ADDICTION

SSA SOCIETY FOR THE STUDY OF ADDICTION

METHODS AND TECHNIQUES

doi:10.1111/add.14938

Development of a computerized adaptive substance use disorder scale for screening and measurement: the CAT-SUD

Robert D. Gibbons[1], Margarita Alegria[2], Sheri Markle[3], Larimar Fuentes[3], Liting Zhang[3], Rodrigo Carmona[4], Francisco Collazos[5,6], Ye Wang[7] & Enrique Baca-García[8,9]

Departments of Medicine and Public Health Sciences, The University of Chicago Biological Sciences, Chicago, IL, USA,[1] Disparities Research Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA,[2] Disparities Research Unit, Massachusetts General Hospital, Boston, MA, USA,[3] Department of Psychiatry, Fundación Jiménez Díaz, Madrid, Spain,[4] Department of Psychiatry and Forensic Medicine, Autonomous University of Barcelona, Barcelona, Spain,[5] Department of Psychiatry, Hospital Universitari Vall d'Hebron, Barcelona, Spain,[6] Disparities Research Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA,[7] Department of Psychiatry, Instituto de Investigación Sanitaria, Fundación Jiménez Díaz, Madrid, Spain[8] and Psychiatry Department, Autonoma University of Madrid, Madrid, Spain[9]

ABSTRACT

Background and aims The focus of this paper is on the improvement of substance use disorder (SUD) screening and measurement. Using a multi-dimensional item response theory model, the bifactor model, we provide a psychometric harmonization between SUD, depression, anxiety, trauma, social isolation, functional impairment and risk-taking behavior symptom domains, providing a more balanced view of SUD. The aims are to (1) develop the item-bank, (2) calibrate the item-bank using a bifactor model that includes a primary dimension and symptom-specific subdomains, (3) administer using computerized adaptive testing (CAT) and (4) validate the CAT-SUD in Spanish and English in the United States and Spain. **Design** Item bank construction, item calibration phase, CAT-SUD validation phase. **Setting** Primary care, community clinics, emergency departments and patient-to-patient referrals in Spain (Barcelona and Madrid) and the United States (Boston and Los Angeles). **Participants/cases** Calibration phase: the CAT-SUD was developed via simulation from complete item responses in 513 participants. Validation phase: 297 participants received the Composite International Diagnostic Interview (CIDI) and the CAT-SUD. **Measurements** A total of 252 items from five subdomains: (1) SUD, (2) psychological disorders, (3) risky behavior, (4) functional impairment and (5) social support. CAT-SUD scale scores and CIDI SUD diagnosis. **Findings** Calibration: the bifactor model provided excellent fit to the multi-dimensional item bank; 168 items had high loadings (> 0.4 with the majority > 0.6) on the primary SUD dimension. Using an average of 11 items (four to 26), which represents a 94% reduction in respondent burden (average administration time of approximately 2 minutes), we found a correlation of 0.91 with the 168-item scale (precision of 5 points on a 100-point scale). Validation: strong agreement was found between the primary CAT-SUD dimension estimate and the results of a structured clinical interview. There was a 20-fold increase in the likelihood of a CIDI SUD diagnosis across the range of the CAT-SUD (AUC = 0.85). **Conclusions** We have developed a new approach for the screening and measurement of SUD and related severity based on multi-dimensional item response theory. The bifactor model harmonized information from mental health, trauma, social support and traditional SUD items to provide a more complete characterization of SUD. The CAT-SUD is highly predictive of a current SUD diagnosis based on a structured clinical interview, and may be predictive of the development of SUD in the future.

Keywords bifactor model, Computerized adaptive testing, item response theory, Latino, mental health, substance use disorder.

Correspondence to: Robert D. Gibbons, Blum-Riese Professor of Biostatistics, Department of Public Health Sciences, The University of Chicago, 5841 South Maryland Ave., MC2000, 60637-1447, Chicago, IL, USA. E-mail: rdg@uchicago.edu

Submitted 27 November 2018; initial review completed 24 May 2019; final version accepted 6 December 2019

INTRODUCTION

Substance use disorders (SUDs) are now considered a public health emergency [1], with urgent need for fast action. Most people with SUD receive no treatment for their behavioral health conditions [2]. A major barrier to receiving treatment is the fast and effective identification of those in need [3]. SUD identification and prevention is predicated on accurate initial risk detection, monitoring changes in risk over time and effective, timely intervention delivery [4]. Instruments are needed that not only assist with initial detection of risk but also provide efficient, accurate quantification of non-negligible risk to assist in clinical decision-making and resource allocation across diverse health-care settings (i.e. emergency departments, in-patient units, out-patient primary care and behavioral health settings). A single computer program that can be used for screening, quantification and monitoring of SUD risk and which promotes just-in-time intervention referral while remaining feasible for trans-setting use would be truly transformative, and might prove a major advance on current practice.

Most people with substance use problems have a higher risk for chronic diseases, yet most receive no treatment for substance use conditions [5–7]. Although the Affordable Care Act (ACA) expanded eligibility for Medicaid, research to date has not found that expansions decreased the substance treatment gap between whites and racial–ethnic minorities [2]. Given that SUD and mental health problems are the largest source of premature disability [8] and that untreated substance use is associated with premature mortality [9,10] and productivity loss [11], these negative outcomes underscore the importance of early identification and treatment. Drug misuse places users at risk for co-occurring health problems such as elevated depression [12,13] or anxiety symptoms [14,15] and poor chronic disease management. Lack of early identification of substance problems can impact functioning and lead to negative outcomes [16,17] for blacks and Latinos, such as homelessness, incarceration [18] and disability [19]. This lack of identification for early referral and treatment is a missed opportunity, given that treatment has been shown to reduce disability days by 40–45% and to improve functioning [20].

Effective detection and measurement, together with clear recommended actions for clinicians (i.e. clinical decision support), are essential for providing effective care for SUDs [4]. Moreover, because SUD is not static, it is equally important to monitor risk over time and enable interventions to be delivered during the moments of greatest need. Psychometrically valid and reliable tools are needed to identify SUD, guide decision-making around care pathways for patients and monitor risk through care transitions and through treatment. Using recent advances in measurement [multi-dimensional item response theory (MIRT) and computerized adaptive testing (CAT): MIRT-based CAT] and information technology (cloud computing, patient portals, electronic health records), it is now possible to develop a SUD monitoring system that will alert clinicians, care managers or other designees, such as parents or significant others, when risk exceeds or escalates beyond a predetermined threshold.

Traditional approaches to SUD screening and measurement suffer from many limitations. First, they are based on fixed sets of symptom items that are often limited to a simple tally of types and frequency of substances used. Secondly, they have limited utility in terms of repeat assessments. Thirdly, many require clinician administration, limiting how widely screening and measurement can be applied. Fourthly, they do not provide uncertainty in estimates of the severity of the underlying SUD. Fortunately, many of the previously noted weaknesses that characterize current approaches to SUD screening and measurement can be addressed through IRT [21] and CAT [22]. Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. In classical test theory [23], a specific counting operation measures severity, the simple sum of the individual item responses (e.g. number of symptoms present). All symptoms or behaviors are treated as if they are equally severe. In IRT, symptoms/behaviors are arranged on a continuum at certain fixed points of increasing severity. This ordering is produced by estimating the parameters of an underlying model of measurement which describe how well each item discriminates between low and high levels of the underlying disorder, and how severely impaired the person must be to endorse the symptom item. Severity of illness is measured by the location on the continuum corresponding to the level of severity of the most severe symptom expressed or behavior manifested. In IRT, severity is measured by a scale point, not a numerical count [24–26].

These two theories of measurement are fundamentally different: changing the symptoms (added or deleted symptom items) produce scores that are no longer comparable on traditional tests. However, this is not the case for IRT-based measurement. If the severity of the symptoms or behaviors is changed, or items are added or deleted, the scores remain comparable; only the precision of measurement at some points on the scale change. This property of scaled measurement, as opposed to counts of events, is the fundamental advantage of IRT over classical methods of measurement [24–26].

CAT takes further advantage of the scaled property of IRT measurement by adaptively administering a subset of symptoms drawn from a much larger ‘bank’ of symptoms or behaviors, targeted to the specific level of severity of each individual [22]. Beginning with an item in the middle of the scale severity distribution, or based on a previous test administration after each item is administered, a provisional score and its uncertainty are computed and, based on the score, the next most informative item in the bank is administered. The process continues until the uncertainty falls below a predefined threshold. The paradigm shift is from short fixed-length tests with varying precision to tests with fixed-precision and varying number of items. Using CAT, we can dramatically increase precision but minimize patient burden and eliminate clinician burden.

In contrast to educational measurement, where IRT-based CAT is used to measure essentially unidimensional constructs such as mathematical ability, MIRT-based CAT is needed to measure complex constructs such as depression, anxiety, suicidality or SUD, where the symptoms are drawn from multiple intercorrelated subdomains [26,27]. A recent study of SUD symptoms showed that MIRT, specifically a bifactor model as used here, was the best-fitting model relative to unidimensional and other multidimensional alternatives [28]. Although there are other computerized SUD scales, for example a computer administered version of the clinician administered Addiction Severity Index [29] or the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST) [30], and some based on unidimensional IRT [31], this is the first example of MIRT-based CAT for SUD measurement. Such an approach allows us to accommodate the multi-dimensionality of SUD symptoms [28], as well as bridge the gap between depression, anxiety, social isolation, post-traumatic stress disorder (PTSD), risk-taking behaviors and traditional SUD items in a psychometrically rigorous way, and to identify symptoms that are the precedents of SUD. This has not been conducted previously, and it cannot be performed in a statistically rigorous way using either classical test theory or unidimensional IRT and related CAT methodology. As such, most traditional SUD measures require that the patient be abusing at least one substance to have a non-zero score. This is not true when using MIRT/CAT methodology, where all related items from mental health dimensions and SUD questions contribute to the estimated score(s) resulting in a continuous dimensional measure of SUD severity.

In this study, we (a) develop a large SUD item bank that cuts across SUD symptoms and the four subdomains, (b) calibrate the item bank using a MIRT bifactor model that accommodates the multi-dimensionality of the item bank, (c) develop the CAT-SUD adaptive test and (d) validate the CAT-SUD using the Composite International Diagnostic Interview (CIDI) as an external standard.

METHODS

Design

This study was conducted in compliance with the ethical principles of the Declaration of Helsinki, the US Food and Drug Administration guidelines and the International Conference on Harmonization's Good Clinical Practices Guidelines. The Institutional Review Boards at Massachusetts General Hospital (for MGH and Boston Medical Center), the University of Chicago, the Fundación Jiménez Díaz (Madrid, Spain), the Hospital Universitari Vall d'Hebron (Barcelona, Spain) and the University of Southern California approved the study, and individuals signed a written informed consent form prior to initiation of study procedures.

The study involved the following steps: (1) development of a comprehensive item bank to cover SUD as well as related psychological, risky behavior, functional impairment and social support subdomains; (2) calibration of the item bank using data collected from a multi-national sample; (3) simulated adaptive testing from complete item–response data to develop an adaptive test, the CAT-SUD; and (4) testing and validation of the CAT-SUD based on the CIDI, a structured clinical interview in a diverse emergency department sample that included minorities taking the test in English and Spanish.

The CAT-SUD design involves adaptive measurement that includes a fixed set of substance-specific questions (administered randomly throughout the testing session) that inquires about use and frequency (0 , 1–10, 11–20, 21–30/31 days) of (1) opiates and analgesics, (2) alcohol, (3) cocaine and amphetamines, (4) heroin or methadone and (5) sedatives (sedatives, hypnotics, tranquilizers, barbiturates), with accompanying lists of specific substances. These fixed items are used to understand more clearly the type and frequency of the substances used but are not directly used in the CAT-SUD scoring, as the absence of use of one substance (e.g. alcohol) is not a sign that the use of another substance (e.g. opiate) is not severe. Rather, the scoring is based on questions related to drug and alcohol use in general; for example, the extent to which drug or alcohol use has led to neglecting one's family.

Item bank

The item bank consisted of 252 items drawn from the following subdomains: substance abuse, depression, anxiety, PTSD, severe mental illness, risky sexual behavior and HIV, functional impairment and social support. The selection of the subdomains was based on a group of experts who selected the existing scales for the original survey (see Supporting information, Appendix S1 for a list of measures). This survey required identifying people with co-occurring mental health and substance use conditions in community-based organizations or clinics. Because the study was offering a transdiagnostic treatment for both mental health and addictions, it needed to cover a broad range of disorders, symptoms and behaviors, including tobacco, alcohol, benzodiazepines, cannabis, cocaine and other illicit drugs. Questions include measures of specific substances of use in the past 30 days, use of multiple drugs including benzodiazepines, ability to stop using substances, injection drug use practices and assessments of the consequences of drug use. Mental health items include common measures of depressive and anxiety symptoms, trauma exposure and symptoms of PTSD. Additional questions asked about use of specialty services and medications, experience of chronic conditions and functioning and disability as well as smoking behaviors. A series of questions address psychosocial stress, context of exit from their country of origin, sense of belonging, barriers to treatment and family relations.

Data collection

Data collection was divided into two stages, calibration and validation.

Calibration sample

Data used in this study were collected as a part of the original study described by Alegria and colleagues [32]. A total sample of 513 participants responded to the battery of questions that was later used to develop the CAT-SUD. These 513 participants completed a baseline interview consisting of 252 items. They were recruited from primary care, community clinics, emergency departments and patient-to-patient referrals in Spain and the United States as part of the International Latino Research Partnership clinical trial. The 513 included three types of respondents. A total of 341 completed the baseline prior to enrollment in the clinical trial, based on screening positive to mental health and substance use problems, and not actively receiving specialty behavioral health care. An additional 145 participants were not eligible for the trial, but were administered the baseline interview to assess accuracy of the screening. A final 27 cases were administered the baseline interview as part of sample of pilot intervention participants. All participants identified either themselves or a parent as being Latino or of Latin American origin. A Consort diagram of the two study phases is displayed in Fig. 1.

Validation sample

An independent validation sample was obtained to compare the CAT-SUD test scores to structured clinical interview (CIDI) diagnoses of SUD. We recruited participants through direct contact in the Emergency Departments from clinics in the United States (Massachusetts General Hospital and Boston Medical Center in Boston, and University of Southern California in California) and Spain (Hospital Fundación Jiménez Díaz in Madrid and Hospital Vall d'Hebron in Barcelona). We began recruitment in March 2017 and the final interviews were conducted in August, 2017. The protocol involved administering the CAT-SUD screening interview, followed-up by the CIDI diagnostic interview approximately 2–4 weeks later. After screening and consent, we administered the CAT-SUD to 424 participants who spoke English or Spanish and were between the ages of 18–70 years. Participants came from the five study sites. Among recruited participants, 99 were lost to follow-up and 28 declined the follow-up CIDI interview. A total of 297 completed the CIDI diagnostic follow-up of the interview. We administered CIDI interviews by telephone and in person, based on the participant's preferences. Reasons for declining the CIDI included not feeling well enough (e.g. severe sickness); admission to the hospital; not having time due to school or work; not being interested in a second interview; or concerns about survey length or not wanting to respond to questions about mental health and substance use.

Study staff administered an informed consent form, together with a short demographic form and the CAT-SUD. Study staff scheduled follow-ups using the 1-hour diagnostic interview, the CIDI, either in person or by telephone. This interview was used to determine the accuracy of the CAT assessment by comparing to a gold standard diagnostic measure, the CIDI. We administered the modules for major depressive episode, dysthymia, mania, generalized anxiety disorder, tobacco, illegal substance abuse and dependence and alcohol abuse and dependence. All interviews were audio-recorded.

Statistical analysis

The bifactor model [33] was used to calibrate the 252-item bank. It allows each item to measure the primary dimension (SUD) and one subdomain (e.g. depression). This approach has computational and interpretational advantages over unrestricted exploratory item factor analytical models [34] and extends CAT to the measurement of multi-dimensional constructs [35]. The bifactor model is uniquely suited to the measurement of multi-dimensional mental health and SUD constructs because it incorporates the multi-dimensionality produced by the sampling of items from pre-identified subdomains, but provides a single overall severity index. This facilitates CAT and minimizes the number of items needed for adaptive measurement. The estimated severity score is useful for both screening and measurement and can be used to assess response to treatment on an underlying continuous scale of measurement. Subdomain scores are also estimable [35], but are not our focus in this paper. The bifactor model was fitted to the data by maximum likelihood using freely available software (www.healthstats.org/bifactor.html). The bifactor model was compared to a unidimensional alternative using a likelihood ratio χ^2 statistic.

Items with loading < 0.4 on the primary dimension were eliminated from the item bank [36]. Based on the final bifactor model, a CAT was developed [36]. The

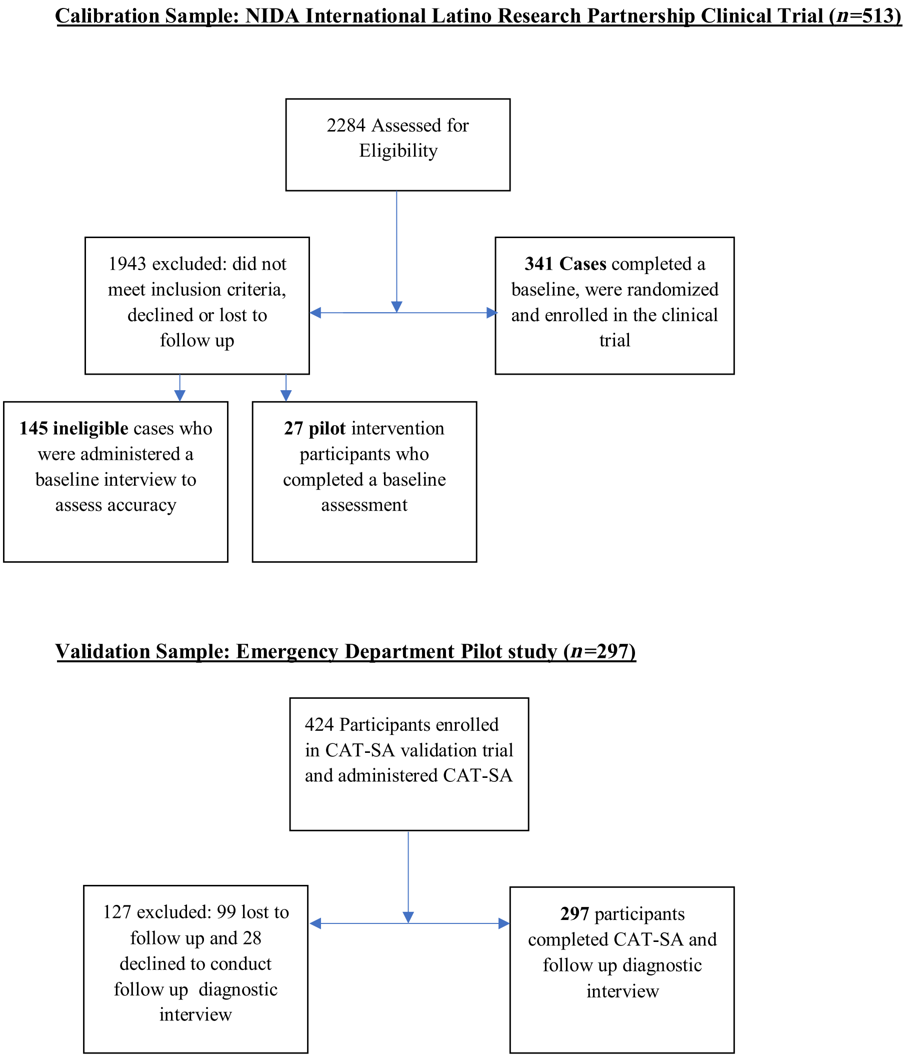
Calibration Sample: NIDA International Latino Research Partnership Clinical Trial (n=513)
2284 Assessed for Eligibility
1943 excluded: did not meet inclusion criteria, declined or lost to follow up
341 Cases completed a baseline, were randomized and enrolled in the clinical trial
145 ineligible cases who were administered a baseline interview to assess accuracy
27 pilot intervention participants who completed a baseline assessment
Validation Sample: Emergency Department Pilot study (n=297)
424 Participants enrolled in CAT-SA validation trial and administered CAT-SA
127 excluded: 99 lost to follow up and 28 declined to conduct follow up diagnostic interview
297 participants completed CAT-SA and follow up diagnostic interview

Figure 1 Consolidated Standards of Reporting Trials (CONSORT) diagram [Colour figure can be viewed at wileyonlinelibrary.com]

properties of the CAT were then determined by simulating CAT from the complete item–response data from the sample of 513 participants. The CAT tuning parameters that minimized the number of items while maintaining a correlation in excess of r = 0.9 with the total item bank score were selected from 1200 different simulations.

For the validation component, we examined the association between the continuous CAT-SUD score (0–100-point scale, transformed from the underlying unit normal score for ease of interpretation by clinicians) and the CIDI SUD diagnosis using a logistic regression model. We computed the probability of SUD as a function of CAT-SUD scale scores and area under the curve (AUC) for the receiver operator curve (ROC), including the CAT-SUD score and self-reported use of drugs, alcohol and opioids. Thresholds for SUD groups were developed based on their ability to differentiate 12-month CIDI SUD diagnoses and self-reported use of drugs, alcohol and opioids. This was performed by examining sensitivity and specificity at various thresholds on the CAT-SUD scale, with low risk having high sensitivity, high risk having high specificity and intermediate risk a balance between sensitivity and specificity.

For more statistical details, the reader is referred to Gibbons, 2016 [26].

RESULTS

Descriptive statistics

Table 1 displays the demographic characteristics of the calibration sample, Table 2 displays the demographic characteristics of the validation sample and Table 3 provides a comparison between the two samples. Both the calibration and the validation samples included most participants in the 18–34-year age group (50.3% calibration, 39.6% validation), although the validation sample had a significantly higher percentage of participants aged 50 and older (15.4% calibration, 30.7% validation). Both had a higher number of females than males. By ethnicity, the calibration

Table 1 Socio-demographic characteristics: CAT-SUD calibration sample (n = 513).

	Total		*Boston*		*Madrid*		*Barcelona*		
	(n = 513)		*(n = 132)*		*(n = 130)*		*(n = 251)*		
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*P-value*
Age (years)									
18–34	258	50.3%	37	28.0%	46	35.4%	175	69.7%	0.001
35–49	164	32.0%	50	37.9%	59	45.4%	55	21.9%	
50+	79	15.4%	45	34.1%	22	16.9%	12	4.8%	
Don't know	12	2.3%	0	0.0%	3	2.3%	9	3.6%	
Gender									
Male	217	42.3%	50	37.9%	48	36.9%	119	47.4%	0.038
Female	284	55.4%	82	62.1%	79	60.8%	123	49.0%	
Don't know	12	2.3%	0	0.0%	3	2.3%	9	3.6%	
Ethnicity									
Non-Latino	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0.095
Latino	500	97.5%	132	100%	126	96.9%	242	96.4%	
Refused or don't know	13	11.6%	0	0.0%	4	20.0%	9	14.5%	
Race									
White	99	19.3%	30	22.7%	16	12.3%	53	21.1%	0.001
Black	24	4.7%	8	6.1%	5	3.8%	11	4.4%	
Indigenous/Native American	35	6.8%	8	6.1%	12	9.2%	15	6.0%	
Asian/Native Hawaiian/Other Pacific Islander	2	0.4%	1	0.8%	1	0.8%	0	0.0%	
Reported no race – only Hispanic/Latino/ Caribbean	54	10.5%	52	39.4%	1	0.8%	1	0.4%	
Mixed	278	54.2%	26	19.7%	90	69.2%	162	64.5%	
Don't know or refuse	21	4.1%	7	5.3%	5	3.8%	9	3.6%	
Education level									
Less than high school	185	36.1%	70	53.0%	47	36.2%	68	27.1%	0.001
HS diploma, GED, vocational school or more	316	61.6%	62	47.0%	80	61.5%	174	69.3%	
Don't know	12	2.3%	0	0.0%	3	2.3%	9	3.6%	
Employment status									
Unemployed	251	48.9%	78	59.1%	55	42.3%	118	47.0%	0.017
Employed	262	51.1%	54	40.9%	75	57.7%	133	53.0%	
Recruitment site									
Primary care	235	45.8%	80	60.6%	84	64.6%	71	28.3%	0.001
Community agency	121	23.6%	37	28.0%	6	4.6%	78	31.1%	
Emergency room	27	5.3%	0	0.0%	27	20.8%	0	0.0%	
Referred	130	25.3%	15	11.4%	13	10.0%	102	40.6%	

Sample includes 341 trial patients, plus 145 cases that were not eligible for the trial but were administered the baseline interview to assess accuracy of the screening (baseline 7 s), and another 27 cases that were pilot intervention participants who completed the baseline assessment. Twelve 12 pilot cases did not complete the screener, therefore missing information on age, gender, race and education level. CAT-SUD = computerized adaptive testing-substance use disorder; GED = General Educational Development.

sample stood out as being almost entirely Latino (97.5%), with only a small percentage of people (2.5%) who did not report ethnicity, given that the parent study focused on Latino immigrants within the United States and Spain. In the validation sample, we broadened the scope to include Latino participants (37.5%), but also most non-Latino (62.3%) respondents, to assess validation in both English and Spanish. We note that 47.6% of the validation sample are from Spain, and among these a majority reported being non-Latino and white. These respondents were native Spanish speakers whose data were included to test the Spanish-language version of the CAT-SUD. By race, the majority (54.2%) of the calibration sample reported being mixed/mestizo, while the majority in the validation sample reported being white (67.5%). Of the white participants, 25% also identified as Latino. In terms of education, most participants had a high school (HS) diploma/general educational development (GED) or higher in both the calibration and validation samples (61.6 and 73.1%).

Calibration

The fit of the bifactor model was significantly improved over the unidimensional alternative [χ^2 = 10 062, degrees

Table 2 Socio-demographic characteristics: CAT-SUD validation sample (*n* = 424).

	Total		*Madrid, FJD*		*Barcelona, Vall d'Hebron*		*LAC USC*		*BMC*		*MGH Boston*		
	(n = 424)		*(n = 109)*		*(n = 93)*		*(n = 62)*		*(n = 24)*		*(n = 136)*		
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*P-value*
Age (years)													
18–34	168	39.6%	59	54.1%	38	40.9%	41	30.1%	9	37.5%	41	30.1%	0.001
35–49	126	29.7%	36	33.0%	30	32.3%	35	25.7%	5	20.8%	35	25.7%	
50+	130	30.7%	14	12.8%	25	26.9%	60	44.1%	10	41.7%	60	44.1%	
Gender													
Male	198	46.7%	43	39.4%	38	40.9%	72	52.9%	11	45.8%	72	52.9%	0.011
Female	224	52.8%	66	60.6%	55	59.1%	64	47.1%	13	54.2%	64	47.1%	
Both	2	0.5%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	
Ethnicity													
Non-Latino	264	62.3%	72	66.1%	69	74.2%	94	69.1%	10	41.7%	94	69.1%	0.001
Latino	159	37.5%	37	33.9%	24	25.8%	42	30.9%	13	54.2%	42	30.9%	
Other	1	0.2%	0	0.0%	0	0.0%	0	0.0%	1	4.2%	0	0.0%	
Race													
White	284	67.0%	92	84.4%	80	86.0%	89	65.4%	9	37.5%	89	65.4%	0.001
Black/African American	32	7.5%	6	5.5%	3	3.2%	8	5.9%	4	16.7%	8	5.9%	
Indigenous/Native American	11	2.6%	0	0.0%	1	1.1%	5	3.7%	3	12.5%	5	3.7%	
Asian/Native Hawaiian/Other Pacific Islander	11	2.6%	2	1.8%	2	2.2%	3	2.2%	4	16.7%	3	2.2%	
Reported no race – only Hispanic/Latino/Caribbean	45	10.6%	2	1.8%	1	1.1%	20	14.7%	2	8.3%	20	14.7%	
Mixed	15	3.5%	2	1.8%	4	4.3%	2	1.5%	0	0.0%	2	1.5%	
Other	26	6.1%	5	4.6%	2	2.2%	9	6.6%	2	8.3%	9	6.6%	
Education level													
Less than high school	111	26.2%	16	14.7%	39	41.9%	24	17.6%	10	41.7%	24	17.6%	0.001
HS diploma, GED, vocational school or more	310	73.1%	92	84.4%	54	58.1%	112	82.4%	12	50.0%	112	82.4%	
Don't know	2	0.5%	0	0.0%	0	0.0%	0	0.0%	2	8.3%	0	0.0%	

Mixed group in Race reported two or more races. CAT-SUD = computerized adaptive testing-substance use disorder; GED = General Educational Development; FJD = Fundación Jiménez Díaz; LAC USC = Los Angeles County University of Southern California; BMC = Boston Medical Center; MGH = Massachusetts General Hospital.

of freedom (d.f.) = 168, $P < 0.0001$]. Of the 252 items in the bank, 168 had strong loadings (≥ 0.4, with the majority > 0.6) on the primary dimension, indicating that there is a strong cross-link between SUD, mental health-related symptomatology and social support.

Simulated CAT

A simulated CAT (i.e. simulating CAT administration from the actual complete item responses) revealed that an average of 11 items (range = 4–26) provided a correlation of $r = 0.91$ with the 168-item scale total score (from the complete test administrations) with precision of 5 points on a 100-point scale metric. These results indicate that the CAT-SUD can reproduce the latent SUD dimension with an average of 11 items (a 94% reduction) in approximately 2 minutes.

Example of CAT-SUD administrations

Table 4 presents an example of a CAT-SUD administration for a patient with high risk of SUD. All sessions include specific abuse questions involving alcohol, sedatives/hypnotics, opioids/analgesics, heroin/methadone and cocaine/amphetamines. This ensures that specific substances of abuse are identified in each adaptive testing session. In all cases, the CAT terminated when the uncertainty was at or below 5 points on the 100-point scale.

Validation study

Figure 2 displays the relationship between the CAT-SUD score and the probability of a CIDI clinician-rated SUD (based on the past 12 months). The logistic regression

Table 3 Comparison of socio-demographic characteristics: calibration sample versus validation sample (*n* = 937).

	Total (n = 937)		*Calibration sample (n = 513)*		*Validation sample (n = 424)*		
	n	*%*	*n*	*%*	*n*	*%*	*P-value*
Age (years)							
18–34	426	45.5%	258	50.3%	168	39.6%	0.001
35–49	290	30.9%	164	32.0%	126	29.7%	
50+	209	22.3%	79	15.4%	130	30.7%	
Don't know	12	1.3%	12	2.3%	0	0.0%	
Gender							
Male	415	44.3%	217	42.3%	198	46.7%	0.003
Female	508	54.2%	284	55.4%	224	52.8%	
Both	2	0.2%	0	0.0%	2	0.5%	
Don't know	12	1.3%	12	2.3%	0	0.0%	
Ethnicity							
Non-Latino	264	28.2%	0	0.0%	264	62.3%	0.001
Latino	659	70.3%	500	97.5%	159	37.5%	
Refused or don't know	14	1.5%	13	2.5%	1	0.2%	
Race							
White	383	40.9%	99	19.3%	284	67.0%	0.001
Black/African American	56	6.0%	24	4.7%	32	7.5%	
Indigenous/Native American	46	4.9%	35	6.8%	11	2.6%	
Asian/Native Hawaiian/Other Pacific Islander	13	1.4%	2	0.4%	11	2.6%	
Reported no race – only Hispanic/Latino/Caribbean	99	10.6%	54	10.5%	45	10.6%	
Mixed (reported 2 or more races or 1 race plus Hispanic/Latino)	293	31.3%	278	54.2%	15	3.5%	
Other or don't know	47	5.0%	21	4.1%	26	6.1%	
Education level							
Less than high school	296	31.6%	185	36.1%	111	26.2%	0.001
HS diploma, GED, vocational school or more	626	66.8%	316	61.6%	310	73.1%	
Don't know	14	1.5%	12	2.3%	2	0.5%	

Difference of the characteristic between the two samples was assessed by χ^2 test. *P*-value of the difference test was reported. GED = General Educational Development.

Table 4 Example CAT-SUD administration: score = 85.3, precision = 5.0, category = high risk.

How many days in the past 30 days have you used opiates/analgesics (including painkillers such as morphine, Dilaudid, Demerol, Percocet, Darvon, Talwin, codeine, fentanyl, OxyContin)?	11–20 days
How many days in the past 30 days did you drink alcohol (including beer, wine and liquor)?	11–20 days
How many days in the past 30 days have you used cocaine (including crack and rock cocaine) or amphetamines (including methamphetamine)?	0 days
How troubled or bothered have you been in the past 30 days by alcohol problems?	Considerably
How troubled or bothered have you been in the past 30 days by drug problems?	Moderately
How many days in the past 30 days have you used heroin or methadone?	1–10 days
During the past year, have you been preoccupied with drinking alcohol and/or using other drugs such as marijuana, or with taking medications without a prescription or more than they were prescribed?	Yes
How many days in the past 30 days have you used sedatives, hypnotics, tranquilizers, or barbiturates (such as Valium, Xanax, Ativan, Seconal, Nembutal)?	1–10 days
In the past 30 days, have you neglected your family because of your use of drugs?	Yes

CAT-SUD = computerized adaptive testing-substance use disorder.

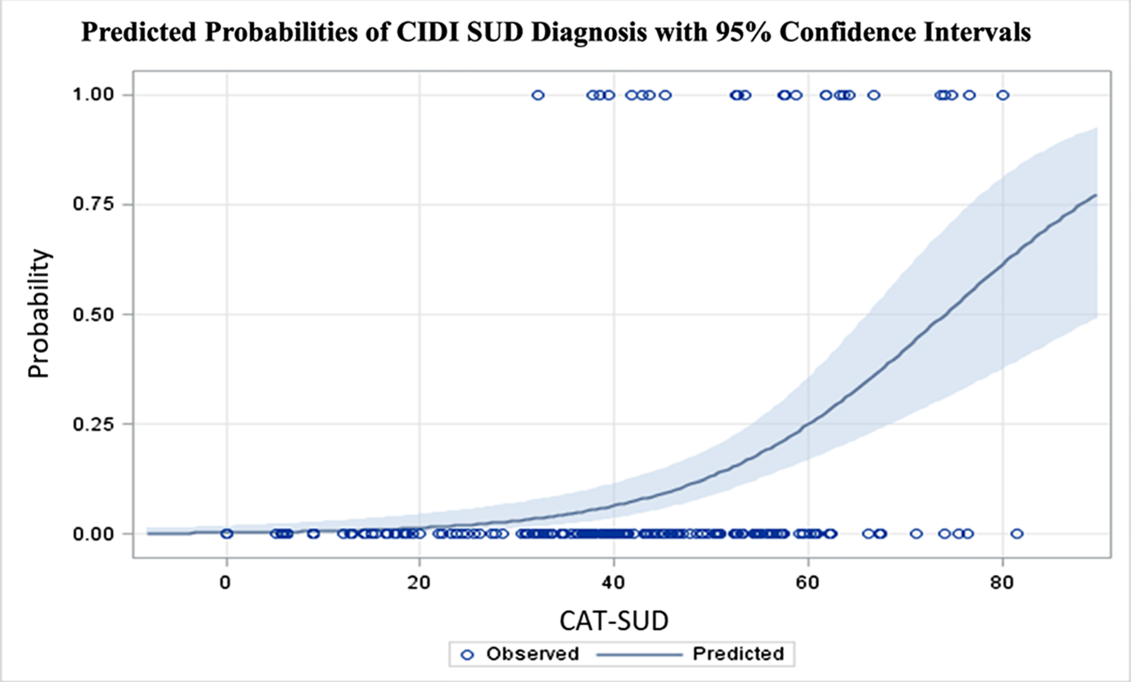

Figure 2 Predicted probabilities of Composite International Diagnostic Interview substance use disorder (CIDI SUD) diagnosis with 95% confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

revealed that for every 10-point increase in CAT-SUD score there is a 2.2-fold increase in the likelihood of a CIDI clinician-rated SUD or a 20-fold increase across the range of the CAT-SUD scale (i.e. odds ratio). Fig. 3 displays the ROC curve for this relationship. The AUC is 0.85 [95% confidence interval (CI) = 0.75, 0.95], confirming the strong relationship between the CAT-SUD test score and the SUD diagnosis.

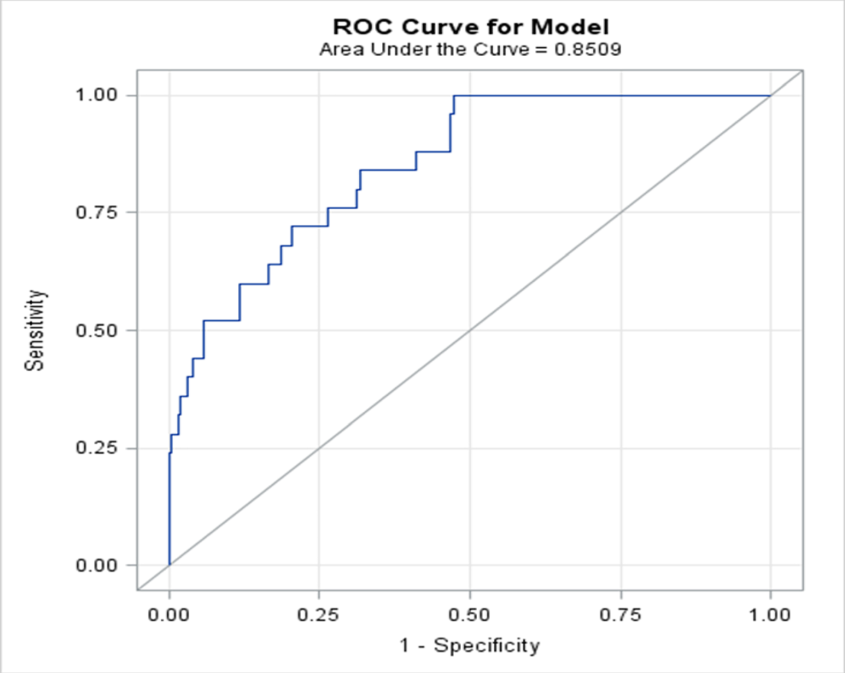

Figure 3 Receiver operating characteristic (ROC) curve [Colour figure can be viewed at wileyonlinelibrary.com]

Severity thresholds

Thresholds were derived based on 12-month CIDI SUD diagnoses and self-reported use of alcohol and drugs. Based on the validation data, thresholds of < 50 (low risk), 50–70 (intermediate risk) and > 70 (high risk) were selected. These thresholds yielded rates of 4, 22 and 50% for SUD diagnoses and 11, 47 and 90% for self-reported alcohol or drug use for low-, intermediate- and high-risk CAT-SUD groups.

DISCUSSION

The CAT-SUD can accurately measure SUD severity with an average of 11 items in approximately 2 minutes. The validation study reveals that the CAT-SUD accurately tracks clinician-rated SUD based on a structured clinical interview in approximately 2 minutes in a very heterogeneous sample taking the tests in two different languages. As such, it can be used to reliably assess SUD in a variety of different health-care settings without clinician burden and minor patient burden. When implemented in a cloud-computing environment [37,38] the CAT-SUD will permit administration in or out of the clinic, anywhere that an internet connection is available on any internet capable device. An advantage of the CAT-SUD over traditional SUD instruments is that it provides quantitative measurement of substance abuse propensity whether or not the individual is actively abusing substances. This is critically important in substance use screening, given the stigma or discomfort attached to reporting substance use [39,40]. Many participants in the calibration sample talked with study clinicians about their hesitance to report substance use in the baseline interview, but their willingness to disclose further as they became comfortable with study staff. Having an instrument that can rapidly and accurately assess substance risk without requiring report of substance use will be a particularly useful addition to the field.

An important limitation of the study is that the calibration and validation samples are drawn from heterogeneous populations that differ significantly on several demographic variables. This is true throughout centers (Boston, Los Angeles, Madrid and Barcelona) and between calibration and validation samples (see Tables 1–3). To the extent that the center heterogeneity during the calibration phase impacts the psychometric properties of our CAT-SUD item bank, this will introduce heterogeneity in the estimated scale scores which will lead to underestimation of the strength of the association with the CIDI SUD diagnoses. Differences between patient characteristics between calibration and validation samples will also limit our ability to validate the results against structured clinical interviews. As we have previously noted [26], for calibration we actually want a heterogeneous sample so that we can fully characterize the latent dimensions of interest. For validation, sample heterogeneity sets an upper bound on the predictive validity that we can achieve while at the same time has the value of increasing generalizability assuming that validity can be demonstrated. As such, the already strong demonstration of validity based on the high AUC for predicting the results of a lengthy structured clinical interview is a lower bound on what we would have observed if the validation and calibration samples were more similar. Our results therefore generalize to patients taking the CAT-SUD in English and Spanish, and between and within American and European cultures, and among immigrants and non-immigrants.

A second limitation is that more than half the participants in the calibration phase were recruited because they screened positive to having mental health and substance use problems. While this may appear to limit generalizability, this was performed to ensure that we have the ability to provide good discrimination throughout the entire continuum of SUD. The validation sample had no such restriction, and the high predictive accuracy indicates that the adaptive test is in fact generalizable to a general patient population.

A third limitation is that the sample was too small to examine differential item functioning (DIF) between item calibrations and diagnostic accuracy in Spain and the United States. Future data collection efforts will address this important issue.

A fourth limitation is that we restricted the list of actual substances being used to five categories to provide a balance between SUD characterization and interview time. Work is under way on the development of the CAT-SUD-E (expanded)- which provides a more comprehensive review of substances used (including marijuana and tobacco), albeit with increased interview time. Use of the CAT-SUD and CAT-SUD-E will depend on the application; however, the adaptive parts of both are identical.

The CAT-SUD can be used for many purposes which include but are not limited to screening, measurement and longitudinal monitoring. Because the CAT-SUD asks different questions upon repeat administration, it can be repeatedly administered at any interval in time without response bias produced by repeated administration of the same items. Previous study of test–retest reliability using CAT for the measurement of depression showed higher reliability ($r = 0.92$) than traditional fixed-length tests despite the use of different items upon repeat testing [41].

A major strength of this study is that the sample is drawn from English and Spanish speakers from the United States and Spain, where the US population is comprised of Latino immigrants. Surveying this subpopulation, which is difficult to reach, is an additional strength of the study.

An additional strength of this study is the development of a cross-link between (1) SUD and related subdomains, including (2) psychological disorders (depression, anxiety, PTSD, severe mental illness), (3) risky sexual behavior and HIV, (4) functional impairment and (5) social support. Including items from the primary SUD domain and each of the four subdomains allows us to determine if they are related in general, and which items from the various subdomains tap the underlying primary SUD dimension. To the extent that they do, using those items in adaptive testing will increase precision of measurement across the SUD continuum.

A further advantage of CAT is that new symptoms can be added to the bank, calibrated and then added to the CAT-SUD once sufficient data are available. We can also determine whether the CAT-SUD is valid in different populations using DIF [42,43]. Finally, the ability to administer the CAT-SUD via the internet using a cloud computing platform further decreases barriers to testing. Where the ability to provide a timely response is available, remote screening for SUD is viable.

More generally, the addition of the CAT-SUD to the CAT-MH, which includes adaptive tests for depression, anxiety, mania/hypomania, PTSD, psychosis, and suicidality will dramatically improve mental health screening and measurement in real-world settings. For example, the CAT-DI (depression test) which was validated in a psychiatric setting was shown to be free of bias when used in an emergency department study [44]. The same test has also been used successfully in primary care [45]. As such, we would expect the CAT-SUD that was validated in an ED setting to work equally well in a primary care setting.

Finally, the validation data were collected via interview, and not a completely computerized self-assessment. However, the interviewer simply read the questions to the subject. We have previously shown that this approach produces similar results to a completely computerized self-assessment [46].

CONCLUSIONS

We have developed a new approach for the screening and measurement of SUD and related severity. Our methodology synthesizes information from multiple related domains from mental health, trauma and social support with traditional SUD questions to provide a more comprehensive measure of SUD. The CAT-SUD is highly predictive of a current SUD diagnosis based on a structured clinical interview, and may be predictive of the development of SUD in the future, in individuals who are currently at high risk of SUD, prior to their development of a SUD.

Declaration of interests

R.D.G. has been an expert witness for the US Department of Justice, Merck, Glaxo-Smith-Kline, Pfizer and Wyeth and is a founder of Adaptive Testing Technologies, which distributes the CAT-MH™ battery of adaptive tests. The terms of this arrangement have been reviewed and approved by the University of Chicago in accordance with its conflict of interest policies.

Acknowledgements

This study was funded through an administrative supplement from the National Institute of Mental Health (Washington DC), R01-MH100155. The general methodology for the use of multi-dimensional item response theory and computerized adaptive testing for the screening and measurement of mental health disorders was preregistered on the NIH website https://projectreporter.nih.gov/project_info_description.cfm?aid=8482531&icde=47628236; however, the specific application to the measurement of substance use disorder was not described; therefore the results should be considered exploratory.

References

1. Haffajee R. L., Frank R. G. Making the opioid public health emergency effective. *JAMA Psychiatry* 2018; **75**: 767–8.
2. Creedon T. B., Lê Cook B. Access to mental health care increased but not for substance use, while disparities remain. *Health Aff* 2016; **35**: 1017–21.
3. Priester M. A., Browne T., Iachini A., Clone S., DeHart D., Seay K. D. Treatment access barriers and disparities among individuals with co-occurring mental health and substance use disorders: an integrative literature review. *J Subst Abuse Treat* 2015; **61**: 47–59.
4. Substance Abuse and Mental Health Services Administration (US); Office of the Surgeon General (US) *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health [internet]. Chapter 4, early intervention, treatment, and management of substance use disorders*. Washington (DC): US Department of Health and Human Services; 2016 Available at: https://www.ncbi.nlm.nih.gov/books/NBK424859/ (Accessed November 28, 2019).
5. Lê Cook B., Alegría M. Racial–ethnic disparities in substance abuse treatment: the role of criminal history and socioeconomic status. *Psychiatr Serv* 2011; **62**: 1273–81.
6. Patel V., Thornicroft G. Packages of care for mental, neurological, and substance use disorders in low- and middle-income countries. *PLOS Med* 2009; **6**: e1000160.
7. Wang P. S., Aguilar-Gaxiola S., Alonso J., Angermeyer M. C., Borges G., Bromet E. J., *et al*. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet* 2007; **370**: 841–50.
8. Compton W. M., Thomas Y. F., Stinson F. S., Grant B. F. Prevalence, correlates, disability, and comorbidity of DSM-IV drug abuse and dependence in the United States: results from the national epidemiologic survey on alcohol and related conditions. *Arch Gen Psychiatry* 2007; **64**: 566–76.
9. Galea S., Ahern M. J., Tardiff K., Leon A., Coffin M. P. O., Derr M. K., *et al*. Racial/ethnic disparities in overdose mortality trends in New York City, 1990–1998. *J Urban Health* 2003; **80**: 201–11.

10. Galvan F. H., Caetano R. Alcohol use and related problems among ethnic minorities in the United States. *Alcohol Res Health* 2003; **27**: 87–94.
11. Office of National Drug Control Policy *The economic costs of drug abuse in the United States: 1992–2002*. Washington, DC, USA: Executive Office of the President; 2004.
12. Rounsaville B. J., Weissman M. M., Crits-Christoph K., Wilber C., Kleber H. Diagnosis and symptoms of depression in opiate addicts: course and relationship to treatment outcome. *Arch Gen Psychiatry* 1982; **39**: 151–6.
13. Daughters S. B., Braun A. R., Sargeant M. N., Reynolds E. K., Hopko D. R., Blanco C., *et al*. Effectiveness of a brief behavioral treatment for inner-city illicit drug users with elevated depressive symptoms: the life enhancement treatment for substance use (LETS act!). *J Clin Psychiatry* 2008; **69**: 122.
14. Grant B. F., Stinson F. S., Dawson D. A., Chou S. P., Dufour M. C., Compton W. Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the national epidemiologic survey on alcohol and related conditions. *Arch Gen Psychiatry* 2004; **61**: 807–16.
15. Brook D. W., Brook J. S., Zhang C., Cohen P., Whiteman M. Drug use and the risk of major depressive disorder, alcohol dependence, and substance use disorders. *Arch Gen Psychiatry* 2002; **59**: 1039–44.
16. Hasin D. S., Stinson F. S., Ogburn E., Grant B. F. Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on alcohol and related conditions. *Arch Gen Psychiatry* 2007; **64**: 830–42.
17. Chung B., Ngo V. K., Ong M. K., Pulido E., Jones F., Gilmore J., *et al*. Participation in training for depression care quality improvement: a randomized trial of community engagement or technical support. *Psychiatr Serv* 2015; **66**: 831–9.
18. Binswanger I. A., Redmond N., Steiner J. F., Hicks L. S. Health disparities and the criminal justice system: an agenda for further research and action. *J Urban Health* 2012; **89**: 98–107.
19. Bing E. G., Burnam M. A., Longshore D., Fleishman J. A., Sherbourne C. D., London A. S., *et al*. Psychiatric disorders and drug use among human immunodeficiency virus–infected adults in the United States. *Arch Gen Psychiatry* 2001; **58**: 721–8.
20. Samet J. H., Friedmann P., Saitz R. Benefits of linking primary medical care and substance abuse services: patient, provider, and societal perspectives. *Arch Intern Med* 2001; **161**: 85–91.
21. Hambleton R. K., Swaminathan H. *Item response theory: principles and applications*. Netherlands: Springer; 1985.
22. Weiss D. J. Adaptive testing by computer. *J Consult Clin Psychol* 1985; **53**: 774–89.
23. Lord F. M., Novick M. R. *Statistical Theories of Mental Test Scores*. Reading MA: Addison-Welsley Publishing Company; 1968.
24. Sani S., Busnello J., Kochanski R., Cohen Y., Gibbons R. D. High frequency measurement of depressive severity in a patient treated for severe treatment resistant depression with deep brain stimulation. *Transl Psychiatry* 2017; **7**: e1207.
25. Gibbons R. D., Perraillon M. C., Kim J. B. Item response theory approaches to harmonization and research synthesis. *Health Serv Outcomes Res Methodol* 2014; **14**: 213–31.
26. Gibbons R. D. Computerized adaptive diagnosis and testing of mental health disorders. *Annu Rev Clin Psychol* 2016; **12**: 83–104.
27. Kessler F., Cacciola J., Alterman A., Faller S., Souza-Formigoni M. L., Cruz M. S., *et al*. Psychometric properties of the sixth version of the addiction severity index (ASI-6) in *Brazil. Rev Bras Psiquiatr* 2012; **34**: 24–33.
28. Kirisci L., Tarter R. E., Reynolds M., Vanyukov M. M. Item response theory analysis to assess dimensionality of substance use disorder abuse and dependence symptoms. *Int J Pers Cent Med* 2016; **6**: 260–73.
29. Butler S. F., Budman S. H., Goldman R. J., Newman F. J., Beckley K. E., Trottier D., *et al*. Initial validation of a computer-administered addiction severity index: the ASI-MV. *Psychol Addict Behav* 2001; **15**: 4–12.
30. Group WA The alcohol, smoking and substance involvement screening test (ASSIST): development, reliability and feasibility. *Addiction* 2002; **97**: 1183–94.
31. Langenbucher J. W., Labouvie E., Martin C. S., Sanjuan P. M., Bavly L., Kirisci L., *et al*. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *J Abnorm Psychol* 2004; **113**: 72–80.
32. Falgas I., Ramoz Z., Herrera L., Qureshi A., Chavez L., Bonal C., *et al*. Barriers to and correlates of retention in behavioral health treatment among Latinos in 2 different host countries: the United States and Spain. *J Public Health Manag Pract* 2017; **23**: e20–e27.
33. Gibbons R. D., Hedeker D. Full information item bi-factor analysis. *Psychometrika* 1992; **57**: 423–36.
34. Bock R. D., Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 1981; **46**: 443–59.
35. Gibbons R. D., Bock D., Hedeker D., Weiss D., Segawa E., Bhaumik D., *et al*. Full-information item bifactor analysis of graded response data. *Appl Psychol Measur* 2007; **31**: 4–19.
36. Gibbons R. D., Weiss D. J., Pilkonis P. A., Frank E., Moore T., Kim J. B., *et al*. Development of a computerized adaptive test for depression. *Arch Gen Psychiatry* 2012; **69**: 1104–12.
37. Rolim C. O., Koch F. L., Westphall C. B., Werner J., Fracalossi A., Salvador G. S. A cloud computing solution for patient's data collection in health care institutions. In eHealth, Telemedicine, and Social Medicine, 2010. ETELEMED'10. Second International Conference on IEEE, 2010, pp. 95–99.
38. Kuo A. M. H. Opportunities and challenges of cloud computing to improve health care services. *J Med Internet Res* 2011; **13**: e67.
39. Stringer K. L., Baker E. H. Stigma as a barrier to substance abuse treatment among those with unmet need: an analysis of parenthood and marital status. *J Fam Issues* 2018; **39**: 3–27.
40. Harnish A., Corrigan P., Byrne T. H., Pinals D., Rodrigues S., Smelson D. A comparison of substance use stigma and health stigma in a population of veterans with co-occurring mental health and substance use disorders. *J Dual Diagn* 2016; **11**: 238–43.
41. Beiser D., Vu M., Gibbons R. D. Test–retest reliability of a computerized adaptive depression test. *Psychiatr Serv* 2016; **67**: 1039–41.
42. 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, NJ: Lawrence Erlbaum Associates, Inc.; 1993, pp. 67–100.
43. Cai L., Yang J. S., Hansen M. Generalized full-information item bifactor analysis. *Psychol Methods* 2011; **16**: 221–48.
44. Beiser D. J., Ward C. E., Vu M., Laiteerapong N., Gibbons R. D. Depression in Emergency Department Patients and

Association with Healthcare Utilization. *Acad Emerg Med* 2019; **26**: 878–88.

45. Graham A. K., Minc A., Staab E., Beiser D. G., Gibbons R. D., Laiteerapong N. Validation of a computerized adaptive test for mental health in primary care. *Ann Fam Med* 2019; **17**: 23–0.
46. Achtyes E. D., Halstead S., Smart L., Moore T., Frank E., Kupfer D., *et al.* Validation of computerized adaptive testing in an outpatient non-academic setting. *Psychiatr Serv* 2015; **66**: 1091–6.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Supporting information.