Data Supplement S1. Supplemental material

Cross-validation and Calibration of the Computerized Adaptive Test-Depression Inventory (CAT-DI) for the Adult Emergency Department Population

INTRODUCTION

The Computerized Adaptive Test-Depression Inventory (CAT-DI)[1] is an adaptive dimensional severity measure for depression that has demonstrated accuracy in psychiatric[2], community mental health[3], primary care[4] and ED setting.[5] In this supplement we describe the process of validation and calibration of this instrument for use in the adult ED population.[6]

We have previously validated the CAT-DI instrument in perinatal women[7] and Spanish speaking Latin American migrants in the United States and Spain.[8] For this we tested for the presence of measurement bias in the form of differential item functioning (DIF).[6] DIF occurs when people from different subgroups with the same underlying level of a latent trait, in this case depression, have different likelihoods of endorsing certain survey items about depression. Such bias may, for example, as a result of translation problems, cultural relevance, or semantic differences. For the current study, we tested for the presence of DIF in the CAT-DI by comparing responses of our adult ED sample to the expected responses based on the original psychiatric outpatient base sample calibration.[1]

METHODS

Subjects

Psychiatric Base 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). 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. Subjects with schizophrenia, schizoaffective disorder, or psychosis; organic neuropsychiatric syndromes (e.g., Alzheimer disease); drug or alcohol dependence within the past 3 months; inpatient treatment status; and individuals who were unable or unwilling to provide informed consent, were excluded. Complete details of the sample have been previously described.[1]

ED Sample: Responses were analyzed from a sample of adult ED patients at an urban, academic tertiary medical center for a non-mental health indication as described in the text of the main manuscript.

Item Bank

The depression item bank contained 389 items selected based on a review of more than 100 existing depression or depression-related rating scales. Items were modified to refer to the previous 2-week period and to have consistent response categories. The majority of items were rated on a 5-point ordinal scale. Example items are provided in the on-line supplement of the previously published paper.[1]

Statistical Methods

This method of DIF estimation have been previously published.[7] Using the original bifactor model calibration for the psychiatric sample[1] we scored the response patterns for depression severity for the ED sample. The ordinal response data were then regressed on the estimated scores for each item using a logistic regression model. A slope of 1.0 is considered to represent the lower bound on good discrimination (factor loading equivalent of 0.5). The beta coefficient for the estimated severity score based on the original psychiatric sample calibration in the logistic regression describes the strength of association between the original calibration-based severity estimate and the probability of a category increase in the response scale for the ED

subjects. This estimate is equivalent to the slope in the multidimensional (bifactor) IRT model for the primary dimension and can also be expressed as an odds ratio (OR) of 2.75 for slope=1.0. As such, items with ORs < 2.75 have evidence of differential item functioning (DIF) and do not discriminate well in ED patients. Differences in the intercepts of the logistic regression between the two populations can be produced by either differences in the underlying means between the two populations or differences in the amount of severity it takes to shift between categories between the two populations. In this analysis our focus is on the key question of differences between the two populations in terms of the items' ability to discriminate between high and low levels of depression, adjusting for differences in overall mean severity at both the item and population levels which are absorbed in the intercept of the regression.

We tested the most commonly administered items (based on CAT-DI) for DIF. These items had a minimum of 50 subjects responding to the item. There were 69 depression items used in the DIF analyses. To examine overall test differences, we scored the ED data using the original bifactor model calibration and a new bifactor model calibration based on the ED data and computed their correlation. Differences in scale can occur if there are differences in the severity level between the two populations, which can be removed by equating the distribution of the severity scores to have mean zero and variance one. We would expect such differences because the majority of the original sample was obtained from psychiatric clinics. This latter test examines the extent to which the optimal calibration for the ED data produces severity estimates which differ from those based on the original calibration.

RESULTS

Differential Item Functioning

Of the 69 depression items evaluated only 1 item exhibited DIF (i.e. failure to discriminate between high and low levels of depression in the ED sample based on the psychiatric sample

calibration parameters: (*In the past 2 weeks, I felt that everything I did was an effort?*). The correlation between the depression expected *a posteriori* (EAP) severity scores based on the original calibration and the ED calibration was r=0.983 (Figure S1).

DISCUSSION

Overall there was very little evidence of DIF in depression symptoms for the CAT-DI in ED patients compared to the original psychiatric calibration sample. This was true for the most commonly administered items for which there were sufficient data to test for DIF. The single exception was "*I felt that everything I did was an effort*." Which is likely biased by medical comorbidities. Even with this item included, the estimated severity scores were almost perfectly correlated based on the original and ED-specific item calibrations (r=0.983) indicating that the CAT-DI scale performs well as is in our emergency department sample.

While there are many different approaches to the analysis of DIF,[9–11] the approach used here has several advantages for determining DIF from CAT-based testing of multidimensional constructs. First, it preserves the multidimensional nature of the underlying IRT model whereas approaches based on multiple-group IRT[10] generally are based on unidimensional IRT and can lead to biased results. Second, the use of the logistic regression model permits DIF analyses where the number of subjects taking any particular item can be small. In our case, we had 1000 subjects taking the CAT-MH; however, our analysis was restricted to the most commonly administered items (symptoms) during adaptive testing (administered to 50 or more patients). Nevertheless, we were able to detect DIF where it existed. Third, our analysis focused on the item's ability to discriminate high and low levels of the underlying traits of interest while holding differences in population means and item parameters related prevalence constant. The key interest here is determining which items should and should not be used in patients with a particular comorbidity. Overall, the major advantage is that this approach provides for

continuous quality improvement where the results of routine adaptive testing in a population of interest can be used to determine DIF once a sufficient number of CAT interviews have been conducted. Here, 1000 interviews produced reasonable results for DIF testing based on large item banks.

CONCLUSIONS

The CAT-DI has been previously validated in a psychiatric population. We found little evidence of item-level DIF within the CAT-DI item bank in our ED population. This demonstrates the validity of the CAT-DI as an appropriate diagnostic screening instrument in the ED.

REFERENCES

1. Gibbons RD, Weiss DJ, Pilkonis PA, et al. Development of a computerized adaptive test for depression. Arch Gen Psychiatry 2012;69(11):1104–12.
2. Bock RD, Gibbons R, Muraki E. Full-Information Item Factor Analysis. Appl Psychol Meas 1988;12(3):261–80.
3. Achtyes ED, Halstead S, Smart L, et al. Validation of Computerized Adaptive Testing in an Outpatient Non-academic Setting: the VOCATIONS Trial. Psychiatr Serv 2015;66(10):1091–6.
4. Graham AK, Minc A, Staab E, Beiser DG, Gibbons RD, Laiteerapong N. Validation of the Computerized Adaptive Test for Mental Health in Primary Care. Ann Fam Med 2019;17(1):23–30.
5. Gibbons RD, Beiser DG, Edwin D, et al. Author ' s Accepted Manuscript Accepted date : 17. J Affect Disord 2016;
6. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland P, Wainer H, editors. Differential item functioning. Hillsdale: Lawrence Erlbaum Associates, Inc.; 1993. p. 67–100.
7. Kim JBJ, Silver RK, Elue R, et al. The experience of depression, anxiety, and mania among perinatal women. Arch Womens Ment Health 2016;19(5):883–90.
8. Gibbons RD, Alegría M, Cai L, et al. Successful Validation of the CAT-MH Scales in a Sample of Latin American Migrants in the United States and Spain. Psychol Assess 2018;
9. Baker FB. The Basics of Item Response Theory. 2001.
10. Weiss DJ. Adaptive testing by computer. J Consult Clin Psychol 1985;53(6):774–89.
11. Breslau J, Javaras KN, Blacker D, Murphy JM, Normand S-LT. Differential item functioning between ethnic groups in the epidemiological assessment of depression. J Nerv Ment Dis 2008;196(4):297–306.

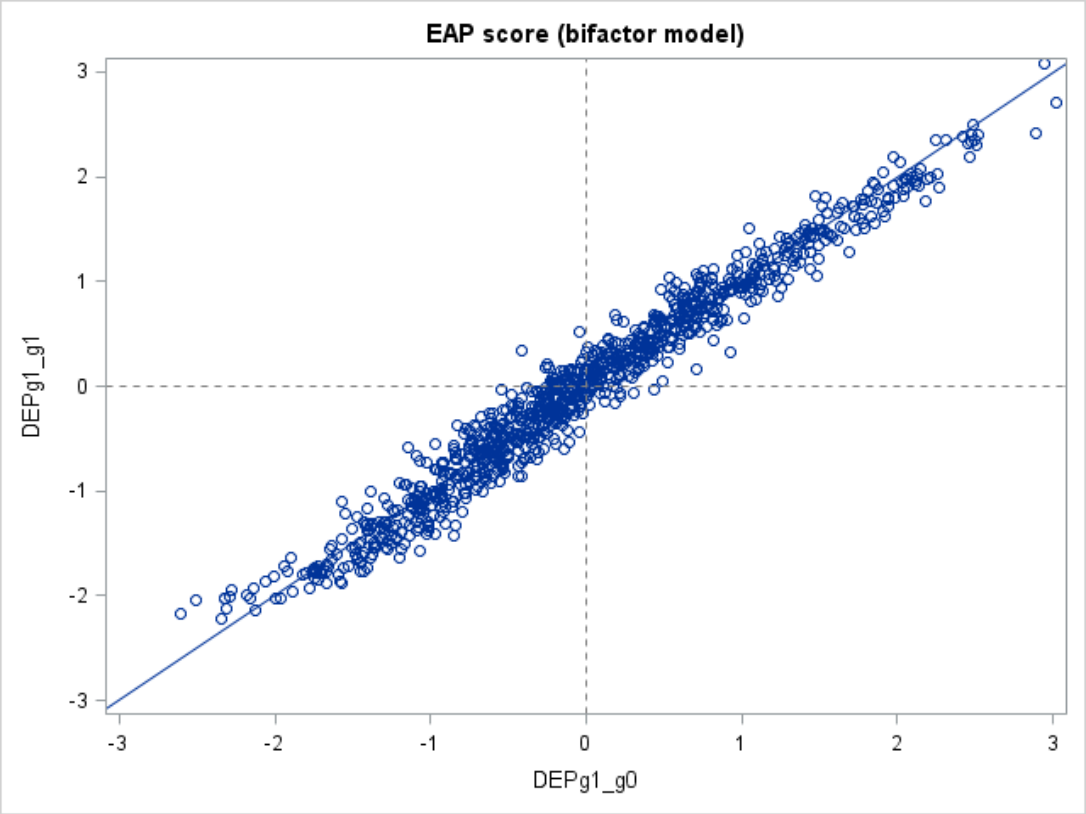
EAP score (bifactor model)
DEPg1_g1
DEPg1_g0
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Figure S1: Correlation between Expected *A Posteriori* (EAP) severity scores for the ED group (g1) based on the original calibration (DEPg1_g0) and the ED calibration (DEPg1_g1)

Table S1. Incident Rate Ratios (95% CIs) for ED Utilization and Hospitalizations: CAD-MDD[a] Full Models

Independent Variable	Number of ED Visits IRR (95% CI)	P-Value	Number of Hospitalizations IRR (95% CI)	P-Value
MDD				
Negative	REF		REF	
Positive	1.61 (1.27-2.03)	<0.0001	1.49 (1.06-2.09)	0.02
CCI	1.09 (1.01-1.18)	0.04	1.23 (1.10-1.38)	<0.001
Age	0.98 (0.97-0.98)	<0.0001	0.98 (0.97-0.99)	<0.0001
Insurance Status				
Commercial	REF		REF	
Medicaid	1.79 (1.32-2.44)	<0.001	2.04 (1.28-3.27)	0.003
Medicare	3.26 (2.35-4.53)	<0.0001	3.58 (2.19-5.84)	<0.0001
Miscellaneous	1.46 (0.70-3.04)	0.31	1.94 (0.69-5.48)	0.21
Uninsured	0.82 (0.45-1.49)	0.51	0.09 (0.01-0.79)	0.03
Gender				
Male	REF		REF	
Female	0.89 (0.71-1.10)	0.28	0.68 (0.50-0.94)	0.02
Ethnicity				
Not Hispanic/Latino	REF		REF	
Hispanic/Latino	0.58 (0.29-1.17)	0.13	0.22 (0.05-0.94)	0.04
Race				
Not White	REF		REF	
White	0.73 (0.52-1.03)	0.08	0.94 (0.56-1.56)	0.80
Has PCP				
No	REF		REF	
Yes	2.18 (1.73-2.76)	<0.0001	2.59 (1.82-3.70)	<0.0001
Current use of alcohol				
No	REF		REF	
Yes	0.88 (0.69-1.12)	0.30	0.89 (0.63-1.26)	0.50
Current smoker				
No	REF		REF	
Yes	1.78 (1.35-2.36)	<0.0001	1.76 (1.18-2.63)	0.006
Current use of drugs				
No	REF		REF	
Yes	1.33 (0.94-1.88)	0.10	1.68 (1.03-2.74)	0.04

[a] CAD-MDD: Computerized Adaptive Diagnostic Test for Major Depressive Disorder

Table S2. Incident Rate Ratios (95% CIs) for ED Utilization and Hospitalizations: CAT-DI[a] Full Models

Independent Variable	Number of ED Visits IRR (95% CI)	P-Value	Number of Hospitalizations IRR (95% CI)	P-Value
CAT-DI	1.10 (1.04-1.16)	<0.001	1.10 (1.02-1.18)	0.02
CCI	1.08 (1.00-1.17)	0.05	1.23 (1.10-1.37)	<0.001
Age	0.98 (0.97-0.98)	<0.0001	0.98 (0.97-0.99)	<0.0001
Insurance Status				
Commercial	REF		REF	
Medicaid	1.85 (1.36-2.52)	<0.0001	2.12 (1.33-3.39)	0.002
Medicare	3.30 (2.37-4.58)	<0.0001	3.67 (2.25-6.01)	<0.0001
Miscellaneous	1.41 (0.68-2.94)	0.35	1.96 (0.69-5.53)	0.21
Uninsured	0.86 (0.47-1.57)	0.62	0.10 (0.01-0.84)	0.03
Gender				
Male	REF		REF	
Female	0.87 (0.69-1.08)	0.21	0.66 (0.48-0.91)	0.01
Ethnicity				
Not Hispanic/Latino	REF		REF	
Hispanic/Latino	0.55 (0.27-1.10)	0.09	0.21 (0.05-0.91)	0.04
Race				
Not White	REF		REF	
White	0.75 (0.53-1.05)	0.10	0.96 (0.58-1.60)	0.87
Has PCP				
No	REF		REF	
Yes	2.18 (1.72-2.76)	<0.0001	2.54 (1.78-3.62)	<0.0001
Current use of alcohol				
No	REF		REF	
Yes	0.85 (0.67-1.08)	0.19	0.87 (0.62-1.23)	0.44
Current smoker				
No	REF		REF	
Yes	1.78 (1.34-2.36)	<0.0001	1.70 (1.14-2.55)	0.01
Current use of drugs				
No	REF		REF	
Yes	1.33 (0.94-1.88)	0.11	1.63 (1.00-2.66)	0.05

[a]CAT-DI: Computerized Adaptive Testing-Depression Inventory (severity classifier)