Three Studies on Self-Report Scales to Detect Bipolar Disorder

By: Christopher J. Miller, Sheri L. Johnson, Thomas R. Kwapil, and Charles S. Carver

Miller, C.J., Johnson, S.L., Kwapil, T.R., & Carver, C.S. (2011). Three studies on self-report scales to detect bipolar disorder. *Journal of Affective Disorders, 128*(3), 199-210. DOI:10.1016/j.jad.2010.07.012.

Abstract:

Background-This study investigated the usefulness of self-report scales for detecting bipolar disorder in several settings.
Methods-Study 1 developed a short form of the Hypomanic Personality Scale (the HPS-6) based on clinic/community and undergraduate samples. Study 2 used this scale for recruiting participants with bipolar disorder from the community. Study 3 administered the full-length Hypomanic Personality Scale, the Mood Disorder Questionnaire, and a short form of the General Behavior Inventory (the GBI-15) to an undergraduate sample. Each study featured a reference standard diagnostic interview.
Results-In Study 2, about half of those responding to the advertisement (based on the HPS-6 developed in Study 1) reported a history of at least one hypomanic episode on a telephone-based SCID. In Study 3, the most robust findings emerged for the GBI-15: about one-third of participants screening positive on that measure met criteria for bipolar disorder using the Structured Clinical Interview for the DSM-IV (SCID).
Limitations-Despite large sample sizes and stratified sampling, this study was limited by a low number of participants with bipolar I disorder.
Conclusions-These three studies produced mixed findings regarding the detection of bipolar disorder via self-report. The HPS-6 was reasonably successful in recruiting participants with a history of at least one manic or hypomanic episode into a study on bipolar disorder. The GBI-15 showed some promise as a screening tool in an undergraduate setting, but there is a need for more sensitive and specific scales. Discussion focuses on potential strategies for developing such scales.

Article:

INTRODUCTION
Bipolar disorder is associated with high health care costs, lost productivity, elevated medical comorbidity, hospitalization, and the risk of suicide (Andlin-Sobocki and Wittchen, 2005, Baldessarini and Tondo, 2003, Kupfer, 2005 and Peele et al., 2003). The lifetime prevalence of bipolar I disorder is reported to be approximately 1% (Judd and Akiskal, 2003, Kessler et al., 1997, Kessler et al., 2005, Kessler et al., 1994, Narrow et al., 2002, Regier et al., 1993 and Weissman et al., 1996). For bipolar II disorder, some authors argue that estimates of 1.1% (Merikangas et al., 2007) are artifactually low, due to poor sensitivity of diagnostic

measures (Angst and Cassano, 2005). Estimates of the prevalence of cyclothymia range from 0.4% (American Psychiatric Association, 2000) up to 4% (Regeer et al., 2004).

Even though bipolar disorder affects many people, it is often misdiagnosed. Studies conducted in the United States and Europe indicate that 26 to 70% of people with bipolar disorder may either receive an incorrect diagnosis or go undiagnosed altogether. Proper diagnosis may not occur for a decade after bipolar symptoms emerge, especially for those with bipolar II disorder (Ghaemi et al., 2005, Lewis, 2004, Lish et al., 1994 and Mantere et al., 2004). A significant percentage of those presenting with unipolar depression may actually suffer from some type of bipolar disorder (Dunner, 2003).

Several factors may contribute to this lack of recognition of the illness (Benazzi et al., 2004). Most people with bipolar disorder who seek treatment do so because of their depression rather than mania (Hirschfeld et al., 2005) and perceive their depression to be more troublesome than their mania (Calabrese et al., 2004). Some people with bipolar disorder lack insight into the presence or consequences of their manic symptoms. Hypomanic episodes, by definition, do not cause severe impairment or distress, so they are less likely to be defined as a problem by people who experience them. When people with bipolar disorder seek treatment for depression, their symptom profile may be virtually indistinguishable from that of unipolar depression (Cuellar et al., 2005). In addition, many clinicians do not screen for bipolar disorder, even when a client presents with depressive symptoms (Brickman et al., 2002).

Unfortunately, the consequences of misdiagnoses are serious. Misdiagnosis as unipolar depression can lead to antidepressant monotherapy, which may trigger mania or rapid cycling in up to 50% of people with bipolar disorder (Ghaemi et al., 2004).

One potential solution to the problem of misdiagnosis could be the use of brief self-report scales to alert providers to the possibility of bipolar disorder. This would allow clinicians to complete a more rigorous diagnostic assessment with patients who seem most likely to have a bipolar disorder. Several such scales have been developed, including the Hypomanic Personality Scale (HPS; Eckblad and Chapman, 1986), the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000), the Temperament Evaluation of Memphis, Pisa, Paris, and San Diego—Autoquestionnaire version (TEMPS-A; Akiskal et al., 2005), the Bipolar Spectrum Diagnostic Scale (BSDS; Ghaemi et al., 2005), the General Behavior Inventory (GBI; Depue et al., 1989), and the Hypomania Checklist (HCL; Angst et al., 2005).

Each of these scales has strengths. For example, they generally have high internal consistency, and each yields significantly higher scores in patient populations than among healthy controls. Several of them have demonstrated at least modest sensitivity and specificity. For instance, the MDQ had a sensitivity of .73 and a specificity of .90 in one study (Hirschfeld et al., 2000). More than three quarters of a sample of students scoring high on the HPS had a history of at least one hypomanic episode; none with low scores had a hypomanic history (Eckblad and Chapman, 1986). The GBI has shown a sensitivity of .78 and a specificity of greater than .98 in two studies (Depue et al., 1981 and Klein et al., 1989). The HCL (32-item version) has more limited specificity in detecting bipolar disorder (Angst et al., 2005); however, this may reflect its goal of detecting hypomania, which is less easily distinguished from normal functioning. The TEMPS-A

has been studied extensively in Europe, and was the subject of an issue of the Journal of Affective Disorders (Akiskal et al., 2005); however, sensitivity and specificity estimates for this measure have not yet been published.

Some measures are too long to be widely used as screeners. For example, although the GBI has been well-validated (Angst et al., 2005), its full 73-item version can take more than 20 min to complete. Shorter versions of the GBI have been developed (Lewinsohn et al., 2003 and Meyer and Hautzinger, 2003), but have not been validated in other samples. In sum, successful detection of bipolar disorder via self-report scales faces difficult challenges. First, it is not clear that existing scales have adequate sensitivity to bipolar disorder for use in routine screening. Second, the length of some scales (e.g., the GBI) limits their utility for screening.

The studies reported here were designed to add to the literature on the detection of bipolar disorder by further testing measures that have shown promising sensitivity and specificity in previous research. Study 1 developed a brief version of the HPS and tested it across samples. The HPS was chosen because it has yet to be applied to clinical samples and may hold promise for detecting bipolar II disorder due to its attention to hypomanic traits. Study 2 used the small set of items derived from Study 1 in an advertising campaign to recruit research participants into a study on bipolar disorder. Study 3 involved administration of screening questionnaires (including the HPS-6 produced in Study 1 and tested in Study 2) to a large sample of undergraduates, yielding a sample enriched for vulnerability to bipolar disorder that was studied in detail. That study included the HPS due to its long-standing application to undergraduate populations, and the MDQ because it has been extensively studied in clinical settings. The only study to use the MDQ in an undergraduate population did not include a diagnostic interview to examine correspondence between the MDQ and bipolar diagnoses (Chandler et al., 2008). The complete GBI is too long to serve as a screening measure, but we included a shortened version, the GBI-15 (Meyer and Johnson, 2003) based on the existing psychometric evidence for the 73-item version.

STUDY 1

The goal of Study 1 was to develop a briefer version of the HPS that could be used to reliably screen for bipolar disorder. The original HPS has 48 true–false items and takes about 7 min to complete. The goal was to create a subset of items that could be completed in less than 2 min, which would be appropriate for advertisements or posters. Such a set of items then could be used to recruit potentially bipolar participants for studies through notices in papers or public transportation.[1]

Methods

Methods for Study 1 were approved by the Institutional Review Boards at the University of Miami and the University of Wisconsin-Madison before data collection.

Participants. This study included participants from two samples. One was an undergraduate sample, studied earlier by Eckblad and Chapman (1986). That sample consisted of 40 college students with *z*-scores of less than .50 on the HPS, and 40 students with *z*-scores above 1.67 (for women) or 1.82 (for men) on the HPS. The undergraduate sample was 50% female and had an average age of 19.41 years; total age range was from 16 to 24 years, and all of the participants were White non-Hispanic. The second sample was a clinic/community sample, and included

participants from a large Day Hospital program for mood disorders (n = 90) as well as participants recruited through community advertising (n = 37). Participants from the Day Hospital had diagnoses of major depression or bipolar disorder, while participants from the community did not have any mood disorder diagnosis. This combined clinic/community sample was 62% female, 13% African-American, 27% Hispanic, 58% White non-Hispanic, and 3% of other or mixed race/ethnicity; age for this subsample was 39.55 ± 11.87 years with a total range of 20 to 65 years.

Measures. As stated earlier, the study focused on the HPS. Bipolar diagnoses (including bipolar I or bipolar II disorder) based on the Schedule for Affective Disorders and Schizophrenia (SADS-L; Endicott and Spitzer, 1978) served as the reference standard for the undergraduate sample, while the Structured Clinical Interview for DSM-IV (First and Gibbon, 2004) served as the reference standard for the clinic/community sample.

Hypomanic Personality Scale (HPS). The HPS (Eckblad and Chapman, 1986) is a 48-item measure designed to assess subsyndromal manic symptoms. Items are in a true/false format; sample items include “Sometimes ideas and insights come to me so fast that I cannot express them all” (keyed true) and “I am usually in an average sort of mood, not too high and not too low” (keyed false). The full 48-item HPS has shown adequate ability to detect hypomanic episodes in undergraduates (Eckblad and Chapman, 1986) and has predicted the onset of hypomania and mania in the same sample over an impressive thirteen-year follow-up (Kwapil et al., 2000). Across studies, persons with high HPS scores have demonstrated higher rates of mania and hypomania than those with low scores, as well as similarities to those with bipolar I disorder ([Bentall and Thompson, 1990], [Johnson et al., 2005] and [Meyer and Johnson, 2003]). The HPS has good reliability, with a Cronbach's alpha of .87 and a test–retest reliability of .81 over a fifteen-week period (Eckblad and Chapman, 1986).

Reference standard (undergraduate sample): Schedule for Affective Disorders and Schizophrenia, Lifetime Version (SADS-L). The SADS-L (Endicott and Spitzer, 1978) was designed to assess Axis I diagnoses. The reliability and validity of the SADS-L has been established across 21 studies (Rogers et al., 2003). The SADS-L has good to excellent reliability for both symptoms and diagnoses, including reliability for manic episodes ([Andreasen et al., 1981], [Coryell et al., 1995] and [Rice et al., 1986]).

Reference standard (clinic/community sample): Structured Clinical Interview for the DSM-IV (SCID). The SCID (First and Gibbon, 2004) is a semi-structured interview designed specifically to yield diagnoses based on DSM-IV criteria. This study involved administration of the mood and psychosis modules, allowing diagnoses of bipolar I and II disorders. It was also possible to obtain a diagnosis of bipolar NOS based on recurrent hypomania without any history of a major depressive episode, in line with DSM criteria (APA, 2000). The SCID was administered by graduate students who had completed detailed training in the SCID (involving written materials, didactic lectures, role plays, and reviews of interviews). SCID interviewers were blind to other study measures. Administration of the SCID was audio recorded, allowing reliability to be established under the direction of Sheri L. Johnson, PhD. Interviewers met weekly to review audiotapes and difficult diagnostic decisions. Interrater reliability for the diagnosis of bipolar was acceptable (κ = .84; Williams et al., 1992).

Procedure. Participants from each sample completed the HPS. As described, HPS scores had been used to select participants for the undergraduate sample. All participants in both samples then completed the reference standard SADS-L or SCID. More details on the procedures for the undergraduate sample can be found in Eckblad and Chapman (1986).

Results

Table 1 contains tetrachoric correlations (used for dichotomous variables) between individual HPS items and a bipolar disorder diagnosis from both the undergraduate sample and the clinic/community sample. Preliminary analyses were conducted to investigate the ability of the full HPS to detect bipolar disorder in both subsamples. The HPS correlated strongly with a bipolar diagnosis in both the undergraduate sample (r = .41, p < .001) and the clinic/community sample (r = .56, p < .001).

Table 1: Tetrachoric correlations between individual HPS items and bipolar disorder diagnosis in the undergraduate and mood clinic samples from Study 1.

Hypomanic Personality Scale items—Study 1	Correlation w/ initial bipolar (undergraduate sample)	Correlation w/ BP I diagnosis (mood clinic sample)	
1. I consider myself to be pretty much an average kind of person.	.29	.08	
2. It would make me nervous to play the clown in front of other people.	.20	−.01	
3. I am frequently so “hyper” that my friends kiddingly ask me what drug I'm taking.	.18	.38	
4. I think I would make a good nightclub comedian.	.13	.22	
5. Sometimes ideas and insights come to me so fast that I cannot express them all.	.14	.13	
6. When with groups of people, I usually prefer to let someone else be the center of attention.	.25	.03	
7. In unfamiliar surroundings, I am often so assertive and sociable that I surprise myself.	.25	.13	
8. There are often times when I am so restless that it is impossible for me to sit still.	.12	.47	
9. Many people consider me to be amusing but kind of eccentric.	.19	.24	
10. When I feel an emotion, I usually feel it with extreme intensity.	.16	.38	
11. I am frequently in such high spirits that I can't concentrate on any one thing for too long.	.29	.38	*
12. I sometimes have felt that nothing can happen to me until I do what I am meant to do in life.	.06	.19	
13. People often come to me when they need a clever idea.	.30	−.10	
14. I am no more self-aware than the majority of people.	.26	.19	
15. I often feel excited and happy for no apparent reason.	.26	.14	
16. I can't imagine that anyone would ever write a book about my life.	.31	.04	
17. I am usually in an average sort of mood, not too high and not too low.	.41	.41	*
18. I often have moods where I feel so energetic and optimistic that I feel I could outperform almost anyone at anything.	.19	.22	
19. I have such a wide range of interests that I often don't know what to do next.	.38	.26	*
20. There have often been times when I had such an excess of energy that I felt little need to sleep at night.	.26	.47	*
21. My moods do not seem to fluctuate any more than most people's do.	.34	.48	*
22. I very frequently get into moods where I wish I could be everywhere and do everything at once.	.31	.33	*
23. I expect that someday I will succeed in several	.23	.04	

different professions.			
24. When I feel very excited and happy, I almost always know the reason why.	.19	.36	
25. When I go to a gathering where I don't know anyone, it usually takes me a while to feel comfortable.	.25	−.05	
26. I think I would make a good actor, because I can play many roles convincingly.	.31	.08	
27. I like to have others think of me as a normal kind of person.	.33	.20	
28. I frequently write down the thoughts and insights that come to me when I am thinking especially creatively.	.20	.17	
29. I have often persuaded groups of friends to do something really adventurous or crazy.	.20	.08	
30. I would really enjoy being a politician and hitting the campaign trail.	.05	.15	
31. I can usually slow myself down when I want to.	.13	.42	
32. I am considered to be kind of a “hyper” person.	.25	.37	*
33. I often get so happy and energetic that I am almost giddy.	.31	.40	*
34. There are so many fields I could succeed in that it seems a shame to have to pick one.	.27	.22	
35. I often get into moods where I feel like many of the rules of life don't apply to me.	.32	.39	*
36. I find it easy to get others to become sexually interested in me.	.16	.09	
37. I seem to be a person whose mood goes up and down easily.	.10	.43	
38. I frequently find that my thoughts are racing.	.33	.38	*
39. I am so good at controlling others that it sometimes scares me.	.18	.24	
40. At social gatherings, I am usually the “life of the party”.	.13	.11	
41. I do most of my best work during brief periods of intense inspiration.	.08	.09	
42. I seem to have an uncommon ability to persuade and inspire others.	.26	.12	
43. I have often been so excited about an involving project that I didn't care about eating or sleeping.	.20	.30	
44. I frequently get into moods where I feel very speeded-up and irritable.	.15	.54	
45. I have often felt happy and irritable at the same time.	.26	.45	*
46. I often get into excited moods where it's almost impossible for me to stop talking.	.29	.46	*
47. I would rather be an ordinary success in life than a spectacular failure.	.17	.04	
48. A hundred years after I'm dead, my achievements will probably have been forgotten.	.08	−.12	

* Asterisks indicate items correlating strongly with bipolar diagnosis in both samples.

A subset of twelve items is marked by asterisks in the table. These items correlated relatively strongly with a bipolar disorder diagnosis in both samples ($r \geq .25$). Two are reverse-coded (item 17, “I usually tend to be in an average sort of mood, not too high and not too low”; and item 21, “My moods do not seem to fluctuate more than most people's do”), making them inappropriate for a simply-worded poster.

To keep the scale brief, only the six items that had the highest average correlation across both samples were retained. These six items appear to have adequate content validity, tapping constructs associated with mania such as high energy, little need for sleep, happy and/or irritable mood, racing thoughts, talkativeness, and a disregard for rules. This six-item version (hereafter the HPS-6; see Table 2) correlated strongly in the undergraduate sample with the full-length HPS ($r = .83$, $p < .001$). The HPS-6 also correlated as robustly as did the full HPS did with a history of at least one manic episode ($r = .43$, $p < .001$). This pattern emerged in the clinic/community

sample as well, with the HPS-6 correlating strongly with the full-length HPS (r = .81, p < .001) and with manic history (r = .63, p < .001). Alpha reliability for the full HPS was .93, as compared to .79 for the HPS-6. This difference can be accounted for by the relative length of the two scales according to the Spearman Brown prophecy equation (Stanley, 1971).

Table 2: The HPS-6.

1) There have often been times when I had such an excess of energy that I felt little need to sleep at night.
2) I often get so happy and energetic that I am almost giddy.
3) I often get into moods where I feel like many of the rules of life don't apply to me.
4) I frequently find that my thoughts are racing.
5) I have often felt happy and irritable at the same time.
6) I often get into excited moods where it's almost impossible for me to stop talking.

Further analyses suggested, however, that the cutoff required to achieve a positive screen would likely differ in the two populations. Table 3 contains sensitivity, specificity, and positive and negative predictive values for the HPS-6 and the full HPS at various cutoffs in each of the subsamples. Sensitivity and specificity in the undergraduate sample should be interpreted with caution, as stratified sampling biases these statistics.

Table 3: Sensitivity, specificity, and positive and negative predictive values for the HPS and HPS-6, Study 1.

	Sensitivity	Specificity	Positive predictive value (PPV)	Negative predictive value (NPV)
Full HPS (48 items)				
Clinic/community sample—cutoff of 25/48 items	.70*	.88*	.77	.83
HPS-6				
Undergraduate sample—cutoff of all six items	.88*	.81*	.56	.96
Clinic/community sample—cutoff of 3/6 items	.89	.73	.67	.92

* Asterisks indicate statistics that are biased due to stratified sampling.

Discussion

In Study 1 we set out to develop a briefer version of the HPS to screen for bipolar disorder. A small subset of items appeared to perform well in two different contexts: an undergraduate sample and a clinic/community sample. The HPS-6 worked as well as or better than the full HPS in detecting bipolar disorder in these samples. This is not surprising, as the items of the HPS-6 were chosen specifically because they were among the best-performing items in the samples themselves. Cutoffs of three items endorsed in the mood clinic or six items endorsed in the undergraduate samples resulted in good sensitivity (.88 or higher) and modest specificity (.73 or higher) in both samples. Given Zimmerman and Mattia's (2001) argument that a measure should maximize sensitivity to be used as a screening tool, it would seem that cutoffs of three out of six (clinic/community) or all six (undergraduate) items represent the best threshold to screen for bipolar disorder in these settings. This discrepancy in cutoffs may reflect the different psychometric properties required in predominantly clinical versus predominantly non-clinical samples (for another example of differential cutoffs, see Langbehn et al., 1999 and Trull and Amdur, 2001).

Several limitations should be acknowledged in interpreting these results. First, the primary outcome—presence of bipolar disorder—was assessed differently in the two subsamples. This

makes comparisons in psychometric properties across samples difficult, but also provides a more robust test of the generalizability of results. Second, the two subsamples (labeled as undergraduate and clinic/community) were collected many years apart, and had significantly different mean ages. It is therefore possible that the requirement for different cutoffs may be due to age or cohort effects, rather than the settings from which the subsamples were drawn. Third, our statistical analyses relied primarily on simple correlations and cross-tabulations of HPS scores and bipolar history. Given the sheer number of correlations investigated (each of 48 items in two samples), it is possible that some of the apparently strong items achieved that strength across samples merely by chance. This highlights the need for replicating findings, as described next.

Despite these limitations, the results from Study 1 suggest that a small subset of HPS items may screen for bipolar disorder more efficiently than the full 48-item HPS. Study 2 tested the application of these items in a real-world setting.

STUDY 2

The goal of Study 2 was to determine whether advertising featuring items derived from Study 1 could be used to recruit research participants with bipolar disorder. Few studies have investigated the usefulness of self-report scales in the general population. What little research exists suggests that self-report screens are likely insensitive to bipolar disorder in the community (Hirschfeld et al., 2003).

Methods

All procedures for this study were approved by the University of Miami Institutional Review Board. The study was conducted between July and September of 2007.

Participants. The sample included 100 people who responded to an advertisement posted on the Metrorail mass transit system in Miami-Dade County. The sample was 28% female, and the average age was 37.24 ± 11.99 years; total age range was from 18 to 59 years. The sample was racially heterogeneous, with 40% describing themselves as Caucasian, 40% as African-American, and 20% as of other or mixed race. In addition to endorsing one of the racial categories described above, 41% of the sample also described themselves as Hispanic or Latino/a.

Measures

Reference standard: Structured Clinical Interview for the DSM-IV (SCID). The SCID (First and Gibbon, 2004) was administered over the telephone to those who called to express interest in the study. Modules were included for past and present depression, mania, and hypomania, allowing diagnosis of bipolar I or II disorder. We did not administer the cyclothymia module over the telephone, given concerns about the low reliability of milder forms of the bipolar spectrum (Isometsä et al., 2003). The psychosis module was administered, allowing for the diagnosis of schizoaffective disorder. The SCID has adequate reliability when administered by telephone (Cacciola et al., 1999). A small subset of participants also completed the same SCID modules in person (see Section 3.2).

Procedure. An advertisement featuring the HPS-6 (developed in Study 1) was posted in 176 rail cars of the Miami-Dade County Metrorail system in the summer of 2007. It contained the items of the HPS-6, along with a heading asking "Does this describe you?" and an indication that persons endorsing three or more items could call to be screened for a separate study paying $25 per hour.

Participants who called to express interest in the study were read a brief script describing the research. Trained graduate students or research assistants asked basic demographic questions, and graduate students with substantial training on the SCID administered the modules described above over the telephone. Participants who appeared to meet criteria for bipolar I disorder based on the telephone interview and who also met other study criteria (i.e. did not have any medical rule-outs or current substance abuse) were invited to the University of Miami to take part in a larger study that included in-person administration of the same SCID modules.

Results

Eighty-four of the 100 respondents completed the SCID interview over the telephone. The remaining sixteen did not complete the telephone SCID for a variety of reasons, the most common of which were not speaking English or being unwilling to spend time answering questions over the telephone. Of the 84 people who completed the SCID, 41(49%) appeared to have no significant bipolar symptoms, 17 (20%) had bipolar I disorder, 12 (14%) had bipolar II disorder, four (5%) had substance-induced mania, three (4%) had schizoaffective disorder, and one (1%) had bipolar disorder NOS based on a history of recurrent hypomania without a history of depression. The remaining six participants (7%) reported a history of at least one hypomanic episode without a history of depression.

Eight participants with a tentative diagnosis of bipolar I disorder without a history of brain trauma, neurological disorder, or current substance abuse based on the telephone SCID were invited to complete an in-person SCID. Five had their diagnosis of bipolar I disorder confirmed, one had the diagnosis downgraded to bipolar II disorder, and two did not appear to have a bipolar spectrum disorder. Of the last two participants, one reported a history of hypomania without depression and one reported a history of an impulse control disorder that did not appear to qualify as a bipolar disorder.

Discussion

Results from Study 2 suggest that a brief (six-item) screener can be used to recruit participants from the community for a study on bipolar disorder. If one defines the bipolar spectrum as consisting solely of bipolar I, bipolar II, and bipolar NOS disorders, then one-third (28/84) of those who called in response to the ad appeared to be on the bipolar spectrum according to a telephone-based SCID. It may be argued, however, that a useful screening tool for bipolar disorder (focused on episodes of mania/hypomania without reference to drugs, psychosis, or depression) would also be expected to identify those with hypomania, schizoaffective disorder, or substance-induced mania. Viewed in this light, 50% of those who called in response to the ad might represent "successes" on the part of the HPS-6. That is, fully half of participants reported at least one manic or hypomanic episode over the phone. In addition, it is unclear whether the remaining participants truly represent "false alarms" on the part of the HPS-6. It is possible that

some or all of these participants had subclinical symptoms representing some type of vulnerability to bipolar disorder.

Some limitations should be kept in mind when considering these results. First, the design of Study 2 did not allow us to consider different thresholds; it remains possible that a higher threshold would have minimized false positive screens. A second limitation was that the HPS-6 was not readministered when participants called to express interest in the study. It is therefore possible that some participants actually endorsed fewer than three items on the advertisement, but called anyway due to misunderstanding of the wording, a desire to take part in paid studies even if they did not meet the exact eligibility criteria, etc. That we did not administer SCID modules for cyclothymia is another limitation, leaving open the possibility that some of our apparently non-bipolar participants would have met criteria for a milder form of bipolar spectrum disorder. A final limitation is that only those who appeared to have a bipolar I diagnosis on the telephone, and did not report any other study rule-outs, were invited to complete an in-person SCID. Unfortunately, bringing in more participants for a face-to-face SCID was not feasible. Past studies have found telephone-based diagnostic assessments to be reasonably reliable (e.g. Cacciola et al., 1999), but even so, results from Study 2 should be considered exploratory.

Taken together, the results from Study 1 and Study 2 paint something of a mixed picture regarding bipolar disorder screening. The HPS and HPS-6 had relatively strong correlations with bipolar diagnosis in Study 1. In Study 2, fully half of those who responded to an ad featuring the HPS-6 reported a history of at least one manic or hypomanic episode on a telephone-based SCID. Although the HPS-6 thus appears to be a helpful tool, its psychometric properties for detecting bipolar disorder leave room for improvement, in that about half of persons with high scores did not appear to meet criteria for any bipolar spectrum disorder. In Study 3 we compared the utility of this brief scale versus two alternative scales as screeners for bipolar disorder diagnoses.

STUDY 3

The first two studies focused on the usefulness of the HPS in detecting bipolar disorder. Few studies have directly compared the sensitivity and specificity of different self-report screening tools for bipolar disorder diagnoses. In one study, Youngstrom et al. (2004) found that adolescents with bipolar disorder scored .7 standard deviations higher on the GBI and the Youth Self-Report Form (Achenbach, 1991) than their non-bipolar counterparts. Sensitivity and specificity were not reported in that study. We are unaware of any studies, in fact, that have compared sensitivity and specificity indices in a young adult sample.

The goal of Study 3 was to assess whether any of three screeners could be used to detect bipolar disorder among undergraduates. In Study 3, participants completed the HPS, the MDQ, and the GBI-15 (the latter two of which are described in more detail in 4.1.2.2 and 4.1.2.3). Scores on those measures were used as potential predictors of SCID diagnoses of bipolar disorder.

Methods

All procedures for this study were approved by the University of Miami Institutional Review Board. This study was conducted between the fall of 2006 and the winter of 2007.

Participants. Participants were University of Miami undergraduates, who took part in the study in partial fulfillment of a course requirement. About 1200 students completed the three self-report screeners during group testing sessions on the first few days of the semester (the "screening sample"). Demographic information for the screening sample is limited, but most students were either eighteen or nineteen years old, and approximately 62% were female. Students who were selected using the stratified random sampling design (described in Section 4.1.3) were contacted via email and invited to participate in the study; those who chose to participate (N = 103) were enrolled. This final sample was 62% female, and the average age was 18.62 ± .83 years (range from 18 to 22 years). Six percent of participants were Asian-American, 9% were African-American, 17% were Hispanic/Latino, 60% were Caucasian, and 9% were of another or mixed ethnicity.

Measures. The primary measures for this study were three self-report screeners for bipolar disorder. For each of the screeners, cutoffs were chosen to create balanced sampling: that is, those who scored above a *z*-score of about 1.55 were given a positive screen. This cutoff aligns closely with those used in similar studies (Eckblad and Chapman, 1986). The Structured Clinical Interview for the DSM-IV (First and Gibbon, 2004) served as the reference standard.

Hypomanic Personality Scale (HPS). Details on the HPS (Eckblad and Chapman, 1986) are described in Study 1. For Study 3, endorsing 32 or more items on the full HPS resulted in a positive screen. Parallel analyses were also conducted using the HPS-6, described earlier.

Mood Disorder Questionnaire (MDQ). The MDQ (Hirschfeld et al., 2005 and Hirschfeld et al., 2003) is a self-report inventory designed to screen for lifetime history of a manic or hypomanic episode based on thirteen yes/no questions. The questions were derived from the DSM-IV criteria for bipolar disorder as well as clinical experience. Additional questions focus on whether the symptoms reported occurred at the same time (the "simultaneity" criterion) and caused at least moderate functional impairment (the "severity" criterion). In its original validation studies, achieving a positive screen required that a participant answer at least seven out of the 13 yes/no questions as "yes," and endorse both the simultaneity and severity criteria (Hirschfeld et al., 2005 and Hirschfeld et al., 2003). The MDQ has good internal consistency (Hirschfeld et al., 2000), and has been used in advertisements meant to promote awareness of bipolar disorder (AstraZeneca, 2006). In an initial study, the MDQ demonstrated a sensitivity of .73 and a specificity of .90 for detecting bipolar spectrum diagnoses (bipolar I, bipolar II, and bipolar NOS) in a mood disorders clinic (Hirschfeld et al., 2000). Subsequent studies, however, have found limited sensitivity to bipolar II disorder in an outpatient sample (Miller et al., 2004) and low sensitivity to bipolar disorders overall in the general population (Hirschfeld et al., 2003). In our sample the severity and simultaneity criteria were retained, but the cutoff for a positive screen was raised to 11 out of the 13 initial items of the MDQ due to the high rate of endorsement for many of the items (e.g. four of the thirteen items had an endorsement rate of at least 75%, and nearly 75% of the screening sample endorsed seven or more of the initial items).

General Behavior Inventory—15-item version (GBI-15). The GBI (Depue et al., 1989 and Depue et al., 1981) was designed to identify lifetime bipolar affective disorders. In its complete 73-item form, the GBI has been validated in both undergraduate (Depue et al., 1989) and outpatient

(Klein et al., 1989) samples. Items are rated on a one to four scale ranging from “never or hardly ever” to “very often or almost constantly.” This study used the 15-item version (the GBI-15; Meyer and Johnson, 2003). The GBI-15 contains three subscales: depression (six items), hypomanic (five items), and biphasic (four items referring to rapid mood fluctuations). In the original development study, each of these subscales correlated highly with its respective subscale in the full-length GBI ($r > .84$ in each case), and demonstrated good internal consistency (Cronbach's alpha >.70 in each case). Classification efficiency (using the full-length GBI as a gold standard) exceeded .92. The current study used only the latter two subscales in assigning a positive screen: endorsing four or more of the nine hypomanic/biphasic items represented a positive screen.

Reference standard: Structured Clinical Interview for the DSM-IV (SCID). As stated earlier, the reference standard for this study was the SCID, administered by highly trained graduate students. Once again, the mood, psychosis, and substance abuse modules were administered. In addition, a custom cyclothymia module (based on the DSM cyclothymia criteria and interview probes developed by Alloy's group; see Shen et al., 2008) was administered to capture this milder form of the bipolar spectrum. That is, a diagnosis of cyclothymia was given if the participant endorsed two years (or one year if under eighteen years old) of numerous high and low periods lasting days at a time. These periods were required to involve at least two manic (for high) and two depressed (for low) symptoms, and participants could not receive a cyclothymia diagnosis if they endorsed any two-month period that was free from such high and/or low periods. Ten percent of tapes were reviewed for reliability in the mania and hypomania modules. Disagreement arose only regarding one diagnosis for one participant, such that the intraclass correlation coefficient for diagnosis was .79 for past hypomania and absolute agreement (1.00) for diagnoses of current hypomania and current and past mania.

Procedure. The screening sample ($n = 1204$) completed the HPS, MDQ, and GBI-15 during the first week of class. Participants with high scores on the screening tools were oversampled. Specifically, all participants who screened positive on the HPS, MDQ, and/or the GBI-15 were invited to participate in an additional two-hour session as part of the final sample. In contrast, only a randomly selected 7% of those with low scores on at least one screener were invited to be a part of the final sample. Forty-eight percent of the high scorers participated in the final study, while 31% of the low scorers who were invited to take part in the study did so.

The final session involved completion of an informed consent form, the SCID, and other measures unrelated to the current report. The final sample included 23 persons with low scores on all three measures, and 80 persons with high scores on at least one screener. (Due to overlap in some high and low scores across measures, high scorers numbered 32 for the HPS, 35 for the MDQ, and 35 for the GBI-15. It was possible for participants to be enrolled as a “high-scorer” on one measure but a “low-scorer” on the others, resulting in control groups that also differed slightly in size across the three screening tools). See Table 4 for means and standard deviations for each of the screening tools. After this session, participants were debriefed and referrals were provided as needed.

Analysis plan. The primary outcome for these analyses was the presence of a bipolar spectrum diagnosis (bipolar I disorder, bipolar II disorder, cyclothymia, or bipolar NOS). Sensitivity and

specificity, reported earlier for Study 1, are biased by stratified sampling designs such as the one used here (Kraemer, 1992). Several existing methodologies can adjust for different types of stratification schemes (Choi, 1992, Sukhatme and Beam, 1994 and Weinstein et al., 1989). Of these methods, that described by Choi (1992) is most useful for easily deriving sensitivity and specificity in the context of stratification.

This study also used positive and negative predictive values, which are less biased by stratified sampling. Positive predictive value (PPV) is defined as the proportion of participants who screen positive on a particular measure who meet the relevant diagnostic criteria. Negative predictive value (NPV), in contrast, is defined as the proportion of participants who screen negative on a particular measure who fail to meet diagnostic criteria.

Results

Results are presented separately below for the screening sample (N = 1204) and the final sample (n = 103).

Screening sample. Table 4 contains information regarding the mean, SD, and scale reliability for each of the self-report screening tools. The Spearman Brown Prophecy Formula (Stanley, 1971) estimates expected reliability when scales are shortened; in this sample the HPS-6 appeared less reliable (alpha = .58) than would be expected (expected alpha = .70) given the reliability of the full HPS (alpha = .85).

Table 4: Mean, standard deviation, and reliability for screeners across different samples, Study 3.

	Sample	*N*	Mean	SD	Reliability
Full HPS	Screening sample	1174	19.43	8.01	.85
Full HPS	Screen-positive group	32	35.31	3.18	*
Full HPS	Screen-negative group	25	18.18	6.11	*
HPS-6	Screening sample	1171	2.22	1.62	.58
MDQ	Screening sample	1196	8.19	3.04	.78
MDQ	Screen-positive group	35	11.71	.79	*
MDQ	Screen-negative group	24	8.02	3.27	*
GBI-15	Screening sample	1199	.94	1.58	.75
GBI-15	Screen-positive group	35	5.20	1.47	*
GBI-15	Screen-negative group	24	.63	1.01	*
GBI-15—cyclothymic subscale	Screening sample	1197	.55	1.00	.70
GBI-15—hyperthymic subscale	Screening sample	1195	.39	.82	.59

*Due to stratified sampling, reliability statistics for these samples would be severely biased, and therefore are not included.

Final sample. Stratified random sampling resulted in a final sample size of 103. Table 4 contains information regarding mean, standard deviation, and reliability for each of the screening tools among both high and low scorers. Eighteen of the 103 participants met diagnostic criteria for a bipolar spectrum disorder, including seven with bipolar I disorder, six with bipolar II disorder, two with cyclothymia, and three with bipolar disorder NOS. This apparently high prevalence is not surprising, given this study's oversampling of participants with high scores on the self-report screeners.

Preliminary analyses tested whether demographic characteristics were related to scores on any of the three screeners in the final sample. Independent samples t-tests (for gender), one-way ANOVAs (for ethnicity), and correlations (for age) revealed no significant relationships with screening scores (all $t < 1.5$, $F < 1.2$, $r < .07$, n.s.; all Cohen's $d < .3$ for t-tests), with one

exception: age was significantly correlated with the GBI-15 (r = .25, p = .03). This result may be an artifact of the number of comparisons undertaken to identify confounds.[2]

Preliminary analyses also examined whether there was bias in which participants accepted invitations to take part in the second part of the study. Independent samples *t*-tests suggested that those who participated in the final sample did not differ significantly on HPS, MDQ, or GBI-15 scores (all *t*s < .6, n.s.) compared to those who declined invitations to be in the final sample.

Diagnostic efficiency statistics

Correlations among the self-report screeners (including the HPS-6) as well as their correlations with a bipolar spectrum diagnosis are presented in Table 5. Correlations among the screeners were generally in the small to moderate range. Point-biserial correlations with bipolar spectrum diagnosis were generally low for all of the screeners, ranging between .09 and .19 (all n.s.).

Table 5: Correlations among total scores for screeners and bipolar spectrum diagnosis, Study 3.

	HPS	MDQ	GBI-15	HPS-6	Bipolar spectrum diagnosis
HPS	–	.37**	.42**	.69**	.14
MDQ	–	–	.31**	.45**	.10
GBI-15	–	–	–	.43**	.19
HPS-6	–	–	–	–	.09

Note. Correlations between self-report screeners are based on the entire screened sample (N = 1204). Correlations involving bipolar spectrum diagnosis are drawn from the final sample (N = 103).
**p < .01.

Table 6 includes positive and negative predictive values, as well as sensitivity and specificity adjusted to account for stratified sampling (Kraemer, 1992). Each scale demonstrated high NPV (.88 to .92) and adjusted specificity values (.93 to .95). That is, most persons who obtained low scores on the screeners did not meet criteria for a bipolar spectrum diagnosis. The HPS and MDQ had relatively low PPV (.19 to .20) and adjusted sensitivity values (.09 to .11). The GBI-15 had somewhat higher, yet still modest, PPV and sensitivity. On the whole, only 20–30% of persons with a high score on a screener met diagnostic criteria for a bipolar spectrum disorder.

Table 6: Sensitivity, specificity, and positive and negative predictive values for each screener, Study 3.

Screener	Cutoff	PPV	NPV	Sensitivity (adjusted)	Specificity (adjusted)
HPS	32 items	.19 (6/32)	.88 (22/25)	.09	.95
MDQ	11/13 items plus simultaneity/severity	.20 (7/35)	.88 (21/24)	.11	.93
GBI-15	4 items	.31 (11/35)	.92 (22/24)	.26	.93

Note: Only 7 participants achieved a positive screen on the HPS-6 using a cutoff of all six items (based on Study 1), and so it has not been included in the table.

The analyses above used cutoffs that differed from those reported in the literature. Parallel analyses were conducted using previously recommended cutoffs, which were generally less stringent than those used above. These analyses were limited, as our stratification scheme had relied on the higher thresholds. Nonetheless, our limited analyses suggested that adjusting cutoffs did not yield marked improvements in sensitivity or specificity.

Discussion

The goal of Study 3 was to investigate the psychometric properties of three brief self-report screeners for detecting bipolar disorder among undergraduates. Sensitivity estimates for all three scales were low. That is, sensitivity to a bipolar spectrum diagnosis was less than .30 for each

screener, suggesting that each would miss more than 70% of bipolar spectrum diagnoses in similar college populations (and more than 85% in the cases of the HPS and MDQ). The adjusted NPV for the screeners in this study was high (above .93 in all three cases). That is, a participant scoring below the cutoff on a screener is unlikely to have a bipolar spectrum condition. These two statements—that the screeners would miss the majority of bipolar diagnoses, but that any given individual with a low screener score is unlikely to have bipolar disorder—may seem contradictory, but can be reconciled by the low base rate of bipolar disorder in typical undergraduate populations. These data illustrate how difficult it can be to detect a disorder with a low prevalence (e.g. suicidality, see Pokorny, 1983): even if specificity is relatively high, the number of false positives will quickly overwhelm the number of true positives.

The specificity for each scale was above .90, meaning that over 90% of those without bipolar disorder would be expected to screen negative using these tools. This number is generally consistent with previous findings on bipolar screening tools (Hirschfeld et al., 2005, Depue et al., 1981 and Klein et al., 1989). PPV for the HPS and MDQ generally fell between .15 and .20, suggesting that only one out of every five to six college students scoring above the cutoff on these measures would be expected to have a bipolar spectrum condition. The GBI-15 performed better, with about one in three high scorers achieving a bipolar spectrum diagnosis according to the SCID. Previous studies suggest that less than one in twenty undergraduates in an unscreened population might be expected to have a bipolar spectrum disorder (Lewinsohn et al., 1995). Accordingly, this PPV for the GBI-15 may be seen as encouraging, but also leaves room for improvement.

In sum, low scores on one of these measures are a fairly good indicator that a person would not meet criteria for a bipolar spectrum disorder. High scores, though, are only a rough indicator of likelihood of diagnosis. One potential solution would be to use higher scores as indicators, but doing so would reduce sensitivity even further, leaving a larger proportion of bipolar spectrum diagnoses undetected.

Overall, the screening tools in this study identified a large ratio of false positives to true positives, while also failing to identify a large proportion of bipolar spectrum diagnoses. Put another way, these tools label as bipolar many people who do not have a bipolar spectrum disorder, while simultaneously letting most of bipolar spectrum diagnoses slip through the cracks. This combination of a large number of false positives and low sensitivity to bipolar disorder is disheartening.

Scale-specific issues. There are several possible explanations for these findings. Some of the issues are specific to a given scale. In reference to the MDQ, the low rates of diagnosis (20% of those with a high score) were especially striking, because we used a higher cutoff (11 out of 13 items) for a positive screen than previous studies have used (7 items). One would expect that this would yield a sample at greater risk for a bipolar spectrum diagnosis than the traditional cut-off, but this was not the case. Many of the MDQ symptom items reference experiences that are quite common among college students (e.g., “have you ever had a time…when you were much more active or did many more things than usual?”). Most college students go through such times routinely for exams and other normative aspects of college life. Indeed, several items on the

MDQ were endorsed by more than three quarters of the screening sample, suggesting that these items may be of limited usefulness in college students.

The items on the MDQ that tap symptoms, however, may be only part of the problem. Previous research suggests that requiring that symptoms be rated as causing moderate impairment can impair the MDQ's diagnostic utility. This may be especially true for bipolar II disorder because by definition, hypomania (the defining characteristic of bipolar II disorder) is not characterized by significant impairment (APA, 2000). Therefore, both the symptom items and the severity criterion of the MDQ appear problematic.

In reference to the HPS, about half of the items refer to fairly constant patterns rather than episodic periods of symptoms (e.g., "I am so good at controlling others that it sometimes scares me." and "Many people consider me to be amusing but kind of eccentric."), whereas the DSM emphasizes the episodic. It is also the case that many items do not directly assess DSM-IV symptoms of mania, and no items assess the duration of symptoms. The lack of content overlap with DSM criteria might have limited the accuracy of this scale in predicting DSM bipolar spectrum diagnoses. More broadly, a focus on episodes of abnormally elevated mood and behavior, versus a focus on consistent or enduring traits, is an important one to consider in bipolar screening tools. The MDQ focuses on episodes, while the HPS and GBI-15 contain a mix of episodic- and personality/temperament-related items; this difference may help explain the modest correlations between total scores for these scales. As such, the HPS and GBI-15 may be better seen as vulnerability markers for bipolar disorder rather than screening tools per se; the performance of the HPS in predicting bipolar diagnoses over a thirteen-year follow-up lends weight to this interpretation (Kwapil et al., 2000).

The GBI-15 has other issues. Many of its items are grammatically complex (e.g., "Have you experienced periods of several days or more when, although you were feeling unusually happy and intensely energetic [clearly more than your usual self], you also were physically restless, unable to sit still and had to keep moving or jumping from one activity to another?"). This complexity presumably interferes with interpretability. On the other hand, the GBI-15 is the only scale of the three that specifically covers duration and frequency for each symptom assessed. This might explain the higher PPV estimates for this scale compared to the HPS or the MDQ.

Comparisons with prior research. Across the scales, these findings compare unfavorably to many previous reports regarding these same scales (e.g. Eckblad and Chapman, 1986, Hirschfeld et al., 2005, Ghaemi et al., 1999 and Depue et al., 1981). Why might this be? One possibility concerns differences among samples. Our sample differs from those used in some of the original validation studies in two ways. First, we did not use an inpatient sample. Differentiating mania from other inpatient disorders, such as schizophrenia, may be less challenging than differentiating mania from absence of disorder. Indeed, at least one other study has obtained similarly low sensitivity and specificity estimates for self-report screeners in the general population (Hirschfeld et al., 2003).

Second, this study focused on a young adult population. This sample is younger than those of most studies on adults, but older than studies of children or adolescents. In a recent meta-analysis focused on youth, Youngstrom et al. (2004) reported low sensitivity and specificity for self-

report scales among adolescents age 11–17. They found that parent report provided better diagnostic utility, suggesting that younger populations may be less accurate in reporting their bipolar symptoms. It is possible that our sample—with an average age of 18.62 years—similarly reflects such inaccuracy. Informant reports might have improved the validity of the screening tools.

Concerns about the young adult sample may apply to our gold standard SCID as well. Beyond the issues of self-awareness of symptoms in a younger sample, age may also influence the rate of symptom expression. The median age of onset for bipolar disorder is estimated to be around age 20–25 (American Psychiatric Association, 2000 and Kennedy et al., 2005). Thus some participants without a bipolar diagnosis at the time of the interview in the current study may develop episodes later. Indeed, Kwapil et al. (2000) found that the HPS predicted bipolar diagnoses at thirteen-year follow-up.

Other aspects of this study should be kept in mind when interpreting these results. Despite administering scales to over 1200 persons and oversampling high scorers on the scales, only seven participants received a diagnosis of bipolar I disorder, with another eleven receiving diagnoses elsewhere on the bipolar spectrum. This represents a difficulty inherent to studying a rare disorder in a non-clinical population. For this reason, study results should be interpreted with caution.

In summary, this study produced mixed results regarding the usefulness of the HPS, MDQ, and GBI-15 in detecting bipolar disorder among undergraduates. In terms of PPV, approximately one in three positive screens on the GBI-15 yielded a bipolar spectrum diagnosis. Despite this modestly encouraging result, none of the screeners showed an ability to successfully detect more than 30% of persons with a bipolar spectrum condition. To be useful in routine screening, a minimum sensitivity of .90 is recommended (Zimmerman and Mattia, 2001); thus none of these screening tools can be recommended for college populations.

GENERAL DISCUSSION AND FUTURE DIRECTIONS

This article presented results from three studies on the detection of bipolar disorder via brief self-report scales. In Study 1, the HPS was administered to an undergraduate and a clinic/community sample. This revealed a small subset of HPS items that correlated particularly well with bipolar diagnosis across samples, allowing creation of the briefer HPS-6. In Study 1, the HPS-6 had good sensitivity and modest specificity in both samples, although different cutoffs were required across samples.

In Study 2, the HPS-6 was used in an advertisement on the public transportation system in Miami to recruit participants with bipolar I disorder for a research study. On the SCID, half of the 84 callers who were interviewed reported a history of at least one manic or hypomanic episode.

Study 3 used stratified sampling among undergraduates, and administered the full-length HPS as well as the MDQ and GBI-15. None of these scales showed adjusted sensitivity to a bipolar spectrum disorder higher than .30. Moreover, PPV for the three screening tools fell in the range of .15–.30. This latter result was roughly comparable to results from Study 2, in which one in

three respondents had a bipolar spectrum disorder. Another potential explanation for the results of Study 3 is that many behaviors that might clearly reflect bipolarity in clinical settings (e.g. periods of intense activity with little sleep) may be normative in college populations.

What do these results suggest for future refinements to screening tools? One thing to consider is the scope of the questions that each screener contains. Of the self-report screeners investigated in these studies, only the GBI-15 references severity, frequency, and duration within each question; this may explain its marginally superior performance in Study 3. There is a need for a scale that captures these dimensions, yet uses simpler wording. In addition, measures that tap temperament rather than episodes may be useful in detecting bipolar disorder; ideally, future studies will subject such measures to analyses of sensitivity and specificity (e.g. the TEMPS-A; Akiskal et al., 2005). In sum, the current findings suggest that there is much room for improvement when it comes to self-report screening tools for bipolar disorder. However, several promising avenues for improving such scales have yet to be fully explored.

ROLE OF FUNDING SOURCE

Funding for these studies was supported by grants from the National Institute of Mental Health (NIMH R01-075021), the National Science Foundation (BCS0544617), the Scottish Rite Schizophrenia Research Program, and University of Miami Seed Funding. None of these organizations played any further role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and/or in the decision to submit the paper for publication.

CONFLICT OF INTEREST

None of the authors have any conflicts of interest.

REFERENCES

Achenbach, T.M., 1991. Manual for the Youth Self-Report Form and 1991 Profile. University of Vermont Department of Psychiatry, Burlington, VT.

Akiskal, H.S., Akiskal, K.K., Haykal, R.F., Manning, J.S., Connor, P.D., 2005. TEMPS-A: progress towards validation of a self-rated clinical version of the Temperament Evaluation of the Memphis, Pisa, Paris, and San Diego Autoquestionnaire. J. Affect. Disord. 85, 3–16.

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders, Text Revision 4th ed. American Psychiatric Publishing, Inc, Washington, DC.

Andlin-Sobocki, P., Wittchen, H.U., 2005. Cost of affective disorders in Europe. Eur. J. Neurol. 12 (Suppl. 1), 34–38.

Andreasen, N.C., Grove, W.M., Shapiro, R.W., Keller, M.B., Hirschfeld, R.M., McDonald-Scott, P., 1981. Reliability of lifetime diagnosis: a multicenter collaborative perspective. Arch. Gen. Psychiatry 38, 400–405.

Angst, J., Cassano, G., 2005. The mood spectrum: improving the diagnosis of bipolar disorder. Bipolar Disord. 7 (4), 4–12.

Angst, J., Adolfsson, R., Benazzi, F., Gamma, A., Hantouche, E., Meyer, T.D., et al., 2005. The HCL-32: towards a self-assessment tool for hypomanic symptoms in outpatients. J. Affect. Disord. 88, 217–233.

AstraZeneca, 2006. Is it more than just depression? Newsweek, CXLVII 9, 2–3.

Baldessarini, R.J., Tondo, L., 2003. Suicide risk and treatments for patients with bipolar disorder. JAMA 290, 1517–1519.

Benazzi, F., Koukopoulos, A., Akiskal, H.S., 2004. Toward a validation of a new definition of agitated depression as a bipolar mixed state (mixed depression). Eur. Psychiatry 19 (2), 85–90.

Bentall, R.P., Thompson, M., 1990. Emotional stroop performance and the manic defense. Br. J. Clin. Psychol. 29, 235–237.

Brickman, A., LoPiccolo, C., Johnson, S.L., 2002. Screening for bipolar disorder [Letter to the editor]. Psychiatr. Serv. 53, 349.

Cacciola, J.S., Alterman, A.I., Rutherford, M.J., McKay, J.R., May, D.J., 1999. Comparability of telephone and in-person Structured Clinical Interview for DSM-III-R (SCID) diagnoses. Assessment 6, 235–242.

Calabrese, J.R., Hirschfeld, R.M.A., Frye, M.A., Reed, M.L., 2004. Impact of depressive symptoms compared with manic symptoms in bipolardisorder: results of a U.S. community-Based sample. J. Clin. Psychiatry 65, 1499–1504.

Chandler, R.A., Wang, P.W., Ketter, T.A., Goodwin, G.M., 2008. A new US–UK diagnostic project: mood elevation and depression in first-year undergraduates at Oxford and Stanford universities. Acta. Psych. Scand. 118, 81–85.

Choi, B.C.K., 1992. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J. Clin. Epidemiol. 45, 581–586.

Coryell, W., Endicott, J., Maswer, J.D., Keller, M.B., Leon, A.C., Akiskal, H.S., 1995. Long-terms stability of polarity distinctions in the affective disorders. Am. J. Psychiatry 152, 385–390.

Cuellar, A.K., Johnson, S.L., Winters, R., 2005. Distinctions between bipolar and unipolar depression. Clin. Psychol. Rev. 25, 307–339.

Depue, R.A., Slater, J.F., Wolfstetter-Kausch, H., Klein, D., Goplerud, E., Farr, D., 1981. A behavioral paradigm for identifying persons at risk for bipolar depressive disorder: a conceptual framework and five validation studies. J. Abnorm. Psychol. 90, 381–437.

Depue, R.A., Krauss, S., Spoont, M.R., Arbisi, P., 1989. General behavior inventory identification of unipolar and bipolar affective conditions in a nonclinical university population. J. Abnorm. Psychol. 98, 117–126.

Dunner, D.L., 2003. Clinical consequences of under-recognized bipolar spectrum disorder. Bipolar Disord. 5, 456–463.

Eckblad, M., Chapman, L.J., 1986. Development and validation of a scale for hypomanic personality. J. Abnorm. Psychol. 95, 214–222.

Endicott, J., Spitzer, R.L., 1978. A diagnostic interview: the schedule for affective disorders and schizophrenia. Arch. Gen. Psychiatry 35, 837–844.

First, M.B., Gibbon, M., 2004. The Structured Clinical Interviewfor DSM-IV Axis I disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II disorders (SCID-II). In: Segal, D.L., Hilsenroth, M.J. (Eds.), Comprehensive Handbook of Psychological Assessment: Personality Assessment, 2. John Wiley & Sons, Inc, Hoboken, NJ, pp. 134–143.

Ghaemi, S.N., Sachs, G.S., Chiou, A.M., Pandurangi, A.K., Goodwin, F.K., 1999. Is bipolar disorder still underdiagnosed? Are antidepressants overutilized? J. Affect. Disord. 52, 135–144.

Ghaemi, S.N., Rosenquist, K.J., Ko, J.Y., Baldassano, C.F., Kontos, N.J., Baldessarini, R.J., 2004. Antidepressant treatment in bipolar versus unipolar depression. Am. J. Psychiatry 161, 163–165.

Ghaemi, S.N., Miller, C.J., Berv, D.A., Klugman, J., Rosenquist, K.J., Pies, R.W., 2005. Sensitivity and specificity of a new bipolar spectrum diagnostic scale. J. Affect. Disord. 84, 273–277.

Hirschfeld, R.M.A., Williams, J.B.W., Spitzer, R.L., Calabrese, J.R., Flynn, L., Keck, P.E.J., et al., 2000. Development and validation of a screening instrument for bipolar spectrum disorder: the mood disorder questionnaire. Am. J. Psychiatry 157, 1873–1875.

Hirschfeld, R.M.A., Holzer, C., Calabrese, J.R., Weissman, M., Reed, M., Davies, M., et al., 2003. Validity of the mood disorder questionnaire: a general population study. Am. J. Psychiatry 160, 178–180.

Hirschfeld, R.M.A., Cass, A.R., Holt, D.C., Carlson, C.A., 2005. Screening for bipolar disorder in patients treated for depression in a family medicine clinic. J. Am. Board Fam. Pract. 18, 233–239.

Isometsä, E., Suominen, K., Mantere, O., Valtonen, H., Leppämäki, S., Pippingsköld, M., Arvilommi, P., 2003. The mood disorder questionnaire improves recognition of bipolar disorder in psychiatric care. BMC Psychiatry 3 (8).

Johnson, S.L., Ruggero, C.J., Carver, C.S., 2005. Cognitive, behavioral, and affective responses to reward: links with hypomanic symptoms. J. Soc. Clin. Psychol. 24, 894–906.

Judd, L.L., Akiskal, H.S., 2003. The prevalence and disability of bipolar spectrum disorders in the US population: re-analysis of the ECA database taking into account subthreshold cases. J. Affect. Disord. 73, 123–131.

Kennedy, N., Everitt, B., Boydell, J., Van Os, J., Jones, P.B., Murray, R.M., 2005. Incidence and distribution of first-episode mania by age: results from a 35-year study. Psychol. Med. 35, 855–863.

Kessler, R.C., McGonagle, K.A., Zhao, S., Nelson, C.B., 1994. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results fromthe national comorbidity study.Arch. Gen. Psychiatry 51, 8–19.

Kessler, R.C., Rubinow, D.R., Holmes, C., Abelson, J.M., Zhao, S., 1997. The epidemiology of DSM-III-R bipolar I disorder in a general population survey. Psychol. Med. 27, 1079–1089.

Kessler, R.C., Chiu, W.T., Demler, O., Walters, E.E., 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. Arch. Gen. Psychiatry 62, 617–627.

Klein, D.N., Dickstein, S., Taylor, E.B., Harding, K., 1989. Identifying chronic affective disorders in outpatients: validation of the General Behavior Inventory. J. Consult. Clin. Psychol. 57, 106–111.

Klein, D.N., Lewinsohn, P.M., Seeley, J.R., 1996. Hypomanic personality traits in a community sample of adolescents. J. Affect. Disord. 38, 135–143.

Kraemer, H.C., 1992. Evaluating Medical Tests. Sage Publications, Newbury Park, CA, US.

Kupfer, D.J., 2005. The increasing medical burden in bipolar disorder. JAMA 293, 2528–2530.

Kwapil, T.R., Miller, M.B., Zinser, M.C., Chapman, L.J., Chapman, J., Eckblad, M., 2000. A longitudinal study of high scorers on the Hypomanic Personality Scale. J. Abnorm. Psychol. 109, 222–226.

Langbehn,D.R., Pfohl, B.M., Reynolds, S., Clark, L.A., Battaglia, M., Bellodi, L., et al., 1999. The Iowa Personality Disorder Screen: development and preliminary validation of a brief screening interview. J. Personal. Disord. 13, 75–89.

Lewinsohn, P.M., Klein,D.N., Seeley, J.R., 1995. Bipolar disorders in a community sample of older adolescents: prevalence, phenomenology, comorbidity, and course. J. Am. Acad. Child Adolesc. Psychiatry 34, 454–463.

Lewinsohn, P.M., Seeley, J.R., Klein, D.N., 2003. Bipolar disorder in adolescents: epidemiology and suicidal behavior. In: Geller, B., DelBello, M. (Eds.), Child and Early Adolescent Bipolar Disorder. Guilford, New York, pp. 7–24.

Lewis, L., 2004. Dual diagnosis: the Depression and Bipolar Support Alliance's patient perspective. Biol. Psychiatry 56, 728–729.

Lish, J.D., Dime-Meenan, S., Whybrow, P.C., Price, R.A., Hirschfeld, R.M., 1994. The National Depressive and Manic-Depressive Association (DMDA) survey of bipolar members. J. Affect. Disord. 31, 281–294.

Mantere, O., Suiminen, K., Leppamaki, S., Arvilommi, P., Isometsa, E., 2004. The clinical characteristics of DSM-IV bipolar I and II disorders: baseline findings from the Jorvi Bipolar Study (JoBS). Bipolar Disord. 6, 395–405.

Meads, D.M., Bentall, R.P., 2008. Rasch analysis and item reduction of the Hypomanic Personality Scale. Pers. Indiv. Differ. 44, 1772–1783.

Merikangas, K.R., Akiskal, H.S., Angst, J., Greenberg, P.E., Hirschfeld, R.M.A., Petukhova, M., Kessler, R.C., 2007. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey Replication. Arch. Gen. Psychiatry 64, 543–552.

Meyer, T.D., Hautzinger, M., 2003. Screening for bipolar disorders using the Hypomanic Personality Scale. J. Affect. Disord. 75, 149–154.

Meyer, B., Johnson, S. L., 2003. Brief screening for manic-depressive vulnerability: a short (15-Item) version of the General Behavior Inventory (GBI). Manuscript in preparation 2003.

Miller, C.J., Klugman, J., Berv, D.A., Rosenquist, K.J., Ghaemi, S.N., 2004. Sensitivity and specificity of the Mood Disorder Questionnaire for detecting bipolar disorder. J. Affect. Disord. 81, 167–171.

Narrow,W.E., Rae,D.S., Robins, L.N., Regier, D.A., 2002. Revised prevalence based estimates of mental disorders in the United States: using a clinical significance criterion to reconcile 2 surveys' estimates. Arch.Gen. Psychiatry 59, 115–123.

Peele, P.B.,Xu, Y., Kupfer, D.J., 2003. Insurance expenditures on bipolar disorder: clinical and parity implications. Am. J. Psychiatry 160, 1286–1290. Pokorny, A.D., 1983. Prediction of suicide in psychiatric patients: report of a prospective study. Arch. Gen. Psychiatry 40, 249–257.

Regeer, E.J., ten Have, M., Rosso, M.L., van Roijen, L.H., Vollebergh, W., Nolen, W.A., 2004. Prevalence of bipolar disorder in the general population: a reappraisal study of the Netherlands mental health survey and incidence study. Acta Psychiatr. Scand. 110, 374–382.

Regier, D.A., Narrow, W.E., Rae, D.S., Manderscheid, R.W., 1993. The de facto US mental and addictive disorders service system: epidemiologic catchment area prospective 1-year prevalence rates of disorders and services. Arch. Gen. Psychiatry 50, 85–94.

Rice, J.P., McDonald-Scott, P., Endicott, J., Coryell, W., Grove, W.M., Keller, M.B., et al., 1986. The stability of diagnosis with an application to bipolar II disorder. J. Psychiatr. Res. 19, 285–296.

Rogers, R., Jackson, R.L., Cashel, M.L., 2003. SADS-L: comprehensive assessment of mood and psychotic disorders. In: Hersen, M., Hilsenroth, M.J., Segal, D.J. (Eds.), The Handbook of Psychological Assessment. Personality Assessment, 2. Wiley, New York, pp. 144–152.

Shen, G.H.C., Alloy, L.B., Abramson, L.Y., Sylvia, L.G., 2008. Social rhythm regularity and the onset of affective episodes in bipolar spectrum individuals. Bipolar Disord. 10, 520–529.

Stanley, J., 1971. Reliability, In: Thorndike, R.L. (Ed.), Educational Measurement, Second edition. American Council on Education, Washington, DC.

Sukhatme, S., Beam, C.A., 1994. Stratification in nonparametric ROC studies. Biometrics 50, 149–163.

Trull, T.J., Amdur, M., 2001. Diagnostic efficiency of the Iowa Personality Disorder Screen items in a nonclinical sample. J. Personal. Disord. 15, 351–357.

Weinstein, M.C., Berwick, D.M., Goldman, P.A., Murphy, J.M., Barsky, A.J., 1989. A comparison of three psychiatric screening tests using receiver operating characteristic (ROC) analysis. Med. Care 27, 593–607.

Weissman,M.M., Bland, R.C., Canino, G.J., Faravelli, C., Greenwald, S., Hwu, H.G., et al., 1996. Cross-national epidemiology of major depression and bipolar disorder. JAMA 276, 293–299.

Williams, J.B.W., Gibbon, M., First, M.B., Spitzer, R.L., Davies, M., Borus, J., et al., 1992. The Structured Clinical Interview for DSM-III-R (SCID): II. Multisite test–retest reliability. Arch. Gen. Psychiatry 49, 630–636.

Youngstrom, E.A., Findling, R.L., Calabrese, J.R., Gracious, B.L., Demeter, C., Behoya, D.D., Price, M., 2004. Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. J. Am. Acad. Child Adolesc. Psychiatry 43, 847–858.

Zimmerman, M., Mattia, J.I., 2001. A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. Arch. Gen. Psychiatry 58, 787–794.