Test-Retest Reliability of a Computerized Adaptive Depression Screener

David Beiser, M.D., M.S., Milkie Vu, M.A., Robert Gibbons, Ph.D.

Objective: Computerized adaptive testing (CAT) provides improved precision and decreased test burden compared with traditional, fixed-length tests. Concerns have been raised regarding reliability of CAT-based measurements because the items administered vary both between and within individuals over time. The study measured test-retest reliability of the CAT Depression Inventory (CAT-DI) for assessment of depression in a screening setting where most scores fall in the normal range.

Methods: A random sample of adults (N=101) at an academic emergency department (ED) was screened twice with the CAT-DI during their visit. Test-retest scores, bias, and reliability were assessed.

Results: Fourteen percent of patients scored in the mild range for depression, 4% in the moderate range, and 3% in the severe range. Test-retest scores were without significant bias and had excellent reliability (r=.92).

Conclusions: The CAT-DI provided reliable screening results among ED patients. Concerns about whether changes in item presentation during repeat testing would affect test-retest reliability were not supported.

Psychiatric Services in Advance (doi: 10.1176/appi.ps.201500304)

Depression is associated with increased mortality, adverse health outcomes, and increased overall treatment-related costs (1,2). The emergency department (ED) is an important safety net for patients with behavioral health problems (3) and thus may be an ideal setting to diagnose and initiate treatment for patients with depression. Current estimates suggest that between 8% and 32% of ED patients present with depression (4–6). However, conducting the detailed assessments of depression severity required to initiate treatment is often infeasible in the ED because of high patient volumes and limited access to behavioral health expertise. Therefore, any strategy that reduces the burden of empirically based assessment of depression has the potential to improve outcomes (7).

Challenges related to depression screening and diagnosis in the ED may be overcome by the considerable progress made recently in the development of computerized adaptive testing (CAT) based in multidimensional item response theory (IRT), a method for the rapid screening and measurement of depression (8–12). The advantages of IRT-based CAT include the use of large item banks (≥1,000 items) that tap every domain, subdomain, and facet of an underlying disorder. From this bank, a small, optimal set of items is adaptively administered for a given patient depending on his or her severity level. Other advantages include a constant level of precision for all patients on all measurement occasions, despite changes in severity level; adaptation across testing sessions, such that the previous depression severity score is used to initiate the next testing session; elimination of response-set bias, in which patients are repeatedly asked the same questions; the use of models for both diagnostic screening and dimensional severity that are based on different statistical approaches; incorporation of the multidimensionality of mental health constructs; and the ability to combine items with different response formats, different severity levels, and different ability to discriminate high and low levels of the construct of interest in the same test. IRT-based CAT represents a paradigm shift away from traditional measurement, which fixes the number of items administered and allows measurement uncertainty (imprecision in the test score) to vary. Instead, CAT fixes measurement uncertainty and allows the content and number of items to vary.

Although CAT promises several practical advantages for depression screening and measurement, concerns about its test-retest reliability (stability) have been raised in the literature (13). Test-retest reliability reflects the variation in measurements for a given person under the same conditions in a short period of time. Because the same test is administered twice, differences between scores should be due solely to measurement error. Determining test-retest reliability is often problematic for psychological testing, given that the construct being measured may change between the two test administrations (14). Repeated administration of classical fixed-length tests within a short time interval is problematic because respondents may repeat answers that they recall giving earlier, leading to inflated test-retest reliability. This outcome is not a concern for CAT administration because different items are administered upon repeat administration even if the underlying trait of interest has not changed. However, it has been suggested that the use of different items upon repeat administration may lead to diminished stability relative to traditional fixed-length tests.

This study evaluated the test-retest reliability of the CAT–Depression Inventory (9) (CAT-DI) in the dynamic environment of an academic ED.

METHODS

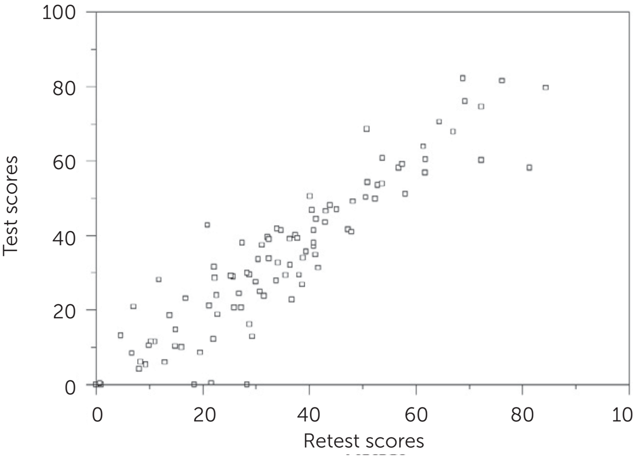
From May 2015 to July 2015, patients (N=101) presenting to the University of Chicago Medical Center ED were screened twice within the course of their ED visit with the CAT-DI. The patients were part of a larger sample (N=1,000). Research assistants randomly selected patients to approach on the basis of a snapshot of the current ED census. Patients who had a critical illness, were age 17 or younger, were non–English speaking, were without decisional capacity, or had a behavioral health–related chief complaint were excluded. After written consent was obtained, the CAT-DI was administered twice by research assistants using tablet computers. The second test was administered within one to three minutes following the end of the first test. All procedures were approved by the University of Chicago Institutional Review Board.

The CAT-DI test is designed to ask different questions on repeated administrations on the basis of changes in severity. It also selects the next two optimal items at each point in the adaptive testing session and randomly selects between them with a .5 probability. In this way, even if the depressive severity level is unchanged, different items are presented during the two testing sessions. Scores are expressed on a 100-point scale with precision equal to 5 points. Results were determined on the basis of categories developed in our original study, with scores of 50–64 indicating mild depression; 65–74, moderate depression; and 75–100, severe depression (9). Pearson product-moment correlation was used to assess test-retest reliability and a paired t test was used to examine bias.

RESULTS

Test-retest reliability was assessed (r=.92). Mean±SD scores for the two testing sessions were 34.60±19.28 (range 0.0–84.4) and 33.81±20.77 (range 0.0–82.1), respectively, with an average difference in overall test score of .83. The paired t test indicated no significant bias between test sessions. Figure 1 shows consistent results between the two testing sessions. Median time to test completion was 93 seconds (interquartile range 67–128 seconds) across the total sample. The sample included 80 (79%) patients in the normal range, 14 (14%) with mild depression, four (4%) with moderate depression, and three (3%) with severe depression.

FIGURE 1. Test-retest correlation of scores on the Computerized Adaptive Testing–Depression Inventory[a]



[a] Scores range from 0 to 100, with 50–64 indicating mild depression, 65–74 indicating moderate depression, and 75–100 indicating severe depression.

DISCUSSION AND CONCLUSIONS

CAT based in multidimensional IRT led to reliable screening results upon repeated testing. Scores were highly correlated between the two occasions, and there was no evidence of bias. Concerns regarding limitations in test-retest reliability due to administration of different items were not supported by our findings. Test-retest reliability for CAT-DI, in fact, exceeded test-retest reliability reported in the literature for the fixed-length PHQ-9 (r=.84) (15). The ED is an ideal setting to test the reliability of CAT because of the dynamic nature of acute conditions, which can lead to greater fluctuations in mood.

Items that provide good discrimination of high and low levels of depression in a psychiatric setting may fail to do so in a general medical ED. In future work we will examine differential item functioning between these two settings and identify specific items—for example, somatic items—that may be less useful for the assessment of depression in the ED. These items can be eliminated from the adaptive administration process in the ED, leading to further increases in precision and decreases in test length in this setting.

AUTHOR AND ARTICLE INFORMATION

Dr. Beiser and Ms. Vu are with the Section of Emergency Medicine, and Dr. Gibbons is with the Center for Health Statistics, University of Chicago, Chicago. Send correspondence to Dr. Gibbons (e-mail: rdg@uchicago.edu).

The work was supported by National Institute of Mental Health grant R01 MH66302. The authors acknowledge the efforts of Dave Patterson, B.S., and of Alexandra Berthiaume, Cody Davis, Annie Hao, and Anna Shin, B.A., who served as volunteer research assistants in the conduct of this study.

Dr. Gibbons is a founder of Adaptive Testing Technologies, which distributes the CAT mental health instruments. The other authors report no financial relationships with commercial interests.

Received July 29, 2015; revision received October 23, 2015; accepted December 18, 2015; published online April 15, 2016.

REFERENCES

1. Stewart WF, Ricci JA, Chee E, et al: Cost of lost productive work time among US workers with depression. JAMA 289:3135–3144, 2003
2. Wang PS, Kessler RC: Global burden of mood disorders; in The American Psychiatric Publishing Textbook of Mood Disorders. Edited by Stein DJ, Kupfer DJ, Schatzberg AF. Arlington, Va, American Psychiatric Publishing, 2005
3. Dolan MA, Mace SE: Pediatric mental health emergencies in the emergency medical services system. Pediatrics 18:1764–1767, 2006
4. Goldberg SE, Whittamore KH, Harwood RH, et al: The prevalence of mental health problems among older adults admitted as an emergency to a general hospital. Age and Ageing 41:80–86, 2012
5. Hoyer D, David E: Screening for depression in emergency department patients. Journal of Emergency Medicine 43:786–789, 2012
6. Kumar A, Clark S, Boudreaux ED, et al: A multicenter study of depression among emergency department patients. Academic Emergency Medicine 11:1284–1289, 2004
7. Gibbons RD, Kupfer DJ, Frank E: Computerized adaptive testing. Annual Review of Clinical Psychology (Epub ahead of print, March 28, 2016)
8. Gibbons RD, Hooker G, Finkelman MD, et al: The Computerized Adaptive Diagnostic Test For Major Depressive Disorder (CAD-MDD): a screening tool for depression. Journal of Clinical Psychiatry 74: 669–674, 2013
9. Gibbons RD, Weiss DJ, Pilkonis PA, et al: Development of a computerized adaptive test for depression. Archives of General Psychiatry 69:1104–1112, 2012
10. Achtyes ED, Halstead S, Smart L, et al: Validation of computerized adaptive testing in an outpatient nonacademic setting: the VOCATIONS trial. Psychiatric Services 66:1091–1096, 2015
11. Gibbons RD, Hedeker DR: Full-information item bi-factor analysis. Psychometrika 57:423–436, 1992
12. Gibbons RD, Bock RD, Hedeker D, et al: Full-information item bifactor analysis of graded response data. Applied Psychological Measurement 31:4–19, 2007
13. Fraley RC, Waller NG, Brennan K: An item response theory analysis of self-report measures of adult attachment. Journal of Personality and Social Psychology 78:350–365, 2000
14. Davidshofer KR, Murphy CO: Psychological Testing: Principles and Applications. Upper Saddle River, NJ, Pearson/Prentice Hall, 2005
15. Kroenke K, Spitzer RL, Williams JBW: The PHQ-9: validity of a brief depression severity measure. Journal of General Internal Medicine 16:606–613, 2001