2008-01-01

Testing the Utility of Self-report Screens to Detect Bipolar Disorder among Undergraduates

Christopher J. Miller
University of Miami, cmiller@psy.miami.edu

UNIVERSITY OF MIAMI

TESTING THE UTILITY OF SELF-REPORT SCREENS TO DETECT BIPOLAR DISORDER AMONG UNDERGRADUATES

By

Christopher J. Miller

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science

Coral Gables, Florida

May 2008

UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

TESTING THE UTILITY OF SELF-REPORT SCREENS TO DETECT BIPOLAR DISORDER AMONG UNDERGRADUATES

Christopher J. Miller

Approved:

_____	_____
Dr. Sheri Johnson	Dr. Terri A. Scandura
Professor of Psychology	Dean of the Graduate School
_____	_____
Dr. Charles Carver	Dr. Malcolm Kahn
Professor of Psychology	Associate Professor

_____
Dr. Eric Youngstrom
Professor of Psychology
University of North Carolina, Chapel Hill

MILLER, CHRISTOPHER (M.S., Psychology)
Testing the Utility of Self-Report Screens to Detect Bipolar Disorder among Undergraduates (May, 2008)

Abstract of a thesis at the University of Miami.

Thesis supervised by Professor Sheri Johnson.
(72 pages)

Background: Bipolar disorders represent a serious mental health problem, but clinicians often fail to detect bipolar diagnoses. The validation of brief and accurate self-report questionnaires may aid in the detection of bipolar disorder, leading to more appropriate treatment and faster recovery. Many such measures exist, but few have been thoroughly tested in undergraduates.

Methods: Three self-report questionnaires used to detect bipolar disorder (the Hypomanic Personality Scale[HPS], Mood Disorder Questionnaire [MDQ], and General Behavior Inventory – 15 item version[GBI-15]) were administered to undergraduate psychology students during the first week of the semester. Participants who were selected based on high and low scores on the self-report screeners completed the Structured Clinical Interview for the DSM-IV, an instrument for diagnosing mental disorders. Participants also completed a battery of self-report measures for constructs previously found to be related to bipolar disorder.

Results: Receiver Operating Characteristic (ROC) curves, sensitivity and specificity, and positive and negative predictive values were used to investigate usefulness of the three screeners in predicting SCID diagnoses of bipolar spectrum disorders. The three

screeners did not demonstrate very good sensitivity or area under the curve for detecting a bipolar spectrum diagnosis, and they generally demonstrated low to moderate predictive values. Of the three, the GBI-15 performed the most adequately in this sample (positive predictive value of approximately .33). All three screeners demonstrated adequate negative predictive values (between .88 and .92).

Discussion: The GBI-15 has some unique features that may help explain its outperformance of the other screeners in undergraduates, but suggestions are provided for the development of better screening tools.

TABLE OF CONTENTS

Chapter 1: Introduction

Bipolar disorders represent a very serious mental health problem. The impact of bipolar disorders can include increased health care costs, lost productivity at work, elevated psychiatric and medical comorbidity, hospitalization, and the risk of suicide (Andlin-Sobocki & Wittchen, 2005; Baldasserini & Tondo, 2003; Kupfer, 2005; Peele, Xu, & Kupfer, 2003).

The lifetime prevalence of bipolar I disorder is reported to be approximately 1% (Judd & Akiskal, 2003; Kessler et al., 1994, 1997, 2005; Narrow et al., 2002; Regier et al., 1993; Weissman et al., 1996). The prevalence rate for bipolar II disorder is less clear. Although some large-scale studies indicate prevalence rates of approximately 1.1% (Merikangas et al., 2007), some authors argue that bipolar II may be much more common. According to this view, previously reported low rates of bipolar II are an artifact of measures that do not adequately assess hypomanic symptoms (Angst & Cassano, 2005). Studies on the prevalence of cyclothymia are rarer; the DSM-IV suggests its prevalence in the community is between 0.4% and 1% (Association, 2000), but another study conducted in the Netherlands found a prevalence of cyclothymia greater than 4% in the community (Regeer et al., 2004). In addition, subthreshold cases – cases that do not meet DSM criteria for bipolar I, bipolar II, or cyclothymia, but that nonetheless constitute a part of the "bipolar spectrum" – may affect as much as 5% of the population (Judd & Akiskal, 2003).

Many people with bipolar disorder will experience an initial onset of symptoms during the typical college years (i.e., between the ages of 18 and 24; Bellivier et al., 2003; Leboyer et al., 2005). Bipolar disorder diagnoses are becoming increasingly common in

adolescence and early adulthood, with some authors reporting a 50-fold increase in visits for such disorders nationally over the past 10 years (e.g. Blader & Carlson, 2007; Moreno et al., 2007), Findings of one carefully conducted epidemiological study suggested that approximately one to two percent of the population will meet criteria for bipolar disorder (either bipolar I, bipolar II, or cyclothymia) during their late teens and early twenties (Lewinsohn et al., 2000), and a further 5% may experience subthreshold bipolar symptoms. Taken together, these findings suggest that many people may experience bipolar symptoms as adolescents or young adults, but that the incidence of bipolar spectrum diagnoses may be somewhat less in this population than in adulthood.

Although the citations above demonstrate that bipolar disorder affects many people, it is frequently misdiagnosed in psychiatric settings. Studies conducted in the United States and Europe indicate that anywhere from 26 to 70 percent of people with bipolar disorder may either receive an incorrect diagnosis or go undiagnosed altogether. Proper diagnosis may not occur for almost a decade after bipolar symptoms emerge, especially for those with bipolar II disorder (Ghaemi et al., 1999; Lewis, 2004; Mantere et al., 2004; Lish et al., 1994).

Several factors may contribute to this lack of recognition of bipolar disorder (Dunner, 2003). Most people with bipolar disorder who do seek treatment do so because of their depressive rather than manic episodes (Hirschfeld et al., 2005) and perceive their depressive symptoms to be more troublesome than their manic symptoms (Calabrese et al., 2004). In some cases people with bipolar disorder may lack insight into their manic or hypomanic symptoms, and do not recognize the extent to which adverse consequences may result from decisions made while manic. 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 do seek treatment for their depression, the clinical presentation of depressive symptoms may be virtually indistinguishable from that of unipolar depression (Cuellar, Johnson, & Winters, 2005). In addition, many clinicians do not screen for bipolar disorder, even when a client presents with depressive symptoms (Brickman, LoPiccolo & Johnson, 2002).

Unfortunately, the consequences of misdiagnoses go beyond the already severe costs of correctly identified bipolar disorder. Many people with bipolar disorder are misdiagnosed with unipolar depression and therefore receive antidepressants; however, antidepressants provided without a mood-stabilizing medication may cause a shift into mania in up to half of bipolar cases. In addition, antidepressant treatment without a mood-stabilizing medication can lead to other adverse effects, such as rapid cycling in up to 30% of bipolar patients (Ghaemi et al., 2004). Costs associated with misdiagnosis are not limited to false negatives, however. People incorrectly diagnosed as bipolar are likely to receive mood stabilizing medications with the potential for negative side effects but little therapeutic value.

Problems with the identification of bipolar disorder may be particularly noteworthy in undergraduate populations. Many symptoms of mania, such as reduced sleep or bursts of high energy or activity, may be seen by students or counselors as normative for college. Insight and awareness of symptoms are less likely earlier in the course of disorder (Peralta & Cuesta, 1998). Because early onset bipolar disorder may lead to particularly poor outcomes (Carter et al, 2003), it is essential to be able to reliably detect bipolar disorders in college populations.

One possible solution to the problem of misdiagnosis in psychiatric and undergraduate populations alike could involve raising the awareness of providers regarding the need to screen for possible bipolar symptoms. This tactic has been used with some success with regards to depression and other mental disorders. In a review of the literature, Kroenke and colleagues (2000) concluded that training can lead to improved diagnosis in the primary care setting. Moreno and colleagues (2003) found similar results for a nurse training program conducted in primary care settings in Panama, that increased recognition of unipolar depression. Despite these successes, however, it appears that few if any training programs have been developed to assist with the detection of bipolar disorder within medical settings (Kroenke et al., 2000).

Bipolar disorders are particularly challenging on this front, because they are much less prevalent than unipolar depression. The relative rarity of the disorder, combined with the time pressure routinely placed on physicians, may contribute to the lack of sufficient investigation of whether bipolar symptoms are present (Brickman et al., 2002).

Another possible solution to the problem of misdiagnosis could involve the use of brief self-report scales to alert providers to the possibility of bipolar disorder. Indeed, this tactic has worked well in detecting other mental disorders, notably unipolar depression. Several scales to detect depression, such as the Inventory to Diagnose Depression (IDD; Zimmerman & Coryell, 1987; Zimmerman, Coryell, Corenthal, & Wilson, 1986), the Beck Depression Inventory (BDI; Beck et al., 1961), and the Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977) have been validated as screeners in psychiatric or community settings (Mulrow, Williams & Gerety, 1995). Many of them

have been subjected to numerous revisions to ensure continued usefulness in detecting depressive symptoms.

Several brief self-report format scales have been developed for detecting bipolar disorder. These include the Hypomanic Personality Scale (HPS; Eckblad & Chapman, 1986), the Mood Disorder Questionnaire (MDQ; Hirschfeld et al., 2000), the Bipolar Spectrum Diagnostic Scale (BSDS; Ghaemi et al., 2005), the General Behavior Inventory (GBI; Depue et al., 1989), the Hypomania Checklist (HCL; Angst et al., 2005), and the Temperament Evaluation of Memphis, Pisa, Paris and San Diego – autoquestionnaire version (TEMPS-A; Mendlowicz et al., 2005). Each of these scales has significant strengths. For example, the scales generally demonstrate high internal consistency, have each been validated against some form of diagnostic measure, and result in significantly higher scores in patient populations when compared to normal controls. With one exception (the HCL; Angst et al., 2005; Meyer et al., 2007), each of the screeners that has been subjected to an analysis of sensitivity and specificity has performed at least moderately well. For instance, the MDQ has demonstrated sensitivity of .73 and specificity of .90 in one study (Hirschfeld et al., 2000). The GBI has demonstrated sensitivity of .78 and specificity of greater than .98 across two studies (Depue et al., 1989; Klein et al., 1989). More recently, it has been used to identify undergraduate participants at high risk for bipolar disorders. In that study, 27 percent of students who scored above a cutoff on the full-length GBI went on to achieve a diagnosis of a bipolar spectrum diagnosis (Grandin, Alloy, & Abramson, 2007). More detailed information on some of the more commonly-studied screeners for bipolar disorder can be found in Tables 1-6.

Table 1

Mood Disorder Questionnaire (MDQ): 15-item inventory meant to assess outpatients for possible bipolarity

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Hirschfeld et al., 2000	198	Bipolar spectrum patients from five outpatient mood disorder clinics: bipolar I (n = 70), bipolar II (n = 26), bipolar NOS (n = 13)	Non-bipolar participants from the same outpatient mood disorder clinics (n = 89). No information available regarding possible diagnoses	Sensitivity: 73% Specificity: 90%	Telephone-administered SCID by experienced psychiatric research social worker	Sensitivity and specificity are based on ROC analyses; ideal cutoff of 7 items plus endorsement of co-occurrence and severity items

Table 1, continued

Mood Disorder Questionnaire (MDQ)

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Miller et al., 2004	72	Bipolar spectrum outpatients from a hospital clinic (n = 36)	Unipolar depressed outpatients from a psychiatric clinic (n = 36)	Sensitivity: 58% Specificity: 67%	SCID mood modules, augmented by diagnostic criteria to detect "soft" bipolar signs and symptoms	Measure was somewhat more sensitive to bipolar I than bipolar II/NOS. Removing the severity criterion item from the MDQ increased psychometric value

Table 1, continued

Mood Disorder Questionnaire (MDQ)

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Hirschfeld et al., 2003	695	Adults in the general population, stratified by MDQ score, with a bipolar diagnosis (n = 78)	Adults in the general population, stratified by MDQ score, without a bipolar diagnosis (n = 617)	Sensitivity: 28% Specificity: 97%	Telephone-administered abbreviated SCID, conducted by Masters- and Doctoral-level research interviewers	Authors note that sensitivity of the MDQ in this study may have been limited by the low test-retest reliability for the SCID in the general population

Table 2

General Behavior Inventory (GBI): 73-item inventory meant to detect both unipolar and bipolar conditions

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Depue et al., 1989	201	Stratified random sampling of GBI scores from a screening sample of 1,068 white university students led to 111 bipolar spectrum cases	957 non-bipolar cases (27 of whom were diagnosed with depression) from same student sample	Sensitivity: 78% Specificity: 99%	SADS-Lifetime	This 73-item version of the GBI may be too long to administer as a brief screener; current study will use shorter version (the GBI-15)

Table 2, continued

General Behavior Inventory (GBI)

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Klein et al., 1989	167	Fifteen bipolar-spectrum patients from stratified random sampling of GBI hypomanic-biphasic scale scores from 492 outpatients at a CMHC and university-based training clinic.	153 patients with various non-bipolar conditions from the same clinics.	Sensitivity: 78% Specificity: 98%	SADS, with additional items to capture chronic subsyndromal affective conditions	

Table 3

Hypomanic Personality Scale (HPS): 48-item inventory that assesses hypomanic personality traits, primarily in undergraduates

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Eckblad & Chapman, 1986	80	Participants selected by stratified random sampling of HPS score from 1,519 undergraduates; 40 subjects scored at least 1.67 SD above the mean (in that sample, cutoff of 34)	Controls (n = 40) were undergraduates from the same initial sample of 1,519, who scored no more than 0.5 SD above the mean on the HPS	Sensitivity: 100% (for hypomanic episodes) Specificity: 69%	SADS-Lifetime was used to determine history of depressive or hypomanic episodes	No participants met criteria for bipolar I, and only hypomania (not depression) was assessed

Table 3, continued

Hypomanic Personality Scale (HPS)

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Kwapil et al., 2000	67	College students with high scores on the HPS (n = 36) were followed up 13 years later (see Eckblad & Chapman, 1986, above)	Controls with relatively lower scores on the HPS (n = 31) were followed up 13 years later (see Eckblad & Chapman, 1986, above)	HPS-positive participants had higher rates of hypomania/bipolar disorders at follow-up (28% and 25%) than did HPS-negative participants (3% and 0%)	SADS-Lifetime, modified to provide diagnoses in line with DSM criteria	Although the HPS did predict bipolar conditions at 13-year follow-up, the authors caution against its use as a clinical or case-finding instrument due to low specificity

Table 3, continued

Hypomanic Personality Scale (HPS)

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Meyer & Hautzinger, 2003	224	German college students: positive screens were composed of top decile of HPS scorers (n = 24)	German college students: negative screens comprised those who scored < 0.5 SD above the mean on the HPS (n = 151)	Sensitivity: 5/8 Specificity: 89%	CIDI, computerized version	Generalizeability limited by German college population, low incidence of bipolar disorder (n = 8), and CIDI's low detection of bipolar disorder

Table 4

Bipolar Spectrum Diagnostic Scale (BSDS): 20-item narrative-based inventory meant to assess bipolarity across the bipolar spectrum

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Ghaemi et al., 2005	95	Bipolar spectrum outpatients (n = 68) from a hospital clinic	Unipolar depressed outpatients (n = 27) from a psychiatric clinic	Sensitivity: 76% Specificity: 85%	SCID mood modules, augmented by diagnostic criteria to detect "soft" bipolar signs and symptoms	Ideal cutoff for positive screen: total score of 13 out of a possible 25 points; severity of symptoms assessed separately from symptoms themselves

Table 5

Temperament Evaluation of Memphis, Pisa, Paris and San Diego – autoquestionnaire version (TEMPS-A): Temperament measure with several subscales potentially relevant to bipolar disorder including the Cyclothymic subscale

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Mendlowicz et al., 2005	177	Bipolars in remission (n = 23) and relatives of bipolar patients (n = 52), originally recruited for a multicenter genetic study	Normal controls (n = 102), recruited by by newspaper, radio, television, flyer, newsletter, and word of mouth	Cyclothymic subscale differentiated bipolar patients from relatives and normal controls	N/A	Although mean scores for groups were presented in this article, cutoffs, sensitivity, and specificity were not calculated

Table 5, continued

Temperament Evaluation of Memphis, Pisa, Paris and San Diego – autoquestionnaire version (TEMPS-A)

Citation	**N**	**Population (positive)**	**Population (comparison)**	**Outcome**	**Reference measure**	**Other information**
Nowakowska et al., 2005	153	Euthymic bipolar patients (n = 49) from a university bipolar disorders clinic	Euthymic unipolar patients (n = 32) from the same university bipolar clinic, artistically creative controls (n = 32) recruited from graduate programs, healthy controls (n = 47; source unclear from the article)	Cyclothymic subscale differentiated bipolar from unipolar patients and healthy controls, but not from creative controls	N/A	Although mean scores for groups were presented in this article, cutoffs, sensitivity, and specificity were not calculated

Table 6

Hypomania Checklist – 32-item (HCL-32): Checklist meant to detect hypomanic components in depressed outpatients

Citation	N	Population (positive)	Population (comparison)	Outcome	Reference measure	Other information
Angst et al., 2005	426	Bipolar adults from outpatient clinic in Italy (n = 124), and university/ psychiatric inpatient affective disorder units in Sweden (n = 142)	Unipolar adults from the same clinics (n = 160)	Sensitivity: 80% Specificity: 51%	SCID, modified to more easily capture bipolar II	Sensitivity and specificity are based on ROC analyses; ideal cutoff of 14 or more for positive screen

A screener must do more than simply detect bipolar disorder, however; it must also be brief enough to be practically administered. Some measures are too long to be widely used as screeners. For example, although the GBI has been validated in numerous samples (Depue et al., 1989), its full 73-item version can take more than 20 minutes to complete. A shortened 15-item version has been developed based on one sample (the GBI-15), but it has not been validated in other samples.

Other researchers have focused on particular subscales of longer or more general measures. For example, Bagby and colleagues (2005) used the MMPI-2 to attempt to distinguish unipolar, bipolar, and schizophrenic patients from one another. Several subscales were able to distinguish unipolar and bipolar patients from schizophrenic patients, but no clinical or content subscales were able to reliably distinguish bipolar from unipolar patients. Those authors concluded that the Restructured Clinical Scales might be of more use in detecting bipolar patients, but this realm has not yet been fully explored.

Focusing on the scales that have at least initially positive validational evidence, there are some other important limitations to be considered. Careful research has demonstrated that the utility and thresholds for scales designed to screen for juvenile bipolar disorder differ greatly across populations (Youngstrom et al., 2005); it is therefore likely that thresholds for college students and adults likewise vary across populations. Some of the scales were developed for use, or have only been validated, in clinics (e.g., the BSDS, Ghaemi et al., 2005). Similarly, the MDQ was initially validated in a psychiatric setting, but has only been the subject of one validational study in the general population (Hirschfeld et al., 2003). Others were developed for use, or have only been validated, in university settings (e.g., the HPS, Eckblad & Chapman, 1986). Overall,

then, it appears that few existing screeners for bipolar disorder have shown themselves as robust measures that can be used across different settings with different populations. This is a crucial issue. Care providers surveying the current literature on bipolar screeners may be faced with scales that appear to have excellent psychometric properties, but that may not be applicable to the population they serve. It cannot be assumed that these scales will fare equivalently in new populations, and inappropriate application of scales without validation can lead to misdiagnosis.

In addition to the concerns outlined above, other methodological issues may limit the applicability of the validational literature for bipolar screeners. There is a broad range of diagnostic instruments that can be used to validate screeners. Two with well-established reliability and validity are the Structured Clinical Interview for the Diagnostic and Statistical Manual – IV (SCID; First & Gibbon, 2004) and the Schedule for Affective Disorders and Schizophrenia – Lifetime Version (SADS-L; Endicott & Spitzer, 1998). Although some scales have been validated against such measures (e.g., the GBI, Depue et al., 1989), some authors have relied on other instruments. The original validation study of the HPS relied on the SADS-L (Eckblad & Chapman, 1986), but Meyer and Hautzinger (2003) validated the HPS using the Composite International Diagnostic Interview (Mezzich & Cranach, 1988). In this latter study the CIDI was administered with the aid of a computer program.

Some evidence suggests, however, that the CIDI may systematically misdiagnose bipolar disorders (Regeer et al., 2004), missing up to 50% of bipolar diagnoses (Kessler et al., 1997). These types of problems may be particularly severe for bipolar II disorder, as several authors have reported very low interrater and test-retest reliability for this

diagnosis despite the use of well-validated measures such as the SADS (Dunner & Tay, 1993; Simpson et al., 2002). In addition, some researchers have relied on telephone-based interviews, which may further limit detection of bipolar disorder.

In the face of this difficulty, some authors have advocated modifying reference-standard semi-structured interviews in important ways to more easily detect bipolar disorder. These adjustments are in some cases well-documented (see Akiskal & Pinto, 1999), but in other cases are not clearly described. For instance, Benazzi and Akiskal (2003) have recommended skipping initial questions about mood and duration when probing for hypomanic episodes with the SCID, instead asking directly about behavioral symptoms. Several bipolar screeners investigated in the current study were validated using modified or telephone-administered versions of the SCID (e.g., the BSDS and MDQ), and others used a modified SADS-L (e.g., the GBI and HPS). When combined with imperfect sensitivity, these adjustments raise the possibility that each screener is actually identifying a different subset of people with bipolar disorder. This may make comparisons between measures difficult or impossible.

Yet another issue plaguing the current literature on bipolar screeners is a large variety of statistical approaches to validation. Even if screeners are compared to a well-validated gold standard measure, reported outcomes could include sensitivity/specificity, positive predictive value/negative predictive value, correlations, chi-square analyses, t-tests for scores between bipolar and non-bipolar groups, or Area Under the Receiver Operating Characteristic (AUROC) curves (Hanley & McNeil, 1982). The lack of consistency in the literature can complicate the process of making comparisons across measures. Sensitivity and specificity are perhaps the most commonly-reported outcomes

and have the benefit of being relatively (though not totally; see Kraemer, 1992) robust to differences across samples in a disorder's baserate. AUROC analyses, however, incorporate more information than sensitivity and specificity at a particular cutpoint, and as such represent a more useful index of a screener's utility in a given population (Youngstrom, 2004). Unfortunately, power analyses for AUROC cannot be easily used to determine the sample size required to detect a significant difference between screeners. When screeners are correlated, however (as would be expected in the current study), AUROC represents a relatively powerful statistical technique, allowing multiple correlations among measures with only a moderate sample size (E. A. Youngstrom, personal communication, 2006).

In sum, despite many years of research and the development of several brief screening measures to detect bipolar disorder, much work still needs to be done in this important domain. Lacks of assessment in comparable populations, and relatively little consistency in "gold standard" instruments, hamper the overall utility of the brief screeners listed above. The present study aims to address some of these problems by administering several screeners to an undergraduate sample simultaneously. Due to practical considerations, only three screeners could be investigated in the current study.

Based on previous research, each of the screeners selected has significant strengths. The HPS (Eckblad & Chapman, 1986) was designed specifically to detect hypomanic symptoms in undergraduates, and is the only screener that has demonstrated an ability to detect bipolarity over a thirteen-year follow-up (Kwapil, 2000). The GBI-15 is based on a longer 73-item version that has demonstrated good psychometric properties in detecting bipolarity across numerous settings and populations (Klein et al., 1989); the

15-item version is itself relatively untested. The MDQ (Hirschfeld et al., 2000) was developed for clinical populations and is designed to closely mirror DSM-IV criteria. It has been widely disseminated, including a full-page advertisement in a recent edition of "Newsweek" (AstraZeneca, 2006), despite a relative absence of data on how it fares outside of clinical settings. Assessing the psychometric value of each of these screeners in an undergraduate population will add significantly to the literature on the detection of bipolar disorder.

Hypotheses

Many of the factors that interfere with the clinical detection of bipolar disorder (low patient insight, a lack of distress during hypomania or mania, limited impairment due to hypomania, etc.) may also hamper the utility of self-report screeners for bipolar disorder. Nonetheless, I hypothesized that the self-report screeners investigated in this study would demonstrate modest sensitivity and good specificity in detecting bipolar disorder. Specifically, I hypothesized that the Hypomanic Personality Scale and General Behavior Inventory – 15 item version, both of which were designed specifically to detect manic vulnerability in undergraduates, would display good psychometric properties. I further hypothesized that the Mood Disorder Questionnaire, having been developed for clinical populations, would demonstrate lower utility in detecting bipolar disorder in an undergraduate sample.

Chapter 2: Methods

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

Participants

Participants were drawn from the University of Miami pretesting pool. This pool is comprised of all students taking an introductory psychology course (typically between 300 and 800 students per semester). Each student taking introductory psychology completes a battery of self-report measures during the first week of class. Students are asked to complete a set of research credits for their introductory psychology class, which can be accomplished by writing about recent empirical articles or taking part in follow-up studies (such as the current study) throughout the course of the semester.

Over 1,200 introductory psychology students at the University of Miami completed the three self-report screeners during pretesting. For the remainder of this paper, this set of approximately 1,200 students will be referred to as the "screening sample." Students were classified according to their scores on each of the three screeners as either high-scoring (a positive screen) or low scoring (a negative screen). A stratified random sampling design (described below) was used to choose students for further follow-up. Students selected according to high and low scores were emailed and invited to take part in a second testing session. Those who chose to take part in the second testing session will be referred to as the "final sample."

Measures

The primary measures for this study were three self-report screeners for detecting bipolar disorder. The Structured Clinical Interview for the DSM-IV (First & Gibbons, 1994) served as the reference standard. Several self-report measures of constructs related to bipolar disorder were used for secondary analyses. More detail on these measures follows.

Brief Self-report Screeners for Bipolar Disorder

A brief description of each of the self-report bipolar screeners administered for this study follows. In each case two sets of cutoffs were used to determine a positive screen: (1) a cutoff determined by previous literature or the designers of the screeners themselves, and (2) a cutoff determined specifically for this screening sample, to ensure that the final dataset included a relatively equal balance of positive screens on all three measures. Information on original cutoffs is presented in this section, and information on revised cutoffs is presented in the results section.

Hypomanic Personality Scale (HPS). The HPS (Eckblad & Chapman, 1986) is an inventory designed to assess hypomanic personality traits. It consists of 48 items, distilled from an initial pool of 97 items that were administered to 768 undergraduates. Questions are presented in a true/false format, and 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). It has demonstrated adequate ability to detect hypomanic episodes in undergraduates (Eckblad & Chapman, 1986), and predicted the onset of hypomania and bipolar disorders in the same sample over a thirteen-year follow-up (Kwapil et al., 2000). Most studies on the

HPS, however, have focused on the detection of manic or hypomanic episodes rather than bipolar diagnoses per se (a diagnosis of bipolar II disorder involves both hypomanic episodes and depressive episodes); the ability of the HPS to detect concurrent bipolar disorder diagnoses that require depressive episodes remains largely unexplored. The HPS has demonstrated good internal consistency with Cronbach's alpha of .87 (Eckblad & Chapman, 1986). It has also demonstrated test-retest reliability of .81 over a fifteen-week period (Eckblad & Chapman, 1986) and split half reliability of .85 (Meyer & Hautzinger, 2003). In previous research (Eckblad & Chapman, 1986), a cutoff of 36 was chosen, corresponding to a *z*-score of approximately 1.75 in their sample. Among University of Miami undergraduates, data from previous semesters indicate that a cutoff of 34 for a positive screen has corresponded to approximately that same z-score.

Mood Disorder Questionnaire (MDQ). The MDQ (Hirschfeld et al., 2000) is a single-page inventory designed to be scored quickly and easily by any trained medical personnel. It screens for lifetime history of a manic or hypomanic episode based on thirteen yes/no questions, derived from the DSM-IV criteria for bipolar disorder as well as clinical experience. Additional questions focus on whether the symptoms reported all co-occurred, as well as the overall level of functional impairment caused by the symptoms. To receive a positive screen, participants must answer at least seven out of the thirteen yes/no questions as "yes." In addition, participants must indicate that the symptoms co-occurred and caused at least moderate problems for them (Hirschfeld et al., 2000, 2003). These additional requirements will be referred to in this paper as the "simultaneity" and "severity" criteria, respectively. In an initial study, the MDQ demonstrated sensitivity of .73 and specificity of .90 for detecting bipolar spectrum

diagnoses (bipolar I, bipolar II, and bipolar NOS) in a mood disorders clinic (Hirschfeld et al., 2000). The diagnoses of the nonbipolar participants in this study were not reported. 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). The MDQ has demonstrated good internal consistency with a Cronbach's alpha of .90 (Hirschfeld et al., 2000).

General Behavior Inventory (GBI). The GBI (Depue et al., 1989) was designed to identify lifetime bipolar affective disorders. In its complete 73-item form, it has been subjected to many validation studies with both clinical and non-clinical populations, and has performed well at detecting milder portions of the bipolar spectrum. The complete GBI has been validated in both undergraduate (Depue et al., 1989) and outpatient (Klein et al., 1989) populations, achieving sensitivity of .76 and specificity of .98 or greater in those two studies. It has also demonstrated good internal consistency (Cronbach's alpha = .94) and test-retest reliability over a fifteen-week period (.73; Depue et al., 1989). For the present study I utilized the 15-item version (the GBI-15), in which participants rate each item on a 1 to 4 scale ranging from "never or hardly ever" to "very often or almost constantly." This 15-item version was distilled from the longer 73-item version (Meyer & Johnson, 2003), but the applicability of these 15 items in detecting bipolar disorder in other samples has not yet been investigated. Participants answering with a "3" or "4" on at least five of the nine hypomanic/biphasic items of the GBI-15 receive a positive screen.

Reference standard: *Structured Clinical Interview for the DSM-IV (SCID).* The SCID (First & Gibbon, 2004) is a semi-structured interview designed specifically to yield

diagnoses based on DSM-IV criteria. This study involved administration of the mood, psychosis, and substance abuse modules, allowing definitive diagnoses of bipolar I disorder, bipolar II disorder, or cyclothymia. It was also possible to obtain a diagnosis of bipolar NOS based on recurrent hypomania without any history of a major depressive episode. To maintain strict adherence to DSM-IV criteria, no adjustments aimed at capturing a wider spectrum of bipolar-like conditions were used.

The SCID was administered by a highly trained graduate-level research assistant who was blind to participants' scores on the self-report screeners described above. Administration of the SCID was audio recorded, allowing reliability to be established under the direction of Sheri Johnson, PhD. Tapes were each rated by between two and five trained graduate-level members of Dr. Johnson's research team, with a special emphasis placed on the ratings of current and past mania and hypomania. 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. .

Additional Measures for Secondary Analyses

In addition to the self-report screeners and the reference-standard SCID, several other measures (tapping constructs potentially related to bipolar disorder) were administered. These measures are described briefly below.

Willingly Approached Set of Statistically Unlikely Pursuits (WASSUP). The WASSUP (Johnson & Carver, 2006) is a 30-item measure comprising statements regarding unlikely and ambitious goals. Example items include "You will have a major role in a movie" and "Someone will write a book about your life." Respondents use a

Likert-type scale to indicate how likely they are to pursue each those goals, with a response of "1" indicating "no chance I will set this goal for myself" and a response of "5" indicating "definitely WILL set this goal for myself." The scale's total score is derived from summing the scores from all itemsThe WASSUP includes several factor-analytically derived subscales, including Popular Fame, Political Power, Idealized Friendships, Idealized Family Relationships, Positive Impact on the World, Financial Success, and Creativity. In five previous samples, persons at risk for hypomania and those diagnosed with mania have demonstrated elevated scores on scales related to extremely high extrinsic goals, such as extreme levels of fame and wealth (Johnson, 2005; Johnson & Carver, 2006; Johnson, Eisner, & Carver, under review; Gruber & Johnson, under review). Alpha reliability for the WASSUP in my sample was .87.

Positive Generalization Scale (POG). The Positive Generalization Scale (Eisner, Johnson, & Carver, in press) is an eighteen-item self-report scale designed to assess the respondent's tendency to become overly confident after small successes. Sample items include "When one thing goes right, it makes me feel I'm good at everything" and "After one date goes well, I know that person is in love with me forever." Answer choices are given on a Likert-type response scale with a response of "1" indicating"very true for me" and "4" indicating "very false for me." A total score is derived by reverse-coding each item score and then summing all items, such that higher scores on the POG represents higher levels of positive generalization. The scale includes three factor-analytically derived subscales: Lateral Generalization, Social Generalization, and Upward Generalization. Previous research has suggested that POG scores, and particularly

Upward Generalization scores, are elevated among persons at risk for mania as measured by the HPS (Eisner et al., in press). Alpha reliability for the POG in my sample was .91.

BMSIBS-CS. The BMSIBS-CS is a new measure designed to assess socially intrusive behaviors that have been clinically observed among persons with bipolar disorder. The scale asks about a range of inappropriately dominant behaviors over the most recent two-week period. It is a 28-item scale, with Likert-style answer choices ranging from "1" ("never") to "4"("nearly always"); a total score is derived by summing responses to each item. Sample items include "I've been insisting that my friends join me in the activities I want to do" and "I've shared my ideas with business owners about how they can improve their business." The scale has been shown to robustly correlate with HPS scores (Siegel et al., 2007). Alpha reliability in my sample was .84.

Personality Research Form (PRF). The Personality Research Form (Jackson, 1967, 1989) is a comprehensive personality measure. In its complete form it measures 20 motivational traits and two test validity scales, with 16 items per scale summing to a total of 352 items. Three of the most-studied scales are those that tap achievement, affiliation, and dominance themes (Moneta & Wong, 2001). For the current study, participants completed each of these three subscales, for a total of 48 items. All items are answered in true-false format, with answers of "true" contributing one point toward total scores. The PRF Social Dominance subscale has been found to correlate robustly with the HPS in previous research (S. L. Johnson, personal communication, November 2007). Alpha reliability for the PRF in my sample was .79.

Narcissistic Personality Inventory (NPI). The Narcissistic Personality Inventory (Raskin & Terry, 1988) is a 37-item inventory to assess narcissistic personality qualities.

Each item asks the respondent to choose between two answer choices, such as "I am assertive" versus "I wish I was more assertive." For each item, one of the answer choices (associated with relatively more narcissism) contributes a point to the total score. The NPI was distilled from a longer, 54-item version by retaining only those items with factor loadings above .35. Subscales include Authority, Self-Admiration, Arrogance, and Exploitiveness (Morf & Rhodewalt, 1993). The scale is one of the most widely used measures of narcissistic traits, and it was included because of the comorbidity of narcissistic personality disorder and mania (Stormberg et al., 1998). In previous studies, the NPI has been correlated with the mania scale of the Minnesota Multiphasic Personality Inventory (e.g. Raskin & Novacek, 1989), and with the HPS (Fulford, Johnson, & Carver, 2007). Alpha reliability for the NPI in my sample was .83.

Barratt Impulsiveness Scale Version 11 (BIS-11). The Barratt Impulsiveness Scale Version 11 (Patton, Stanford, & Barratt, 1995) is a 34-item scale that is one of the most commonly used indices of impulsivity. Factor analysis has demonstrated three major factors : Attentional Impulsiveness, Motor Impulsiveness, and Nonplanning Impulsiveness. Answer choices fall on a Likert-type scale, with "1" indicating "rarely/never" and "4" indicating "always/almost always." Sample questions include "I say things without thinking" and "I buy things on impulse"; the total score for the scale is derived by summing scores from each item. Previous studies have suggested that the scale correlates with current mania, and that it is elevated among persons with a history of mania compared to controls (Peluso et al., 2007; Swann et al., 2003). Alpha reliability for the BIS-11 in my sample was .78.

Procedure

Participants were scheduled for a two-hour session to complete the study. When they came to the appointment, they completed an informed consent form, followed by administration of the SCID modules. After completing the SCID, participants were administered two brief verbal tasks: Reverse Digit Span from the Wechsler Adult Intelligence Scale (Wechsler, 1997), and a verbal fluency task similar to Word Retrieval from the Woodcock Johnson Test of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001). After completing these tasks, participants completed three computer-based tasks not discussed in this report: the Picture Story Exercise, the Reward Discounting Task, and the Iowa Gambling Task. Following the completion of these computer-based tasks, participants completed an additional packet of self-report measures. After completion of these tasks, participants were debriefed, received a research credit slip to apply toward their introductory research requirement, and were dismissed.

Stratified Random Sampling

An overarching goal of this study was to maximize the useful information obtained from the final sample regarding each of the three self-report screeners. To achieve this, however, "screen-positive" and "screen-negative" subsamples needed to be constructed independently for each screener. For the HPS, the standard cutoff of 34 would likely have resulted in 55 screen-positives out of 1,172 screened over the course of the three semesters in which the study was run. Considering that fewer than half of those eligible to take part in the study chose to do so, this would not have yielded a sufficient number of high scorers on the HPS. Thus I adopted a cutoff of 32 on the HPS. Similarly,

the recommended cut-off of at least 5 out of the 9 hypomanic/biphasic items on the GBI-15 (Meyer & Johnson, 2003) would likewise have resulted in too few screen-positives (55 out of 1,199 screened), and so I adopted a cutoff of four or more items on the GBI-15.

Previous research on the MDQ had established that a cutoff of seven or more out of thirteen initial items, plus endorsement of the simultaneity (item 14) and severity (item 15) criteria, resulted in the best psychometric properties (Hirschfeld et al., 2000). Almost 75% of those who completed the MDQ during pretesting, however, endorsed seven or more of the initial thirteen items. Of those, 84% reported that many of the symptoms they experienced had occurred simultaneously (item 14). Thus, using the original cutoff for the MDQ would have basically reduced the MDQ to a one-item screening measure, whereby most participants would achieve a positive or negative screen based on their response to item 15 ("How much of a problem did any of these [symptoms] cause you…?"). To assess the usefulness of the MDQ as a whole, rather than simply testing the utility of one specific item, I adopted a cutoff for the MDQ consisting of (1) endorsing *eleven* or more of the initial thirteen items and (2) endorsing the simultaneity and severity criteria.

To construct the screen-positive group for each screener, all students from the screening sample meeting the cutoffs described above were contacted via email and invited to participant in the study. Detailed results from this sampling procedure are provided in the flowchart (Figure 1). Table 7 contains total scores and standard deviations for each of the screeners in question for the entire screening sample (ignoring the simultaneity and severity criteria for the MDQ). In total, 32 participants were enrolled in

the study as screen-positives for the HPS, 35 were screen-positives for the MDQ, and 35 were screen-positives for the GBI-15. There was some overlap between high scorers on each screener (for instance, six participants screened positive on both the HPS and the GBI-15).

A separate set of participants (n = 23), selected randomly from among all students who scored below the cutoff on at least one screener, formed the screen-negative groups. These participants served as the "control group" for each screener. Because sampling for each screener was conducted independently, however, there were two participants who served as screen-positives for one screener, but screen-negatives for the other two screeners. Thus the screen-negative comparison groups consisted of either 24 (for the MDQ and GBI-15) or 25 (for the HPS) randomly selected participants.

Table 7

Mean and standard deviations of screener scores for screen-positive groups, screen-negative groups, and total screening sample

	N	Mean	Standard Deviation
HPS – screen-positive group	32	35.31	3.18
HPS – screen-negative group	25	18.18	6.11
HPS – entire screened sample	1,172	19.43	8.01
MDQ – screen-positive group	35	11.71	0.79
MDQ – screen-negative group	24	8.02	3.27
MDQ – entire screened sample	1,197	8.19	3.04
GBI-15 – screen-positive group	35	5.20	1.47
GBI-15 – screen-negative group	24	0.63	1.01
GBI-15 – entire screened sample	1,199	0.94	1.58

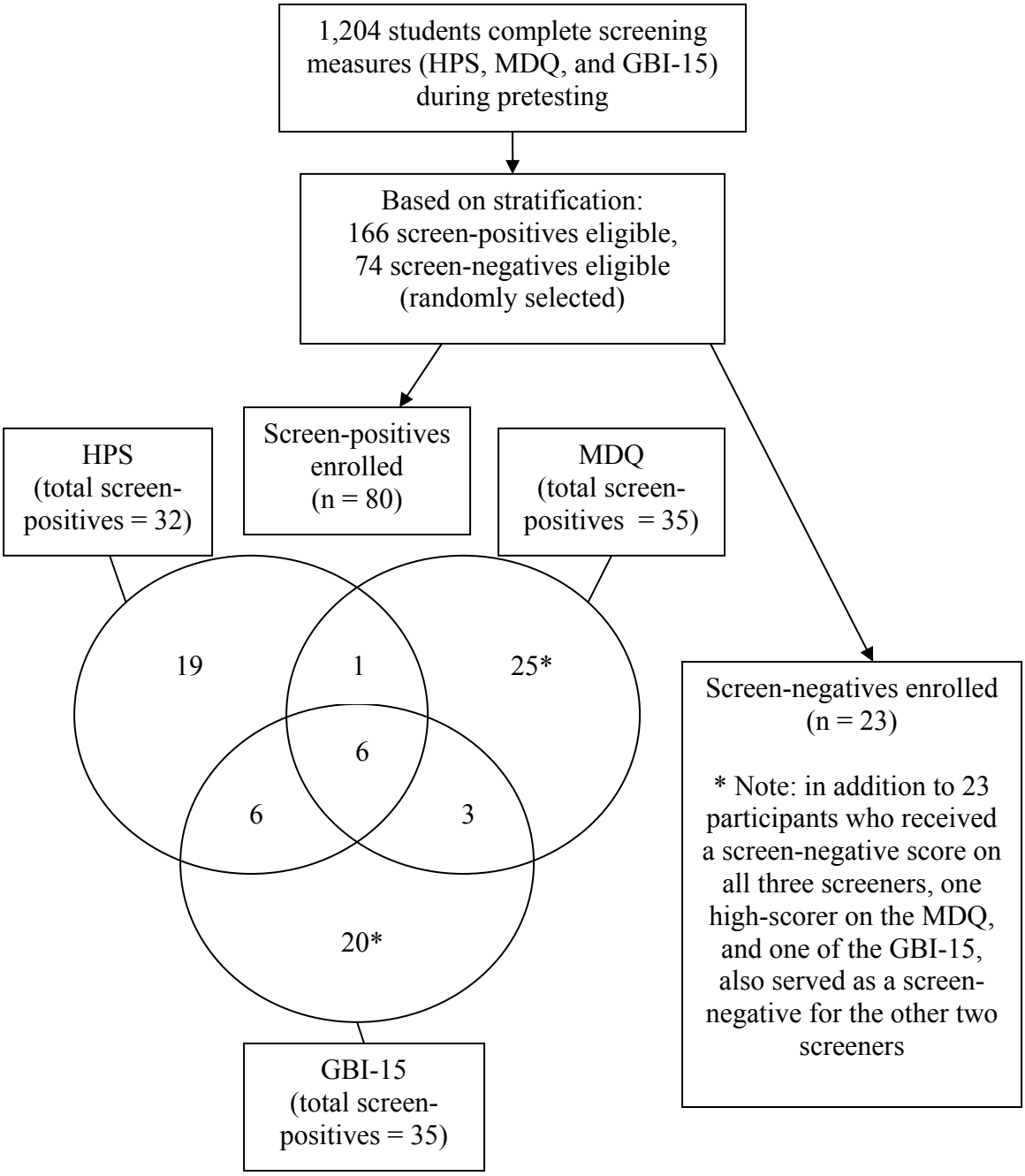

Figure 1. Flowchart of participant enrollment.

Analysis Plan

The primary outcome for these analyses was presence or absence of a bipolar spectrum condition (bipolar I disorder, bipolar II disorder, cyclothymia, or bipolar NOS) based on the SCID. There are several ways to investigate the alignment between a participant's score on a screening measure and the diagnosis assigned to that participant by the SCID. Results for several of these outcomes will be presented to give as complete a picture as possible of the screeners investigated in the current study.

To date, the most common measures of a screener's usefulness are sensitivity and specificity. Sensitivity is defined as the proportion of participants with a target diagnosis (in this case any bipolar spectrum diagnosis) who screen positive on a particular measure. Specificity is defined as the proportion of participants without the target diagnosis who screen negative on a particular measure.

Positive predictive value (PPV) and negative predictive value (NPV) are additional ways to rate screening measures. PPV is defined as the proportion of participants who screen positive on a particular measure that meet the relevant diagnostic criteria. NPV is defined as the proportion of participants who screen negative on a particular measure that do not meet diagnostic criteria.

The methods described above, while useful, are based heavily on the particular cutpoint used to differentiate low and high scorers. That is, changing the threshold to obtain a positive screen changes sensitivity, specificity, PPV, and NPV. Area under the receiver operating characteristics (AUROC) curve, in contrast, uses more information about how the screener performs across all possible thresholds, and thus provides a more complete picture of the screener's utility (Hanley & McNeil, 1982). AUROC analyses are

based on sensitivity and specificity across the entire range of possible cutoff scores for a screener. These points are then plotted on a two-dimensional space, with sensitivity as one axis (typically the vertical) and one minus specificity plotted on the other axis. When these points are connected, it yields a visual representation of the screener's accuracy across its entire range. The area beneath this curve is an index of the screener's overall usefulness. The AUROC methodology also allows the calculation of cutoff scores that maximize the screener's overall efficiency in a given population (Kraemer, 1992).

Chapter 3: Results

The paragraphs below describe the characteristics of both the screening sample and the final sample used in the current study. I then consider potential confounds, correlations among the screeners, diagnostic efficiency statistics, and results from measures of constructs related to bipolar disorder in previous studies.

Characteristics of the sample

The three self-report screeners were completed by 1,204 students taking an introductory psychology course at the University of Miami during the spring and fall semesters of 2006 and 2007. Limited demographic information is available for this complete sample, but most students were either eighteen or nineteen years old, and approximately 62% were female. Stratified random sampling resulted in a final sample size of 103. Of this final sample, 62% were female, and the average age $\pm$ SD was 18.62 $\pm$ .83 years. Based on participants' completion of a demographic questionnaire, 6% of the sample were Asian-American, 9% were African-American, 17% were Hispanic/Latino, 60% were White, and 9% were of another or mixed ethnicity. Of those who listed themselves as Hispanic, approximately half were of Cuban descent. A total of 18 of the 103 participants achieved a bipolar spectrum diagnosis, including seven with bipolar I disorder, six with bipolar II disorder, two with cyclothymia, and three with bipolar disorder NOS. Given our strategy of oversampling those with high scores on the three self-report screeners, this apparently high incidence rate is not surprising.

Table 8

Correlations among total scores for self-report bipolar screeners and bipolar spectrum diagnosis

	HPS	MDQ	GBI-15	Bipolar Spectrum Diagnosis
HPS	-	.37**	.42**	.14
MDQ	-	-	.31**	.10
GBI-15	-	-	-	.19

Note. Correlations between self-report screeners are drawn from the entire screened sample. Correlations involving bipolar spectrum diagnosis are drawn from the final sample.

** $p < .01$

Table 7 provides descriptive statistics for the low and high scorers for each screener. Table 8 shows correlations among the three screeners. These correlations were generally in the moderate range ($r = .31$ to .42), and each achieved statistical significance at the $p = .01$ level. Figure 2 depicts histograms of the screener scores for the screen-positive and screen-negative groups for each measure in the final sample.

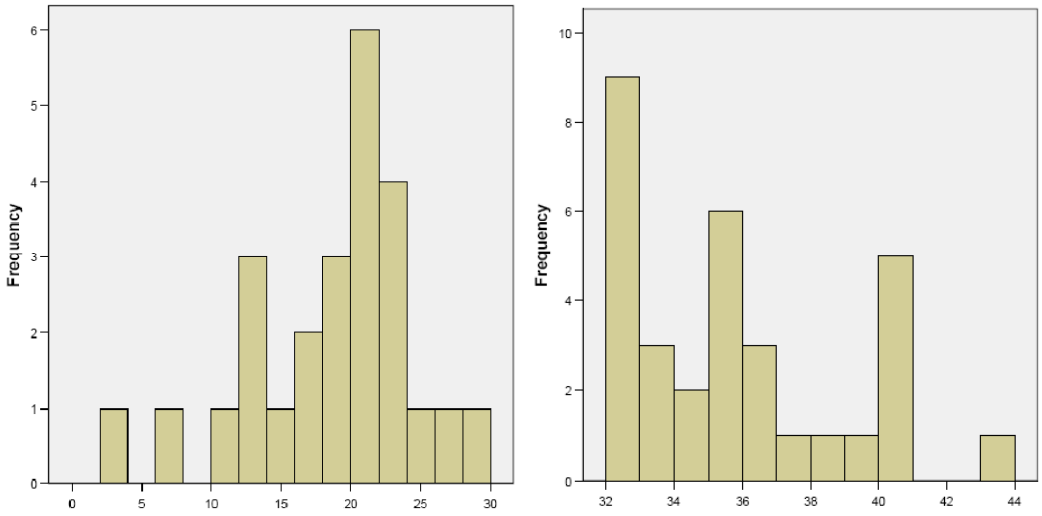

Figure 2a. Total scores on the HPS for those in the screen negative group on the HPS (left) and those in the screen positive group on the HPS (right)

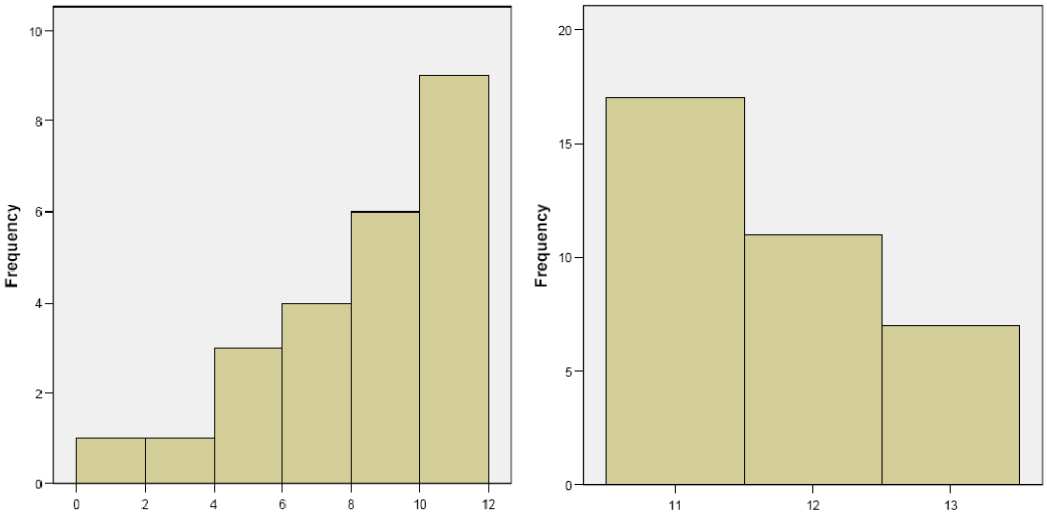

Figure 2b. Total scores on the MDQ (first thirteen items) for those in the screen negative group on the MDQ (left) and the screen positive group on the MDQ (right)

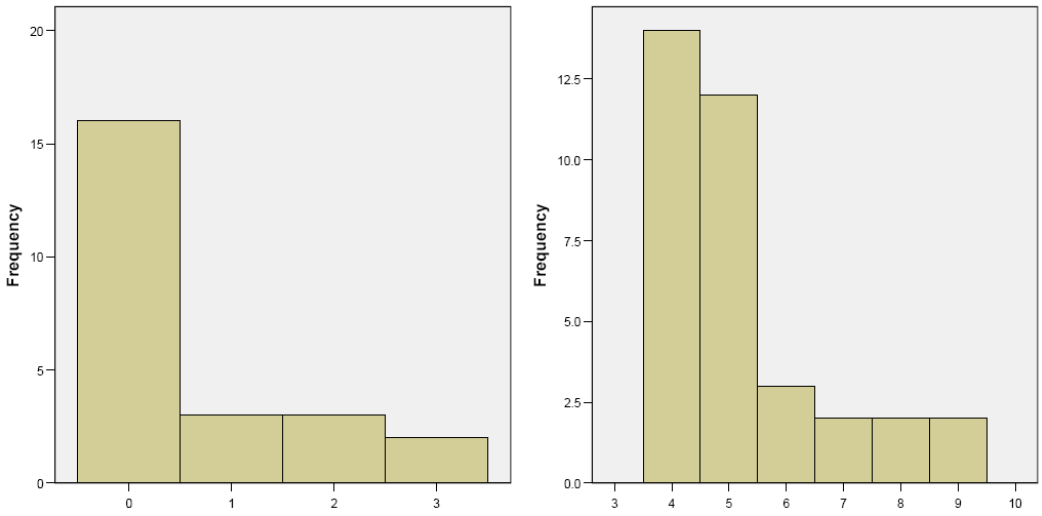

Figure 2c. Total scores on the GBI-15 (nine hypomanic/biphasic items) for those in the screen negative group on the GBI-15 (left) and the screen positive group on the GBI-15 (right)

Analyses of Potential Confounds

Preliminary analyses were undertaken to determine if demographic characteristics (such as age, gender, or ethnicity) were related to scores on any of the three screeners in the final sample of participants. Results in this domain were largely negative. Males and females did not differ significantly in terms of total scores on any of the three screeners according to three independent samples *t*-tests ($p > .15$ in each case). Three one-way ANOVAs indicated that ethnic groups did not differ on screener scores, ($p > .30$ in each case). Age was not significantly correlated with the HPS ($r = .012$, $p = .917$) or the MDQ ($r = -.07$, $p = .56$). Age was significantly correlated with the total score for the GBI-15, however ($r = .25$, $p = .03$). I have no specific explanation for this result, although it may be an artifact of the number of comparisons undertaken while investigating these confounds.

Among each of the three screen-positive subgroups, about two-thirds of those contacted to participate in the study did not complete the second session (i.e. chose not to participate). To investigate potential bias in who decided to take part in the study, I conducted three independent samples *t* tests. Within the screen-positive subgroup contacted for each screener, I compared total scores on that screener for those who did versus those who did not choose to participate in the study. In each case, there was no evidence of bias in who took part in the study or not (each $p > .2$). For instance, I invited all students who achieved a score of 32 or above on the HPS to participate in this study. The average score for those who took part in the study was 35.3, while the average score for those who did not was 35.4, $t(84) = .16$, $p = .87$. Taken together, these findings suggest that these results will likely generalize to students with similar scores on the HPS, MDQ, and GBI-15.

Diagnostic efficiency statistics

Sensitivity, specificity, positive predictive value, and negative predictive value for each of the screeners using the revised and original cutoffs are presented in Table 9. The study's methods allowed these statistics to be calculated under both the original and revised cutoffs for each screener. There were two caveats related to this methodology, however, the first of which was related to the MDQ. To achieve balanced sampling of the three screeners, the threshold for a positive screen on the MDQ needed to be *raised* from endorsing *seven* items to endorsing *eleven* items out of the first thirteen (thresholds for screening positive on the HPS and MDQ needed to be *lowered* to achieve balanced sampling). Because of this, the final sample included almost no participants who met the original cutoff for the MDQ (seven of thirteen items) but not the adjusted cutoff (eleven

of thirteen items). Thus results related to the MDQ's performance under its original scoring algorithm should be interpreted with caution.

The second caveat to the sensitivity and specificity results was related to our stratification design (oversampling high scorers). Several existing methodologies can adjust for stratification (Choi, 1992; Sukhatme & Beam, 1994; Weinstein et al., 1989). These techniques weight the accuracy of scales according to the expected distribution of screener scores in the population;sensitivity and specificity estimates using two of these methods are presented in Table 10. As expected, weighted sensitivity estimates were much lower than the unweighted estimates, and the weighted specificity estimates were much higher than the unweighted estimates

Table 11 includes mean screener scores for three different diagnostic groups: those with a bipolar I diagnosis, those with another bipolar spectrum diagnosis, and those with no bipolar spectrum diagnosis. In general, those with a bipolar spectrum diagnosis had higher average scores on the screeners than those without.

Receiver Operating Characteristic (ROC) curves are presented in Figure 2. An AUC of .50 represents detection of a disorder at the chance level, and AUC of 1.00 represents perfect detection. Area under the curve (AUC) for the HPS was .602, for the MDQ was .606, and for the GBI-15 was .690. These results are generally weak for the HPS and MDQ, and weak/moderate for the GBI-15 (Fischer, Bachman & Jaeschke, 2003). In terms of statistical significance, the AUC for the MDQ and HPS total scores did not differ significantly from chance ($p > .15$ in both cases), but the GBI-15's AUC was significantly different from chance ($p = .01$). The areas under the curve for each screener were not statistically significantly different from one another, however.

There was one caveat to the AUC analysis, also related to the MDQ. A positive screen on the MDQ is based on the sum of one group of items, as well as endorsement of symptom simultaneity and severity. Because of the nature of these analyses, AUC analyses for the MDQ do not integrate the criteria of simultaneity or severity. This inability to incorporate these two "critical items" into certain outcomes for the MDQ is consistent with previous analyses (e.g. Hirschfeld et al., 2000)

Table 9

Sensitivity, specificity, positive predictive value, and negative predictive value with original and revised cut-off values under two scoring schemes

Original cutoffs					
Screener	Cutoff	Sensitivity	Specificity	PPV	NPV
HPS	34 items	.50 (3/6)	.56 (22/39)	.15 (3/20)	.88 (22/25)
MDQ	7/13 items plus simultaneity/severity	.70 (7/10)	.41 (21/52)	.18 (7/38)	.88 (21/24)
GBI-15	5 items	.70 (7/10)	.61 (22/36)	.33 (7/21)	.88 (22/25)

Revised cutoffs					
Screener	Cutoff	Sensitivity	Specificity	PPV	NPV
HPS	32 items	.67 (6/9)	.46 (22/48)	.19 (6/32)	.88 (22/25)
MDQ	11/13 items plus simultaneity/severity	.70 (7/10)	.43 (21/49)	.20 (7/35)	.88 (21/24)
GBI-15	4 items	.85 (11/13)	.48 (22/46)	.31 (11/35)	.92 (22/24)

Table 10

Estimated sensitivity/specificity from stratification

Weighting based on Weinstein et al., 1989		
Measure	**Sensitivity**	**Specificity**
HPS (revised cutoff)	.07	.95
MDQ (revised cutoff)	.11	.93
GBI-15 (revised cutoff)	.13	.96
Weighting based on Choi et al., 1992		
HPS (revised cutoff)	.09	.95
MDQ (revised cutoff)	.11	.93
GBI-15 (revised cutoff)	.26	.93

Table 11

Mean scores (and standard deviation) on each screener for different diagnostic groups across the entire sample

	HPS	MDQ	GBI-15
No bipolar spectrum diagnosis (n = 85)	25.03 (8.75)	9.78 (2.45)	2.27 (2.42)
Bipolar II, NOS, or cyclothymia (n = 11)	29.36 (6.23)	9.82 (2.93)	3.36 (1.86)
Bipolar I diagnosis (n = 7)	26.29 (9.12)	11.43 (2.23)	3.57 (1.51)

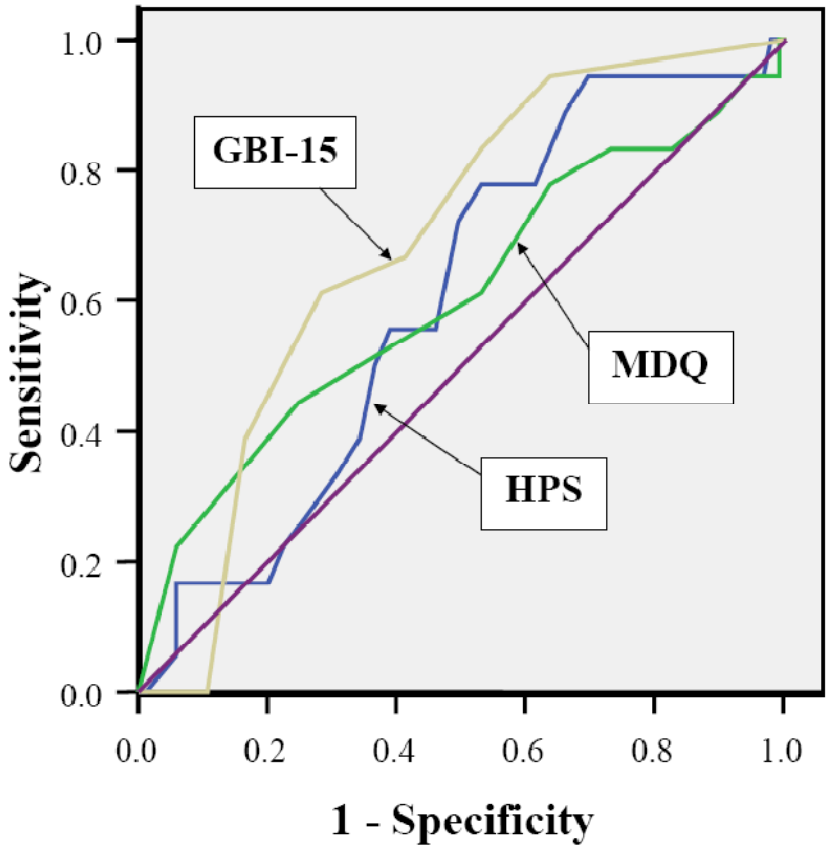

Figure 2. Receiver Operating Characteristic curves for each of three screeners. Area under the curve for the HPS was .602, for the MDQ was .606, and for the GBI-15 was .690.

Secondary Analyses

Table 12 presents correlations among each of the screeners and several self-report measures of constructs that have been tied to bipolar disorder in previous studies. The HPS correlated strongly and significantly with total scores and/or subscales of all measures other than the Barratt Impulsivity Scale, whereas the MDQ and GBI-15 correlated significantly with only the Barratt Impulsivity Scale. Thus, it appears that the HPS relates to different aspects of the potential risk for bipolar disorder than do the MDQ and GBI-15.

Table 12 also includes correlations between a bipolar spectrum diagnosis and scores on these measures potentially related to bipolar disorder. Only one subscale – the Affiliation subscale of the PRF – correlated significantly with diagnosis (r = -.27), and it correlated negatively rather than positively. That is, within this sample, bipolar spectrum diagnoses did not correlate with expected constructs. Parallel analyses were conducted to examine whether persons with bipolar I diagnoses (n = 7) were differentiated from healthy controls (n = 85) on these measures. Again, diagnostic groups differed significantly on only one subscale – the Positive Impact on the World subscale of the WASSUP, $t(84) = 2.27$, $p = .03$. This statistically significant difference may represent an artifact of the number of comparisons conducted. With that exception results were comparable, in that diagnostic status was unrelated to these measures.

Table 12

Correlations of measures potentially related to bipolar disorder with bipolar screeners and bipolar spectrum diagnosis (n = 103)

	HPS	MDQ	GBI-15	Bipolar Spectrum
NPI	.59**	.07	.09	.13
WASSUP				
Popular fame	.40**	-.03	-.01	-.01
Idealized friendships	.24*	.10	.04	-.03
Idealized family relationships	.05	.19	-.02	-.05
Positive impact on the world	.13	.13	.04	.11
Financial success	.26*	.11	-.10	.09
Political power	.28*	-.03	.11	.04
Creativity	.33**	.13	.14	.12
POG				
Lateral	.24*	.047	-.11	-.10
Social	.176	-.02	-.12	.04
Upward Generalization	.35**	.05	.00	.01
BIS-11	.04	.28*	.33*	.17
BMSIBS-CS	.43**	.19	.14	-.16
PRF				
Achievement	.11	-.09	-.09	-.03
Affiliation	.10	.12	-.14	-.27*
Dominance	.49**	-.01	.05	.06

Table 12, continued

Note. Correlations are drawn from the final sample.

NPI = Narcissistic Personality Inventory

WASSUP = Willingly Approached Set of Statistically Unlikely Pursuits

POG = Positive Generalization Scale

BIS-11 = Barratt Impulsiveness Scale Version 11

BMSIBS-CS = Social Dominance Scale

PRF = Personality Research Form

* $p < .05$ ** $p < .01$

Chapter 4: Discussion

The goal of this study was to investigate the psychometric properties of three brief self-report screeners for detecting bipolar disorder among undergraduates. The final sample (n = 103) was chosen via stratification, such that undergraduates with high scores on the three screeners were oversampled. Overall, the screeners demonstrated low to moderate utility for detecting bipolar I, bipolar II, bipolar NOS, and cyclothymia. Areas under the curve (AUC) generally fell between .60 and .70. Perfect detection of bipolar disorder and rejection of non-bipolar diagnosis would be represented by an AUC of 1.00, whereas detection of bipolar disorder and rejection of non-bipolar diagnosis at the chance level would be represented by an AUC of .50. Area under the curve analyses suggest that only the GBI-15 performed significantly better than chance at detecting bipolar disorder in this population.

Sensitivity and specificity bore out the finding of low to moderate utility for each of the screeners. Regardless of whether the original cutoffs or adjusted cutoffs (chosen to create balanced sampling in the current sample) were used, the GBI-15 generally demonstrated the best sensitivity (.85 in the sample using the adjusted cutoff).On the other hand, the HPS and MDQ demonstrated lower sensitivity in the range of .50 to .70, which meant that they were capturing approximately one half to two thirds of participants with a SCID-diagnosed bipolar spectrum condition. In terms of specificity, approximately half of those without a bipolar spectrum diagnosis in the final sample achieved a positive screen (specificity values ranging from .41 to .61).

The results presented above, however, have not been adjusted to account for the oversampling of high scorers I used in this study. Previous validational studies (e.g.

Depue et al., 1989; Klein et al., 1989; Meyer & Hautzinger, 2003) did not systematically oversample high scorers on the screening measures being tested when calculating sensitivity and specificity. By oversampling those with high scores on at least one screener, however, I minimized inclusion of participants with low scores, hence the aforementioned biases in sensitivity and specificity (Kraemer, 1992). Given practical limitations regarding the number of participants I could enroll in the study, this approach (oversampling high scorers) was necessary to enroll enough participants with a bipolar spectrum condition. Nonetheless, readers should keep in mind that results from AUC and sensitivity/specificity presented above are biased.

There are ways to adjust sensitivity and specificity outcomes that account for the stratification scheme used in this study (Table 10). Unfortunately, these results were not particularly encouraging: if used in a general college population similar to the one used in this study, these screeners would be expected to miss between 74 and 93 percent of bipolar spectrum diagnoses. These results compare unfavorably to most previous reports in the literature (see Tables 1-6; Angst et al., 2005; Depue et al., 1989; Klein et al., 1989; Ghaemi et al., 2005). My results are somewhat similar, however, to those found by Hirschfeld and colleagues (2003) in testing the MDQ in the general population. Reliance on the original cutoffs for each of the screeners (rather than the adjusted cutoffs) would not have fundamentally changed these results. It is important to note, however, that these adjustments to sensitivity and specificity relied heavily on the relatively small screen-negative groups for each screener. They should thus be interpreted with caution.

Fortunately, positive and negative predictive values provide additional information on a screener's usefulness that are unbiased by the study's sampling

procedure. Positive predictive value (PPV) is defined as the proportion of participants who score above the cutoff on a screener who have a bipolar spectrum condition according to the SCID. The PPV of a screening measure is a useful statistic if a researcher or clinician wants to know how fruitful it would be to administer more detailed diagnostic assessments to those people who scored above the cutoff on a particular screener. A screener's PPV is therefore subtly different from its sensitivity: PPV is the proportion of those with a positive screen who achieve the diagnosis in question, while sensitivity is the proportion of those with the diagnosis in question who screen positive. 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 approximately one in three high-scorers achieving a bipolar spectrum diagnosis according to the SCID. Given that approximately one in twenty undergraduates in an unscreened population might be expected to have a bipolar spectrum disorder based on previous studies, the positive predictive value for the GBI-15 is encouraging.

Negative predictive value (NPV) is defined as the proportion of participants scoring below the cutoff on a screener (i.e., who screened negative) who do not have a bipolar spectrum condition according to the SCID. (This is in contrast to specificity, which is the proportion of participants without a diagnosis who screened negative.) In general, the NPV for the screeners in this study were high (between .88 and .92 depending on the particular screener and cutoff chosen). This means that a participant scoring below the cutoff on one of these three screeners is generally unlikely to have a bipolar spectrum condition. This rate of "false negative" screens might lead to

unacceptable costs in certain clinical settings (Matza et al., 2005). In research or other settings, however, this could represent an acceptably high NPV, as costs of misdiagnosis can vary depending upon the setting.

Taken together, the findings summarized above present a somewhat mixed picture regarding the overall utility of the HPS, MDQ, and GBI-15 for detecting bipolar disorder among undergraduates. Areas under the curve, as well as sensitivity and specificity, generally fell in the weak to moderate range. These estimates were likely biased downward, however, by the current method of oversampling high scorers on each screener. In contrast, PPV and NPV suggest that administration of these screeners (especially the GBI-15) could help clinicians or college counselors detect students in need of more detailed follow-up assessments.

Several issues might help explain the differential performance of these three screeners. One area to consider is the unique construction of the MDQ, which features both a summing of item scores (items 1-13) and two critical items (the simultaneity and severity criteria). A full three quarters of the total screening sample endorsed the original cutoff of seven or more of the initial thirteen items. Many of the items on the scale appear to tap constructs that are perceived as normative by college students. For example, one item asks if the participant "has ever had a time… when you were much more active or did many more things than usual." Many, if not most, college students likely go through such times routinely (e.g. for exams, final projects, etc.), and in fact several items on the MDQ were endorsed by more than three quarters of the screening sample. Using the original cutoff, then, meant that the severity criterion (where participants rate how much of a problem their bipolar-related symptoms caused them) carried most of the burden for

determining whether participants received a positive screen. Previous research suggests, however, that requiring that symptoms cause at least moderate impairment (according to the participant) can be detrimental to 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 (Isometsä et al., 2003; Miller et al., 2004). For the current study, then, a much higher threshold of eleven out of thirteen items was required for a positive screen. Even this high cutoff did not result in particularly good psychometric properties in this sample, and it has not been validated in any other sample. Its generalizability therefore remains unknown.

The MDQ's reliance on the severity and simultaneity criteria may have hurt its performance in this study in more concrete way. Specifically, it was impossible to incorporate the simultaneity and severity criterion into calculations of the AUC for the MDQ. Nonetheless, the other results (sensitivity, specificity, PPV, and NPV) were unaffected by this caveat and also suggested that the MDQ did not perform very well. These results, especially when combined with previous reports on the use of the MDQ in the general population (e.g. Hirschfeld et al., 2003), suggest that the MDQ is not a good choice for bipolar screening in college settings.

There are other differences among the screeners that might help explain the differential results in this sample. One important area to consider is item content. All thirteen of the initial items of the MDQ can be more or less directly tied to symptoms of a manic or hypomanic episode listed in the DSM-IV. Some HPS items also align closely with DSM-IV criteria (e.g. "I frequently find that my thoughts are racing"). About half of

the HPS items, however, do not focus on episodes of symptoms, but rather ask about more enduring trait-like qualities. For instance, consider the following HPS items:

> I am so good at controlling others that it sometimes scares me.
> Many people consider me to be amusing but kind of eccentric.
> I am no more self-aware than the majority of people.

None of these items can be tied directly to DSM-IV symptoms of bipolar disorder, and in addition, none of them hint at the episodic nature of bipolar disorder. More broadly, the HPS was the only screener that correlated significantly with a host of measures tapping trait-like constructs previously related to bipolar disorder, but that do not play an explicit role in the DSM definition of mania. These constructs include dominance, ambitious goals, and narcissistic traits. It is noteworthy, however, that no HPS items specifically address impulsivity, while items tapping impulsivity appear in both the MDQ and GBI-15.

Almost all of the GBI-15 items can be directly tied to DSM-IV mania symptoms. It contains some items that ask about manic symptoms alone ("up" periods), as well as other items that ask about a combination of manic and depressive symptoms (both "up" and "down"). Among the screeners studied, this format is unique to the GBI-15. The DSM-IV does not require depressive symptoms to achieve a diagnosis of bipolar I disorder, but does require at least one depressive episode to achieve a diagnosis of bipolar II disorder. Unfortunately the current study only enrolled six participants with bipolar II disorder. Thus, I was unable to detect whether the inclusion of depressive symptoms in the GBI-15 made it especially sensitive to that particular diagnosis. It remains possible that this unique feature of the GBI-15 may have had something to do with its success in the current sample.

Another difference among the screeners studied is the breadth or complexity of the items. Specifically, all of the GBI-15 items are multidimensional. Consider the item below:

> 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?

The answer choices for this item range from "never or hardly ever" to "very often or almost constantly." Thus this singular item taps the duration, severity, and frequency of physical restlessness, all in the context of happy mood and high energy. The HPS and MDQ items, in contrast, are much simpler. With the exception of the lone severity criterion of the MDQ, no items from it or the HPS explicitly assess the severity of the symptoms in question. They also do not assess the duration of the symptoms experienced. The lack of attention to duration, in particular, may account for some of the false positive screens for those measures, as many participants experienced symptoms for only a few hours or days at a time.

In addition to the item content, the format of the answer choices may help explain the relative success of the GBI-15 when compared to the other screeners. All HPS items are answered as simply "true" or "false". With the exception of the severity item, all MDQ items are answered as simply "yes" or "no," referring to whether the respondent has had a period of time like the one described in the item. As was mentioned above, however, the GBI-15 answers range from "never or hardly ever" to "very often or almost constantly" (four answer choices per item in total). This built-in ability to assess the frequency of such symptoms set the GBI-15 apart from the other screeners studied. In sum, the GBI is distinguished from other measures by its clear and explicit alignment

with DSM-IV symptoms (of bipolar disorder and unipolar depression alike), the complex and multidimensional nature of its items, its specific attention to symptom duration and frequency, and its more detailed answer format.

The paragraphs above have sought to address the issue of why the GBI-15 appeared to outperform the other two screeners. A more general question remains, though: why did total scores from all of the screeners relate at such a low level with diagnosis? The screeners demonstrated only small correlations with diagnostic status, and the overall AUC results from this study were similarly low. These findings paralleled those from a meta-analysis conducted by Youngstrom and colleagues (2007) on detecting bipolar disorder in youth. They found that parent report provided better diagnostic utility, suggesting that children and adolescents may be unable to report their bipolar symptoms accurately. It is possible that for the current sample – with an average age of 18.62 years, barely out of adolescence themselves – parental report on bipolar symptoms might have proven more in informative than self-report. This argument could be applied either to the screeners or to the reference standard diagnostic interview itself (the SCID). Even though this sample was technically composed of adults, supplementing the SCID with parental report might have led to different results.

The results from secondary analyses also raise the possibility that the SCID diagnoses, not the screeners themselves, may be the source of the weak correlations observed in this study. Previous research, for instance, has demonstrated high comorbidity between bipolar disorder and narcissistic personality disorder (e.g.; Brieger, Ehrt, & Marneros, 2003; Garno et al., 2005; George, Miklowitz, Richards, Simoneau, & Taylor, 2003; Stormberg et al., 1998). Scores on impulsivity scales have also been

demonstrated to be higher among those diagnosed with bipolar disorder than among normal controls (Peluso et al., 2007; Swann et al., 2003; Swann, Steinberg, Lijffijt, & Moeller, 2008). In addition to narcissism and impulsivity, existing research also supports a strong connection between bipolar diagnoses and unrealistic goal pursuit (Johnson, Eisner, & Carver, under review). The current study did not replicate any of these previously demonstrated correlations between bipolar diagnosis and measures of narcissism, impulsivity, or goal pursuit. Although reliability analyses suggested that the SCID was administered properly and bipolar spectrum diagnoses assigned correctly, the low correlations between SCID diagnosis and all other measures completed during the study is a troubling issue. One possible explanation for this result, however, is related to current symptoms. Other studies, for instance, have found that current depressive symptoms can suppress expected elevations in reward sensitivity among those with bipolar disorder (Meyer, Johnson, & Winters, 2001). Along a similar vein, manic symptoms may intensify levels of impulsivity (Moeller et al., 2001; Swann et al., 2008) and narcissism (Stormberg et al., 1998). Given the links between these constructs and symptom severity levels, the absence of current symptom severity indices in this study may have limited the ability to detect correlations with diagnosis. Without more data, however, this hypothesis remains entirely speculative.

Another possible explanation for the low correlations between bipolar spectrum diagnosis and measures such as impulsivity and narcissism is that participants were enrolled before the typical age of onset of bipolar disorder. Given a sufficient follow-up, it is likely that some participants without a current bipolar diagnosis may develop one. Indeed, Kwapil and colleagues (2000) found that the HPS predicted bipolar diagnoses

even at thirteen year follow-up. Other studies conducted with college populations, though, have still found robust correlations between bipolar spectrum diagnoses and many of the measures administered in the current study, so this represents, at best, a partial explanation. Statistical power should also be considered. Formal power analyses were not conducted for the study. Although many correlations were in the expected direction, they were very small overall. Only five scale or subscale scores correlated with diagnosis at a level higher than $r = \pm.10$, and two of these correlations were not in the expected direction. It thus seems unlikely that low power by itself can account for these results. Despite the oversampling of high scorers on the self-report screeners, only seven participants achieved a diagnosis of bipolar I disorder, and eleven additional participants achieved diagnoses elsewhere on the bipolar spectrum. Given these low numbers, results from this study should be interpreted with caution.

In summary, the current study produced mixed results regarding the usefulness of the three screeners studied. Areas Under the Curve for the screeners were lower than expected given previous reports in the literature, but were negatively biased by the study's sampling procedure. Sensitivities, adjusted to control for the study's sampling procedure, were very low for each screener (ranging from .06 to .24), although adjusted specificities were generally good (ranging from .93 to .95). PPV and NPV are not biased by the current design, and The GBI-15 performed moderately well in terms of PPV, with approximately one in three positive screens achieving a bipolar spectrum diagnosis on the SCID. The study also revealed lower-than-expected correlations between diagnosis and related constructs such as impulsivity, social dominance, and reward sensitivity. The study was limited by enrollment of few participants with a bipolar spectrum condition,

and it is possible that college students may be poor informants on the SCID interview itself.

Future directions

The overarching goal of the current study was to evaluate the potential usefulness of three screening measures to detect bipolar disorder in undergraduates (the HPS, MDQ, and GBI-15). Results revealed that, of the three, the GBI-15 appeared to be the most effective, but there is still room for improvement in this domain.

Perhaps the most striking result in this sample was that many participants with high scores on the screeners did not have a bipolar spectrum diagnosis according to the SCID. This was especially striking for the MDQ, as I had to *raise* the cutoff for a positive screen on that measure. One would expect this would result in a "purified" sample, highly likely to be suffering from a bipolar spectrum diagnosis, but this did not appear to be the case.

What do these results suggest for future refinements to screening tools? It is worth noting that the current screeners do not mirror DSM-IV criteria for symptom duration or severity, and that improvements in this domain could prove helpful. As noted above, the GBI-15 items provide more coverage of duration. Qualitatively, it seemed that many participants had experienced bipolar symptoms, but were ruled out of a bipolar spectrum diagnosis specifically because their symptoms did not last four days (the minimum duration to be rated as a hypomanic episode on the SCID). There are several potential ways to incorporate duration criteria into a screening measure. First, it is possible to make each item multidimensional (assessing severity, frequency, and/or duration

simultaneously as the GBI-15 did). A potential drawback to this method is that multidimensional questions are difficult to understand.

Another option would be to include a "critical item" that specifically assesses the duration of previously-reported symptoms. For instance, at the end of the measure , a simple yes/no item could be added asking "did the experiences you endorsed above last at least four days?" The MDQ uses a similar methodology to assess the simultaneity and severity of symptoms. A potential drawback of this method is that it makes calculation of area under the curve (and therefore comparison to other measures) difficult. Another concern is that this approach presumes that symptoms last a similar duration; participants whose symptoms differ in duration may have a difficult time providing an accurate response.

Yet another option for addressing the duration of symptoms within a screener would be to add a set of answer choices to each question. For instance, a question might ask about a manic symptom: "have you ever had a period of time when you felt you needed very little sleep?" In addition to answering yes or no to this item, an additional item could ask "did this period last at least four days?" Only people who answered yes to *both* items would be considered to have the manic symptom of "little need for sleep." This type of answer format allows duration to be assessed without (a) resorting to potentially confusing multidimensional items or (b) adding "critical items" that make psychometric analysis difficult. This type of strategy has been used successfully in depression scales such as the Inventory to Diagnose Depression, Lifetime Version (IDDL, Zimmerman & Coryell, 1987).

Beyond designing screeners that more accurately capture the DSM criteria, another approach is to supplement symptom questions with measures of trait-like characteristics believed to increase risk for bipolar disorder. For example, Kwapil and colleagues (2000) found that the HPS interacted with a measure of impulsivity to predict diagnoses of bipolar spectrum disorder among undergraduates. Numerous personality traits and behavioral constructs, such as narcissism, social dominance, and goal-seeking, have been correlated with bipolar disorder in previous studies, suggesting that questions about such qualities might aid in bipolar screening. Unfortunately, unlike previous studies (e.g. Johnson & Carver, 2006; Eisner et al., in press; Stormberg et al., 1996; Peluso et al., 1997; Swann et al., 2003), these constructs were uncorrelated with bipolar spectrum diagnoses in the current sample. The small correlations do not provide hope that these measures will perform well as screening measures. .

Beyond the hypothesized risk constructs measured in this study, other risk variables are worth considering. The MDQ in its original form, for instance, includes a question about family history of bipolar disorder. This question is not technically included in calculation of a positive or negative screen for the MDQ, and so it was not administered for this study. It remains to be seen whether such questions – questions that go beyond the DSM-IV to tap family history , or other risk factors for bipolar disorder – can prove helpful in a screening measure.

In sum, larger samples will be needed to assess the replicability of current findings. Nonetheless, findings of this study suggest that current screening measures for bipolar disorder may identify a large number of false positive cases in college samples while also missing a large proportion of bipolar cases. There is a need for screeners that

more carefully assess the severity and duration of symptoms, as well as a need to consider a broader range of constructs that might help identify persons at risk for bipolar disorder.

References

Akiskal, H. S., & Pinto, O. (1999). The evolving bipolar spectrum: Prototypes I, II, III and IV. *Psychiatric Clinics of North America, 22*, 517-534.

American Psychiatric Assn, Washington, DC, US. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision)*. Washington, DC, US: American Psychiatric Publishing, Inc.

Andlin-Sobocki, P., & Wittchen, H. U. (2005). Cost of affective disorders in Europe. *European Journal of Neurology, 12*(s1), 34-38.

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. *Journal of Affective Disorders, 88*, 217-233.

Angst, J., & Cassano, G. (2005). The mood spectrum: Improving the diagnosis of bipolar disorder. *Bipolar Disorders, 7*, 4-12.

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: Journal of the American Medical Association, 290*, 1517-1519.

Bagby, M. R., Marshall, M. B., Basso, M. R., Nicholson, R. A., Bacchiochi, J., & Miller, L. S. (2005). Distinguishing bipolar depression, major depression, and schizophrenia with the MMPI-2 clinical and content scales. *Journal of Personality Assessment, 84*, 89-95.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry,* (4), 561-571.

Bellivier, F., Golmard, J. L., Rietschel, M., Schulze, T. G., Malafosse, A., & Preisig, M. et al. (2003). Age at onset in bipolar I affective disorder: Further evidence for three subgroups. *American Journal of Psychiatry, 160*, 999-1001.

Benazzi, F., & Akiskal, H. S. (2003). Refining the evaluation of bipolar II: Beyond the strict SCID-CV guidelines for hypomania. *Journal of Affective Disorders, 73*, 33-38.

Blader, J. C., & Carlson, G. A. (2007). Increased rates of bipolar disorder diagnoses among U.S. child, adolescent, and adult inpatients, 1996-2004. *Biological Psychiatry, 62*, 107-114.

Brickman, A., LoPiccolo, C. & Johnson, S. L. (2002). Screening for Bipolar Disorder. *Psychiatric Services, 53,* 349. Letter to the editor.

Brieger, P., Ehrt, U., & Marneros, A. (2003). Frequency of comorbid personality disorders in bipolar and unipolar affective disorders. *Comprehensive Psychiatry, 44*, 28-34.

Calabrese, J. R., Hirschfeld, R. M. A., Frye, M. A., & Reed, M. L. (2004). Impact of depressive symptoms compared with manic symptoms in bipolar disorder: Results of a U.S. community-based sample. *Journal of Clinical Psychiatry, 65*, 1499-1504.

Carter, T. D., Mundo, E., Parikh, S. V., & Kennedy, J. L. (2003). Early age at onset as a risk factor for poor outcome of bipolar disorder. *Journal of Psychiatric Research, 37*, 297-303.

Choi, B. C. K. (1992). Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *Journal of Clinical Epidemiology*, 45, 581-586.

Cuellar, A. K., Johnson, S. L., & Winters, R. (2005). Distinctions between bipolar and unipolar depression. *Clinical Psychology Review, 25*, 307-339.

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. *Journal of Abnormal Psychology, 98*, 117-126.

Dunner, D. L. (2003). Clinical consequences of under-recognized bipolar spectrum disorder. *Bipolar Disorders, 5*, 456-463.

Dunner, D. L., & Tay, L. K. (1993, September/October). Diagnostic reliability of the history of hypomania in bipolar II patients and patients with major depression. *Journal of Comprehensive Psychiatry, 34,* 303-307.

Eckblad, M., & Chapman, L. J. (1986). Development and validation of a scale for hypomanic personality. *Journal of Abnormal Psychology, 95*, 214-222.

Eisner, L. R., Johnson, S. L., & Carver, C. S. (in press). Cognitive responses to failure and success relate uniquely to bipolar depression versus mania. *Journal of Abnormal Psychology.*

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of General Psychiatry, 35*, 837-844.

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

Fischer, J. E., Bachman, L. M., Jaeschke, R. A. (2003). A reader's guide to the interpretation of diagnostic test properties: Clinical examples of sepsis. *Intensive Care Medicine, 29*, 1043-1051.

Fulford, D. C., Johnson, S. L., & Carver, C. S. (2007). Commonalities and distinctions in risk factors for subclinical narcissism and mania: Three forms of dysregulation. Manuscript in preparation.

Garno, J. L., Goldberg, J. F., Ramirez, P. M., Ritzler, B. A. (2005). Bipolar disorder with comorbid cluster B personality disorder features: impact on suicidality. *Journal of Clinical Psychiatry, 66*, 339-345.

George, E. L., Miklowitz, D. J., Richards, J. A., Simoneau, T. L. & Taylor, D. O. (2003). The comorbidity of bipolar disorders and axis II personality disorders: prevalence and clinical correlates. *Bipolar Disorders, 5,* 115-122.

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. *Journal of Affective Disorders, 84*, 273-277.

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. *American Journal of Psychiatry, 161*, 163-165.

Ghaemi, S. N., Sachs, G. S., Chiou, A. M., Pandurangi, A. K., & Goodwin, F. K. (1999). Is bipolar disorder still underdiagnosed? are antidepressants overutilized? *Journal of Affective Disorders, 52*, 135-144.

Grandin, L. D., Alloy, L. B., & Abramson, L. Y. (2007). Childhood stressful life events and bipolar spectrum disorders. *Journal of Social and Clinical Psychology, 26*, 460-478.

Gruber, J., & Johnson, S. L. (under review). Positive emotional traits and ambitious goals among people putatively at risk for mania: The need for specificity.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

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. *Journal of the American Board of Family Practice, 18*, 233-239.

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. *American Journal of Psychiatry, 160*, 178-180.

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. *American Journal of Psychiatry, 157*, 1873-1875.

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*.

Jackson, D. N. (1967). *Personality Research Form*. Goshen, NY: Research Psychologists Press.

Jackson, D. N. (1989). *Personality Research Form Manual, 4th Edition*. Port Huron, MI: Sigma Assessment Systems, Inc.

Johnson, S. L. (2005). Mania and dysregulation in goal pursuit: A review. *Clinical Psychology Review, 25*, 241-262.

Johnson, S. L., & Carver, C. S. (2006). Extreme goal setting and vulnerability to mania among undiagnosed young adults. *Cognitive Therapy and Research, 30*, 377-395.

Johnson, S. L., Eisner, L., & Carver, C. S. (under review). Unrealistic goal setting among persons diagnosed with bipolar disorders.

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. *Journal of Affective Disorders, 73*, 123-131.

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. *Archives of General Psychiatry, 62*, 617-627.

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 from the national comorbidity study. *Archives of General 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. *Psychological Medicine, 27*, 1079-1089.

Klein, D. N., Dickstein, S., Taylor, E. B., & Harding, K. (1989). Identifying chronic affective disorders in outpatients: Validation of the General Behavior Inventory. *Journal of Consulting and Clinical Psychology, 57*, 106-111.

Kraemer, H. C. (1992). *Evaluating Medical Tests*. Newbury Park, CA, US: Sage Publications.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Medical Care, 41*, 1284-1292.

Kupfer, D. J. (2005). The increasing medical burden in bipolar disorder. *JAMA: Journal of the American Medical Association, 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. *Journal of Abnormal Psychology, 109*, 222-226.

Leboyer, M., Henry, C., Paillere-Martinot, M. L., & Bellivier, F. (2005). Age at onset in bipolar affective disorders: A review. *Bipolar Disorders, 7*, 111-118.

Lewinsohn, P. M., Rohde, P., Seeley, J. R., Klein, D. N., & Gotlib, I. H. (2000). Natural course of major depressive disorder in a community sample: Predictors of recurrence in young adults. *American Journal of Psychiatry, 157*, 1584-1591.

Lewis, L. (2004). Dual diagnosis: The Depression and Bipolar Support Alliance's patient perspective. *Biological Psychiatry, 56*, 728-729.

Lish, J. D., Dime-Meenan, S., Whybrow, P. C., Price, R. A., et al. (1994). The National Depressive and Manic-Depressive Association (DMDA) survey of bipolar members. *Journal of Affective Disorders, 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 Disorders, 6*, 395-405.

Matza, L. S., Rajagopalan, K., Thompson, C. L., & de Lissovoy, G. (2005). Misdiagnosed patients with bipolar disorder: Comorbidities, treatment patterns, and direct treatment costs. *Journal of Clinical Psychiatry, 66*, 1432-1440.

Mendlowicz, M. V., Jean Louis, G., Kelsoe, J. R., & Akiskal, H. S. (2005). A comparison of recovered bipolar patients, healthy relatives of bipolar probands, and normal controls using the short TEMPS-A. *Journal of Affective Disorders, 85*, 147-151.

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. *Archives of General Psychiatry, 64*, 543-552.

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.

Meyer, T. D., Hammelstein, P., Nilsson, L. G., Skeppar, P., Adolfsson, R., & Angst, J. (2007). The Hypomania Checklist (HCL-32): Its factorial structure and association to indices of impairment in German and Swedish nonclinical samples. *Comprehensive Psychiatry, 48*, 79-87.

Meyer, T. D., & Hautzinger, M. (2003). Screening for bipolar disorders using the Hypomanic Personality Scale. *Journal of Affective Disorders, 75*, 149-154.

Mezzich, J. E., & Cranach, M. v. (Eds.). (1988). *International classification in psychiatry: Unity and diversity* (xxi ed.). New York, NY, US: Cambridge University Press.

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

Moeller, F. G., Barratt, E. S., Dougherty, D. M., Schmitz, J. M., & Swann, A. C. (2001). Psychiatric aspects of impulsivity. *American Journal of Psychiatry, 158*, 1783-1793.

Moneta, G. B., & Wong, F. H. Y. (2001). Construct validity of the Chinese adaptation of four thematic scales of the Personality Research Form. *Social Behavior and Personality, 29*, 459-476.

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

Morf, C. C., & Rhodewalt, F. (1993). Narcissism and self-evaluation maintenance: Explorations in object relations. *Personality and Social Psychology Bulletin, 19*, 668-676.

Mulrow, C. D., Williams Jr., J. W., Gerety, M. B., Ramirez, G., Montiel, O. M., & Kerber, C. (1995). Case-finding instruments for depression in primary care settings. *Annals of Internal Medicine, 122*, 913-921.

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. *Archives of General Psychiatry, 59*, 115-123.

Nowakowska, C., Strong, C. M., Santosa, C. M., Wang, P. W., & Ketter, T. A. (2005). Temperamental commonalities and differences in euthymic mood disorder patients, creative controls, and healthy controls. *Journal of Affective Disorders, 85*, 207-215.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*, 768-774.

Peele, P. B., Xu, Y., & Kupfer, D. J. (2003). Insurance expenditures on bipolar disorder: Clinical and parity implications. *American Journal of Psychiatry, 160*, 1286-1290.

Peluso, M. A. M., Hatch, J. P., Glahn, D. C., Monkul, E. S., Sanches, M., Najt, P., Bowden, C. L., et al. (2007). Trait impulsivity in patients with mood disorders. *Journal of Affective Disorders, 100*, 227-231.

Peralta, V., Cuesta, M. J. (1998). Lack of insight in mood disorders. *Journal of Affective Disorders, 49*, 55-58.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Raskin, R., & Novacek, J. (1989). An MMPI description of the narcissistic personality. *Journal of Personality Assessment, 53*, 66-80.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54*, 890-902.

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 Psychiatrica Scandinavica, 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. *Archives of General Psychiatry, 50*, 85-94.

Siegel, R., Johnson, S. L., Miller, C. J. (2007). Social dominance and bipolar disorder. Manuscript in preparation.

Simpson, S. G., McMahon, F. J., McInnis, M. G., MacKinnon, D. F., Edwin, D., Folstein, S. E., DePaulo, & J. R. (2002). Diagnostic reliability of bipolar II disorder. *Archives of General Psychiatry, 59*, 746-740.

Stormberg, D. Ronningstam, E., Gunderson, J., & Tohen, M. (1998). Pathological narcissism in bipolar disorder patients. *Journal of Personality Disorders, 12*, 179-185.

Sukhatme, S., & Beam, C. A. (1994). Stratification in nonparametric ROC studies. *Biometrics*, 50, 149-163.

Swann, A. C., Pazzaglia, P., Nicholls, A., Dougherty, D. M., & Moeller, F. G. (2003). Impulsivity and phase of illness in bipolar disorder. *Journal of Affective Disorders, 73*, 105-111.

Swann, A. C., Steinberg, J. L., Lijffijt, M., & Moeller, F. G. (2008). Impulsivity: differential relationship to depression and mania in bipolar disorder. *Journal of Affective Disorders, 106*, 241-248.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third Edition*. San Antonio, TX: The Psychological Corporation.

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. *Medical 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. *Journal of the American Medical Association, 276*, 293-299.

Woodcock, R. V., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside Publishing.

Youngstrom, E. A. (2004). Improving diagnosis and measuring clinically significant change. *Association for the Advancement of Behavior Therapy,* New Orleans.

Youngstrom, E. A., Findling, R. L., Youngstrom, J. K., & Calabrese, J. R. (2005). Toward an evidence-based assessment of pediatric bipolar disorder. *Journal of Clinical Child and Adolescent Psychology, 34*, 433-448.

Youngstrom, E. A., Joseph, M. F., Miller, C. J., Frazier, T., Meyers, O. I., & Findling, R. L. (2007, August). Validity of parent report of bipolar symptoms in youth: A meta-analysis. Poster presentation at the American Psychological Association 115th Annual Meeting, San Francisco, CA.

Zimmerman, M., & Coryell, W. (1987). The inventory to diagnose depression, lifetime version. *Acta Psychiatrica Scandinavica, 75*, 495-499.

Zimmerman, M., Coryell, W., Corenthal, C., & Wilson, S. (1986). A self-report scale to diagnose major depressive disorder. *Archives of General Psychiatry, 43*, 1076-1081.