Using Computerized Adaptive Testing to Reduce the Burden of Mental Health Assessment

Robert D. Gibbons, Ph.D.
David J. Weiss, Ph.D.
David J. Kupfer, M.D.
Ellen Frank, Ph.D.
Andrea Fagiolini, M.D.

Victoria J. Grochocinski, Ph.D.
Dulal K. Bhaumik, Ph.D.
Angela Stover, M.A.
R. Darrell Bock, Ph.D.
Jason C. Immekus, Ph.D.

Objective: This study investigated the combination of item response theory and computerized adaptive testing (CAT) for psychiatric measurement as a means of reducing the burden of research and clinical assessments. Methods: Data were from 800 participants in outpatient treatment for a mood or anxiety disorder; they completed 616 items of the 626-item Mood and Anxiety Spectrum Scales (MASS) at two times. The first administration was used to design and evaluate a CAT version of the MASS by using post hoc simulation. The second confirmed the functioning of CAT in live testing. Results: Tests of competing models based on item response theory supported the scale’s bifactor structure, consisting of a primary dimension and four group factors (mood, panic-agoraphobia, obsessive-compulsive, and social phobia). Both simulated and live CAT showed a 95% average reduction (585 items) in items administered (24 and 30 items, respectively) compared with administration of the full MASS. The correlation between scores on the full MASS and the CAT version was .93. For the mood disorder subscale, differences in scores between two groups of depressed patients—one with bipolar disorder and one without—on the full scale and on the CAT showed effect sizes of .63 (p<.003) and 1.19 (p<.001) standard deviation units, respectively, indicating better discriminant validity for CAT. Conclusions: Instead of using small fixed-length tests, clinicians can create item banks with a large item pool, and a small set of the items most relevant for a given individual can be administered with no loss of information, yielding a dramatic reduction in administration time and patient and clinician burden. (Psychiatric Services 59:361–368, 2008)

Dr. Gibbons, Dr. Bhaumik, and Dr. Bock are affiliated with the Center for Health Statistics, University of Illinois at Chicago, Psychiatric Institute 457, M/C 912, Chicago, IL 60680-6998 (e-mail: rdgib@uic.edu). Dr. Gibbons and Dr. Bhaumik are also with the Departments of Psychiatry and Biostatistics at the university. Dr. Weiss is with the Department of Psychology, University of Minnesota, Minneapolis. Dr. Kupfer, Dr. Frank, Dr. Fagiolini, Dr. Grochocinski, and Ms. Stover are with the Department of Psychiatry, Western Psychiatric Institute and Clinic, University of Pittsburgh School of Medicine, where Dr. Frank is also with the Department of Psychology. Dr. Immekus is with the Kremen School of Education and Human Development, California State University at Fresno.
A complication of applying item response theory to psychiatric measurement problems is that unlike traditional ability testing (for example, mathematics achievement), for which approximately unidimensional scales are used, psychiatric measurement scales are generally multidimensional. A primary reason for the multidimensionality of psychiatric measurement scales is that items are often sampled from multiple domains (for example, various mood disorders). Bock and colleagues (11,12) extended the item response theory model to the multidimensional case, where each item is related to multiple underlying factors. Despite the availability of these item response theory models (11–14), they have not been well studied in the context of CAT. A major limitation of multidimensional CAT is that it involves selection of items that will increase precision of measurement simultaneously on all dimensions of interest. This difficulty arises from the complication of obtaining accurate item parameter and trait estimates across a large number of dimensions (11,12).

A plausible alternative factor structure is the “bifactor” model (15). The bifactor solution constrains each item to have a nonzero loading on the primary dimension (for example, depression) and a secondary loading on no more than one of the domain factors (for example, sleep disturbance) (16). The bifactor structure is plausible in psychiatric measurement, in which symptom items that are related to a primary dimension of interest are often selected from underlying measurement subdomains.

Gibbons and Hedeker (15) derived a bifactor model for binary response data, and Gibbons and colleagues (17) extended it for analysis of graded response data. Their estimation method permits the items to be sampled from any number of subdomains. In the context of psychiatric measurement, the advantage of the bifactor model is that it yields a measure of overall impairment that can be the focus of CAT. [Statistical details of the bifactor model are provided in an online appendix to this article at ps.psychiatryonline.org.]

In settings in which CAT is used, items are selected during the process of test administration for each individual. Specifically, CAT allows the test administrator to control measurement precision and to maximize the efficiency of the testing process. Characteristics of an adaptive test include a precalibrated item bank, an item selection procedure, a scoring method, and a criterion for terminating the test. Research since the 1970s has shown that these four characteristics are most easily achieved by using item response theory (18–20). Commercially available software for implementing CAT on the basis of item response theory has been developed (21,22), and research shows that adaptive tests are more efficient than conventional tests; for example, CAT reduces testing time and the number of administered items (20,23). Adaptive tests can result in a 50%–90% reduction in the number of items administered, with no decrease in measurement quality (23).

In educational testing, more than one million tests were administered by using CAT in 1999; the number was projected to rise to ten million in the next few years (2).

Measurement models based on item response theory also include models that are applicable to personality instruments that are not dichotomously scored (24–28). In terms of psychiatric measurement, research has demonstrated that both item response theory and CAT (29) can be meaningfully applied to the measurement of attitudes and personality variables (30,31).

Notwithstanding the enormous utility of DSM-IV in advancing our understanding of the clinical course, biological bases, and treatment of mood disorders, there is increasing recognition that these disorders rarely come in the pure and seemingly isolated prototypes described in the current nomenclature. To this end Cassano and colleagues (32,33) defined the full “spectrum,” or domain, of clinical features of mood and anxiety disorders.

The Mood and Anxiety Spectrum Scales (MASS) consist of 626 items and assess four domains: mood (161 items), panic-agoraphobia (114 items), obsessive-compulsive disorder (183 items), and social phobia (168 items). Traditional psychometric properties of the MASS have been reported (34,35). Test-retest and interrater reliability ranged from .89 to .99, and measures for each domain showed concurrent validity compared with established measures of comparable constructs. Excellent levels of agreement have been shown between structured clinical interviews and the more efficient self-report measures: mood, r=.97 (36); panic-agoraphobia, r=.94 (37); obsessive-compulsive, r=.96 (38); and social phobia, r=.97 (39). The instruments were co-developed in Italian and English, and cross-cultural validity has also been demonstrated (39).

The MASS has been shown to have substantial clinical utility, displaying associations with treatment outcome in unipolar depression (40), with functional impairment (37), and with treatment outcome in bipolar disorder (32,41) that persist after traditional DSM diagnostic comorbidity is controlled for (32,42–44). Thus the aim of the study reported here was to expand on this research by applying item response theory to test the scale’s dimensionality and to investigate the utility of MASS scores derived in a CAT environment on the basis of the scale’s empirical factor structure. More broadly, the aim of this article is to show that CAT can provide a new paradigm for conceptualizing and operationalizing psychiatric measurement.

Methods
Participants
Data were from 800 participants (565 women, or 71%) in outpatient treatment for a mood or anxiety disorder at the Western Psychiatric Institute and Clinic in Pittsburgh from January 2003 through December 2005. The mean age was 38.5±11.6 years (range 18–66 years), with a mean education level of 14.3±2.70 years. Medical charts from all patients were reviewed for diagnostic information. This study was conducted with institutional review board approval, and informed consent was obtained from all participants.

The sample was representative of
patients seen at adult outpatient clinics at the Western Psychiatric Institute and Clinic and generalizable to the larger U.S. population of persons with depression. Across 14 published studies of population samples (that is, nontreatment samples), the median lifetime rate of major depression was found to be 10.8%, ranging from 9.0% to 24.4% (45). The average age of onset of recurrent unipolar major depression is between 30 and 35 years, whereas single-episode major depression usually begins some years earlier (46,47). Compared with men, women have an approximately two-fold elevation in lifetime rates of depression, panic disorder, generalized anxiety disorder, agoraphobia, and simple phobia (45–49). The mean age of our sample is consistent with epidemiological findings regarding the age of patients with recurrent depression and anxiety.

Of the 859 participants who consented, 59 (7%) were excluded from analyses because they met exclusion criteria after a medical history review following study entry. Of these, 59, 12 were excluded because they had schizophrenia or psychotic symptoms, 21 because of current drug or alcohol dependence, 19 because medical records were missing or there was no medical record to verify diagnoses, and five because they had an organic affective diagnosis; one person was excluded because of an antisocial personality diagnosis, and one was excluded because data were missing from the computer test.

Participants were recruited via flyers and clinician referrals from the Western Psychiatric Institute and Clinic. After a telephone screen to determine eligibility, ten participants declined to schedule an appointment. Thirty-six participants were screened and met eligibility criteria but did not show up for scheduled appointments.

The bifactor model uses all available item responses from each individual and can therefore handle missing item responses. However, because of the small number of patients with missing data (one patient), our analyses were restricted to patients with complete data.

Data analysis
The MASS factor structure was tested by comparing goodness of fit of a unidimensional item response theory model and a bifactor item response theory model. Comparison of –2 times the log-likelihood values were used to select the model that most accurately described the data (12), which was used for CAT administration of the MASS. Specifically, if the difference in log-likelihood value between the unidimensional and bifactor model was statistically significant, then it was concluded that the bifactor model more accurately described the scale’s multidimensional factor structure. Model-data fit was also judged in terms of correspondence between the observed and expected proportion of item categorical frequencies (17).

The item parameters used in the CAT administration of the MASS were based on data collected from an optimal balanced incomplete block design (50) that maximized the number of pairings of 616 of the 626 individual items while minimizing the number of items administered to each person. The FastTEST Professional Testing System (22) was used to create 36 different test forms, each consisting of 154 items extracted from the four MASS subscales. These binary-scored items (1, yes; 0, otherwise; 9, not administered) were administered by computer to 36 randomly assigned subgroups of research participants. Sample sizes for the 36 groups varied from 17 to 28, with a median of 22 participants per group.

In addition to analyses of the total sample of 800, we also analyzed data from a previous study conducted in Italy (36) (sample A), as well as two subsamples of the 800 participants (samples B and C) that were used to test the efficiency of CAT administration of the MASS (testing time and reduction in administered items). Sample A refers to complete responses (paper-and-pencil administration) to all 626 MASS items from 148 patients with depression in an earlier study conducted jointly by the University of Pittsburgh (N=90) and the University of Pisa (N=58) (36). Of the 148 participants with depression, 100 were women (68%), 136 were Caucasian (92%), six were Hispanic (4%), 56 were married (38%), and 80 were working full-time (54%). The mean±SD age was 38.11±12.15, and the mean educational level was 13.93±3.47 years (some post–high school education). Sample B refers to complete responses (computer administration) from 204 depressed patients from the study reported here. These data were used to compare administration of the full test and a simulated CAT (51)—that is, the comparison examined what would have happened if only a subset of responses had been administered with CAT. Sample C consisted of a subset of 156 research participants from sample B who were tested with live CAT administration (not simulated) of the bifactor-based CAT (mean of 5.5±1.3 months after administration of the full test).

Two analyses were used to investigate the validity of scores obtained with the CAT version of the MASS. First, CAT scores were correlated to full-scale MASS scores. This analysis provided a measure of the comparability of CAT scores to full-scale MASS scores, with values above .90 desired. Second, the utility of scores on the CAT version of the MASS for differentiating clinical groups was investigated on the basis of a contrasting-groups study. Specifically, differences in scores between two groups of patients with depression—one with bipolar disorder and one without—on the CAT version and the full-scale version of the mood disorder subscale were compared to identify significant differences and effect sizes.

Results
Demographic and clinical characteristics of the sample are shown in Table 1. The proportion of African Americans in the study was twice the proportion in the greater Pittsburgh area (26%, or 209 participants, in the study compared with 13% in the Pittsburgh–Allegheny County area). The diagnostic categories shown in Table 1 are not mutually exclusive because some participants had comorbid diagnoses. Participants had a
of administration

Results of the post hoc simulation analyses revealed that on average, 25 items in sample A (range 18–55 items) and 24 items in sample B (range 18–77 items) were needed to achieve convergence. This represents a savings of 96% (or approximately 591 items) over administration of the complete test. The overall correlation between the scores on the CAT version and the full-scale MASS was .92 for sample A and .93 for sample B.

Similar results were obtained with live CAT administration. The average number of items required for sample C was 30, representing a 95% reduction (586 items) in test length. Actual test administration time was 115 minutes for the 616 items administered to sample B and 22 minutes for the same group of patients in the CAT administration to sample C—a time savings of one and one-half hours. Figure 1 illustrates an item-by-item CAT administration of the MASS for a single participant for the primary dimension of impairment. The session begins with an impairment estimate of zero. With ad-

**Comparison of types of administration**


![Figure 1](image-url)

*Item-by-item administration to a study participant of the Mood and Anxiety Spectrum Scales by use of computerized adaptive testing*
administration of each new item the standard error shrinks, until after 26 items the impairment estimate is 1.26 with an SE of .297, just under the termination criterion of SE=.30. Figure 2 presents a frequency distribution of impairment estimates for the entire sample of 800 participants. Figure 2 shows good coverage of the entire impairment distribution, which should typically range from –4 to 4 since the impairment estimates are scaled to have a unit-normal distribution (that is, mean=0 and variance=1).

**Clinical utility of CAT**

A chart review was performed for six patients with mood disorders (three had major depressive disorder and three had bipolar disorder) who were patients of one psychiatrist, and an interview was conducted with the psychiatrist. Most of the CAT items that were endorsed positive were not documented in the six patients’ psychiatric evaluation report, progress notes, and Structured Clinical Interview for DSM-IV. These items included clinically important information, such as a history of manic symptoms, potentially risky behaviors, sexual dysfunction, and agoraphobic traits.

Table 2 reports the endorsed items of the first randomly selected participant, a 51-year-old African-American man with a diagnosis of major depressive disorder. Most of the lifetime symptoms endorsed by this patient on the MASS were not documented in his chart. The treating psychiatrist confirmed that he was unaware of a number of undocumented symptoms detected from the computerized assessment and acknowledged the importance of those symptoms.

**Subdomain scoring**

CAT was used to assess individuals on both the primary and the secondary dimensions of the bifactor model. Each of the four MASS subscales on the CAT version correlated above .90 with impairment estimates from the full scales (r=.92 to r=.97). The average number of required items increased to 98 for sample A and to 118 for sample B—an approximately 85% reduction (524 items) from the full scale.

**External validity**

Scores from the CAT version and the full scale for patients with depression, with or without bipolar disorder, were compared on the mood disorder subdomain (161 items). For administration of the full scale, significant differences were found between the two diagnostic groups (with and without bipolar disorder) (t=3.20, df=154, p<.003; effect size of .63 SD units). Conversely, a much larger between-group difference was found in CAT scores (t=6.00, df=154, p<.001; effect size of 1.19 SD units). The effect size of the between-group difference in scores from the CAT version was almost double that from administration of the full test, with an 83% reduction (134 items) in items administered. This finding supports the conclusion of increased external validity of the CAT scores compared with scores on the full scale.

**Discussion**

The findings supported the utility and flexibility of multidimensional item response theory and CAT for the delivery of psychiatric measurements. Data on the MASS were used to demonstrate how a construct or constructs can be accurately measured by administering only items that are relevant to a given individual. This paradigm represents a distinct shift from the standard approach of administering fixed-length tests in psychiatric measurement. In the context of mental health services research, CAT may result in a drastic reduction in the time required to administer a collection of psychiatric instruments for mental health measurement and diagnostic purposes. Although there have been some applications of unidimensional item response theory in the context of psychiatric measurement (52–56), this is one of the first applications of multidimensional item response theory (that is, the bifactor model). Simulated CAT administration for research participants who had completed the full scale revealed huge time savings—a 96% reduction in the number of items administered. The correlation between scores obtained from CAT and from administration of the full scale was high (.90) in two independent samples, indicating no significant loss of infor-
and from administration of the full treatment. Specifically, the correlation primary measure of overall impairment levels in addition to the MASS could obtain domain-specific with other measures (3,4). Be noted that similar results were obtained for live CAT testing. It should be noted that similar results were obtained for live CAT testing. It should be noted that similar results were obtained for live CAT testing.

The study reported here also subjected to traditional psychometric analysis. A test of the scale’s factor structure supported conceptualizing it in terms of the bifactor model, with the primary dimension representing a composite of four factors (mood, panic-agoraphobia, obsessive-compulsive, and social phobia). Of the 626 items, 161 had primary factor loadings in excess of .50. Of these items, 28 were mood, 33 were panic-agoraphobia, 45 were obsessive-compulsive, and 55 were social phobia items. As such, there are an ample number of good discriminating items from each of the four domains, and their jointly high loadings on the primary dimension validates the existence of this composite dimension, referred to as "mood anxiety spectrum."

There are three noteworthy limitations of the bifactor model. First, model specification relies on a priori information (for example, theory) to indicate the relationships between items and factors. This should not be a problem in mental health research because the sampling of items from well-established psychological domains is an established method of scale construction. Second, a primary dimension must exist. If the items have a simple factor structure in which subsets of items are related to a single factor alone, then the bifactor model will not be useful. This is empirically testable by comparing the fit of the bifactor model to corresponding unidimensional models and unrestricted mul-

### Table 2

<table>
<thead>
<tr>
<th>Subscale and item number</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mood</strong></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>You were warm, extroverted, and sociable, and it was very easy to introduce yourself to others or to make new friends.</td>
</tr>
<tr>
<td>38</td>
<td>You were the kind of person to whom others were attracted because of your confidence, enthusiasm, and energy.</td>
</tr>
<tr>
<td>39</td>
<td>You did a lot of entertaining either at home or in restaurants.</td>
</tr>
<tr>
<td>40</td>
<td>You enjoyed being the center of attention or were particularly seductive or flirtations, as if you were playing a role.</td>
</tr>
<tr>
<td>41</td>
<td>You had a particularly intense romantic life.</td>
</tr>
<tr>
<td>47</td>
<td>You found it very pleasurable and exciting to get involved in dangerous, risky, challenging, or emotionally intense activities.</td>
</tr>
<tr>
<td>48</td>
<td>You tended to do the opposite of what people wanted you to do or to play devil's advocate.</td>
</tr>
<tr>
<td>75</td>
<td>You spent a lot of time on social, political, or religious causes.</td>
</tr>
<tr>
<td>77</td>
<td>You felt vigorous, much livelier than usual, and full of energy.</td>
</tr>
<tr>
<td>79</td>
<td>You were constantly active and had the pleasant sensation of never getting tired, and your energy was so high that it exhausted or irritated others.</td>
</tr>
<tr>
<td>109</td>
<td>You thought you were very artistic and creative.</td>
</tr>
<tr>
<td>110</td>
<td>You had bursts of inspiration or creativity.</td>
</tr>
<tr>
<td>116</td>
<td>You felt self-assured, charismatic, or tended to assume a leadership role.</td>
</tr>
<tr>
<td>121</td>
<td>You felt particularly strong and invulnerable, resistant to illness and accidents.</td>
</tr>
<tr>
<td>122</td>
<td>You were unusually spiritual or mystical.</td>
</tr>
<tr>
<td>127</td>
<td>You did things such as spending too much money, driving recklessly or speeding, or making foolish business decisions.</td>
</tr>
<tr>
<td><strong>Obsessive-compulsive</strong></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Felt anxious if you didn’t have a written list when doing things like going shopping, going to the doctors, or meeting with someone.</td>
</tr>
<tr>
<td>35</td>
<td>Had difficulty throwing things away, because they might be useful someday.</td>
</tr>
<tr>
<td>111</td>
<td>Saved things, even though they might not have any value.</td>
</tr>
<tr>
<td>112</td>
<td>Liked to have a complete set of anything, once you had gotten the first one.</td>
</tr>
<tr>
<td>114</td>
<td>Kept the address and phone number of anyone you know, so that you could always have the number close at hand.</td>
</tr>
<tr>
<td><strong>Panic-agoraphobia</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>You were separated or anticipated separation from home or loved ones.</td>
</tr>
<tr>
<td>2</td>
<td>You experienced a lot of distress because of thoughts that you might lose someone close to you or some harm might come to them.</td>
</tr>
<tr>
<td>42</td>
<td>Awakened in a panic for no reason.</td>
</tr>
</tbody>
</table>
tidimensional models by using either item response theory or confirmatory factor analysis methods. Third, the model requires each item to load on a primary dimension and on no more than one subdomain. If items are related to multiple subdomains, they will not be appropriate for the bifactor model. Nevertheless, research has shown the benefits of using the bifactor model to conceptualize health outcome measurements (15,17,59).

Conclusions

The importance of performing research in real-world clinical settings is widely recognized, as is the importance of measurement-based care outside the bounds of clinical research. However, in busy psychiatric practices and mental health clinics, the feasibility of conducting the kind of extensive evaluations typical of clinical research is questionable. Therefore, any strategy that reduces the burden of empirically based assessment has the potential to improve outcomes through measurement-based clinical decision making. Thus, apart from its research utility, CAT may complement routine clinical assessments and save clinicians time that may be more productively used for other activities. As we demonstrated in this study, CAT permits a clinician to gather important clinical information via self-report in a way that does not impose an excessive burden on the patient and that does not require the clinician to select items that are pertinent to the individual patient. CAT strategies may therefore be used to personalize virtually any kind of questionnaire. For instance, they can facilitate the creation of child mental health measures that are better adapted to a child’s developmental stage than more traditional questionnaires. They can be used to assess the entire array of psychopathology and make a reliable dimensional diagnosis practical and feasible in almost any clinical setting.

Acknowledgments and disclosures

This work was supported by grants R01-MH-66302 and R01-MH-30915 from the National Institute of Mental Health. The authors report no competing interests.

References

31. Reise SP, Waller NG: Fitting the two-parameter model to personality data. Ap


