New York, NY: Churchill Livingstone.

Van Meter, A. R., Burke, C., Kowatch, R. A., Findling, R. L., & Youngstrom, E. A. (2016). Ten-year updated meta-analysis of the clinical characteristics of pediatric mania and hypomania. *Bipolar Disorders, 18,* 19-32. doi:10.1111/bdi.12358

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening Efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A Systematic Review. *Child and Adolescent Mental Health, 13,* 140-147. doi:10.1111/j.1475-3588.2007.00461.x

Yee, A. M., Algorta, G. P., Youngstrom, E. A., Findling, R. L., Birmaher, B., Fristad, M. A., &

Group, L. (2015). Unfiltered Administration of the YMRS and CDRS-R in a Clinical Sample of Children. *Journal of Clinical Child & Adolescent Psychology, 44*, 992-1007. doi:10.1080/15374416.2014.915548

Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: Reliability, validity, and sensitivity. *British Journal of Psychiatry, 133*, 429-435.

Youngstrom, E. A. (2020). Introduction to Evidence-Based Assessment: A Recipe for Success. In E. A. Youngstrom, M. J. Prinstein, E. J. Mash, & R. Barkley (Eds.), *Assessment of Disorders in Childhood and Adolescence* (5th ed.). New York, NY: Guilford Press.

Youngstrom, E. A., Egerton, G. A., Genzlinger, J., Freeman, L. K., Rizvi, S. H., & Van Meter, A. (2018). Improving the global identification of bipolar spectrum disorders: Meta-analysis of the diagnostic accuracy of checklists. *Psychological Bulletin, 144*, 315-342. doi:10.1037/bul0000137

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., DelPorto Bedoya, D., & Price, M. (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*, 847-858. doi:10.1097/01.chi.0000125091.35109.1e

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*, 267-276.

Youngstrom, E. A., Frazier, T. W., Findling, R. L., & Calabrese, J. R. (2008). Developing a ten item short form of the Parent General Behavior Inventory to assess for juvenile mania and hypomania. *Journal of Clinical Psychiatry, 69*, 831-839. doi:10.4088/JCP.v69n0517

Youngstrom, E. A., Genzlinger, J. E., Egerton, G. A., & Van Meter, A. R. (2015). Multivariate meta-analysis of the discriminative validity of caregiver, youth, and teacher rating scales for pediatric bipolar disorder: Mother knows best about mania. *Archives of Scientific Psychology, 3*, 112-137. doi:10.1037/arc0000024

Youngstrom, E. A., Gracious, B. L., Danielson, C. K., Findling, R. L., & Calabrese, J. (2003). Toward an integration of parent and clinician report on the Young Mania Rating Scale. *Journal of Affective Disorders, 77*, 179-190. doi:10.1016/s0165-0327(02)00108-8

Youngstrom, E. A., Halverson, T. F., Youngstrom, J. K., Lindhiem, O., & Findling, R. L. (2018). Evidence-Based Assessment from simple clinical judgments to statistical learning: Evaluating a range of options using pediatric bipolar disorder as a diagnostic challenge. *Clinical Psychological Science, 6*, 234-265. doi:10.1177/2167702617741845

Youngstrom, E. A., Joseph, M. F., & Greene, J. (2008). Comparing the psychometric properties of multiple teacher report instruments as predictors of bipolar disorder in children and adolescents. *Journal of Clinical Psychology, 64*, 382-401.

Youngstrom, E. A., Meyers, O. I., Demeter, C., Kogos Youngstrom, J., Morello, L., Piiparinen, R., . . . Calabrese, J. R. (2005). Comparing diagnostic checklists for pediatric bipolar disorder in academic and community mental health settings. *Bipolar Disorders, 7*, 507-517. doi:10.1111/j.1399-5618.2005.00269.x

Youngstrom, E. A., Meyers, O. I., Youngstrom, J. K., Calabrese, J. R., & Findling, R. L. (2006). Diagnostic and measurement issues in the assessment of pediatric bipolar disorder: Implications for understanding mood disorder across the life cycle. *Development and Psychopathology, 18*, 989-1021. doi:10.1017/S0954579406060494

Youngstrom, E. A., Morton, E., & Murray, G. (2020). Bipolar disorder. In E. A. Youngstrom, M. J. Prinstein, E. J. Mash, & R. Barkley (Eds.), *Assessment of Disorders in Childhood and Adolescence* (5th ed.). New York, NY: Guilford Press.

Youngstrom, E. A., Murray, G., Johnson, S. L., & Findling, R. L. (2013). The 7 Up 7 Down Inventory: A 14-item measure of manic and depressive tendencies carved from the General Behavior Inventory. *Psychological Assessment, 25*, 1377-1383. doi:10.1037/a0033975

Youngstrom, E. A., Salcedo, S., Frazier, T. W., & Perez Algorta, G. (2019). Is the finding too good to be true? Moving from "more is better" to thinking in terms of simple predictions and credibility. *Journal of Clinical Child and Adolescent Psychology, 48,* 811-824. doi:10.1080/15374416.2019.1669158

Youngstrom, E. A., & Van Meter, A. (2016). Empirically supported assessment of children and adolescents. *Clinical Psychology: Science and Practice, 23,* 327-347. doi:10.1111/cpsp.12172

Youngstrom, E. A., Van Meter, A., Frazier, T. W., Youngstrom, J. K., & Findling, R. L. (2018). Developing and validating short forms of the Parent General Behavior Inventory Mania and Depression Scales for rating youth mood symptoms. *Journal of Clinical Child & Adolescent Psychology*, 1-16. doi:10.1080/15374416.2018.1491006

Youngstrom, E. A., Youngstrom, J. K., Freeman, A. J., De Los Reyes, A., Feeny, N. C., & Findling, R. L. (2011). Informants are not all equal: predictors and correlates of clinician judgments about caregiver and youth credibility. *Journal of Child and Adolescent Psychopharmacology, 21,* 407-415. doi:10.1089/cap.2011.0032

Youngstrom, E. A., Zhao, J., Mankoski, R., Forbes, R. A., Marcus, R. M., Carson, W., . . . Findling, R. L. (2013). Clinical significance of treatment effects with aripiprazole versus placebo in a study of manic or mixed episodes associated with pediatric bipolar I disorder. *Journal of Child & Adolescent Psychopharmacology, 23*, 72-79. doi:10.1089/cap.2012.0024

Table 1

Demographics and Clinical Characteristics by Clinic Setting

	Academic Clinic *(N=427)*	*Community Clinic* *(N=313)*	*Effect* *Size*[a]
Youth Demographics			
Female, % (*n*)	47%	48%	-.005
Age, *M* (*SD*)	14.2 (1.9)	13.4 (1.9)	$.42^{***}$
White, % (*n*)	75%	6%	$.68^{***}$
Family Income ($1000s)[b]	$36.1 ($22.9)	$18.4 ($14.6)	$.95^{***}$
Clinical Characteristics			
Number Axis I KSADS Diagnoses	2.1 (1.3)	2.7 (1.4)	$-.45^{***}$
Any mood disorder diagnosis	70%	52%	$-.19^{***}$
Unipolar depressive disorder	33%	39%	.06
Bipolar spectrum diagnosis	38%	13%	$-.27^{***}$
Any attention-deficit/hyperactivity	47%	53%	.05
Any oppositional defiant disorder	26%	34%	$.09^{*}$
Any conduct disorder	9%	19%	$.14^{***}$
Any anxiety disorder	15%	30%	$.18^{***}$
Any posttraumatic stress disorder	4%	12%	$.17^{***}$
Mania Severity (YMRS Interview of both)	10.27 (11.58)	6.02 (7.90)	$.42^{***}$
Depression Severity (CDRS-R Interview of both)	38.85 (15.93)	32.87 (14.11)	$.39^{***}$
Youth Self Report			
YRS Externalizing *T*	59.22 (11.60)	58.43 (11.49)	.07
YRS Internalizing *T*	57.13 (13.22)	56.46 (12.02)	.06
AGBI – Hypo/Biphasic Raw	24.85 (16.09)	24.16 (15.17)	.04
AGBI – Depression Raw	41.53 (30.61)	39.53 (27.21)	.07
A-7 Up	5.74 (4.28)	5.90 (4.31)	-.04
A-7 Down	6.48 (5.98)	5.40 (4.88)	.20**
A-10M Raw	9.29 (6.98)	8.81 (6.14)	.07
A-10Da Raw	9.74 (7.54)	9.05 (5.56)	.10
A-10Db Raw	9.79 (7.86)	9.02 (6.83)	.10
Parent/Caregiver Report			
CBCL Externalizing *T*	65.10 (11.67)	69.39 (9.62)	$-.40^{***}$
CBCL Internalizing *T*	64.65 (11.16)	63.88 (10.35)	.07
PGBI – Hypo/Biphasic Raw	*23.42 (16.64)*	*19.45 (13.83)*	*$.26^{***}$*
PGBI – Depression Raw	*39.89 (26.28)*	*28.65 (23.66)*	*$.45^{***}$*

Note. [a]*phi* for categorical variables (sex, race, diagnostic group), Cohen's *d* for continuous variables (age, number of diagnoses, rating scales). A positive coefficient means the effect was larger in the Academic sample, and a negative coefficient means that the effect was larger in the Community – the academic parameter would underestimate the corresponding value in the community. [b]Income assessed via ranked bands. $^{*}p<.05$, $^{**}p<.005$, $^{***}p<.0005$, two-tailed.

Table 2

Projected and empirical estimates of internal consistency reliability, correlation with full-length scale, and length reduction for short forms

	Hypomanic/Biphasic	*Depression*
Full Length		
Items	28	46
Alpha (Academic)	.939	.973
Mean inter-item correlation	.355	.439

Short Form	*Mania (10M)*	*7 Up*	*Depression-A (10Da)*	*10Db*	*7 Down*
Items	10	7	10	10	7
Projected alpha	.846	.794	.886	.886	.846
Observed alpha-Academic	.899	.812	.913	.923	.924
Observed alpha-Community	.851	.800	.856	.883	.861
Projected correlation with full	.794	.745	.863	.863	.823
*Observed correlation – Academic	.952	.859	.963	.961	.934
*Observed correlation – Community	.943	.853	.935	.944	.901
Savings in Length (%)	64%	75%	78%	78%	85%
Projected validity reduction (%)	21%	25%	14%	14%	18%
Standard Error of Measurement	2.29	1.63	2.36	2.26	1.74
Standard Error of Difference	3.24	2.30	3.33	3.19	2.46
90% Critical Change	5.35	3.80	5.50	5.27	4.06
95% Critical Change	6.36	4.51	6.53	6.26	4.83

*Observed correlations are based on embedded item administration.

Table 3

Criterion correlations for 10 item Mania scale, 7 Up, and full length Hypomanic/Biphasic scale

		Academic (*N*=427)			Community (*N*=313)		
Expected		Full			Full		
Rank	Criterion Variable	Length	10M	*7 Up*	Length	10M	*7 Up*
1	YSR Externalizing *T* Score	.54	.52	.41****	.56	.52*	.43****
2	YMRS Total (interview)	.23	.24	.19	.20	.22	.21
3	YSR Internalizing *T* Score	.62	.59*	.44****	.58	.55*	.41****
4	Parent GBI Hypo/Biphasic Score	.29	.28	.23*	.27	.29	.26
5	Bipolar Spectrum Diagnosis	.20	.21	.19	.16	.16	.17
6	CBCL Externalizing *T* Score	.30	.29	.26	.17	.18	.16
7	Parent GBI Depression Score	.27	.24*	.16***	.24	.25	.18
8	CBCL Internalizing *T* Score	.26	.20**	.19*	.18	.17	.09*
9	CDRS Total (interview)	.24	.22	.13***	.21	.21	.11**
10	Any Mood Disorder Diagnosis	.18	.17	.12*	.13	.13	.07*
11	Count of Comorbid Diagnoses	.11	.08	.14	.28	.29	.27
13	ADHD Diagnosis	-.07	-.10	.02**	.07	.06	.09
13	ODD Diagnosis	.01	.01	.03	.03	.03	.03
13	Conduct Disorder Diagnosis	.06	.04	.07	.18	.17	.17
16	PTSD Diagnosis	.05	.05	.01	.10	.12	.01**
16	Any Anxiety Diagnosis	.03	.02	.04	.21	.22	.18
16	Female Youth	.26	.30*	.10****	.03	.07*	-.01
19	Youth Age (Years)	.08	.07	-.01**	.02	.02	-.06*
19	White Youth	-.06	-.05	-.01	.09	.06	.02*
19	Any Unipolar Depression	-.04	-.05	-.08	.03	.02	-.04*

Note. Coefficients are point-biserial correlations for dummy-coded categorical variables, and Pearson correlations for continuous variables. Steiger's test of dependent correlations tested difference between full length and short form criterion correlations. Differences, where significant, favored short form validity unless noted otherwise. Criterion variables 95.5% complete for Academic, 98.6% complete for Community sample. Correlations larger than .11 significant $p<.05$ in both samples, and >.15 are $p<.01$.

[a]Discriminant validity of 10M was slightly worse than full length scale in Community sample.

Table 4

Criterion correlations for 10 item Depression Forms A & B and full length Depression scale

		Academic (*N*=427)				Community (*N*=313)			
Expected		Full				Full			
Rank	Criterion Variable	Length	10Da	10Db	7 Down	Length	10Da	10Db	*7 Down*
1	YSR Internalizing *T* [a]	.74	$.71^{*}$	.73	.73	.69	$.65^{*}$	.67	$.61^{****}$
2	CDRS total (interview)	.41	.41	.40	.40	.38	.40	.41	$.43^{*}$
3	YSR Externalizing *T*	.49	$.45^{**}$	.46	$.44^{*}$	.53	$.48^{**}$	.51	$.45^{**}$
4	Parent GBI Depression	.13	.15	.14	.13	.18	.16	.20	.21
5	Any unipolar depression	.25	.25	.26	.26	.25	.24	$.29^{*}$	.27
6	Any Mood Disorder	.41	.40	.39	.38	.33	.35	.32	.32
7	Bipolar Spectrum	.10	.08	.10	.10	.11	.13	.14	.09
8	CBCL Internalizing *T* [a]	.33	.32	.31	.30	.27	.28	.27	.25
9	Female youth	.42	.40	.42	.44	.14	.16	.13	.17
10	YMRS total (Interview)	.12	.09	.12	.10	.16	.17	.16	.15
11	Youth age (Years)	.20	$.23^{*}$	.22	.20	.13	.17	$.18^{*}$	.17
12	Parent GBI Hypo/Biphasic	.21	$.17^{*}$	$.18^{*}$	$.15^{**}$	.26	.25	.25	.24
14	CBCL Externalizing *T*	.18	$.15^{*}$	$.15^{*}$	$.13^{*}$	.10	$.06^{*}$	.08	$.05^{*}$
14	Count of comorbid diagnoses	.05	$.02^{*}$	.02	.03	.30	$.26^{*}$	.28	.26
14	Any anxiety disorder	.10	.11	.11	.12	.29	.27	.31	.27
16	Any PTSD	.11	.11	.10	.13	.21	.21	.20	.19
17	White youth	-.08	-.10	-.09	$-.04^{*}$	.14	.11	.14	.14
19	Any ADHD diagnosis	-.22	-.23	-.24	-.25	-.04	$-.08^{*}$	-.07	-.09
19	Any ODD diagnosis	-.05	-.07	-.07	-.07	-.03	-.03	-.05	-.06
19	Any conduct disorder	.03	$.00^{*}$	.02	.02	.12	.08	.09	.09

Note. Coefficients are point-biserial correlations for dummy-coded categorical variables, and Pearson correlations for continuous variables. Steiger's test of dependent correlations tested difference between full length and short form criterion correlations. Differences, where significant, favored short form validity unless noted otherwise. Criterion variables 95.5% complete for Academic, 98.6% complete for Community sample. Correlations larger than .11 significant $p<.05$ in both samples, and >.15 are $p<.01$.

[a] Convergent validity of both short forms with Internalizing was significantly lower than for full length depression in both Academic and Community samples, largest difference $r=.05$.

Table 5

Receiver Operating Characteristic (ROC) analysis of the discriminative validity of the AGBI-10M, 7 Up, Hypomanic/Biphasic scale (full length) and the CBCL Externalizing score for discriminating cases with bipolar spectrum disorder from all other cases at clinic

	Academic		Community	
Predictor	AUC	95% CI	AUC	95% CI
Diagnosis ROC – Any Bipolar	.93	--	.93	--
Hypo/Biphasic Full Length	.63	(.57 to .69)	.66	(.57 to .75)
7 Up	.62	(.56 to .67)	.63	(.54 to .72)
10 item Mania	.62	(.56 to .68)	.65	(.56 to .73)
Externalizing	.67	(.61 to .72)	.60	(.51 to .70)
Diagnosis ROC – Any Mood	.93	--	.93	--
Depression Full Length	.69	(.63 to .75)	.64	(.58 to .70)
7 Down	.68	(.62 to .74)	.66	(.60 to .72)
10 Item Depression - Form A	.69	(.63 to .75)	.63	(.56 to .69)
10 Item Depression - Form B	.68	(.63 to .74)	.67	(.61 to .73)
7 Down Internalizing	.66	(.60 to .72)	.65	(.59 to .72)

The diagnosis ROC is dictated by the inter-rater reliability of the LEAD diagnoses (kappa=.85), and sets an upper border for the AUC that could be empirically observed (Kraemer, 1992). None of the short forms performed significantly differently than the full length version, largest DeLong test value 1.19, smallest unadjusted *p* >.25.

Table 6

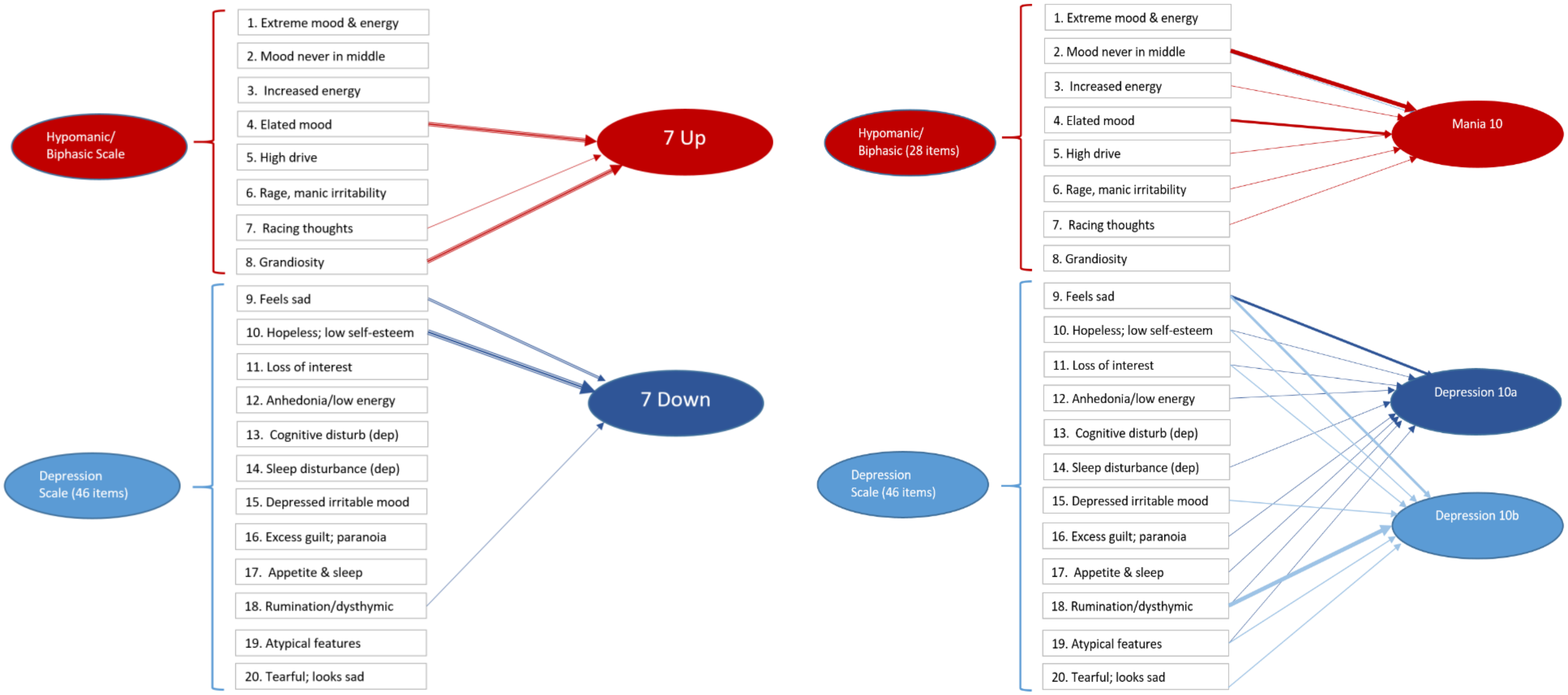
Multilevel diagnostic likelihood ratios (DiLR) for short forms, using 10M to predict bipolar spectrum disorders, and 10 item Depression Forms A & B to predict any mood disorder

	Risk Change Label	Low	Neutral	High	Very High
10M for Bipolar	Score Range	0 to 9.99	10 to 14.99	15+	NA
	DiLR	0.71	1.04	2.05	
7 Up for Bipolar	Score Range	0 to 1.99	2 to 13.99	14+	NA
	DiLR	0.55	1.02	2.68	
10DepA for Any Mood	Score Range	0 to 5.99	6 to 15.99	16 to 21.99	22+
	DiLR	0.54	1.23	1.76	4.74
10DepB for Any Mood	Score Range	0 to 5.99	6 to 16.99	17 to 20.99	21+
	DiLR	0.52	1.19	2.55	5.33
7 Down for Any Mood	Score Range	0 to 2.99	3 to 11.99	12 to 15.99	16+
	DiLR	0.51	1.22	2.20	6.50

Note. Segments defined by quintiles in the Academic sample, splitting the top quintile to examine potential value of extremely high scores (Youngstrom et al., 2004), and then adjusted to avoid degenerate distributions in either sample (Pepe, 2003).

Figure 1

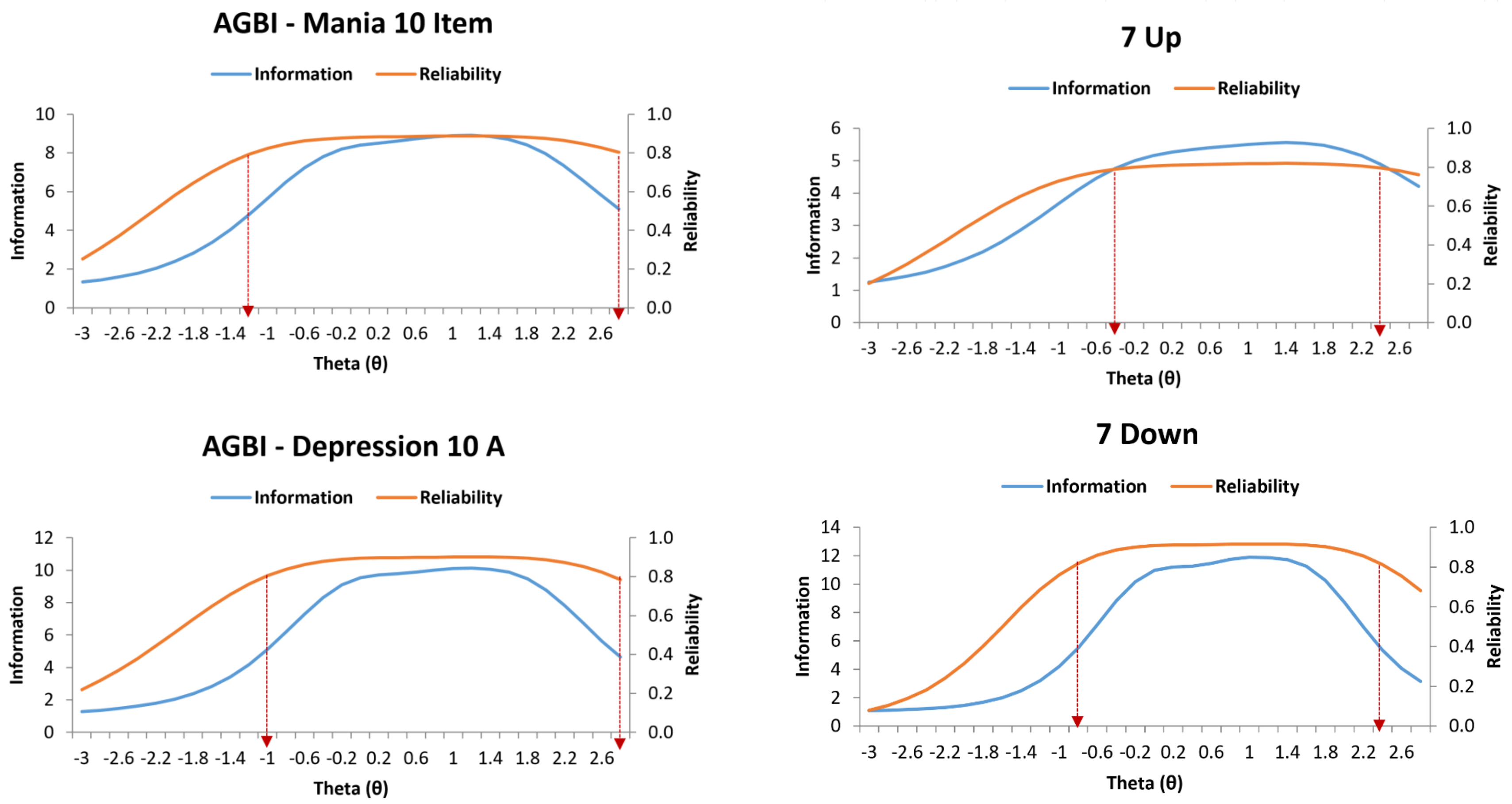
Content coverage showing which item parcels are represented in the seven and ten item short forms. Line thickness denotes whether 1, 2, or 3 items come from the parcel.



Note. Reproduced from https://en.wikiversity.org/wiki/Evidence_based_assessment/Instruments/General_Behavior_Inventory

Figure 2

Information and reliability estimates from IRT analysis of short forms

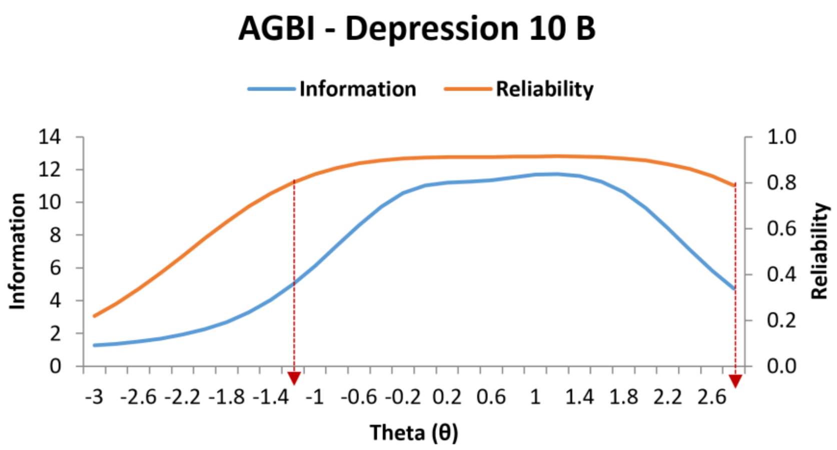


AGBI - Mania 10 Item

7 Up

AGBI - Depression 10 A

7 Down



Note. Reproduced from https://en.wikiversity.org/wiki/Evidence_based_assessment/Instruments/General_Behavior_Inventory, CC-BY 4.0, Eric Youngstrom, PhD & Thomas Frazier, PhD

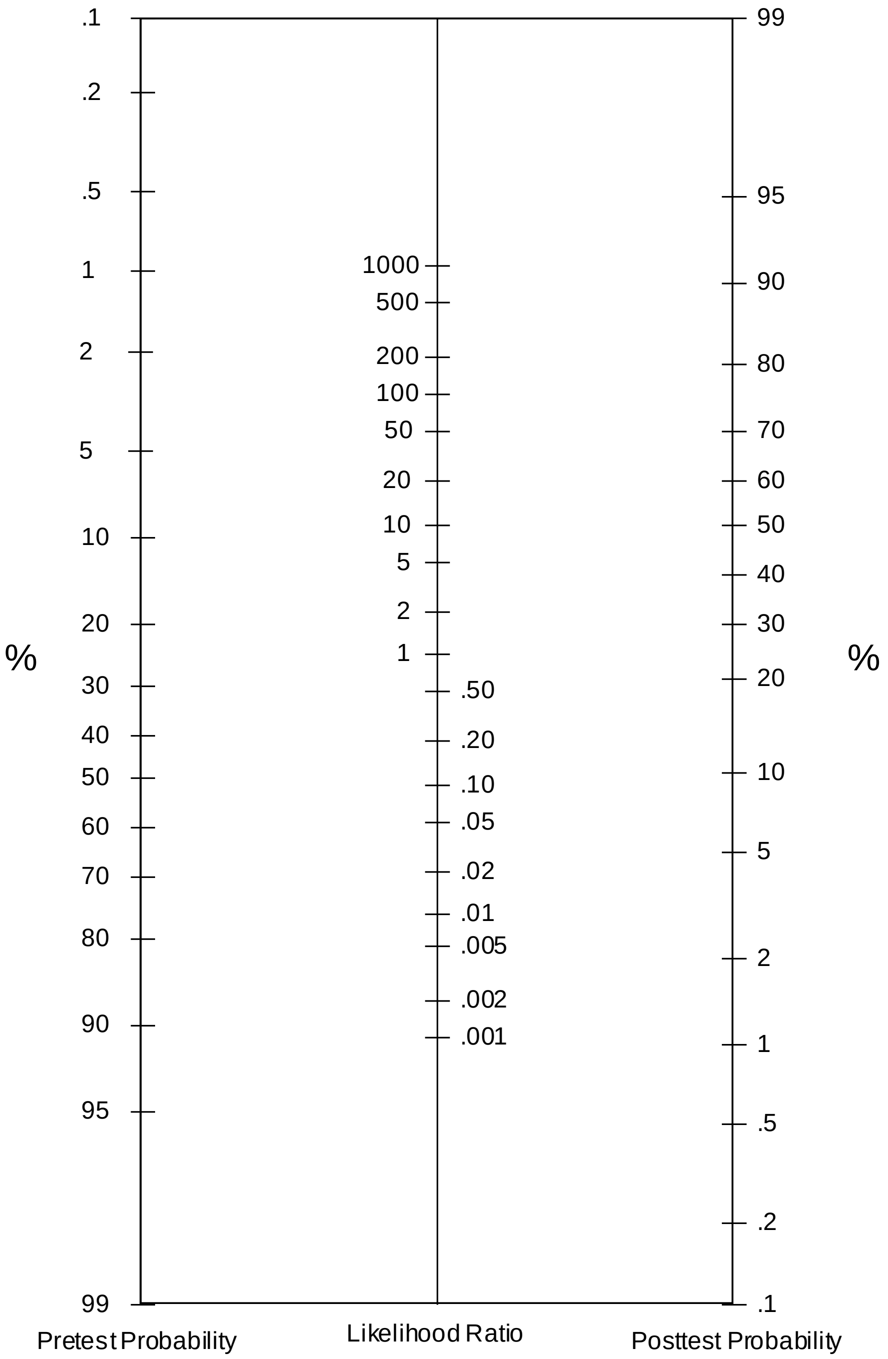


Figure 3. Probability nomogram for combining diagnostic likelihood ratios with other information about cases to revise probability estimates.