Published in final edited form as:

J Affect Disord. 2016 November 15; 205: 73–80. doi:10.1016/j.jad.2016.06.062.

An Item Response Theory Evaluation of the Young Mania Rating Scale and the Montgomery-Asberg Depression Rating Scale in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD)

James J. Prisciandaro* and Bryan K. Tolliver

Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

Abstract

Background—The Young Mania Rating Scale (YMRS) and Montgomery-Asberg Depression Rating Scale (MADRS) are among the most widely used outcome measures for clinical trials of medications for Bipolar Disorder (BD). Nonetheless, very few studies have examined the measurement characteristics of the YMRS and MADRS in individuals with BD using modern psychometric methods. The present study evaluated the YMRS and MADRS in the Systematic Treatment Enhancement Program for BD (STEP-BD) study using Item Response Theory (IRT).

Methods—Baseline data from 3,716 STEP-BD participants were available for the present analysis. The Graded Response Model (GRM) was fit separately to YMRS and MADRS item responses. Differential item functioning (DIF) was examined by regressing a variety of clinically relevant covariates (e.g., sex, substance dependence) on all test items and on the latent symptom severity dimension, within each scale.

Results—Both scales: 1) contained several items that provided little or no psychometric information, 2) were inefficient, in that the majority of item response categories did not provide incremental psychometric information, 3) poorly measured participants outside of a narrow band of severity, 4) evidenced DIF for nearly all items, suggesting that item responses were, in part, determined by factors other than symptom severity.

Limitations—Limited to outpatients; DIF analysis only sensitive to certain forms of DIF.

*Corresponding author. Medical University of South Carolina, Department of Psychiatry and Behavioral Sciences, 67 President Street, P.O. Box 250861, Charleston, SC 29425 or send e-mail to priscian@musc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Contributors

James J. Prisciandaro and Bryan K. Tolliver, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

Conflict of Interest

None.

Ethical Standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Conclusions—The present study provides evidence for significant measurement problems involving the YMRS and MADRS. More work is needed to refine these measures and/or develop suitable alternative measures of BD symptomatology for clinical trials research.

Keywords

Item Response Theory; Young Mania Rating Scale; Montgomery-Asberg Depression Rating Scale; STEP-BD

INTRODUCTION

Discovery of improved treatments for bipolar disorder is an urgent public health priority (Sachs et al., 2003). Unfortunately, progress in medication development is impeded by a number of methodological issues that uniquely affect clinical trials for bipolar disorder. Central to the problem is the lack of accurate, valid, and objective assessment instruments by which a medicine's efficacy may be reliably measured (Bagby, Ryder, Schuller, & Marshall, 2004; Bech, 2006). Still lacking pathophysiologic markers of disease or treatment response, subjective appraisal of clinical symptoms has defined all primary outcome measures for over 50 years in mood disorders treatment research. Using clinician-rated itemized mood rating scales (e.g. Montgomery Asberg Depression Rating Scale (MADRS), Young Mania Rating Scale (YMRS)), a medicine's efficacy is supported by statistically superior reduction in mood scale scores or in delayed time to a recurrence of mood episodes relative to placebo (Hamilton, 1960; Montgomery & Asberg, 1979; Young, Biggs, Ziegler, & Meyer, 1978). Investigations of the classical psychometric properties of the MADRS and YMRS have been generally supportive, demonstrating good interrater reliability, convergent validity, and sensitivity to change (Hamilton, 1960; Montgomery & Asberg, 1979; Young, Biggs, Ziegler, & Meyer, 1978).

Unfortunately, however, itemized rating scales are compromised in numerous ways (Bagby et al., 2004; Bech, 2006; Gelenberg et al., 2008). This is particularly true for bipolar disorder (Baldassano, 2005; Berk et al., 2007; Berk et al., 2004; Keck, 2004; Licht, 2001; Martinez-Aran et al., 2008; Nolen, 2002; Picardi, 2009; Rush et al., 2000) due in part to inherent characteristics of the illness: (i) Because depression rating scales like the MADRS were designed to assess unipolar depression, which differs qualitatively from bipolar depression (Berk et al., 2007; Berk et al., 2004; Perlis, Brown, Baker, & Nierenberg, 2006; Weinstock, Strong, Uebelacker, & Miller, 2009), they are insensitive to 'atypical' symptoms (e.g. hypersomnia, hyperphagia) common in bipolar depression and therefore routinely underestimate depressive severity in bipolar patients. (ii) As even clinician-rated scales depend heavily on patient responses, symptom ratings can be influenced by impairments in patient insight that commonly accompany bipolar mood episodes (Dell'Osso et al., 2002). (iii) Differential item weighting, integral to scales like the YMRS, can correspond poorly to the clinical importance of given domains in individual patients. For example, sleep disruption is a core feature of mania but is underweighted on YMRS, and symptoms such as impulsivity, high-risk behavior (e.g. indiscriminant spending or sex), or suicidality are not addressed at all by YMRS but can have catastrophic consequences in manic/mixed episodes. Conversely, rating scales that lack differential weighting assign equal value to all items in

the overall score, essentially treating mood episodes as unitary phenomena. Overall scores, however, can be identical in two patients for widely divergent reasons, potentially obscuring subsets of symptoms (e.g. suicidality, hopelessness) that are critical to clinical decision-making. Similarly, total scores obtained from serial repeated ratings provide limited information about an individual's clinical status because symptoms may not change in unison over time (P. D. Harvey, Endicott, & Loebel, 2008). (iv) Psychotic symptoms confer profound functional impairment in some patients with bipolar disorder, but are minimally assessed by scales like MADRS / YMRS (Montgomery & Asberg, 1979; Young et al., 1978).

Despite these concerns, the MADRS and YMRS remain among the most widely used outcome measures in bipolar disorder clinical trials. Given this, it is perhaps surprising that the MADRS and YMRS have yet to be evaluated using modern psychometric methods (item response theory; IRT) in individuals with bipolar disorder. Unlike previous classical test theory investigations of the psychometric properties of the MADRS and YMRS, IRT methods allow for a focused evaluation of the individual items of the MADRS and YMRS, as opposed to their total scores (Embretson & Reise, 2000), which is needed in order to properly evaluate the concerns and limitations noted above. The present study investigated the psychometric properties of the YMRS and MADRS using IRT in STEP-BD, the largest available sample of outpatients with bipolar disorders.

METHOD

Study Overview

STEP-BD was a multicenter effectiveness study that prospectively evaluated clinical outcomes in individuals with bipolar disorder from 1999 to 2005 (Sachs et al., 2003). Participation was offered to all patients meeting DSM-IV criteria for bipolar disorder (i.e., bipolar I disorder, bipolar II disorder, bipolar disorder not otherwise specified, cyclothymia, schizoaffective disorder bipolar type) at participating STEP-BD centers. All participants provided written informed consent, and 15- to 17-year-old participants additionally provided parental assent.

Assessments

As part of the STEP-BD study, participants completed extensive baseline assessment interviews. Bipolar diagnoses and retrospective course (e.g., past year rapid cycling, age of bipolar disorder onset) were assessed by study clinicians at baseline using the Affective Disorders Evaluation (which included the mood and psychosis modules of the Structured Clinical Interview for DSM-IV); comorbid Axis I diagnoses (e.g., substance use disorders, current anxiety disorders) were assessed using the Mini-International Neuropsychiatric Interview (Sachs et al., 2003; Sheehan et al., 1998). Current (i.e., past-week) manic and depressive symptoms were assessed using the YMRS and the MADRS, respectively. Demographic information (e.g., age, sex) was collected via a study-specific questionnaire.

Statistical Analysis

Prior to fitting separate unidimensional item response models (IRMs) to the YMRS and MADRS data, exploratory factor analyses (EFAs) were conducted to ensure that each scale was essentially unidimensional. EFAs were estimated using a weighted least squares mean and variance adjusted (WLSMV; (B. O. Muthen, 1989)) estimator to properly account for the ordinal nature of the item data. The number of factors to extract for each scale was determined through parallel analyses (Horn, 1965) with 1000 sets of random data. Specifically, only factors with eigenvalues exceeding the 95th percentile of the distribution of eigenvalues derived from random data were retained. Once the essential unidimensionality of each scale was confirmed, the Graded Response Model (GRM; (Samejima, 1968)), a generalization of the two-parameter logistic IRM for ordinal data, was fit separately to YMRS and MADRS item responses. The GRM was chosen over alternative models because item response categories were ordered, items were expected to vary in terms of discrimination, and because the number of available response options varied across YMRS items (Embretson & Reise, 2000). According to the GRM, each item within a given scale is described by an item slope (i.e. discrimination, α_i) parameter and $k-1$ between-category threshold (i.e., difficulty, $\beta_{i,k-1}$) parameters, where k = the item's number of response options. Discrimination refers to an item's ability to discriminate between different latent levels of manic (YMRS) or depressive (MADRS) symptom severity (i.e., theta, θ). Descriptive guidelines for discrimination suggest that $\alpha < 0.65$ = low, $0.65-1.34$ = moderate, and > 1.34 = high (Baker, 2001). Difficulty parameters reflect the standardized latent level of symptom severity (e.g., 0 = average, -1 = 1SD below average, $+1$ = 1SD above average) at which subsequent response options become more probable than the previous option. A “good” test has difficulty parameters spread across a range of symptom severity. All IRM parameters were estimated in a logistic modeling framework; Mplus generated response option thresholds were subsequently transformed to IRT parameterization via $\alpha_i =$ unstandardized item loading, $\beta_{i,k-1} =$ unstandardized response option threshold/ α_i , (L. K. Muthen & Muthen, 2011). Category response curves, which represent the probability of a participant at a given latent symptom severity level responding in a particular response category, were calculated for each item (Embretson & Reise, 2000). Category response curves were subsequently transformed into item information curves, which were summed to provide test information curves for each scale; information curves indicate the amount of psychometric information (i.e., the reciprocal of the standard error of measurement) at each point along a latent severity dimension. Following estimation of the YMRS and MADRS GRMs, we investigated whether scale items were “biased” (i.e., exhibited differential item functioning) by simultaneously regressing a variety of clinically relevant covariates (i.e., age, sex, past year rapid cycling, current anxiety disorder, substance dependence, bipolar episode status, bipolar subtype, age of bipolar disorder onset) on all test items and on the latent symptom severity dimension, within each scale; this model specification is referred to as a Multiple Indicators Multiple Causes (MIMIC) model (B. O. Muthen, 1985). The model provides evidence for “item bias” if individuals with the same level of a given symptom severity score differently ($p < .01$) on one or more items based on their status on one or more covariates. In other words, the MIMIC model reveals differences in item response probabilities that remain once participants are equated on manic or depressive symptom severity; differences in item response probabilities that are unlikely to be due to the

underlying construct of interest. DIF estimates were additionally evaluated for “clinical significance” by converting odds ratios (OR) of covariates predicting individual test items to Cohen's D estimates (via: $\ln(OR) * \frac{\sqrt{3}}{\pi}$ (Hasselblad & Hedges, 1995)). As per Weinstock and colleagues, Cohen's D estimates > 0.20 were deemed “clinically significant” (Weinstock et al., 2009). IRT and MIMIC models were estimated using maximum likelihood estimation with robust standard errors (MLR; (L. K. Muthen & Muthen, 2011)) which uses all available data and returns unbiased parameter estimates under the assumption that data are missing at random. All models were estimated using MPlus 6.1 software (L. K. Muthen & Muthen, 2011); study site was included as a clustering variable for all models.

RESULTS

Demographics and clinical characteristics of the sample ($n = 3,721$) have been described previously (Kogan et al., 2004). Patient characteristics concerning covariates included in MIMIC (DIF) models were as follows: Age ($M = 39.6$, $SD = 12.8$), gender (57.8% female), past year rapid cycling (35.1%), current anxiety disorder (30.3%), substance dependence (34.8%), bipolar episode status (Major Depressive Episode = 30.4%, Manic Episode = 2.5%, Mixed Episode = 7.2%, Hypomanic Episode = 3.7%), bipolar subtype (Bipolar I Disorder = 64.1%, Bipolar II Disorder = 27.2%, Bipolar NOS = 7.0%, other = 1.7%), age of bipolar disorder onset ($< 13 = 38.8\%$, $13-18 = 34.7\%$, $> 18 = 26.5\%$).

Of the evaluable sample, between 7 (0.2%) and 23 (0.6%) participants were missing data on individual MADRS items; 5 participants (0.1%) were missing data on all MADRS items and were therefore not included in MADRS IRT models. Between 9 (0.2%) and 25 (0.7%) participants were missing data on individual YMRS items; 9 participants (0.2%) were missing data on all YMRS items and were therefore not included in YMRS IRT models. For MIMIC (DIF) models, 255 participants (6.9%) were missing data on at least one covariate and were therefore not included in MIMIC models. This left an effective sample size of 3,460 and 3,463 for DIF models involving YMRS and MADRS, respectively. Across models, missing data were $< 10\%$, which is considered non problematic (Kline, 2005).

Parallel analyses suggested that the YMRS was best characterized by up to two factors (Comparative Fit Index [CFI] = 0.99, Root Mean Square Error of Approximation [RMSEA] = 0.025). However, only two items (“irritability” and “disruptive-aggressive behavior”), similar in content, loaded substantially (i.e., > 0.40) on the second factor of the two-factor model, suggesting that a one-factor model (M loading = 0.59, CFI = 0.91, RMSEA = 0.06) with the assumption of local independence relaxed for “irritability” and “disruptive-aggressive behavior” was the most appropriate and substantively meaningful option. Parallel analyses suggested that the MADRS was best characterized by one factor (M loading = 0.69, CFI = 0.98, RMSEA = 0.06; two factor model CFI = 0.99, RMSEA = 0.04).

Results from the YMRS GRM, with the assumption of local independence relaxed for “irritability” and “disruptive-aggressive behavior,” are presented in Table 1 (model parameter estimates) along with Figures 1 (category response curves) and 2 (item and test information curves). The YMRS GRM findings demonstrated several problems with the measure. First, as depicted in Figure 1, although items contained between 5 and 9 response

categories, most response categories were non-informative. Even for relatively well behaving items (e.g., speech), for most response categories (e.g., 1-3, 5, 7-8) there was no point along the trait continuum at which participants were more likely to choose that category (e.g., 2) relative to neighboring categories (e.g., 0, 4). This suggests that most or all items could employ 2-3 category response scales without any loss of test information. Second, as depicted in Figure 2, several items provided virtually zero (e.g., appearance, insight) or very little (e.g., sexual interest, sleep, irritability) test information, suggesting that eliminating these items would not appreciably reduce the amount of psychometric information provided by the test. Third, as depicted in Figure 2, most items provided test information in a narrow band along the trait continuum, centered around 1 SD above the mean, suggesting that the YMRS does a poor job of assessing individuals with below average (< 0 SD) or substantially above average (e.g., > 2 SD) manic symptom severity.

Results from the MADRS GRM are presented in Table 1 (model parameter estimates) as well as Figures 3 (category response curves) and 4 (item and test information curves). As can be seen in Figure 3, similar to the YMRS, most or all of the items could have used 2-3 category response scales without any appreciable loss of test information. As can be seen in Figure 4, and also similar to the YMRS, several items (e.g., inner tension, reduced sleep, reduced appetite) provided very little psychometric information, suggesting that they could be eliminated without negatively impacting the test. Relative to the YMRS, the MADRS provided psychometric information along a wider band of the trait continuum. However, in an absolute sense, test information for participants with substantially below average (e.g., < -1 SD) or above average (e.g., > 2 SD) depressive symptom severity was very low.

Regarding differential item functioning, MIMIC models demonstrated that the majority of test items evidenced some degree of item bias. For the YMRS, past year rapid cycling, current anxiety disorder, younger age, younger age of onset of bipolar disorder, and male sex each uniquely predicted increased manic symptom severity. Controlling for these associations, responses across all YMRS items, with exception of “Increased Motor Activity-Energy” (YMRS 2), were significantly associated with at least one covariate, providing evidence for DIF across items (see Table 2). However, only 58% (11/19) of statistically significant DIF findings were determined to also be clinically significant (i.e., Cohen's $D > 0.20$; see Table 2 and Supplementary Table 1). At equivalent levels of manic symptom severity, individuals in a mixed episode and individuals with co-occurring anxiety disorder were more likely to endorse Irritability (YMRS 5). At equivalent levels of manic symptom severity, individuals in a major depressive episode were less likely to endorse Elevated Mood (YMRS 1), Sexual Interest (YMRS 3), Content (YMRS 8), and Insight (YMRS 11), but more likely to endorse Sleep (YMRS 4), Irritability, Disruptive-Aggressive Behavior (YMRS 9), and Insight. See Table 2 for a complete listing. For the MADRS, past year rapid cycling, current anxiety disorder, current major depressive, manic, or mixed episode, younger age, younger age of onset of bipolar disorder, and male sex each uniquely predicted increased depressive symptom severity. Controlling for these associations, each MADRS item was significantly predicted by at least one covariate, providing evidence for DIF across all items (see Table 3). However, only 55% (22/40) of statistically significant DIF findings were determined to also be clinically significant (i.e., Cohen's $D > 0.20$; see Table 3 and Supplementary Table 2). At equivalent levels of depressive symptom severity,

women were more likely to endorse Apparent (MADRS 1) and Reported (MADRS 2) Sadness, Concentration Difficulties (MADRS 6), and Lassitude (MADRS 7) relative to men. Co-occurring anxiety disorder was associated with increased endorsement of Inner Tension (MADRS 3). Individuals in a manic episode were more likely to endorse Inner Tension, Reduced Sleep (MADRS 4), and Reduced Appetite (MADRS 5), but less likely to endorse Lassitude and Pessimistic Thoughts (MADRS 9) and individuals in a mixed episode were more likely to endorse Inner Tension, Reduced Sleep, Concentration Difficulties, Inability to Feel (MADRS 8), and Suicidal Thoughts (MADRS 10). See Table 3 for a complete listing.

DISCUSSION

The present study investigated the psychometric properties of the YMRS and MADRS in the largest available sample of outpatients with bipolar disorders: STEP-BD. Item response modeling and differential item functioning analyses demonstrated that both scales: 1) contained several items that provided little or no psychometric information, 2) were inefficient, in that the majority of item response categories did not provide incremental psychometric information, 3) poorly measured participants outside of a narrow band of symptom severity (centered around the sample mean [i.e., MADRS] or 1 SD above the sample mean [i.e., YMRS]), 4) evidenced differential item functioning (i.e., item bias) for nearly all items, suggesting that item responses were, in part, determined by a variety of factors (particularly gender, co-occurring anxiety disorder, and current bipolar episode status) other than manic or depressive symptom severity. Differential item functioning involving bipolar episode status, for example, may suggest that individuals with equivalent levels of depressive symptom severity may experience depression substantially differently depending on their current mood state (e.g., major depressive vs. mixed episode). In sum, the results suggest that there is much room for improvement in the assessment of manic and depressive symptoms in outpatients with bipolar disorders. Although the literature has long raised concerns about relying on measures like the YMRS and MADRS as primary outcomes in clinical trials for individuals with bipolar disorder, the present study is the first to our knowledge to investigate the psychometric properties of these measures using modern psychometric methods in a sample of individuals with bipolar disorder.

There are several potential ways to improve the MADRS and YMRS, and other similar mood symptom assessment measures, as outcome measures in clinical trials for bipolar disorder. Suggestions include: 1) Remove items that do not appear to significantly inform individuals' latent levels of manic or depressive severity (e.g., MADRS: reduced sleep, reduced appetite; YMRS: appearance, insight). Although these items do not inform participants' levels of symptom severity they still contribute to participants' summed scores thereby making them a source of error. The respective sleep items for the MADRS and YMRS provide an illustrative example of specific item characteristics (wording, response scale distribution) that make the items unsuitable to assess severity in this population. In the case of the MADRS, the item is not only unidirectional (omitting hypersomnia as a response) but also identifies reduction of sleep of "at least two hours" as the criterion for maximal score. In contrast, the YMRS 0-4 point sleep item awards one point for sleep disruption of less than one hour, an additional point for sleep disruption of over one hour (i.e. up to 7-8 hours), and additional points only if the patient reports decreased need for

sleep (Young et al. 1978). Thus it is unsurprising that both of these items were found to perform poorly in our analysis, but this is unfortunate given that disruptions of sleep are cardinal neurovegetative symptoms of both depression and (hypo)manic episodes in bipolar disorder. 2) Introduce additional items to provide appropriate alternatives (e.g., hypersomnia, hyperphagia) to removed items or to provide coverage of previously neglected symptom content areas (e.g., psychotic symptoms) (Berk et al., 2007; Berk et al., 2004). The approach taken by the Beck Depression Inventory II (Beck, Steer, & Brown, 1996), whereby change in sleep and change in appetite are each assessed via one item that allows for either increases or decreases to be recorded, would be one option for improving assessment of vegetative and reverse vegetative depressive symptoms on the MADRS. 3) Replace simple summed scale scores with weighted scale scores that take into account the relative associations between individual symptoms/items and the latent dimensions of manic and depressive symptom severity. Such computed scores could also provide correction factors for observed differential item functioning/item bias and thereby remove the undue influence of extraneous factors on item responses. 4) Remove non-informative response options (i.e., response categories for which there is no point along the latent severity continuum at which participants are more likely to choose that category relative to neighboring categories). For example, our results demonstrated that the non-anchored, intermediary response options of the MADRS (i.e., “1”, “3”, “5” across all items) and YMRS (i.e., “1”, “3”, “5,” “7” on the Irritability, Speech, Content, and Disruptive-Aggressive items) could be eliminated with little to no loss of test information. 5) Introduce response options, and additional items, that are designed to flesh out the lower end of the severity spectrum. Given that both the MADRS and, particularly, the YMRS, are not capable of reliably measuring “average” (i.e., defined as the average level of mood symptom severity in outpatients with bipolar disorder) or below average levels of symptom severity, it is perhaps not surprising that symptom improvement and normalization are difficult to detect in clinical trials of bipolar disorder.

Another potential option is to develop alternatives to symptom severity scales like the YMRS and MADRS; outcome measures that are not as affected by observer and participant influence (response biases, memory capacity, etc). Possibilities include actigraphy, sociometry, and neuroimaging (e.g., functional Magnetic Resonance Imaging). Although these methods have provided invaluable insights into the nature of bipolar disorder they have not been extensively evaluated as outcomes in clinical trials for bipolar disorder (A. G. Harvey, Schmidt, Scarna, Semler, & Goodwin, 2005; Phillips & Swartz, 2014). Regardless of whether subjective or objective alternative outcome measures are pursued, it is important to consider that manic and depressive symptomatology is complex and multifaceted and that these phenomena may require more than a handful of test items to adequately assess. For example, take the core feature of bipolar disorder, sleep disturbance; the YMRS and MADRS each devote a single item to this phenomenon. The existence of the Pittsburgh Sleep Quality Index (PSQI; (Buysse, Reynolds, Monk, Berman, & Kupfer, 1989)), actigraphy, and polysomnography, among other assessment measures and techniques, suggest that a test item or two may not be enough to adequately capture and quantify bipolar sleep disturbance. Our results confirm this suspicion; sleep related items from both the YMRS and MADRS performed very poorly in the present study. Overall, it may be that

bipolar symptomatology may be more accurately captured via a collection of tests, each devoted to a particular symptom or neurobehavioral domain.

The preceding discussion should be tempered by the limitations of the present study. First, the present study only examined the psychometric properties of the YMRS and MADRS in outpatients with bipolar disorder, thereby limiting the applicability of our parameter estimates and conclusions to this population. In order to generalize our findings to the larger population of all individuals with bipolar disorder, replication and extension of our results is needed in community and inpatient samples. Second, we implemented the MIMIC method of evaluating differential item functioning because it allowed us to simultaneously examine the impact of a number of covariates (including continuous covariates) on item responses. Although the MIMIC approach has a number of benefits, it is limited in the types of differential item functioning it can identify. Specifically, it can only identify uniform (i.e., occurring uniformly at all levels along the latent severity dimension) differential item functioning in difficulty (response option threshold) parameters. As such, it is quite likely that the YMRS and MADRS evidence substantially more bias than we were able to capture. Further research is needed to, first replicate and extend our findings of differential item functioning to other samples, and subsequently, to examine the impact of this differential item functioning on substantive study results (e.g., treatment outcome in randomized medication trials).

In summary, the present study provides initial evidence for significant measurement problems involving the YMRS and MADRS as applied to outpatients with bipolar disorder. Given that the YMRS and MADRS are widely implemented as primary outcome measures in clinical trials for bipolar disorder, more work is needed to refine these measures, and to develop alternative measures that will improve our assessment of bipolar phenomenology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

None.

Role of Funding

STEP-BD Data Use Certification for Public Release Dataset version 4.1 provided 4/27/11 (Tolliver, Recipient PI). Data used in the preparation of this article were obtained from the limited access datasets distributed from the NIH-supported “Systematic Treatment Enhancement Program for Bipolar Disorder” (STEP BD). This is a multisite, clinical trial studying the current treatments for bipolar disorder, including medications and psychosocial therapies. The study was supported by NIMH Contract # N01MH80001 to Massachusetts General Hospital and the University of Pittsburgh. The [ClinicalTrials.gov](https://clinicaltrials.gov) identifier is NCT00012558. This manuscript reflects the views of the authors and may not reflect the opinions or views of the STEP BD Study Investigators or the NIH.

Dr. Prisciandaro is funded by K23 AA020842.

REFERENCES

- Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry*. 2004; 161(12):2163–2177. doi:10.1176/appi.ajp.161.12.2163. [PubMed: 15569884]
- Baker, F. *The Basics of Item Response Theory*. University of Maryland; College Park, MD: 2001.
- Baldassano CF. Assessment tools for screening and monitoring bipolar disorder. *Bipolar Disord*. 2005; 7(Suppl 1):8–15. doi:10.1111/j.1399-5618.2005.00189.x. [PubMed: 15762864]
- Bech P. Rating scales in depression: limitations and pitfalls. *Dialogues Clin Neurosci*. 2006; 8(2):207–215. [PubMed: 16889106]
- Beck, AT.; Steer, RA.; Brown, GK. *Manual for the Beck Depression Inventory II*. Psychological Corporation; San Antonio, TX: 1996.
- Berk M, Malhi GS, Cahill C, Carman AC, Hadzi-Pavlovic D, Hawkins MT, et al. The Bipolar Depression Rating Scale (BDRS): its development, validation and utility. *Bipolar Disord*. 2007; 9(6):571–579. doi:10.1111/j.1399-5618.2007.00536.x. [PubMed: 17845271]
- Berk M, Malhi GS, Mitchell PB, Cahill CM, Carman AC, Hadzi-Pavlovic D, et al. Scale matters: the need for a Bipolar Depression Rating Scale (BDRS). *Acta Psychiatr Scand Suppl*. 2004; (422):39–45. doi:10.1111/j.1600-0447.2004.00412.x. [PubMed: 15330937]
- Buysse DJ, Reynolds CF 3rd, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res*. 1989; 28(2):193–213. [PubMed: 2748771]
- Dell'Osso L, Pini S, Cassano GB, Mastrocinque C, Seckinger RA, Sacttoni M, et al. Insight into illness in patients with mania, mixed mania, bipolar depression and major depression with psychotic features. *Bipolar Disord*. 2002; 4(5):315–322. [PubMed: 12479664]
- Embretson, SE.; Reise, SP. *Item Response Theory for Psychologists*. Erlbaum; Mahwah, NJ: 2000.
- Gelenberg AJ, Thase ME, Meyer RE, Goodwin FK, Katz MM, Kraemer HC, et al. The history and current state of antidepressant clinical trial design: a call to action for proof-of-concept studies. *J Clin Psychiatry*. 2008; 69(10):1513–1528. [PubMed: 19192434]
- Hamilton M. A RATING SCALE FOR DEPRESSION. *Journal of Neurology, Neurosurgery, and Psychiatry*. 1960; 23(1):56–62. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC495331/>.
- Harvey AG, Schmidt DA, Scarna A, Semler CN, Goodwin GM. Sleep-related functioning in euthymic patients with bipolar disorder, patients with insomnia, and subjects without sleep problems. *Am J Psychiatry*. 2005; 162(1):50–57. doi:10.1176/appi.ajp.162.1.50. [PubMed: 15625201]
- Harvey PD, Endicott JM, Loebel AD. The factor structure of clinical symptoms in mixed and manic episodes prior to and after antipsychotic treatment. *Bipolar Disord*. 2008; 10(8):900–906. doi:10.1111/j.1399-5618.2008.00634.x. [PubMed: 19594505]
- Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull*. 1995; 117(1):167–178. [PubMed: 7870860]
- Horn JL. A RATIONALE AND TEST FOR THE NUMBER OF FACTORS IN FACTOR ANALYSIS. *Psychometrika*. 1965; 30:179–185. [PubMed: 14306381]
- Keck PE Jr. Defining and improving response to treatment in patients with bipolar disorder. *J Clin Psychiatry*. 2004; 65(Suppl 15):25–29. [PubMed: 15554793]
- Kline, RB. *Principles and Practice of Structural Equation Modeling*. Guilford Press; 2005.
- Kogan JN, Otto MW, Bauer MS, Dennehy EB, Miklowitz DJ, Zhang HW, et al. Demographic and diagnostic characteristics of the first 1000 patients enrolled in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). *Bipolar Disord*. 2004; 6(6):460–469. doi:10.1111/j.1399-5618.2004.00158.x. [PubMed: 15541061]
- Licht RW. Limitations in randomised controlled trials evaluating drug effects in mania. *Eur Arch Psychiatry Clin Neurosci*. 2001; 251(Suppl 2):i66–71. [PubMed: 11824841]
- Martinez-Aran A, Vieta E, Chengappa KN, Gershon S, Mullen J, Paulsson B. Reporting outcomes in clinical trials for bipolar disorder: a commentary and suggestions for change. *Bipolar Disord*. 2008; 10(5):566–579. doi:10.1111/j.1399-5618.2008.00611.x. [PubMed: 18657241]

- Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979; 134:382–389. [PubMed: 444788]
- Muthen BO. A Method for Studying the Homogeneity of Test Items with Respect to Other Relevant Variables. *Journal of Educational Statistics*. 1985; 10(2):121–132. doi:10.2307/1164839.
- Muthen BO. Dichotomous factor analysis of symptom data. *Sociological Methods and Research*. 1989; 18:19–65.
- Muthen, LK.; Muthen, BO. *Mplus User's Guide*. Sixth Edition. Muthen & Muthen; Los Angeles, CA: 2011.
- Nolen WA. Outcome measures in treatment trials in bipolar disorder. *Bipolar Disord*. 2002; 4(Suppl 1): 64–65. [PubMed: 12479681]
- Perlis RH, Brown E, Baker RW, Nierenberg AA. Clinical features of bipolar depression versus major depressive disorder in large multicenter trials. *Am J Psychiatry*. 2006; 163(2):225–231. doi: 10.1176/appi.ajp.163.2.225. [PubMed: 16449475]
- Phillips ML, Swartz HA. A critical appraisal of neuroimaging studies of bipolar disorder: toward a new conceptualization of underlying neural circuitry and a road map for future research. *Am J Psychiatry*. 2014; 171(8):829–843. doi:10.1176/appi.ajp.2014.13081008. [PubMed: 24626773]
- Picardi A. Rating scales in bipolar disorder. *Current Opinion in Psychiatry*. 2009; 22(1):42–49. doi: 10.1097/YCO.0b013e328315a4d2. [PubMed: 19122534]
- Rush AJ, Post RM, Nolen WA, Keck PE Jr. Suppes T, Altshuler L, et al. Methodological issues in developing new acute treatments for patients with bipolar illness. *Biol Psychiatry*. 2000; 48(6): 615–624. [PubMed: 11018232]
- Sachs GS, Thase ME, Otto MW, Bauer M, Miklowitz D, Wisniewski SR, et al. Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (STEP-BD). *Biol Psychiatry*. 2003; 53(11):1028–1042. [PubMed: 12788248]
- Samejima F. ESTIMATION OF LATENT ABILITY USING A RESPONSE PATTERN OF GRADED SCORES1. *ETS Research Bulletin Series*. 1968; 1968(1):i–169. doi:10.1002/j. 2333-8504.1968.tb00153.x.
- Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*. 1998; 59(Suppl 20): 22–33. quiz 34–57. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9881538>. [PubMed: 9881538]
- Weinstock LM, Strong D, Uebelacker LA, Miller IW. Differential item functioning of DSM-IV depressive symptoms in individuals with a history of mania versus those without: an item response theory analysis. *Bipolar Disord*. 2009; 11(3):289–297. doi:10.1111/j.1399-5618.2009.00681.x. [PubMed: 19419386]
- Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry*. 1978; 133:429–435. [PubMed: 728692]

HIGHLIGHTS

- The YMRS and MADRS contained several items that provided little test information
- Most item response categories did not provide incremental information.
- Both tests poorly measured participants outside of a narrow band of severity.
- Nearly every item of the YMRS and MADRS evidenced differential item functioning.

Figure 1.

Category Response Curves for each YMRS item. Curves indicate the probability of individuals at varying levels of manic symptom severity choosing each response option.

Figure 2.

Item and test information curves for the YMRS. Curves indicate the amount of psychometric information (i.e., the reciprocal of the standard error of measurement) provided by each item, as well as the overall test, at each level of manic symptom severity.

Figure 3. Category Response Curves for each MADRS item. Curves indicate the probability of individuals at varying levels of depressive symptom severity choosing each response option.

Figure 4.

Item and test information curves for the MADRS. Curves indicate the amount of psychometric information (i.e., the reciprocal of the standard error of measurement) provided by each item, as well as the overall test, at each level of depressive symptom severity.

Table 1
Item Response Theory model parameters from MADRS and YMRS Graded Response Models.

Item	Discrimination				Difficulty							
	α	B_M	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8		
Apparent Sadness	MADRS 1	2.64	3.00	-1.27	0.38	2.41	3.41	5.66	7.39			
Reported Sadness	MADRS 2	3.23	2.52	-1.80	-0.66	1.34	2.87	5.65	7.74			
Inner Tension	MADRS 3	1.31	1.78	-1.28	-0.75	0.71	2.04	4.24	5.73			
Reduced Sleep	MADRS 4	0.68	1.45	-0.05	0.18	0.77	1.26	2.67	3.86			
Reduced Appetite	MADRS 5	0.91	2.64	0.82	1.02	1.85	2.85	4.04	5.23			
Concentration Difficulties	MADRS 6	1.57	1.45	-1.32	-0.87	0.33	1.34	3.86	5.22			
Lassitude	MADRS 7	2.14	2.08	-1.24	-0.38	1.05	1.74	4.45	6.87			
Inability to Feel	MADRS 8	2.60	2.70	-0.49	0.25	1.67	2.67	5.27	6.81			
Pessimistic Thoughts	MADRS 9	2.24	2.55	-1.37	-0.51	1.12	2.48	5.46	8.10			
Suicidal Thoughts	MADRS 10	1.87	3.64	0.69	1.48	2.70	3.58	5.50	7.89			
Elevated Mood	YMRS 1	2.47	4.45	1.23	2.88	5.46	8.21					
Increased Motor Activity-Energy	YMRS 2	2.83	4.19	1.44	3.25	4.32	7.75					
Sexual Interest	YMRS 3	1.26	4.75	2.22	2.95	4.83	9.00					
Sleep	YMRS 4	0.99	3.10	0.65	1.23	3.20	7.32					
Irritability	YMRS 5	1.38	4.04	-0.89	-0.18	1.48	2.87	4.58	5.72	8.01	10.69	
Speech	YMRS 6	2.18	4.36	1.18	1.72	2.71	3.22	4.68	5.30	7.54	8.51	
Language-thought Disorder	YMRS 7	1.89	4.25	0.46	2.00	5.25	9.28					
Content	YMRS 8	1.88	4.67	1.39	2.43	3.79	4.61	5.35	5.73	6.84	7.18	
Disruptive-Aggressive Behavior	YMRS 9	2.15	7.04	1.49	2.90	4.99	6.55	8.33	9.25	10.95	11.88	
Appearance	YMRS 10	0.37	4.68	1.43	3.11	5.88	8.28					
Insight	YMRS 11	0.54	4.05	2.78	4.09	4.58	4.74					

Note. Discrimination (α) refers to an item's ability to discriminate between different latent levels of manic (YMRS) or depressive (MADRS) symptom severity (i.e., theta, θ). Difficulty parameters (β) reflect the standardized latent level of symptom severity at which subsequent response options become more probable than the previous option.

Table 2

Standardized regression coefficients for covariates predicting manic symptom severity and odds ratios for covariates predicting individual YMRS items.

Outcomes	Covariates												
	Age	Sex (M=0)	Substance Dependence	Anxiety Disorder	MDE	ME	MIXE	HYPME	Age of BP Onset	Rapid Cycling	BP II	NOS	Other
Latent Trait (β)													
Manic Symptom Severity	−0.06	−0.05	0.04	0.11	0.02	0.16	0.19	0.19	−0.08	0.08	0.01	0.01	0.02
YMRS items (OR)													
Elevated Mood	1.01	1.13	1.03	0.90	0.40*	1.23	0.55	1.48	1.04	1.27	0.97	0.85	2.05
Increased Activity	1.00	1.16	1.21	1.00	0.59	1.29	1.56	2.08	0.95	1.06	1.10	0.78	1.25
Sexual Interest	1.00	0.81	1.22	1.18	0.54*	1.15	0.66	0.96	1.02	1.32	0.88	0.90	1.71
Sleep	1.01	1.26	1.01	1.22	1.51*	2.52	2.38	2.59	0.97	1.19	1.22	1.33	1.05
Irritability	0.99	1.25	1.30	1.63*	2.60*	1.94	4.80*	1.85	0.87	1.38	1.35	1.23	1.05
Speech	1.02	1.34	1.06	0.88	0.76	2.05	1.71	1.99	0.96	0.96	1.13	1.21	0.66
Language-Thought	1.00	1.29	0.94	1.02	1.53	1.92	2.08	1.71	0.94	1.11	1.07	1.43	1.08
Content	1.01	1.15	1.02	1.07	0.63*	1.21	0.91	1.26	0.97	1.07	0.82	0.82	1.69
Disruptive Behavior	0.99	1.19	1.45	1.50	1.91*	2.51	2.90	1.98	0.92	1.31	1.47	1.48	1.05
Appearance	1.01	1.02	1.21	1.33	2.27*	1.36	1.66	1.20	1.03	1.12	0.67*	0.89	0.63
Insight	1.01	0.76	1.32	1.54	0.60*	1.89	0.37	0.91	1.09	0.98	1.17	3.15	2.28

Note. Statistically significant ($p < 0.01$) coefficients are bolded. Statistically significant and clinically significant (i.e., estimated Cohen's $D > 0.20$) coefficients are starred. ref = reference group for a series of dummy coded covariates. BP = Bipolar. M = Male. MDE = Major Depressive Episode. ME = Manic Episode. MIXE = Mixed Episode. HYPME = Hypomanic Episode. Increased Activity = Increased Motor Activity-Energy. Language-Thought = Language-Thought Disorder. Disruptive Behavior = Disruptive-Aggressive Behavior.

Table 3

Standardized regression coefficients for covariates predicting depressive symptom severity and odds ratios for covariates predicting individual MADRS items.

Outcome	Covariates												
	Age	Sex (M=0)	Substance Dependence	Anxiety Disorder	Episode Status (ref=Euthymic)			Age of BP Onset	Rapid Cycling	BP Subtype (ref=BPI)			
					MDE	ME	MIXE	HYPME			BPII	NOS	Other
Latent Trait (β)													
Depressive Symptom Severity	-0.08	-0.07	0.02	0.13	0.46	0.04	0.20	0.02	-0.06	0.05	0.00	-0.01	-0.03
MADRS items (OR)													
Apparent Sadness	1.02	1.59*	1.20	1.27	2.27*	0.84	1.27	0.54*	1.18	0.99	0.86	0.98	0.79
Reported Sadness	1.02	1.47*	1.26	0.84	2.74*	0.98	1.34	0.69	1.15	1.13	1.16	1.20	0.80
Inner Tension	1.00	1.24	1.17	1.55*	1.34	2.59*	1.87*	1.65	0.93	1.16	1.13	1.39	1.58
Reduced Sleep	1.01	1.24	1.03	1.32	1.20	3.72*	2.77*	3.13*	0.92	1.29	1.13	1.05	0.96
Reduced Appetite	1.00	1.25	1.29	1.23	1.28	1.81*	1.40	1.72	0.99	1.17	0.92	1.07	0.78
Concentration	1.01	1.56*	1.14	1.19	1.85*	1.69	2.42*	2.34*	0.97	1.12	1.08	1.15	1.00
Lassitude	1.01	1.72*	1.17	1.05	1.79*	0.39*	1.02	0.78	1.10	1.16	1.01	0.92	1.11
Inability to Feel	1.01	1.12	1.03	0.98	2.21*	0.64	1.66*	0.74	1.14	1.19	1.20	1.39	0.98
Pessimistic Thgts.	1.01	1.19	1.10	1.07	1.37	0.63	1.25	0.76	1.01	1.11	1.13	1.08	0.85
Suicidal Thoughts	1.01	1.17	1.15	1.16	1.69*	1.28	1.33	0.87	0.97	1.13	1.09	1.06	1.44

Note. Statistically significant ($p < 0.01$) coefficients are bolded. Statistically significant and clinically significant (i.e., estimated Cohen's $D > 0.20$) coefficients are starred. ref = reference group for a series of dummy coded covariates. BP = Bipolar. M = Male. MDE = Major Depressive Episode. ME = Manic Episode. MIXE = Mixed Episode. HYPME = Hypomanic Episode. Concentration = Concentration Difficulties. Thgts. = Thoughts.