

# Development and Validation of a Computerized Adaptive Assessment Tool for Discrimination and Measurement of Psychotic Symptoms

Daniel Guinart<sup>\*,1-3,6</sup>, Renato de Filippis<sup>1,4,6</sup>, Stella Rosson<sup>1,5</sup>, Bhagyashree Patil<sup>1</sup>, Lara Prizgint<sup>1</sup>, Nahal Talasazan<sup>1</sup>, Herbert Meltzer<sup>6</sup>, John M. Kane<sup>1-3</sup>, and Robert D. Gibbons<sup>7</sup>

<sup>1</sup>Department of Psychiatry Research, The Zucker Hillside Hospital, New York, NY; <sup>2</sup>Center for Psychiatric Neuroscience, Feinstein Institutes for Medical Research, Manhasset, NY; <sup>3</sup>Department of Psychiatry, The Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Manhasset, NY; <sup>4</sup>Psychiatric Unit, Department of Health Sciences, University Magna Graecia of Catanzaro, Catanzaro, Italy; <sup>5</sup>Department of Neuroscience, University of Padova, Padova, Italy; <sup>6</sup>Department of Psychiatry, Northwestern University, Chicago, IL; <sup>7</sup>Departments of Medicine, Public Health Sciences (Biostatistics), Psychiatry, Comparative Human Development, and the Committee on Quantitative Methods, Center for Health Statistics, University of Chicago, Chicago, IL

\*To whom correspondence should be addressed; 7559 263rd St, Glen Oaks, 11004 NY, USA; tel: 718-47-4139, fax: 718-343-1659, e-mail: [dguinart@northwell.edu](mailto:dguinart@northwell.edu)

## Abstract

**Objective:** Time constraints limit the use of measurement-based approaches in research and routine clinical management of psychosis. Computerized adaptive testing (CAT) can reduce administration time, thus increasing measurement efficiency. This study aimed to develop and test the capacity of the CAT-Psychosis battery, both self-administered and rater-administered, to measure the severity of psychotic symptoms and discriminate psychosis from healthy controls. **Methods:** An item bank was developed and calibrated. Two raters administered CAT-Psychosis for interrater reliability (IRR). Subjects rated themselves and were retested within 7 days for test-retest reliability. The Brief Psychiatric Rating Scale (BPRS) was administered for convergent validity and chart diagnosis, and the Structured Clinical Interview (SCID) was used to test psychosis discriminant validity. **Results:** Development and calibration study included 649 psychotic patients. Simulations revealed a correlation of  $r = .92$  with the total 73-item bank score, using an average of 12 items. Validation study included 160 additional patients and 40 healthy controls. CAT-Psychosis showed convergent validity (clinician:  $r = 0.690$ ; 95% confidence interval [95% CI]: 0.610–0.757; self-report:  $r = .690$ ; 95% CI: 0.609–0.756), IRR (intraclass correlation coefficient [ICC] = 0.733; 95% CI: 0.611–0.828), and test-retest reliability (clinician ICC = 0.862; 95% CI: 0.767–0.922; self-report ICC = 0.815; 95% CI: 0.741–0.871). CAT-Psychosis could discriminate psychosis from healthy controls (clinician: area under the receiver operating characteristic curve [AUC] = 0.965, 95% CI: 0.945–0.984; self-report AUC = 0.850, 95% CI: 0.807–0.894). The

median length of the clinician-administered assessment was 5 minutes (interquartile range [IQR]: 3:23–8:29 min) and 1 minute, 20 seconds (IQR: 0:57–2:09 min) for the self-report. **Conclusion:** CAT-Psychosis can quickly and reliably assess the severity of psychosis and discriminate psychotic patients from healthy controls, creating an opportunity for frequent remote assessment and patient/population-level follow-up.

**Key words:** schizophrenia/schizoaffective disorder/psychosis/bipolar disorder/testing

## Introduction

Psychotic disorders rank among the most debilitating disorders in all areas of medicine.<sup>1</sup> Clinical outcomes for people with schizophrenia and other psychotic disorders leave much to be desired,<sup>2</sup> due to a complex and multifactorial etiology<sup>3</sup> and pathophysiology.<sup>4</sup> Furthermore, unlike other areas of medicine, psychiatry is highly dependent on patient/family reports to assess the presence and severity of disease.<sup>5</sup> Not only the experience of a subjective event such as psychopathological phenomena but also the recollection, report, and interpretation by the rater of such an experience can be subject to multiple biases.<sup>6</sup> Hence, achieving further objectification has become a challenge and a goal for evidence-based mental health research. While measurement-based approaches are frequently used in schizophrenia research, they are not widely used in the routine clinical management. Time constraints and lack of familiarity with and/or training in validated

assessment tools limit their frequent clinical use.<sup>7</sup> Unfortunately, the measurement of psychopathology does not utilize many recent technological innovations that could help increase its clinical application. So far, the presence or severity of symptoms on a given assessment tool is determined by a total score, which requires that the same items be administered to all respondents.<sup>8</sup> One alternative to traditional assessments is adaptive testing, in which a person's initial item responses are used to determine a provisional estimate of his or her standing on the measured trait, which is then used for the selection of subsequent items,<sup>8</sup> allowing a precise measurement while only requiring the administration of a small subset of questions targeted to the person's specific severity or impairment level.<sup>9</sup> When computational algorithms automatically match questionnaire takers with the most relevant questions for them, this is called computerized adaptive testing (CAT). CAT relies on item response theory (IRT),<sup>10</sup> which models the relationship between a patient's responses to a series of items in terms of one or more latent variables that the test was designed to measure.<sup>11</sup> Hence, CAT automatically selects the items which are the most informative for the candidate taking the assessment, increasing measurement precision and efficiency, thus allowing assessments to be considerably shorter than traditional, fixed-length assessments.

Evidence suggests that one can create item banks with a large item pool, a small relevant set of which can be administered for a given individual with no or minimal loss of information, yielding a dramatic reduction in patient and clinician burden while maintaining high correlation with well-established symptom severity assessments, as well as high sensitivity and specificity for discriminating patients from healthy individuals.<sup>11–15</sup> CAT tools can be, therefore, used as dimensional symptom severity measures to inform stepped-based care and provide measurement-based care assessments not only in the clinic and remotely but also potentially as first-stage symptom screeners.<sup>16</sup> However, such an assessment tool is currently lacking in psychosis. Thus, this study aimed to develop and calibrate a CAT-Psychosis assessment battery, both self-administered and rater-administered versions, and test its psychometric properties and capacity to measure the severity of psychotic symptoms as well as to discriminate patients from healthy controls.

## Methods

### *Development*

We developed the CAT-Psychosis scale using the general methodology introduced by Gibbons and coworkers.<sup>12</sup> This methodology has 5 stages: (1) item bank development, (2) calibration of the multidimensional IRT model using complete data from a sample of subjects, (3) simulated CAT from the complete item response patterns, (4)

development of the live-CAT testing program, and (5) validation. In this section, we describe the first 4 steps used in this process.

*Item Bank Development.* The item bank was constructed from clinician-rated items on the Schedule for Affective Disorders and Schizophrenia (SADS),<sup>17</sup> the Scale for the Assessment of Positive Symptoms (SAPS),<sup>18</sup> the Brief Psychiatry Rating Scale (BPRS),<sup>19</sup> and the Scale for the Assessment of Negative Symptoms (SANS).<sup>20</sup> Items were classified into the positive, negative, disorganized, and manic subdomains for analysis using the bifactor IRT model.<sup>21,22</sup> This yielded 144 items for which existing clinician ratings were available for 649 subjects (535 schizophrenia, 43 schizoaffective, 54 depression, and 17 mania) drawn from an independent sample from inpatient and outpatient units treating psychotic disorders at Case Western Reserve University. Items were reworded to make them appropriate for patient self-report in addition to clinician administration.

*Calibration.* A full-information item bifactor model<sup>21,22</sup> was fitted to the 144 items from the 649 subjects. The bifactor model is the first confirmatory item factor analysis model and one of the first examples of a multidimensional IRT (MIRT) model. All items load on the primary dimension (which is the focus of the adaptive test) and one subdomain (the subdomain from which the item was drawn, which is determined in advance by clinicians). If the choice is not obvious, then different models can be fitted to the data using different subdomain structures and model selection criteria (Bayesian Information Criterion<sup>23</sup>) can be used to select the best-fitting structure. Gibbons and Hedeker<sup>21</sup> developed the bifactor model for binary item response data, and Gibbons<sup>22</sup> generalized it for ordinal item response data. The bifactor model provides a parameter related to each item's ability to discriminate high and low levels of the underlying primary and secondary latent variables and severity parameters for the  $k-1$  thresholds between the  $k$  ordinal response categories. The bifactor model produces a score and uncertainty estimate on the primary dimension for each subject, that is informed by items drawn from each of the subdomains. Items with loadings on the primary dimension of  $<0.3$  were removed due to lack of discrimination.

*Computerized Adaptive Testing.* Once we have calibrated the entire bank of items, we have estimates of each item's associated severity and we can adaptively match the severity of the items to the severity of the person. We do not know the severity of the person in advance of testing, but we learn it as we adaptively administer items. Beginning with an item in the middle of the severity distribution, we administer the item, obtain a categorical response, estimate the person's severity and the uncertainty in that estimate, and select

the next maximally informative item.<sup>12</sup> This process continues until the uncertainty falls below a predefined threshold, in our case, 5 points on a 100-point scale. The CAT has several tuning parameters,<sup>12</sup> which we select by simulating CAT from the complete response patterns; 1200 simulations are conducted, and we select the tuning parameters that minimize the number of items administered and maximize the correlation with the total item bank score.

*Development of the Live-CAT Testing Program.* Once the item bank has been calibrated and the CAT tuning parameters have been selected, a graphical user interface is developed and the CAT is added to a library of CATs, the CAT-Mental Health<sup>24</sup> in a cloud computing environment for a routine test administration on internet-capable devices, such as smartphones, tablets, notebooks, and computers. To accommodate literacy issues, audio is added to the self-report questions. Audio can be muted anytime during the administration. For the clinician version, questions are drawn from the same item bank, but the wording is adapted for a third person to rate. Additionally, we developed an indexed semi-structured script containing suggested optional wording to help inquire about the most common symptoms ([supplementary material 1](#)). Due to the adaptive nature of the test, the script was not designed as a structured guideline to follow sequentially, but rather as a tool to assist the rater in gathering information to rate the most common symptoms.

### Validation

After the assessment tool was developed and calibrated, we designed a validation study in an independent sample. To be considered for inclusion in the validation study, patients had to be English-speaking, 18–80 years of age, and with a diagnosis of schizophrenia, schizoaffective disorder, schizophreniform disorder, delusional disorder or brief psychotic disorder, as well as bipolar disorder, major depressive disorder, and healthy controls. Patients were recruited from inpatient and outpatient units from The Zucker Hillside Hospital, a large, private, non-for-profit semi-urban, psychiatric hospital providing both tertiary and routine care services that draws a representative racial/ethnic and sociodemographic mixture of eligible patients. Healthy controls were referred from other active research studies in our department. This investigation was carried out in accordance with the Declaration of Helsinki,<sup>25</sup> and all participants provided written informed consent after reviewing of the procedures as approved by the local Institutional Review Board (IRB#180626). Participants who did not understand English or were unable to sign informed consent, as well as patients with an active substance use disorder, structural brain disease, or any serious, acute, or chronic

medical illnesses that could pose a danger to self and others and could interfere with the patient's ability to comply with the study procedures, were excluded.

Basic demographic information, family history, current treatment, and diagnoses assigned by the current clinical treatment team were obtained from the patient and from the medical record. The CAT-Psychosis battery was administered using tablet computers with touch screens. Patients rated themselves with the self-administered CAT-Psychosis. Research staff were available to assist the participants if they had difficulty answering the questions.

### Psychometric Properties

Convergent validity of CAT-Psychosis, both for self and rater-administered versions, were tested comparing severity scores obtained with the CAT-Psychosis battery to the Brief Psychosis Rating Scale in its anchored version (BPRS-A),<sup>26</sup> a rater-administered 18-item scale ranging from 1 (“absent”) to 7 (“very severe”). Assessment order (BPRS first vs CAT-Psychosis first) was assigned in an alternating manner. To assess inter-rater reliability (IRR), subjects were requested to be consecutively but independently assessed by 2 raters using the clinician version of the CAT-psychosis. To assess test-retest reliability, both self-administered and rater-administered versions of CAT-Psychosis were administered again within 1–7 days of the initial administration. Psychosis discriminant validity was tested against chart diagnosis and the Structured Clinical Interview (SCID),<sup>27</sup> conducted by trained departmental raters at the time of the CAT-Psychosis assessment, and validated in diagnostic consensus conferences. SCID was not readministered if it had been conducted during the previous 12 months.

### Statistics

The bifactor IRT models were fitted with the POLYBIF program that is freely available at [www.healthstats.org](http://www.healthstats.org). Generalized linear mixed-effects regression models and Pearson product-moment correlation coefficients were used to examine associations between continuous measures (eg, CAT-Psychosis severity scores and the BPRS). Intraclass correlation coefficients (ICCs) computed using linear mixed-models were used to test for inter-rater and test-retest reliability.<sup>28</sup> Mixed-effects logistic regression was used to estimate psychosis discrimination capacity (lifetime ratings) and to estimate sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve with 10-fold cross-validation. We also tested the extent to which gender modified psychosis discrimination capacity. With a sample size of 160 schizophrenia spectrum and mood disorder patients with psychotic symptoms, we expected to have more than 80% power (2-tailed  $\alpha = .05$ ) to determine a Pearson's  $r$  of at least 0.22 for continuous scores and an AUC with 95% confidence interval (CI) with a width of 10% assuming a null

hypothesis value of  $AUC = 0.8$  and power of 80%. The data were analyzed with SuperMix Version 2.1, Stata Version 15.1, and JMP software v.14 (JMP, Version 14, 1989–2007).  $P$ -values lower than .05 were considered statistically significant.

## Results

### Development

Data from 649 subjects were used to calibrate the 144 items. Following the removal of items with poor discrimination on the primary dimension, 73 items remained (supplementary material 2). The bifactor model significantly improved the fit over a unidimensional IRT alternative (chi-square = 2204,  $df = 73$ ,  $P < .00001$ ). Simulated adaptive testing from the clinician ratings ( $n = 420$  patients with <10% missing data) revealed that using an average of 12 items per subject (range 3–21 items), we maintained a correlation of  $r = .92$  with the 73-item total bank score.

### Validation

In total, 200 subjects were enrolled in the validation study and completed the assessment. Of those, 160 subjects were affected by a affective and non-affective psychoses ( $n = 76$  schizophrenia,  $n = 35$  schizoaffective disorder,  $n = 26$  bipolar disorder,  $n = 11$  major depressive disorder,  $n = 7$  psychosis not otherwise specified, and  $n = 5$  schizophreniform disorder) and 40 were healthy controls. Consolidated Standards of Reporting Trials (CONSORT)<sup>29</sup> flow diagram is available in figure 1. Sociodemographic characteristics of the sample are described in table 1.

### Psychometric Properties

CAT-Psychosis showed convergent validity against total BPRS scores for both clinician-administered ( $r = 0.690$ ; 95% CI: 0.610–0.757; marginal maximum likelihood Estimate [MMLE] = 0.057, standard error [SE] = 0.004,  $P < .00001$ ) and self-administered ( $r = .690$ ; 95% CI: 0.609–0.756; MMLE = 0.061, SE = 0.004,  $P < .00001$ ) versions (figure 2). We also found that CAT-Psychosis clinician version and CAT-Psychosis self-rated version showed a significant positive correlation ( $r = .594$ , 95% CI: 0.495, 0.667). CAT-Psychosis clinician version's IRR was strong (ICC = 0.733; 95% CI: 0.611–0.828), and test-retest reliability was strong for both self-report (ICC = 0.815; 95% CI: 0.741–0.871) and clinician (ICC = 0.862; 95% CI: 0.767–0.922) versions.

### Psychosis Discrimination

The CAT-Psychosis clinician version was able to successfully discriminate psychosis vs healthy controls (area under the ROC curve [AUC] = 0.965, 95% CI:

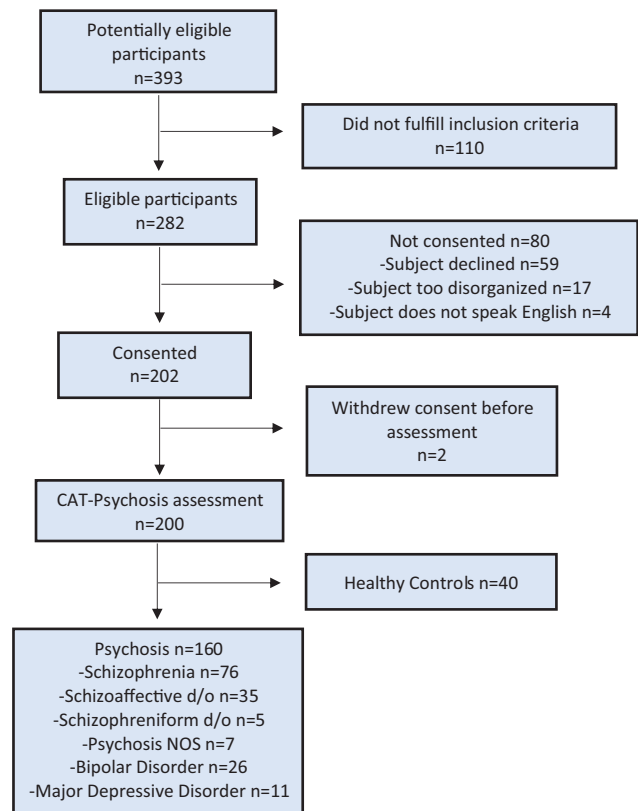


Fig. 1. CONSORT diagram.

0.945–0.984; MMLE = 1.42, SE = 0.12,  $P < .00001$ ), with an effect size (total variance) of 2.23 SD units. The CAT-Psychosis self-report version yielded similar results (AUC = 0.850, 95% CI: 0.807–0.894; MMLE = 1.28, SE = 0.16,  $P < .00001$ ), with an effect size (total variance) of 1.60 SD units (figure 3). We conducted an additional analysis including only patients with a SCID ( $n = 79$ ), showing similar results despite the smaller sample size. CAT-Psychosis clinician version was able to successfully discriminate psychosis vs healthy controls (AUC = 0.968, 95% CI: 0.943–0.993; MMLE = 1.33, SE = 0.103,  $P < .00001$ ), with an effect size (total variance) of 2.522 SD units. The CAT-Psychosis self-report version yielded similar results (AUC = 0.796, 95% CI: 0.730–0.862; MMLE = 1.125, SE = 0.157,  $P < .00001$ ), with an effect size (total variance) of 1.535 SD units. Males and females showed no significant differences (CAT-Psychosis by gender interaction: clinician Odds Ratio (OR) = 1.43, 95% CI: 0.52, 3.94,  $P < .48$ ; self-report OR = 1.11, 95% CI: 0.02, 55.65,  $P = .93$ ).

### Testing Times and Number of Items

The median length of assessment was 5 minutes and 2 seconds (interquartile range [IQR]: 3:23–8:29 min) for the clinician-administered version. Since we used a script that includes a brief introduction (supplementary material 1), an additional 3–8 minutes were used before the

**Table 1.** Sociodemographic Characteristics of the Validation Sample

	Total ( <i>n</i> = 200)	Psychosis ( <i>n</i> = 160)	Healthy Controls ( <i>n</i> = 40)	<i>P</i> -value
Age; median (Q1, Q3)	30 (25;42.8)	30 (24;42)	31.5 (27,45.8)	.1908
Sex				.0458
Male, <i>n</i> (%)	113 (56.5)	96 (60.0)	17 (42.5)	
Female, <i>n</i> (%)	87 (43.5)	64 (40.0)	23 (57.5)	
Race				.5484
White, <i>n</i> (%)	62 (31.0)	48 (30.0)	14 (35.0)	
Black/African American, <i>n</i> (%)	71 (35.5)	60 (37.5)	11 (27.5)	
Asian, <i>n</i> (%)	34 (17.0)	25 (15.6)	9 (22.5)	
Mixed/Other, <i>n</i> (%)	33 (16.5)	27 (16.9)	6 (15.0)	
Marital status				.0610
Single, <i>n</i> (%)	152 (76.0)	123 (76.9)	29 (72.5)	
Married, <i>n</i> (%)	23 (11.5)	15 (9.4)	8 (20.0)	
Divorced/separated, <i>n</i> (%)	8 (4.0)	5 (3.1)	3 (7.5)	
Widowed, <i>n</i> (%)	3 (1.5)	3 (1.9)	0 (0)	
Unknown/unreported, <i>n</i> (%)	14 (7.0)	14 (8.8)	0 (0)	
Adopted				.1658
Yes, <i>n</i> (%)	8 (4.0)	7 (4.4)	1 (2.5)	
No, <i>n</i> (%)	180 (90.0)	141 (88.1)	39 (97.5)	
Unknown/unreported, <i>n</i> (%)	12 (6.0)	12 (7.5)	0 (0)	
English as a primary language				.1639
Yes, <i>n</i> (%)	161 (80.5)	128 (80.0)	33 (82.5)	
No, <i>n</i> (%)	27 (13.5)	20 (12.5)	7 (17.5)	
Unknown/unreported, <i>n</i> (%)	12 (6.0)	12 (7.5)	0 (0)	
Education level				<.0001
Less than high school, <i>n</i> (%)	15 (7.5)	15 (9.4)	0 (0)	
High school diploma, <i>n</i> (%)	37 (18.5)	34 (21.3)	3 (7.5)	
Some college/Associate's, <i>n</i> (%)	50 (25.0)	41 (25.6)	9 (22.5)	
College graduate, <i>n</i> (%)	49 (24.5)	40 (25.0)	9 (22.5)	
Doctorate/PhD, <i>n</i> (%)	11 (5.5)	10 (25.0)	1 (0.6)	
Unknown/unreported, <i>n</i> (%)	21 (10.5)	21 (13.1)	0 (0)	
Current occupation				<.0001
Homemaker/unemployed, <i>n</i> (%)	97 (48.5)	92 (57.5)	5 (12.5)	
Unskilled employment, <i>n</i> (%)	53 (26.5)	40 (25.0)	13 (32.5)	
Skilled employment, <i>n</i> (%)	26 (13.0)	5 (3.1)	21 (52.5)	
Unknown/unreported, <i>n</i> (%)	24 (12.0)	23 (14.4)	1 (2.5)	

Note: *N*, number; %, percentage; Q1, first quartile; Q3, third quartile.

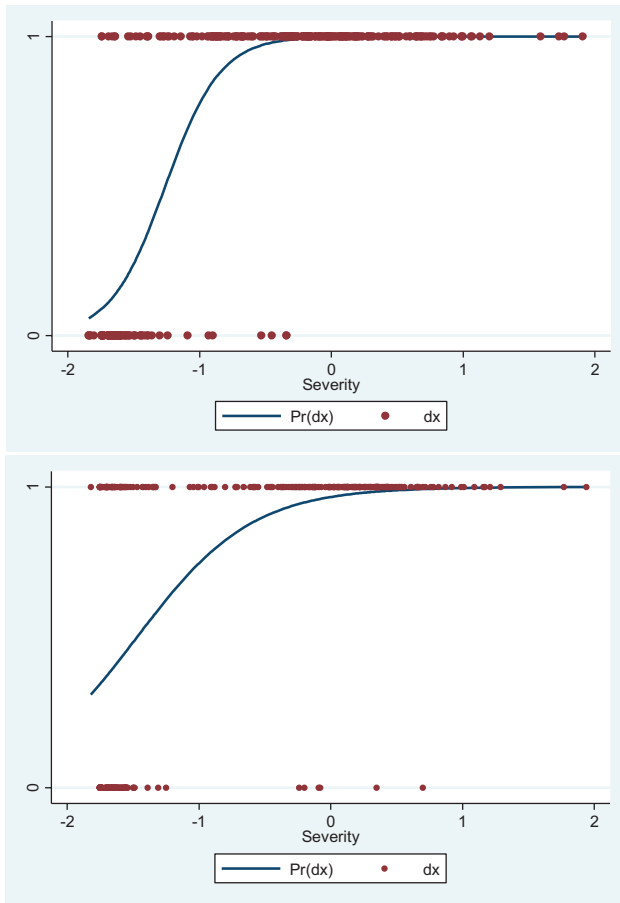
computerized assessment started. The median length of assessment of the self-administered CAT-Psychosis was 1 minute and 20 seconds (IQR: 0:57–2:09 min). The median number of questions needed to elicit a valid severity rating using the clinician-administered version was 12 (IQR: 8–15 items), same as for the patient self-report version (IQR: 8–13). Table 2 presents an example of a CAT-Psychosis self-assessment in a patient with severe psychotic symptoms and an example of a CAT-Psychosis rater-administered assessment in a patient with moderate psychotic symptoms.

## Discussion

We demonstrate that the CAT-Psychosis battery, not only in its rater-assessed but also, notably, in its self-report version, can accurately measure the severity of psychosis in less than 2 minutes, in the case of the self-administered version, or 5 minutes, in the case of the rater-assessed

version. Further, CAT-Psychosis can discriminate psychotic patients from healthy controls.

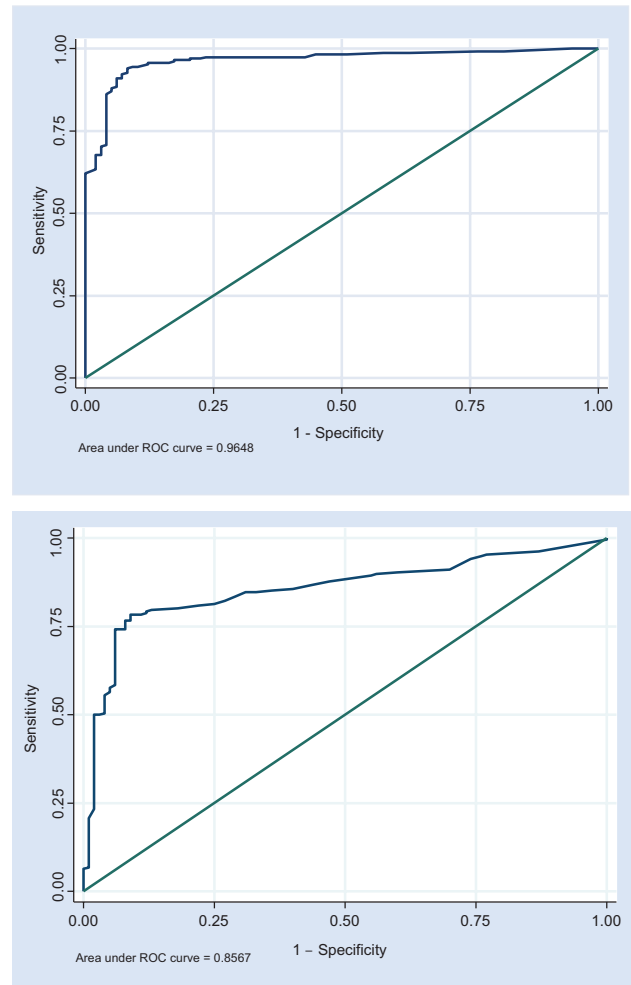
A fundamental advantage of the rater version of the CAT-Psychosis battery is the rapid administration time. Since questions are tailored to every patient and the test is dynamically modified based on the answers provided, the time of administration is dramatically reduced. To our knowledge, there are currently no available assessment tools that can reliably allow a rater to measure the severity of psychotic symptoms in such a brief administration time. Some tools, such as the 4-item Negative Symptom Assessment (NSA-4),<sup>30</sup> a brief version of the NSA-16, a validated tool for evaluating negative symptoms in schizophrenia,<sup>31</sup> as well as the Brief Negative Symptom Scale (BNSS)<sup>32</sup> lasting about 15 minutes, have proven to be quick and reliable, but they are limited to negative symptoms. When positive symptoms are included in the assessment, administration time and rater and patient burden increase substantially. The most



**Fig. 2.** Predicted probability of diagnosis based on severity score for clinician-assessed (above) and self-assessed (below) CAT-Psychosis. Pr(dx), probability of a positive diagnosis; dx, Diagnosis of psychosis.

commonly used rater-based assessment tools, such as the Positive and Negative Syndrome Scale (PANSS), the SAPS, or the BPRS, require a variable but significant amount of time to complete, ranging from 20 to 50 minutes.<sup>18,19,33,34</sup> Recently, a valid and scalable 6-item version of the PANSS was developed,<sup>35</sup> but it appears to require the guidance of the Simplified Negative and Positive Symptom Interview (SNAPSI) a 15-to-25-minute structured interview<sup>36</sup> that, despite significantly optimizing the 30-item PANSS, may still not be rapid enough for routine clinical use. CAT-Psychosis rater-assessed version, however, generates a quick assessment that could facilitate routine screening and measurement of psychotic symptoms and whose potential integration with electronic health records would allow its widespread use in mainstream clinical care, thus generating reliable, real-world data at a scale that is currently not imaginable, while aiming to keep clinician burden to a minimum.

Unlike in other areas of mental health, self-report measures in psychosis are rare, and their use in normal clinical care is unusual,<sup>37</sup> probably due to concerns that self-report questionnaires may not be appropriate for



**Fig. 3.** Receiver operating characteristic (ROC) curve for clinician-assessed (above) and self-assessed (below) CAT-Psychosis, with estimated area under the curve. The 45-degree line represents an chance association between the continuous test score and the binary psychosis diagnosis.

evaluating psychotic disorders because of a risk of minimization and/or denial of symptoms due to stigma and/or impaired awareness of the illness.<sup>38,39</sup> However, the advantages of self-assessments over rater-based evaluations could be substantial, as assessments are not limited to the availability of highly trained interviewers, eliminating interviewer bias and reducing costs, thus enhancing scalability.

In a context of greater availability of digital tools and devices that allow collecting information from patients remotely and repeatedly, the need for valid, rapid, reliable, and easy-to-administer self-report tools for psychosis is greater than ever. Some self-reports for psychosis have been made available,<sup>40-43</sup> either circumscribed to a specific area of evaluation, such as hallucinations,<sup>40</sup> apathy,<sup>41</sup> and negative symptoms,<sup>42</sup> or smartphone based,<sup>43</sup> all of which rely on traditional scoring systems. Unfortunately, the repeated administration of fixed-length, identical tests in short time spans may lead to response bias, limiting the

**Table 2.** Examples of CAT-Psychosis Administrations

Questions	Replies	Severity Score	Precision
<b>(A) CAT-Psychosis self-assessment</b>			
I feel that I am a particularly important person or that I have special powers or abilities.	Not at all	41.5	14.7
I feel that I have committed a crime or have done some terrible things and deserve punishment.	Sometimes	53.5	12.3
I feel emotionally withdrawn, lack spontaneity, and am isolated from others.	Rarely	52.2	12.0
I have trouble communicating with others.	Always	81.2	7.2
I am having trouble concentrating on this interview.	Not at all	77.8	6.8
I am confused, disconnected, disorganized, and/or disrupted.	Rarely	73.6	6.2
I say things that are unrelated to each other, for example, “I’m tired. All people have eyes.”	Sometimes	73.4	5.7
When I speak, my ideas slip off track into unrelated topics.	Always	79.4	4.9
<b>(B) CAT-Psychosis rater-assessment</b>			
Claims power, knowledge or identity beyond the bounds of credibility.	Absent	41.5	14.7
Severity of delusions of any type—consider conviction in delusion, preoccupation, and effect on her actions.	Mild	39.3	12.4
Lack of spontaneous interaction, isolation, deficiency in relating to others.	Moderate	43.0	12.0
Extent to which the patient’s ability to communicate is affected.	Mild	53.2	6.4
This rating should assess the patient’s overall lack of concentration, clinically and on tests.	Questionable	51.1	6.5
Thought processes confused, disconnected, disorganized, disrupted.	Mild	51.1	6.5
Repeatedly saying things in juxtaposition, which lack a readily understandable relationship (eg, “I’m tired; all people have eyes”).	Mild	53.2	6.1
The patient appears uninvolved or unengaged. She may seem “spacey.”	Moderate	55.4	5.6
A pattern of speech in which conclusions are reached that do not follow logically.	Mild	58.5	5.0

*Note:* (A) shows a self-administered test for a patient with severe psychotic symptoms (category = severe, severity = 79.4, and precision = 4.9). (B) shows a rater-administered test for a patient with moderate psychotic symptoms (category = moderate, severity = 58.5, and precision = 5.0). In both cases, the CAT terminated when the uncertainty was at or below 5 points on the 100-point scale.

ability of such a test to be used regularly in longitudinal studies. CAT-Psychosis, however, produces a tailored set of questions at each administration, allowing for frequent and even remote longitudinal monitoring of symptom severity, which could have very relevant implications in the field of mental health research, shortening assessments significantly, minimizing patient burden, and allowing for more frequent evaluations. More importantly, the availability of a valid self-assessment tool that can be used to quickly measure the severity of symptoms not only in the clinic but also remotely can assist in determining the effectiveness of interventions or detecting clinical worsening. Bridging the current information blackout between clinical visits with frequent adaptive tests that can be administered electronically could help bring significant advances to routine clinical care. Nonetheless, the optimization of measurement-based care aims to assist clinicians and does not call into question the need for trained psychologists and psychiatrists.

This study has some limitations. First, the CAT-Psychosis battery provides the investigator with an overall severity score, including all symptom domains, and is not designed to monitor specific individual symptoms (eg, auditory hallucinations), to the extent that they may not always be adaptively administered. While it is possible to develop an adaptive test to score the primary

dimension and each of the subdomains, this would increase the number of items adaptively administered and add to the subject burden. In this study, our objective is to adaptively score the primary severity dimension, to which all of the subdomains contribute. In future work, we will explore the adaptive scoring of subdomain scores. Second, the validation sample size did not allow for specific diagnostic predictions within all the psychoses that we included as well as the detailed study of the influence of sociodemographic factors. Future studies, including larger samples, will address these issues. Third, this study was conducted exclusively in English. Independent replication of our findings in other patient populations, settings, devices, and languages is needed and will be the subject of further research. Once developed, the CAT-Psychosis scale can be tested for suitability in different populations, using differential item functioning (DIF).<sup>44</sup> As an example, CATs developed for depression, anxiety, mania/hypomania, suicidality, and substance use disorder have been studied for DIF in diverse Spanish-speaking populations,<sup>45</sup> perinatal women,<sup>46</sup> emergency department patients,<sup>47</sup> and in subjects within the criminal justice system.<sup>48</sup>

Fourth, the adoption of an exclusively digital self-assessment tool may be limited in non-digital native populations. However, while some studies have suggested that

younger patients would be more willing to participate in digital-based interventions,<sup>49</sup> others show opposite results,<sup>50</sup> and rates of access of schizophrenia patients to technology and smartphones have progressively merged with that of the general population.<sup>51</sup> In fact, in our validation sample,  $n = 32$  subjects were between the ages of 50 and 70 years old and no test was interrupted or incomplete.

Until recently, the use of a valid self-administered psychopathology assessment tool has been missing in schizophrenia and other psychoses research, partly because of lack of trust in the ability of patients to self-report psychotic symptoms. Recent literature, including our study, demonstrates that this is not the case. In fact, we show that self-reports of psychotic symptoms made by psychotic patients can lead to rapid, reliable, and valid measurement of psychotic symptoms, which can have relevant implications for research and clinical care.

## Conclusion

CAT-Psychosis, both rater-administered and as a self-report, provides valid psychosis severity scores and can also discriminate psychotic patients from healthy controls, requiring only a brief administration time.

## Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin*.

## Funding

None.

## Acknowledgments

Dr Guinart has been a consultant for and/or has received speaker honoraria from Otsuka America Pharmaceuticals and Janssen Pharmaceuticals. Dr Kane has been a consultant and/or advisor for or has received honoraria from Alkermes, Allergan, LB Pharmaceuticals, H. Lundbeck, Intracellular Therapies, Janssen Pharmaceuticals, Johnson and Johnson, Merck, Minerva, Neurocrine, Newron, Otsuka, Pierre Fabre, Reviva, Roche, Sumitomo Dainippon, Sunovion, Takeda, Teva, and UpToDate and is a shareholder in LB Pharmaceuticals and Vanguard Research Group. Dr Gibbons 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 in which CAT-Psychosis is included. Drs de Filippis, Rosson, and Patil, as well as Nahal Talasazan and Lara Prizgint, have nothing to disclose.

## References

- Salomon JA, Vos T, Hogan DR, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *Lancet*. 2012;380(9859):2129–2143.
- Fusar-Poli P, McGorry PD, Kane JM. Improving outcomes of first-episode psychosis: an overview. *World Psychiatry*. 2017;16(3):251–265.
- Radua J, Ramella-Cravaro V, Ioannidis JPA, et al. What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry*. 2018;17(1):49–66.
- Owen MJ, Sawa A, Mortensen PB. Schizophrenia. *Lancet*. 2016;388(10039):86–97.
- Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century. *Front Psychol*. 2016;7:1933.
- Guinart D, Kane JM, Correll CU. Is Transcultural psychiatry possible? *JAMA*. 2019. Epub ahead of print. doi:10.1001/jama.2019.17331.
- Correll CU, Kishimoto T, Nielsen J, Kane JM. Quantifying clinical relevance in the treatment of schizophrenia. *Clin Ther*. 2011;33(12):B16–B39.
- Gibbons RD, Weiss DJ, Kupfer DJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv*. 2008;59(4):361–368.
- Wainer H. CATs: whither and whence. *Psychologica*. 2000;21:121–133.
- Lord F, Novick M. *Statistical Theories of Mental Test Scores*. New York, NY: Information Age Publishing Inc; 1968.
- Gibbons RD, Kupfer D, Frank E, Moore T, Beiser DG, Boudreaux ED. Development of a computerized adaptive test suicide scale—the CAT-SS. *J Clin Psychiatry*. 2017;78(9):1376–1382.
- Gibbons RD, Weiss DJ, Pilkonis PA, et al. Development of a computerized adaptive test for depression. *Arch Gen Psychiatry*. 2012;69(11):1104–1112.
- Gibbons RD, Weiss DJ, Pilkonis PA, et al. Development of the CAT-ANX: a computerized adaptive test for anxiety. *Am J Psychiatry*. 2014;171(2):187–194.
- Gibbons RD, Kupfer DJ, Frank E, et al. Computerized adaptive tests for rapid and accurate assessment of psychopathology dimensions in youth. *J Am Acad Child Adolesc Psychiatry*. 2019;59(11):1264–1273.
- Gibbons RD, Alegria M, Markle S, et al. Development of a computerized adaptive substance use disorder scale for screening and measurement: the CAT-SUD. *Addiction*. 2020;115:1382–1394.
- Substance Abuse and Mental Health Services Administration (SAMHSA) MDPS National Survey. Individual grant awards. <https://www.samhsa.gov/grants/awards/2019/FG-19-003>. Accessed April 29, 2020.
- Endicott J, Spitzer RL. A diagnostic interview: the schedule for affective disorders and schizophrenia. *Arch Gen Psychiatry*. 1978;35(7):837–844.
- Andreasen NC. *The Scale for the Assessment of Positive Symptoms (SAPS)*. Iowa City, IA: Iowa University; 1984.
- Overall JE, Gorham DR. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. *Psychopharmacol Bull*. 1988;24(1):97–99.
- Andreasen NC. *The Scale for the Assessment of Negative Symptoms (SANS)*. Iowa City, IA: Iowa University Iowa; 1983.



21. Gibbons R, Hedeker D. Full-information item bifactor analysis. *Psychometrika* 1992;157(3):423–436.
22. Gibbons RD, Bock RD, Hedeker D, et al. Full-information item bifactor analysis of graded response data. *Appl Psychol Meas*. 2007;31(1):4–19.
23. Schwarz GE. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461–464.
24. Gibbons RD, Weiss DJ, Frank E, Kupfer D. Computerized adaptive diagnosis and testing of mental health disorders. *Annu Rev Clin Psychol*. 2016;12:83–104.
25. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191–2194.
26. Woerner MG, Mannuzza S, Kane JM. Anchoring the BPRS: an aid to improved reliability. *Psychopharmacol Bull*. 1988;24(1):112–117.
27. First MB, Williams JBW, Karg RS, Spitzer RL. *Structured Clinical Interview for DSM-5 --Research Version*. Arlington, VA: American Psychiatric Association; 2015.
28. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.
29. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.
30. Alphs L, Morlock R, Coon C, Cazorla P, Szegedi A, Panagides J. Validation of a 4-item Negative Symptom Assessment (NSA-4): a short, practical clinical tool for the assessment of negative symptoms in schizophrenia. *Int J Methods Psychiatr Res*. 2011;20(2):e31–e37.
31. Axelrod BN, Goldman RS, Alphs LD. Validation of the 16-item Negative Symptom Assessment. *J Psychiatr Res*. 1993;27(3):253–258.
32. Kirkpatrick B, Strauss GP, Nguyen L, et al. The Brief Negative Symptom Scale: psychometric properties. *Schizophr Bull*. 2011;37(2):300–305.
33. Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull*. 1987;13(2):261–276.
34. Kumari S, Malik M, Florival C, Manalai P, Sonje S. An assessment of five (PANSS, SAPS, SANS, NSA-16, CGI-SCH) commonly used symptoms rating scales in schizophrenia and comparison to newer scales (CAINS, BNSS). *J Addict Res Ther*. 2017;8(3):324. doi:10.4172/2155-6105.1000324.
35. Østergaard SD, Lemming OM, Mors O, Correll CU, Bech P. PANSS-6: a brief rating scale for the measurement of severity in schizophrenia. *Acta Psychiatr Scand*. 2016;133(6):436–444.
36. Østergaard SD, Opler MGA, Correll CU. Bridging the measurement gap between research and clinical care in schizophrenia: Positive and Negative Syndrome Scale-6 (PANSS-6) and other assessments based on the Simplified Negative and Positive Symptoms Interview (SNAPSI). *Innov Clin Neurosci*. 2017;14(11-12):68–72.
37. Rabinowitz J, Levine SZ, Medori R, Oosthuizen P, Koen L, Emsley R. Concordance of patient and clinical ratings of symptom severity and change of psychotic illness. *Schizophr Res*. 2008;100(1-3):359–360.
38. Doyle M, Flanagan S, Browne S, et al. Subjective and external assessments of quality of life in schizophrenia: relationship to insight. *Acta Psychiatr Scand*. 1999;99(6):466–472.
39. Selten JP, Wiersma D, van den Bosch RJ. Clinical predictors of discrepancy between self-ratings and examiner ratings for negative symptoms. *Compr Psychiatry*. 2000;41(3):191–196.
40. Kim SH, Jung HY, Hwang SS, et al. The usefulness of a self-report questionnaire measuring auditory verbal hallucinations. *Prog Neuropsychopharmacol Biol Psychiatry*. 2010;34(6):968–973.
41. Faerden A, Lyngstad SH, Simonsen C, et al. Reliability and validity of the self-report version of the apathy evaluation scale in first-episode psychosis: concordance with the clinical version at baseline and 12 months follow-up. *Psychiatry Res*. 2018;267:140–147.
42. Dollfus S, Mach C, Morello R. Self-evaluation of negative symptoms: a novel tool to assess negative symptoms. *Schizophr Bull*. 2016;42(3):571–578.
43. Palmier-Claus JE, Ainsworth J, Machin M, et al. The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. *BMC Psychiatry*. 2012;12:172.
44. Holland P, Wainer H. Differential item functioning. *Psicothema* 1995;7:237–241.
45. Gibbons RD, Alegria 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;30(10):1267–1276.
46. Kim JJ, Silver RK, Elue R, et al. The experience of depression, anxiety, and mania among perinatal women. *Arch Womens Ment Health*. 2016;19(5):883–890.
47. Beiser DG, Ward CE, Vu M, Laiteerapong N, Gibbons RD. Depression in emergency department patients and association with health care utilization. *Acad Emerg Med*. 2019;26(8):878–888.
48. Gibbons RD, Smith JD, Brown CH, et al. Improving the evaluation of adult mental disorders in the criminal justice system with computerized adaptive testing. *Psychiatr Serv*. 2019;70(11):1040–1043.
49. Gay K, Torous J, Joseph A, Pandya A, Duckworth K. Digital technology use among individuals with schizophrenia: results of an online survey. *JMIR Ment Health*. 2016;3(2):e15.
50. Ben-Zeev D, Scherer EA, Gottlieb JD, et al. mHealth for Schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. *JMIR Ment Health*. 2016;3(3):e34.
51. Gitlow L, Abdelaal F, Etienne A, Hensley J, Krukowski E, Toner M. Exploring the current usage and preferences for everyday technology among people with serious mental illnesses. *Occup Ther Ment Health*. 2017;33(1):1–14.