ORIGINAL ARTICLE

# Development of a Computerized Adaptive Test for Depression

*Robert D. Gibbons, PhD; David J. Weiss, PhD; Paul A. Pilkonis, PhD; Ellen Frank, PhD; Tara Moore, MA, MPH; Jong Bae Kim, PhD; David J. Kupfer, MD*

**Context:** Unlike other areas of medicine, psychiatry is almost entirely dependent on patient report to assess the presence and severity of disease; therefore, it is particularly crucial that we find both more accurate and efficient means of obtaining that report.

**Objective:** To develop a computerized adaptive test (CAT) for depression, called the Computerized Adaptive Test–Depression Inventory (CAT-DI), that decreases patient and clinician burden and increases measurement precision.

**Design:** Case-control study.

**Setting:** A psychiatric clinic and community mental health center.

**Participants:** A total of 1614 individuals with and without minor and major depression were recruited for study.

**Main Outcome Measures:** The focus of this study was the development of the CAT-DI. The 24-item Hamilton Rating Scale for Depression, Patient Health Questionnaire 9, and the Center for Epidemiologic Studies Depression Scale were used to study the convergent validity of the new measure, and the Structured Clinical Interview for *DSM-IV* was used to obtain diagnostic classifications of minor and major depressive disorder.

**Results:** A mean of 12 items per study participant was required to achieve a 0.3 SE in the depression severity estimate and maintain a correlation of $r=0.95$ with the total 389-item test score. Using empirically derived thresholds based on a mixture of normal distributions, we found a sensitivity of 0.92 and a specificity of 0.88 for the classification of major depressive disorder in a sample consisting of depressed patients and healthy controls. Correlations on the order of $r=0.8$ were found with the other clinician and self-rating scale scores. The CAT-DI provided excellent discrimination throughout the entire depressive severity continuum (minor and major depression), whereas the traditional scales did so primarily at the extremes (eg, major depression).

**Conclusions:** Traditional measurement fixes the number of items administered and allows measurement uncertainty to vary. In contrast, a CAT fixes measurement uncertainty and allows the number of items to vary. The result is a significant reduction in the number of items needed to measure depression and increased precision of measurement.

*Arch Gen Psychiatry. 2012;69(11):1104-1112*

**Author Affiliations:** Center for Health Statistics, Departments of Medicine and Health Studies, University of Chicago, Chicago, Illinois (Drs Gibbons and Kim); Department of Psychology, University of Minnesota, Minneapolis (Dr Weiss); and Western Psychiatric Institute, University of Pittsburgh, Pittsburgh, Pennsylvania (Drs Pilkonis, Frank, and Kupfer and Ms Moore).

IMAGINE A 1000-ITEM MATHEMATICS test with items ranging in difficulty from basic arithmetic to advanced calculus. Consider 2 examinees: a fourth grader and a graduate student in mathematics. Most questions will be uninformative for both examinees (too difficult for the first and too easy for the second). To decrease examinee burden, we could create a short test of 10 items, equally spaced along the mathematics difficulty continuum. Although this test would be quick to administer, it would provide imprecise estimates of these 2 examinees' ability because only 1 or 2 items would be appropriate for either examinee. A better approach would be to begin by administering an item of intermediate difficulty and, based on the response scored as correct or incorrect, select the next item at a level of difficulty either lower or higher. This process would continue until the uncertainty in the estimated ability is smaller than a predefined threshold. This process is called computerized adaptive testing (CAT). To use CAT, we must first calibrate a bank of test items using an item response theory (IRT) model that relates properties of the test items (eg, their difficulty and discrimination) to the ability (or other trait) of the examinee. The paradigm shift is that rather than administering a fixed number of items that provide limited information for any given ex-




Corrected on November 26, 2012

aminee, we adaptively administer a varying number of items that are targeted to the examinee's specific level of ability or impairment. CAT allows us to adaptively select a small set of items for each examinee out of a much larger bank of test items, targeting precision by selecting items based on prior ability, trait, or impairment estimates.

Although use of CAT and IRT has been widespread in educational measurement, it has been less widely used in mental health measurement.[1,2] First, large item banks are generally unavailable for mental health constructs. Second, mental health constructs (eg, depression) are inherently multidimensional, and CAT has primarily been restricted to unidimensional constructs, such as mathematics achievement. Application of unidimensional models to multidimensional data can result in biased trait estimates (severity or impairment) and underestimates of uncertainty.[3] Multidimensional IRT-based CAT has been previously used in analysis of the 626-item Mood and Anxiety Spectrum Scales.[4] To our knowledge, this was the first study of mental health CAT using a large item bank and multidimensional IRT.[5,6] CAT required a mean of 24 items per examinee, yet maintained a correlation of $r=0.93$ with the full 626-item score. In this article, we apply multidimensional CAT to the measurement of depression using the CAT–Depression Inventory (CAT-DI).

Unlike other areas of medicine, psychiatry is almost entirely dependent on patient report (either self-report or reports to evaluators) to assess the presence and severity of disease; therefore, it is particularly crucial that we find more accurate and efficient means of obtaining that report. We use the assessment of depression as an example of what CAT might offer toward that goal. The goal of this report is to describe a new tool for the measurement of depression. With the exception of the previously cited demonstration study[4] with the Mood and Anxiety Spectrum Scales, the statistical foundation of the new tool has not been used in any other area of instrument development and provides a statistical advance over other approaches that are based on unidimensional models and much smaller item banks. Although we will explore the extent to which these theoretical advantages translate to gains in measurement precision, reliability, and validity in a future statistical article, the results of this study clearly demonstrate the improvement in fit of the multidimensional bifactor model over traditional unidimensional alternatives.

## METHODS

### THE BIFACTOR MODEL

The bifactor model[5-7] is a multidimensional IRT model that allows each item to measure the primary dimension (eg, depression) and a single subdomain (eg, somatization), hence, the term *bifactor*. The bifactor model has major computational and interpretational advantages over unrestricted exploratory item factor analytic models[8-10] and extends CAT to the measurement of multidimensional constructs. For example, traditional item factor analysis[8] is generally restricted to 5 dimensions, and the results may be interpreted differently, depending on the rotation of the solution (eg, varimax vs promax rotations). The bifactor model has no limitation in terms of the number of dimensions that can be included under the restriction that each item loads on the primary dimension and 1 subdimension. The solution is also rotationally invariant, leading to more direct interpretation. Technical details are provided in the eAppendix (http://www.archgenpsychiatry.com).

### COMPUTERIZED ADAPTIVE TEST

Within CAT, items are selected during the process of test administration for each individual, allowing the test administrator to control measurement precision and to maximize efficiency. CAT includes (1) a precalibrated item bank, (2) an item selection procedure, (3) a scoring method, and (4) a criterion for terminating the test. In the current study, we used maximum item information,[11] appropriately modified for the bifactor model (eAppendix), to select items for CAT administration and a dual termination criterion of 0.3 for the SE of the estimated primary dimension score or maximum information remaining in the bank of less than 1.25. In this way, CAT terminates if at an examinee's currently estimated specific severity level (eg, extreme score) there is insufficient information for any item to achieve the intended level of precision. Relaxing the SE termination criterion in the 2 extremes (floor and ceiling) reduces respondent burden without compromising our ability to rank patients in terms of severity. Technical details are provided in the eAppendix.

### THE ITEM BANK

The total item bank consisted of 452 depression items. We organized the items into conceptually meaningful categories using a hierarchical approach informed by previous empirical (eg, factor analytic) work. Previous work documents that these constructs can be partitioned usefully into subdomains (eg, mood, cognition, and somatic indicators) and factors, and it is informative about the best manifest indicators (items) for operationalizing such factors.[12-15] Our hierarchy included domains (eg, depression), subdomains (eg, mood, cognition, and behavior), factors (eg, within depressed mood, factors included increased negative affect and decreased positive affect), and facets (eg, within increased negative affect, facets included sadness, irritability, moodiness, and others). The total number of facets was 46 for depression (eTable 1). Example items from each domain and subdomain are presented in **Table 1**.

A key step in editing the item bank was qualitative review of the items done by consensus among the members of the Pittsburgh research site (see the article by DeWalt et al[16] for a description of the qualitative procedures used by the Patient-Reported Outcomes Measurement Information System [PROMIS] network and adapted here). This process involved identification of redundant items (so that they are not administered in the same testing session), items that were too narrow (often by virtue of being disease specific), items that were confusing or vague, and items that were poorly written.

Most items were rated on a 5-point ordinal scale. The items were selected based on a review of more than 100 existing depression or depression-related rating scales (eTable 2). Items were modified to refer to the previous 2-week period and to have consistent response categories.

The CAT-DI has a mandatory suicide screening question, which is presented at the end of the CAT session if not previously administered. If positively endorsed, a suicide alert report is generated and the test administrator notified. In the final version of the CAT-DI, an option will be available to have 4 suicide items administered to achieve an even more complete evaluation of suicide risk.




Corrected on November 26, 2012

**Table 1. Example Items From Each Domain and Subdomain**

| Domain or Subdomain | Example Item (In the Past 2 Weeks) |
|---|---|
| Mood–negative affect | Were you serious, introverted, or gloomy? |
| | How much did any feelings of depression bother you? |
| Mood–positive affect | How much were you able to relax and enjoy yourself? |
| | I was able to reach down deep into myself for comfort. |
| Cognition–information deficits | Did you drift in and out of conversations? |
| | I had difficulty concentrating. |
| Cognition–information unproductive | Did you find that silly or unreasonable thoughts kept recurring in your mind? |
| | Did you see the future as very bleak? |
| Cognition–impaired view | You felt constantly afraid of doing something wrong. |
| | How much have you felt inferior to others? |
| Cognition–social cognition | Have you felt lonely? |
| | You felt as if others were causing all of your problems. |
| Cognition-hopelessness | To what extent did you feel your life was meaningful? |
| | I felt a sense of purpose in my life. |
| Cognition-helplessness | Did you feel defeated? |
| | I felt like I was at the end of my rope. |
| Cognition-guilt | I felt I should be punished. |
| | I was concerned about being forgiven for my sins. |
| Behavior–low activity | Did you have difficulty starting to do anything? |
| | I felt emotionally drained from my work. |
| Behavior–low energy | Did you have a lot of trouble getting out of bed in the morning? |
| | I felt that everything I did was an effort. |
| Behavior-interpersonal | How much have you felt like being alone? |
| | How much have you felt withdrawn from others? |
| Behavior-agitation | I felt restless as if I had to always be on the move. |
| | Did you find it difficult to sit still or to lie down, or you needed to pace the room or be in motion? |
| Somatic–sleep problems | How satisfied were you with your sleep? |
| | My sleep was restless. |
| Somatic–eating changes | I did not feel like eating; my appetite was poor. |
| | I found food unappealing. |
| Somatic-gastrointestinal | Did you repeatedly have distressing physical symptoms, for instance, you had nausea or other stomach or bowel problems? |
| | Did you repeatedly have distressing physical symptoms, for instance, you were constipated? |
| Somatic–increased pain | Were you more sensitive or less sensitive than usual to heat, cold or pain? |
| | Did you repeatedly have distressing physical symptoms, for instance, frequent headaches? |
| Somatic–general somatic | Did you feel as if your body were diseased or somehow transformed? |
| | Did your mood become depressed when you had a medical problem such as the flu or a cold? |
| Somatic–diurnal variation | Morning was when I felt the best. |
| | Has there been any time of day when you felt slower and less energetic? |
| Suicidal ideation | Did you think that life was not worth living? |
| | Did you think about taking your own life? |

## PARTICIPANT SAMPLE

Study participants were male and female treatment-seeking outpatients between 18 and 80 years of age. Patients were recruited from 2 facilities, the Western Psychiatric Institute and Clinic (WPIC) at the University of Pittsburgh and a community clinic at DuBois Regional Medical Center (DuBois RMC). DuBois RMC is one of the leading health centers in western Pennsylvania. The Center for Behavioral Health Services at DuBois RMC provides comprehensive inpatient and outpatient psychiatric care. Participants were screened at both the WPIC and Dubois RMC for eligibility. If they had been in psychiatric treatment within the past 2 years, they were considered a psychiatric participant. If they had not had psychiatric treatment within the past 2 years, they were considered a nonpsychiatric control. Psychiatric diagnoses were confirmed by medical records and their treating physician or clinician. Patients with and without a lifetime diagnosis of major depressive disorder (MDD) were included. For psychiatric participants, exclusion criteria included the following: history of schizophrenia, schizoaffective disorder, or psychosis; organic neuropsychiatric syndromes (eg, Alzheimer disease or other forms of dementia, Parkinson disease, and so on); drug or alcohol dependence within the past 3 months (however, patients with episodic abuse related to mood episodes were not excluded); inpatient treatment status; and individuals who were unable or unwilling to provide informed consent. For nonpsychiatric controls, exclusion criteria included the following: any psychiatric diagnosis within the past 12 months; treatment for a psychiatric problem within the past 12 months; positive responses to telephone screen questions; history of schizophrenia, schizoaffective disorder, or psychosis; and individuals who were unable or unwilling to provide informed consent.

We report on the analysis of data from 798 individuals used to calibrate the IRT model (WPIC) and 816 individuals who received the live CAT-DI (414 WPIC and 402 DuBois RMC study participants). For simulated adaptive testing, 308 participants (of the 798) took all of the 389 items in the bank after 63 items were deleted from the original set of items, permitting computation of the correlation between results of CAT and total test score; these participants were also part of the calibration sample. The other 490 calibration participants took a subsample of 252 items based on a balanced incomplete block design that maximized the pairings of all items.[17]

To study the validity of the CAT-DI, using the calibration and simulated CAT phases described, 292 consecutive psychiatric participants received a full clinician-based diagnostic interview via the Structured Clinical Interview for *DSM-IV* (SCID)[18] and the live CAT-DI. Participants received increased compensation ($50 rather than $25) to complete this phase of the study. Demographic characteristics and SCID-based diagnostic prevalence rates are presented in **Table 2**. The CAT-DI was also administered to 100 nonpsychiatric controls. Nonpsychiatric controls were defined as having no psychiatric treatment within the past 2 years. Controls were recruited via flyers, print media, and Audix messages through the University of Pittsburgh Medical Center.

To examine convergent validity, data were also obtained for the Patient Health Questionnaire 9 (PHQ-9),[19] 24-item Hamilton Rating Scale for Depression (HAM-D),[20] and the Center for Epidemiologic Studies Depression Scale (CES-D).[21] The HAM-D was administered by a trained clinician, and the PHQ-9 and CES-D were self-reports. Major depression, minor depression (*DSM-IV* appendix B), and dysthymia were defined according to *DSM-IV* criteria. The screening information was available to the clinician.




Corrected on November 26, 2012

**Table 2. Demographic Characteristics and Diagnostic Prevalence Rates**

| Characteristic | Study Participants, % |
|---|---|
| Sex | |
| Male | 30.0 |
| Female | 70.0 |
| Age, y | |
| 18-29 | 21.2 |
| 30-39 | 17.1 |
| 40-49 | 23.0 |
| 50-59 | 26.7 |
| ≥60 | 12.0 |
| Educational level | |
| Some high school or <12th grade | 5.1 |
| High school diploma or GED | 21.9 |
| Some college | 39.8 |
| College graduate | 20.2 |
| Graduate or professional degree | 13.0 |
| Annual income, $ | |
| <24 999 | 54.1 |
| 25 000-49 999 | 25.9 |
| 50 000-74 999 | 9.2 |
| 75 000-99 999 | 3.9 |
| >100 000 | 3.9 |
| Not reported | 3.0 |
| Prevalence rates | |
| Major depression | 46.7 |
| Minor depression | 5.1 |
| No depression | 45.1 |
| Dysthymia | 3.1 |

Abbreviation: GED, graduate educational development.

### STATISTICAL ANALYSIS

Calibration was performed using the bifactor model and a unidimensional alternative, both based on graded IRT models.[6] A likelihood ratio $\chi^2$ statistic was used to determine whether the bifactor model improved fit over the simpler unidimensional alternative. The CAT-DI scores were based on expected a posteriori estimates.[22] The CAT-DI scores were then used in a logistic regression to predict a physician-based *DSM* diagnosis of MDD so that the CAT-DI scores could be related to the probability of meeting *DSM-IV* criteria for MDD.

The empirical distribution of the CAT-DI scores was resolved into a mixture of 2 normal distributions[23] and compared with models with 1 and 3 component distributions using the Bayesian information criterion (BIC) to determine the best-fitting model. Sensitivity and specificity of the CAT-DI in predicting SCID diagnostic group classification (MDD and MDD plus minor depression) were determined for 3 thresholds based on the estimated mixture distribution (ie, posterior probability of being in the elevated component distribution of 0.50, 0.80, and 0.95).

## RESULTS

### CALIBRATION

Results of the calibration study revealed that the bifactor model with 5 subdomains (mood, cognition, behavior, somatic, and suicide) significantly improved fit over a unidimensional IRT model ($\chi^2_{389} = 6825$, $P < .001$). A total of 389 items were retained in the model based on having a primary factor loading of 0.3 or greater (96% >0.4 and 79% >0.5).

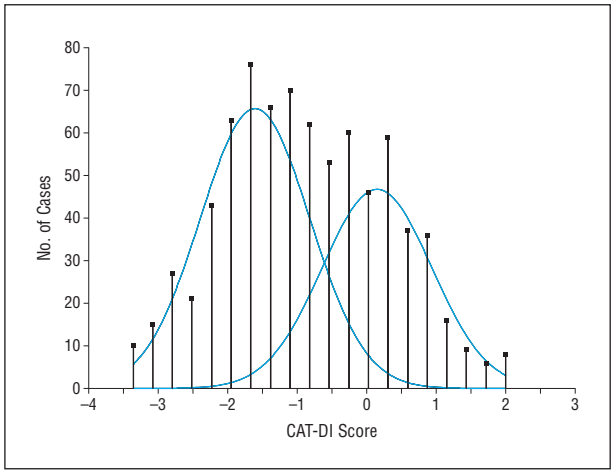


**Figure 1.** Observed and estimated frequency distributions of the Computerized Adaptive Test–Depression Inventory (CAT-DI) depression scale scores.

### SIMULATED CAT

Results of simulated CAT revealed that for an SE of 0.3, a mean of 12.31 items per participant (range, 7-22) were required. The correlation between the 12-item mean length CAT and the total 389-item score was $r = 0.95$. For an SE of 0.4 (less precise), a mean of 5.94 items were required (range, 4-16), but a strong correlation with the 389-item total score ($r = 0.92$) was maintained. The median length of time required to complete the 12-item (mean) CAT was 2.29 minutes (interquartile range, 1.72-2.97 minutes) compared with 51.66 minutes for the 389-item test. Faster times should be achievable using the final platform (a touch screen device) instead of the Windows-based mouse interface currently used.

Mean precision was 0.31, and CAT was terminated for insufficient item information in 15.0% of the cases. In all but 2 cases, the estimated CAT-DI score was less than −2.0, indicating no evidence of depression (too few symptoms to precisely measure). In the other 2 cases, the scores were greater than +2.5, indicating extreme severity (too many symptoms to precisely measure).

### EMPIRICAL DISTRIBUTION OF THE CAT-DI SCORES

**Figure 1** reveals that our sample can be resolved into 2 discrete distributions of depressive severity, with the lower component representing the absence of clinical depression and the higher component representing severity levels associated with clinical depression. The BIC indicated best fit (ie, smallest value) for a mixture of 2 normal distributions (BIC = 2406) relative to a single normal distribution (BIC = 2439) or a mixture of 3 normal distributions (BIC = 2414). The means of the 2 distributions are well separated. A total of 40.8% of the sample is in the elevated component distribution that has a mean of 0.17 vs the lower component distribution that has a mean of −1.61. The pooled estimate of the SD is 0.73, making the 2 component distributions approximately 2.5 SDs apart. Threshold scores of −0.61, −0.19, and 0.28 have a .50, .80, and .95 probability of being in the elevated component distribution, respectively.




Corrected on November 26, 2012

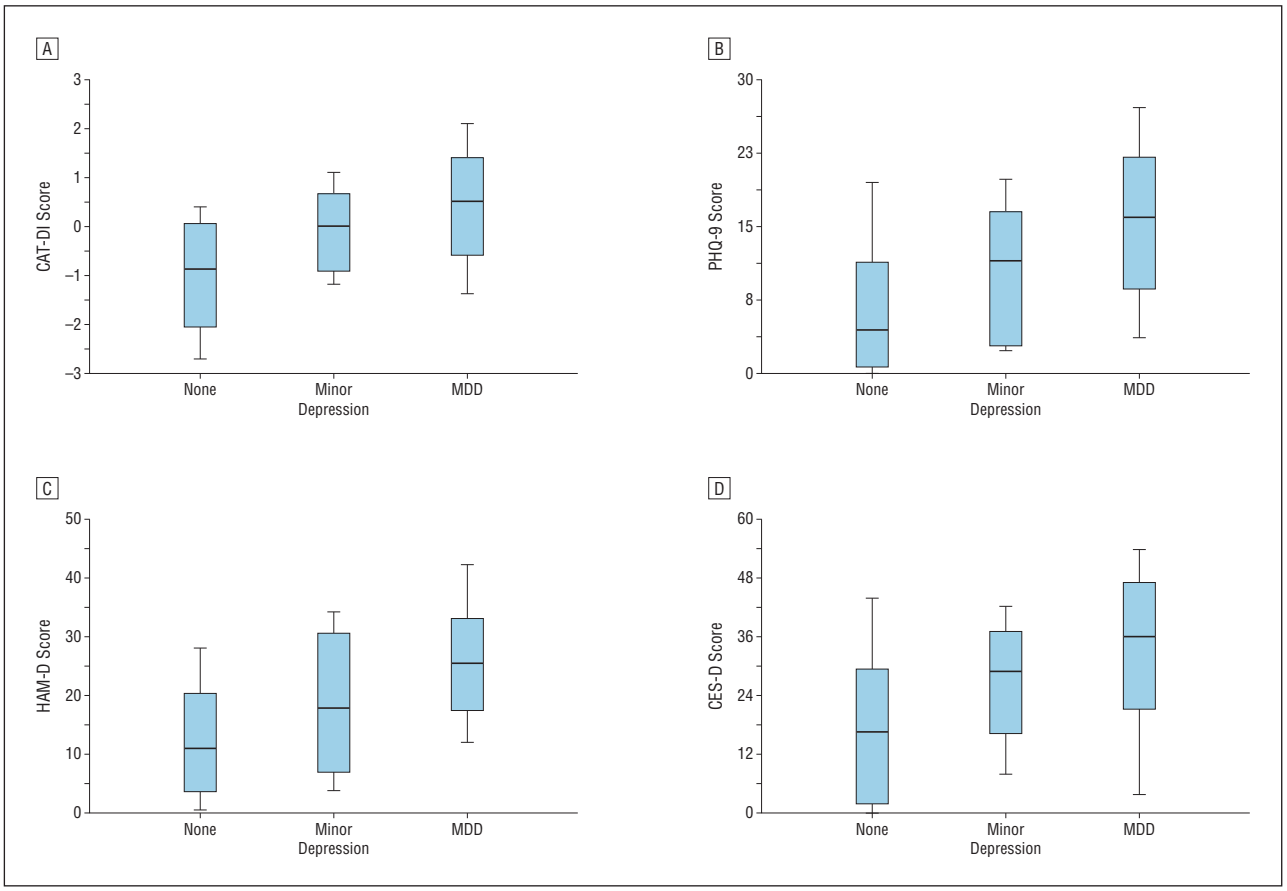


**Figure 2.** Distributions of the Computerized Adaptive Test–Depression Inventory (CAT-DI) scores for patients who met diagnostic criteria for minor depression (including dysthymia) and major depression disorder (MDD) vs those who did not meet the criteria. A, CAT-DI depression scale score. B, Patient Health Questionnaire 9 (PHQ-9) score. C, 24-Item Hamilton Rating Scale for Depression (HAM-D) score. D, Center for Epidemiologic Studies Depression Scale (CES-D) score. Error bars indicate the range; horizontal lines, the 10th, 50th, and 90th percentile points, respectively.

## RELATIONSHIP TO DIAGNOSIS OF MINOR AND MAJOR DEPRESSION

**Figure 2**A displays the distributions of the CAT-DI scores for patients who met the diagnostic criteria for minor depression (including dysthymia) and MDD vs those who did not meet the criteria. There is a clear linear progression between the CAT-DI depression severity scores and the SCID diagnostic categories of none, minor depression (including dysthymia), and MDD. The distributions were also reasonably symmetric within each diagnostic group. Means (SDs) and sample sizes were –0.93 (0.75) (n = 117) for no depression, –0.02 (0.58) (n = 29) for minor depression, and 0.47 (0.693) (n = 146) for MDD. Statistically significant differences between no depression and minor depression ($t_{144} = 6.121$, $P < .001$), no depression and MDD ($t_{261} = 15.736$, $P < .001$), and minor depression and MDD ($t_{173} = 3.558$, $P < .001$) were found, with corresponding effect sizes of 1.271, 1.952, and 0.724 SDs, respectively.

## COMPARISON WITH OTHER DEPRESSION SCALES

Convergent validity of the CAT-DI was assessed by comparing results of the CAT-DI with the PHQ-9, HAM-D, and CES-D results. Correlations were $r = 0.81$ with the PHQ-9, $r = 0.75$ with the HAM-D, and $r = 0.84$ with the CES-D. In general, the distribution of scores among the diagnostic categories showed greater overlap (ie, less diagnostic specificity particularly for no depression vs minor depression), greater variability, and greater skewness for these other scales relative to the CAT-DI (Figure 2, B-D).

## DIAGNOSTIC SCREENING

With the 100 healthy controls as a comparator, sensitivity and specificity for predicting MDD using the 50% probability threshold (CAT-DI score = –0.61) were 0.92 and 0.88, respectively. Results were similar for the combination of major and minor depression (0.90 sensitivity and 0.88 specificity). Using patients treated for depression in the past 2 years who did not currently meet *DSM-IV* criteria for minor depression or MDD as a comparator yielded sensitivity and specificity of 0.92 and 0.64 for MDD and 0.90 and 0.64 for minor depression and MDD combined, respectively. The lower specificity is due to elevated depressive symptoms in patients currently or recently in treatment for depression who did not meet *DSM-IV* criteria for MDD and/or minor depression. Using the 95% probability threshold of +0.28 increased specificity to 0.98 but decreased sensitivity to 0.63. The increased threshold is rarely (2%) exceeded by patients without MDD, but only 63% of patients with MDD exceeded




Corrected on November 26, 2012

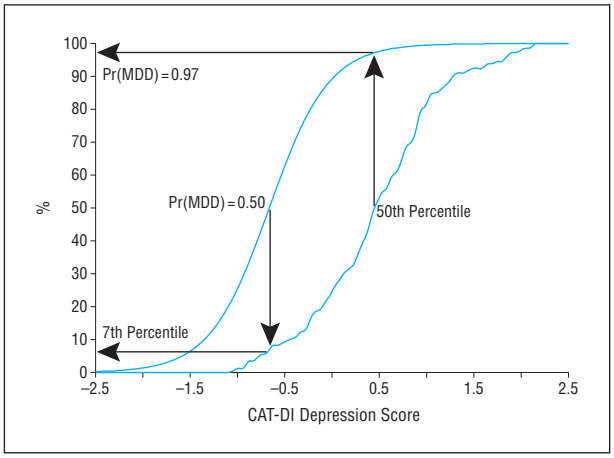


**Figure 3.** Percentile rank (among patients with a major depressive disorder diagnosis) and probability (expressed as percentage) of a major depressive disorder diagnosis. The *y*-axis refers to both of the curves portrayed on the graph. CAT-DI indicates Computerized Adaptive Test–Depression Inventory and Pr(MDD), probability of major depressive disorder.

this threshold. Depending on the application, different thresholds can be used. A reasonable balance for application in clinical (depression) samples is to use an 80% classification probability threshold (CAT-DI score of –0.19), which produces sensitivity of 0.82 and specificity of 0.85. As expected, the rate of positive CAT-DI scores (using the 50% threshold of –0.61) was higher (51.7%) at the university psychiatric clinic than at the community mental health clinic (31.1%).

The CAT-DI scores were strongly related to MDD diagnosis (odds ratio = 24.19; 95% CI, 10.51-55.67; $P < .001$). A unit increase in the CAT-DI score has an associated 24-fold increase in the probability of meeting criteria for MDD. This relationship is shown in **Figure 3**. Figure 3 also presents the CAT-DI score percentile ranking for patients with *DSM-IV*–diagnosed MDD. For example, a patient with a CAT-DI score of –0.6 has a .50 probability of meeting criteria for MDD but would be at the lower seventh percentile of the distribution of the CAT-DI scores among patients who met criteria for MDD. By contrast, a patient with a CAT-DI score of 0.5 would have a .97 probability of meeting criteria for MDD and would be at the 50th percentile of patients meeting criteria for MDD.

## EXAMPLE CAT ADMINISTRATIONS

**Table 3** presents item-by-item results for 2 CAT administrations: 1 study participant with low severity and 1 with high severity. The participant with low severity required 11 items to achieve an SE less than 0.3, and the participant with high severity required 12 items. The first participant had a score of –0.892, which corresponds to a probability of .33 of meeting criteria for MDD and a percentile of 3.4% among patients with MDD. As such, we would not consider this to be a patient who is currently depressed. By contrast, the second participant had a score of 1.028, which corresponds to a probability of .99 of meeting criteria for MDD and a percentile of 83.9% among patients with MDD. As such, we would consider this to be a patient who is highly likely to be currently depressed. If we had used a less stringent SE termination criterion of 0.4, we would have terminated item administration for both participants after 6 items and would have obtained severity scores within 5% of the final values.

**Table 3. Results of 2 Computerized Adaptive Test Administrations**

| Item (In the Past 2 Weeks) | Response | Score (SE) | Information Score |
|---|---|---|---|
| Patient 1 (low severity)[a] | | | |
| I felt depressed. | A little of the time | –0.706 (0.615) | 4.056 |
| Have you been in low or very low spirits? | A little of the time | –0.761 (0.530) | 3.490 |
| How much were you distressed by feelings of worthlessness? | A little bit | –0.679 (0.452) | 3.361 |
| I had difficulty sleeping. | A little bit | –0.722 (0.431) | 3.283 |
| How much have you felt discouraged? | A little bit | –0.803 (0.403) | 2.720 |
| Did fatigue interfere with your mood? | Occasionally | –0.845 (0.375) | 2.534 |
| How often has feeling depressed interfered with what you do? | No more than usual | –0.812 (0.353) | 2.393 |
| How much were you distressed by feeling everything was an effort? | A little bit | –0.799 (0.335) | 2.401 |
| Have you had problems accomplishing less than you would like with your work or other regular daily activities as a result of emotional problems (such as feeling depressed or anxious)? | No | –0.877 (0.324) | 2.319 |
| How much of the time have you been moody or brooded about things? | A little of the time | –0.905 (0.308) | 2.303 |
| Did you feel isolated from others? | A little of the time | –0.892 (0.298) | 2.207 |
| Patient 2 (high severity)[b] | | | |
| I felt depressed. | Most of the time | 0.474 (0.621) | 4.056 |
| Have you felt that life was not worth living? | Quite a bit | 0.810 (0.551) | 3.815 |
| Have you been in low or very low spirits? | Most of the time | 0.900 (0.485) | 3.547 |
| I felt gloomy. | Quite a bit | 0.917 (0.437) | 3.320 |
| How much have you felt that nothing was enjoyable? | Quite a bit | 0.951 (0.424) | 2.855 |
| How much were you distressed by blaming yourself for things? | Quite a bit | 0.973 (0.384) | 2.640 |
| How much were you distressed by feeling everything was an effort? | Quite a bit | 0.996 (0.353) | 2.502 |

*(continued)*




Corrected on November 26, 2012

**Table 3. Results of 2 Computerized Adaptive Test Administrations (continued)**

| Item (In the Past 2 Weeks) | Response | Score (SE) | Information Score |
|---|---|---|---|
| How often did you have negative feelings, such as blue mood, despair, anxiety, or depression? | Often | 0.961 (0.342) | 2.468 |
| How much difficulty have you been having in the area of mood swings or unstable moods? | Quite a bit of difficulty | 0.994 (0.322) | 0.322 |
| I could not get going. | Most of the time | 1.017 (0.313) | 2.400 |
| How much were you distressed by feelings of guilt? | Quite a bit | 1.029 (0.302) | 2.365 |
| I was unhappy. | Often | 1.028 (0.299) | 2.341 |

[a] For patient 1: score, –0.892; SE, 0.298; probability of major depressive disorder, .33; and percentile among patients with major depressive disorder, 3.4%.

[b] For patient 2: score, 1.028; SE, 0.299; probability of major depressive disorder, .99; and percentile among patients with major depressive disorder, 83.9%.

## COMMENT

Results of this study reveal that we can extract most of the information ($r$ = 0.95) from a bank of 389 depression items using a mean of only 12 items (median of 2 minutes 17 seconds) per study participant. The paradigm shift is that rather than using a fixed number of items and allowing measurement uncertainty to vary, we fix measurement uncertainty to an acceptable level for a given application and allow the number and specific items administered to vary from participant to participant. As an example, changing our termination threshold from an SE of 0.30 to 0.40 decreased the mean number of items administered from 12 to 6, with only a small corresponding decrease in correlation with the total 389-item score ($r$ = 0.95 to $r$ = 0.92). Such efficiency would permit depression screening of large populations necessary for conducting studies of psychiatric epidemiology and determining phenotypes for large-scale molecular genetic studies.

The ability to administer the CAT-DI in a few minutes via the Internet, without clinician assistance, makes routine depression screening of patients in primary care possible because the results of the test can be transmitted directly to the medical record and discussed with the patient by the physician at the time of their visit. We note that depressed patients are particularly difficult to assess with a long scale, and the benefits of CAT administration are therefore particularly important in this setting.

From a taxonomic point of view, the distribution of the CAT-DI scores in this study sample suggests a mixture of 2 component distributions, one pathological and one not. However, both distributional components appear to have normal distributions, demonstrating that there is considerable interindividual variability in patients with and without clinical depression. The 2 component distributions are well separated by approximately 2.5 SDs, making classification straightforward.

A somewhat unexpected result is the jointly high levels of sensitivity (0.92) and specificity (0.88) for the CAT-DI for predicting MDD when using general medical sample controls as the comparison group. A different threshold can be used such that even in samples of patients actively or recently being treated for depression, the CAT-DI accurately predicts MDD. The use of a finite mixture distribution model to empirically derive thresholds for studying sensitivity and specificity for diagnostic classification appropriate for the particular sample evaluated in this study is another unique methodologic feature of this study.

The CAT-DI exhibits strong correlation ($r$ = 0.75 to $r$ = 0.84) with other established depression rating scales, both self-rated and clinician rated. Despite this strong association, the CAT-DI appears to be better able to differentiate patients with minor depression from patients who did not meet criteria for minor or major depression. This is not at all surprising because the CAT-DI is adaptively selecting items from a large bank of possible items that are targeted to the specific level of depressive severity of each patient. By contrast, the fixed-length scales consist of a relatively small number of items, for which only a few may be discriminating at any specific level of depressive severity. As a result, they may do well in the extremes (eg, MDD) but have less information for intermediate severity levels. This is one of the major advantages of CAT.

As applied in this study, CAT does not solve all measurement problems. For example, if we were interested simply in diagnostic classification, we would be better off adaptively selecting items that maximized measurement precision in the region between healthy and depressed. Furthermore, although the CAT-DI may be extremely helpful in many settings, it is not able to assess functional impairment or assess the immediacy of treatment necessity.

There are numerous areas for further study. As previously demonstrated,[4] CAT can be applied to a wide variety of areas in mental health measurement, with similar reductions in patient and clinician burden. As a part of the current research program, we have also developed CAT instruments for anxiety and bipolar disorder, which we will soon report. Application to the assessment of depression in children is also viable; however, different item banks would be required and methods for dealing with developmental shifts must also be incorporated. The measurement of depression in different populations (eg, Hispanics) can also be accommodated by translating the CAT-DI item bank into different languages (eg, Spanish) and then testing for differential item functioning between our current sample and a culturally different sample. It is likely that items that may provide excellent discrimination of high and low depressive severity in one culture or race may not perform as well in another. Identifying the most informative items for a given population may provide further benefits over traditional measurement.




Corrected on November 26, 2012

These same opportunities for future research represent the limitations of the current study. It is unclear how the CAT-DI will function in other cultures. It is unlikely that the CAT-DI is currently suitable for children and adolescents, and further study is also required in elderly populations. In this article, we focus exclusively on the primary depressive dimension; however, there may be interest in the subdomains as well. The results of this study must be cross-validated. The number of patients with minor depression in our study is small.

Our estimates of sensitivity and specificity are specific to the population sampled and may be different in other populations, for example, screening depression in an inpatient or outpatient general medical population. The estimated parameters of the mixture distribution (ie, differences in means, SDs, and the proportion in the elevated component distribution) are also sample dependent, and we would expect them to vary, for example, if we looked at a primary care setting where the incidence of depression was lower. Our estimates are based on the combination of an outpatient mental health clinic and a community mental health center in which we would expect a relatively high incidence of patients in the elevated component distribution (ie, high depressive severity). The inclusion of healthy controls, however, allows us to better characterize the lower component distribution. Evaluation of this mixture distribution in other populations (eg, primary care) is important for testing the generalizability of the cut points we have derived for classifying patients as depressed and for assessing sensitivity and specificity. In addition, further statistical research is under way for directly estimating the mixture distribution as a part of the bifactor model and using it to obtain expected a posteriori estimates of the scores.

The National Institutes of Health–funded PROMIS initiative has also studied patient-reported outcomes using CAT and IRT, including depression.[24,25] The primary differences between PROMIS and the CAT-DI for the measurement of depression are (1) the PROMIS item bank consists of 28 items, whereas the CAT-DI bank consists of 389 items; (2) PROMIS has relied on unidimensional IRT models, whereas the CAT-DI incorporates multidimensionality produced by the sampling of items from distinct subdomains of depression; and (3) the CAT-DI has been developed for measuring the severity of depression and screening for depression, whereas PROMIS has focused on measurement of depressive severity. The theoretical advantage of the large item bank and multidimensional approach to calibration is that the CAT-DI is highly discriminating across the entire depressive severity continuum, and the adaptively administered items are representative of several different underlying domains of depressive symptoms. Forcing unidimensionality results in small item banks that may limit generalizability. Further study of this important issue is under way.

The CAT-DI may also have special relevance for longitudinal studies, where the score on the previous measurement occasion can be used as the starting value for the next CAT, leading to even fewer items administered. It would also be of interest to compare the CAT-DI with the HAM-D in terms of identifying treatment effects in randomized clinical trials. Increased precision may be obtained by decreasing the termination SE. This conjecture will be tested in a future study.

The CAT-DI depression scale currently exists as a research instrument; however, the Windows-based and web-based versions of the program will be completed at the end of 2012 (see www.healthstats.org for details). The programs will be made available in both standalone and cloud computing environments and will be fully supported. Further work continues on establishing the validity of the CAT-DI anxiety and bipolar scales.

**Submitted for Publication:** August 19, 2011; accepted January 4, 2012.
**Correspondence:** Robert D. Gibbons, PhD, Center for Health Statistics, Departments of Medicine and Health Studies, University of Chicago, 5841 S Maryland Ave, MC 2007 Office W260, Chicago, IL 60637 (rdg@uchicago.edu).
**Conflict of Interest Disclosures:** None reported.
**Funding/Support:** This work was supported by grant R01-MH66302 from the National Institute of Mental Health.
**Online-Only Material:** The eAppendix and eTables 1 and 2 are available at http://www.archgenpsychiatry.com.
**Additional Information:** The CAT-DI will ultimately be made available for routine administration, and its development as a commercial product is under consideration.
**Additional Contributions:** We acknowledge the outstanding support of R. Darrell Bock, PhD, Scott Turkin, MD, Damara Walters, BA, Suzanne Lawrence, MA, and Victoria Grochocinski, PhD. We are also indebted to the reviewers for many excellent comments and suggestions.

## REFERENCES

1. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a Computer-Adaptive Test for Depression (D-CAT). *Qual Life Res*. 2005;14(10):2277-2291.
2. Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, Buysse D, Frank E. Computerized adaptive measurement of depression: a simulation study. *BMC Psychiatry*. 2004;4:13-23.
3. Gibbons RD, Immekus J, Bock RD. Multi-dimensional Models for Outcomes Measurement.http://outcomes.cancer.gov/areas/measurement/multi-dimensional.html. Accessed August 29, 2012.
4. Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, Bhaumik DK, Stover A, Bock RD, Immekus JC. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv*. 2008;59(4):361-368.
5. Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. *Psychometrika*. 1992;57:423-436.
6. Gibbons RD, Bock RD, Hedeker D, Weiss D, Segawa E, Bhaumik DK, Kupfer D, Frank E, Grochocinski V, Stover A. Full-Information item bi-factor analysis of graded response data. *Appl Psychol Meas*. 2007;31:4-19.
7. Holzinger KJ, Swineford F. The bifactor method. *Psychometrika*. 1937;2:41-54.
8. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*. 1981;46:443-459.
9. Gibbons RD, Rush J, Immekus JC. On the psychometric validity of the diagnostic domains of the PDSQ: an illustration of the bi-factor item response theory model. *J Psychiatr Res*. 2009;43(4):401-410.
10. Cai L. A two-tier full-information item factor analysis model with applications. *Psychometrika*. 2010;75:581-612.
11. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*. 1969;17:1-68.
12. Quilty LC, Zhang KA, Bagby RM. The latent symptom structure of the Beck Depression Inventory–II in outpatients with major depression. *Psychol Assess*. 2010;22(3):603-608.
13. Santor DA, Gregus M, Welch A. Eight decades of measurement in depression. *Measurement*. 2006;4:135-155.




Corrected on November 26, 2012

14. Shafer AB. Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *J Clin Psychol*. 2006;62(1):123-146.
15. Simms LJ, Grös DF, Watson D, O'Hara MW. Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depress Anxiety*. 2008; 25(7):E34-E46.
16. DeWalt DA, Rothrock N, Yount S, Stone AA; PROMIS Cooperative Group. Evaluation of item candidates: the PROMIS qualitative item review. *Med Care*. 2007; 45(5)(suppl 1):S12-S21.
17. Cochran WG, Cox GM. *Experimental Designs*. New York, NY: Wiley; 1957.
18. First MB, Spitzer RL, Gibbon M, Williams JB. *Structured Clinical Interview for the DSM-IV Axis I Disorders Clinician Version (SCID-CV)*. Washington, DC: American Psychiatric Press, Inc; 1996.
19. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606-613.
20. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960; 23:56-62.
21. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas*. 1977;1:385-401.
22. Bock RD, Gibbons RD. Factor analysis of categorical item responses. In: Nering M, Ostini R, eds. *Handbook of Polytomous Item Response Theory Models: Development and Applications*. Florence, KY: Lawrence Erlbaum; 2010.
23. Gibbons RD, Dorus E, Ostrow DG, Pandey GN, Davis JM, Levy DL. Mixture distributions in psychiatric research. *Biol Psychiatry*. 1984;19(7):935-961.
24. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D; PROMIS Cooperative Group. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. *Assessment*. 2011;18(3):263-283.
25. Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, Lai JS, Choi SW, Hays RD, Reeve BB, Reise SP, Pilkonis PA, Cella D. Analysis of differential item functioning in the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS): an item response theory approach. *Psychol Sci Q*. 2009;51(2):148-180.




Corrected on November 26, 2012