10.4414/sanp.2024.1314098815 | 2024;175(1):12–16 | **Swiss Arch Neurol Psychiatr Psychother**

Original Article | Peer-Reviewed Article

Paradox of Self-Stigma Scale

Shorter Yet Accurate Measurements with Computerized Adaptive Testing in Psychiatry

Philippe Golay[a,b,c], Debora Martinez[a,d], Charlotte Bonalumi[a], Benedetta Silva[a,e], Stéphane Morandi[a,e], Charles Bonsack[a]

[a] Community Psychiatry Service, Department of Psychiatry, Lausanne University Hospital and University of Lausanne, Switzerland; [b] Treatment and Early Intervention in Psychosis Program, General Psychiatry Service, Department of Psychiatry, Lausanne University Hospital and University of Lausanne, Switzerland; [c] Institute of Psychology, Faculty of Social and Political Science, University of Lausanne, Switzerland; [d] La Source School of Nursing Sciences, University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland; [e] Cantonal Medical Office, Directorate General for Health of the Canton of Vaud, Department of Health and Social Action, Lausanne, Switzerland.

Abstract

Background: Short, accurate, rapidly completed psychometric tests are in high demand in psychiatry. Time is always limited. Researchers often try not to overburden patients included in studies measuring many variables. However, shorter psychometric tests may demonstrate poorer reliability or lower sensitivity. Computerised adaptive testing (CAT) could be a solution to this important trade-off.

Aim: This study aimed to evaluate whether the Paradox of Self-Stigma scale (PaSS-24), a pen and paper test evaluating three dimensions of self-stigma, could be transformed into a computerised adaptive test to reduce the number of items administered.

Method: The PaSS-24 items were calibrated using the item response theory. A Monte-Carlo simulation was performed to evaluate the number of items needed to ensure the same test reliability as the initial scale. We simulated 50,000 participants with various levels of self-stigma.

Results: Results showed that two of PaSS-24's three subscales could be substantially shortened while maintaining similar reliability. The correlations between simulated and estimated scores were close to unity, indicating that the CAT did not sacrifice accuracy for brevity.

Conclusions: Simulated data suggested that shortened psychometric CAT could achieve similar reliability to the initial PaSS-24 despite the latter already being brief. CAT provides an opportunity to give in to the pressure of using short scales in psychiatry without compromising on reliability and accuracy.

Keywords: Psychiatric assessment; Item Response Theory; computerised adaptive testing; selfstigma; reliability

Background

Short, accurate, rapidly completed psychometric tests are in high demand in psychiatry. While researchers often try not to overburden their patients, they nevertheless want to measure as many variables as possible [1-3]. Time is not the only issue. Patients can be exposed to inadequate items (e.g., items that are too easy or too hard to respond to or items that refer to symptoms that are either too mild or too severe compared to the patient's condition). An experienced clinician will do much better than a psychometric test, because they use their expertise and patients' previous answers to select the optimal next question and they naturally avoid unnecessary questions [4]. Standardised pen and paper tests are not usually this flexible and are mostly administered linearly. Such assessments can tire, demotivate or even shock patients, which could eventually prejudice the assessment's validity [5].

Shortened tests, although highly desirable, may demonstrate poorer reliability and lower sensitivity. Indeed, in classical test theory, there is a monotonic relationship between the number of items and the reliability of the test's total score [6]. In research contexts, when conclusions are made at the interindividual level, this may still be acceptable. When comparing two large groups of individuals, for instance, random errors tend to cancel each other out. Using structural equation modelling it is possible, under certain conditions, to account for measurement errors and create latent variables with perfect reliability. However, because they are based on interindividual variations, drawing conclusions is only possible at the group level.

Mental health professionals face a different challenge in clinical contexts. They must make decisions based on one individual score and each score comes with a level of uncertainty: the standard error of measurement (SEM), which is usually represented using confidence intervals. Confidence interval ranges are inversely proportional to the test score's reliability. With a poor reliability test score the range of possible values for the true score can rapidly become too broad to enable reliable clinical decisions (e.g., a patient's IQ measured between 65 and 135). Thus, although a poor measurement scale may sometimes be sufficient to test different hypotheses at the group level, the same tool may be useless for making decisions at the individual or clinical level. In the worst-case scenario, inference at the group level will also be demonstrated in the form of reduced statistical power and an increased rate of type II errors.

Computerised adaptive testing (CAT) can provide an elegant solution to this important time versus accuracy trade-off. CAT is based on the item response theory (IRT), which at-

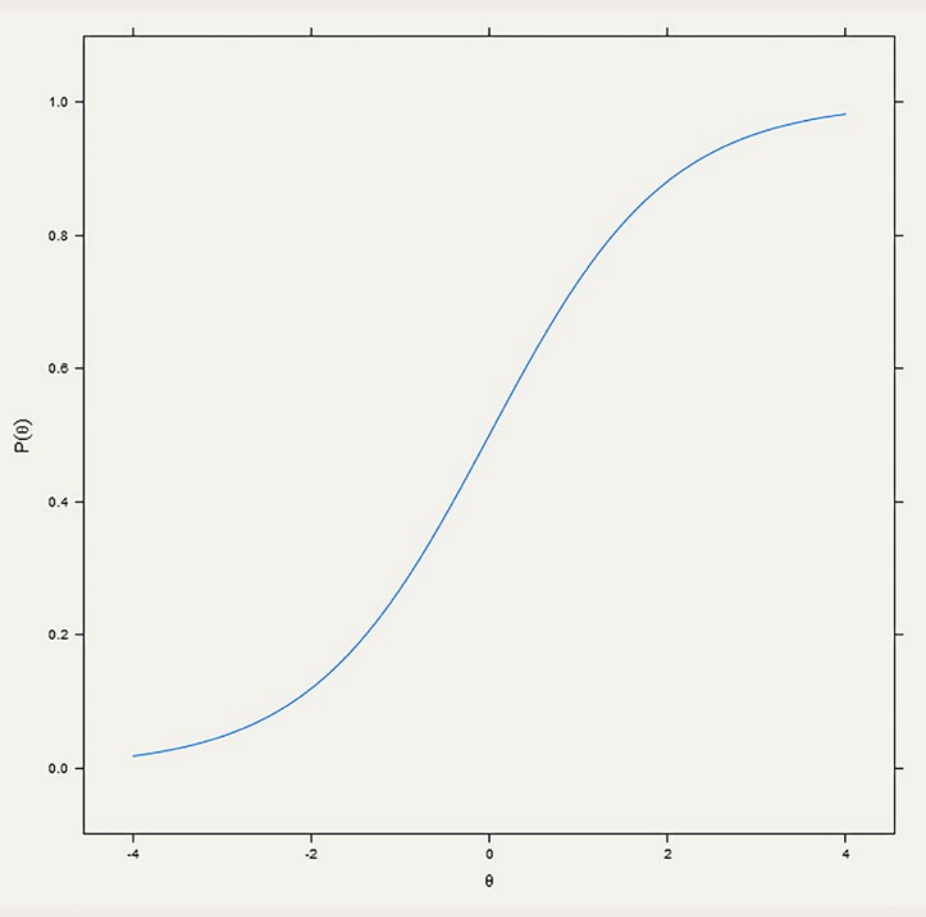

Figure 1: Item characteristic curve describing the relationship between the respondent's ability (theta) and the probability of a correct response.

tempts to describe or explain the relationships between psychological traits, that are not directly observable (latent variables) and patients' test responses (manifest variables). Item responses are considered the observable manifestation of these latent traits. An important characteristic of the IRT is, that individuals and items are located along the same unidimensional continuum [7]. IRT assumes, that the latent trait of interest is continuous. An item's role is to help differentiate between different respondents located at different places along this continuum. The patient's response reduces the uncertainty about their precise location along the continuum. The model describes the relationship between the ability to endorse or give the correct response to an item (or the position along the latent trait continuum, which we will subsequently call the theta) and the probability of doing so (fig. 1).

A great variety of item response models exists and a detailed description of them is beyond the present study's scope. The simplest IRT model is the one-parameter model (usually referred to as the Rasch model), which depicts the relationship between the theta and the response probabilities to items using a series of logit regression lines with identical slopes [7]. The slope is also called the discrimination parameter, whereas the locations of the slopes along the continuum refer to an item's difficulty or severity parameters. Using the pattern of responses and item parameters, it is possible to estimate the respondent's theta score and thus approximate its location along the latent continuum. IRT is considered one of the best methods not only to select items, but also to accurately understand the relationship between patients' probability of responses and their position on the latent variable of interest. IRT also makes it possible to accommodate different item formats (dichotomous or Likert-type) in a statistically appropriate way without discarding information.

CAT is based on the same principles relying on the items' parameters (that form an item bank). The difference is that with CAT, the test will not be based on all the items [8], but will be tailored using the responses to three important questions (9): 1) How do we choose an item to start the test? 2) How do we choose the next item to be administered after having seen the patient's response to the current one? and 3) How do we know when to stop the assessment? (fig. 2).

Several criteria are involved when choosing the first item. At each step, the CAT procedure will try to select the item that will provide the greatest gain in information compared to what we already have. This rationale is also valid for selecting the first item. As the theta distribution is known, its mean can be used as an initial guess. Once we have the answer to the first item, estimation of the theta (*interim estimation*) can be updated and the SEM can be estimated. If this standard error has reached an acceptable threshold, or we have administered a predefined number of items, the procedure can be stopped. If this stopping rule has not been reached, the next best item to maximise the information gathered over and above what we already know will be selected. Given the patient's response, we will update the theta and estimate the SEM again. When the stopping rule is finally reached, the test is stopped and the final theta and SEM are estimated. This procedure closely mimics the approach taken by experienced clinicians. Questions do not follow each other in a fixed order; instead, the

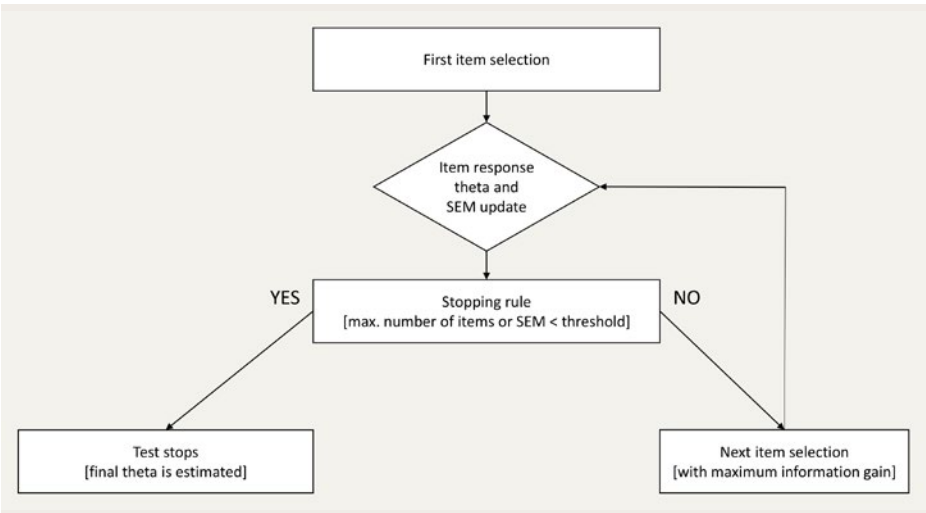

Figure 2: Computerised adaptive testing (CAT) procedure.

next question is selected using all the information gathered from responses to previous questions. Unnecessary questions are avoided. Every question is supposed to be a "good question" in the sense that it maximises the increase in knowledge. When the required level of precision is reached, the interview is stopped.

Historically, CAT was predominantly used in ability or personality testing, but in recent years its use in clinical contexts has been encouraged. Cooperative projects such as the Patient-Reported Outcomes Measurement Information System (PROMIS) [10], funded by the National Institutes of Health of the United States, have revealed the shift to implementing CAT in mental health facilities [1, 3, 4, 11, 12]. CAT has been used successfully in mental health contexts several times: Gibbons and colleagues developed two CAT measures - one for general anxiety disorder and one for depression - to be used for screening patients in primary care settings [13, 14]; Simms and colleagues created an adaptive measure for personality disorder traits [15]; Forbey and colleagues designed an adaptive version of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) to identify psychological difficulties in military veterans [16] and Moore and colleagues developed a CAT version of the Schizotypal Personality Questionnaire (SPQ) [17]. CAT requires a digital device (smartphone, tablet or computer) and, thankfully, open-source, online adaptive testing platforms, such as Concerto, are freely available [18]. Given the move toward CAT in psychiatry, the present study aimed to assess whether the Paradox of Self-Stigma scale (PaSS-24), an existing pen and paper test evaluating three dimensions of self-stigma, could be transformed into a shorter yet equally reliable computerised adaptive test.

Methods

Measures

The Paradox of Self-Stigma Scale

The PaSS-24 is a short but psychometrically robust tool designed in collaboration with patients to measure self-stigma and related constructs in French. PaSS-24's theoretical framework is aligned with Corrigan's social-cognitive model of internalised stigma and puts an emphasis on paradoxical empowerment [19-21]. Another important prerequisite in the design of PaSS-24 was to create a tool, that could be applied across different groups of stigmatised persons (defined as any attribute that could be viewed as different from the norm such as gender, sexual orientation, race, religion, mental or physical health). Using focus groups involving mental health professionals and people living with mental illness, a total of 72 items were initially developed to measure various aspects of self-stigma. PaSS-24's final version used 24 items answered on a five point Likert scale: *1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree and 5 = strongly agree*. Items were divided over three eight-item subscales (stereotype endorsement, non-disclosure and righteous anger). This version was evaluated on a sample of psychiatric patients during their hospitalisation in different psychiatric hospitals or residential psychiatric facilities and then cross-validated on an independent sample. PaSS-24 demonstrated good internal validity. Internal consistency, test–retest reliability and convergent validity estimates were also good [22].

The free online adaptive version of the PaSS-24 is available for clinicians or researchers at the following address: https://adaptest.ch/test/pass24.

Procedure

PaSS-24's items were calibrated using the data from the validation study [22]. Next, to simulate the CAT procedure, estimated item parameters were introduced into Firestar software version 1.3.2 [23].

Firestar is software designed to simulate CAT with polytomous items. It allows to perform Monte-Carlo simulations and provides various item selection procedures, stopping rules and theta estimators as well as an array of output files for secondary analyses [23]. Because our item responses used Likert scales, a graded response model was selected. A large number of participants (50,000) were simulated to achieve accurate estimates under reasonable computing time. It is important to note that the number of participants on which the initial validation study was performed (N=202) is not related to the length of the adaptive procedure and only influences the accuracy of the item parameters estimation. The simulated thetas were sampled from a normal distribution with a mean of 0 and a standard deviation of 1, which corresponds to the level and dispersion of the original sample. Minimum and maximum thetas ranged between minus and plus four with increments of 0.05. The maximum number of items to administer was set to eight and the minimum was two. Each subscale's stopping rule was set to a standard error corresponding exactly to the reliability of the pen and paper version of the PaSS-24: 0.810 for the stereotype endorsement subscale, 0.832 for the rightful anger subscale and 0.879 for the non-disclosure subscale [22]. The first item of each subscale was selected using the prior mean. Interim theta estimations were made using Expected A Posteriori estimations. The next items were selected using the Fisher maximum information method.

Statistical Analysis

PaSS-24 items were calibrated using the Multidimensional Item Response Theory package for R software (version 4.1.1) and a graded response model for the validation study data [22]. During the Firestar simulation, we recorded the minimum, maximum, mean and median numbers of items administered before the stopping procedure. Pearson correlations were estimated between simulated and estimated thetas and the mean reliability was based on the final standard errors.

Results

Items descriptions and parameters are presented in table 1.

A mean of 5.017 items (SD=1.908) was administered for the eight-item stereotype endorsement subscale, with the number of items needed varying between two and eight. The median number of items was four. Average reliability was 0.813 and the correlation between the simulated and estimated thetas was close to unity (r=0.980).

A mean of 6.679 items (SD=1.684) was administered for the eight-item rightful anger subscale, with the number of items needed varying between three and eight. The median number of items was eight. Average reliability was 0.782 and the correlation between the simulated and estimated thetas was close to unity (r=0.991).

A mean of 5.245 items (SD=2.503) was administered for the eight-item non-disclosure subscale, with the number of items needed varying between three and eight. The median number of items was five. Average reliability was 0.853 and the correlation between the simulated and estimated thetas was close to unity (r=0.989).

Discussion

This study aimed to evaluate whether the PaSS-24, an existing pen and paper scale evaluating three dimensions of self-stigma, could be transformed into a computerised adaptive test. Our goal was to reduce the number of items administered while maintaining a similar level of reliability. Results showed that two out of the three scales could be substantially shortened. Correlations between the simulated and estimated scores were close to unity, indicating that estimated scores closely matched the simulated true scores and that CAT did not sacrifice accuracy for brevity.

These results suggest that a substantial reduction in the number of items necessary for psychometric tests can be achieved using CAT. This is even more remarkable as the initial sub-

Table 1: English language version of the PaSS-24 scale items and its parameters using the graded response model

Item		Scale*	a	b1	b2	b3	b4
1	People with my condition are less useful to society.	SE	1.705	-1.168	.181	.901	2.111
2	The restricted rights of people with my condition is scandalous.	RA	1.045	-2.615	-.880	.147	1.547
3	Because of people's ignorance about my condition, I do not speak to anybody about the problems linked to it.	ND	1.577	-1.376	-.086	.339	1.235
4	I tell myself, "What is the point of struggling to have the same rights?"	SE	1.599	-.515	.375	.738	1.860
5	I am really fed up with preconceived ideas about my condition.	RA	1.723	-2.090	-1.171	-.403	.918
6	Because of people's preconceptions, I do not speak to anybody about the problems linked to my condition.	ND	1.943	-1.308	-.144	.428	1.245
7	People with my condition should be banned from certain jobs.	SE	1.145	-.984	.004	.560	2.279
8	The public's lack of knowledge about my condition outrages me.	RA	1.853	-2.135	-1.238	-.317	.909
9	To stop myself from getting into trouble, I avoid situations where my condition might be revealed.	ND	1.779	-1.816	-.817	-.070	.971
10	Why bother making any effort when I am inferior to others?	SE	2.314	-.177	.738	1.108	1.789
11	The lack of accurate information about my condition is scandalous.	RA	1.577	-2.249	-1.299	-.594	.958
12	I use strategies to avoid talking about my condition.	ND	2.383	-1.302	-.476	-.048	.986
13	People with my condition should not be allowed to carry out certain activities.	SE	2.560	-.371	.431	.958	1.457
14	The stereotypes about my condition make me angry.	RA	2.419	-1.957	-1.063	-.471	.615
15	To avoid being discriminated against, I use strategies not to have to talk about my condition.	ND	4.168	-1.355	-.522	-.057	.855
16	People with my condition will never have a happy life.	SE	1.115	-.942	.495	.945	2.484
17	The media's lack of knowledge about my condition is appalling.	RA	1.765	-2.131	-1.033	-.248	.947
18	To avoid disagreeable remarks, I use strategies not to have to talk about my condition.	ND	3.353	-1.214	-.506	-.139	.878
19	People with my condition should stay among themselves.	SE	1.554	-.673	.249	.727	1.588
20	I am angry about the way my condition is caricatured on television.	RA	1.313	-2.077	-1.088	.013	1.075
21	To avoid any prejudice, I choose who I talk to about my condition.	ND	1.506	-1.981	-1.350	-.812	.773
22	I have come to terms with the idea that I will never be able to have a satisfying social life.	SE	2.073	.212	1.344	1.658	2.366
23	Certain people's attitudes towards my condition appal me.	RA	2.311	-2.042	-1.220	-.665	.583
24	I do not reveal my condition to anybody to avoid being judged.	ND	2.166	-1.205	-.094	.290	1.137

Note. Scale*: SE = Stereotype endorsement / RA = Righteous anger / ND = Non-disclosure.

scales were already very short (eight items) and demonstrated high reliability, putting our study in a rather challenging context. We, therefore, hypothesise that our results may be rather conservative and that greater gains could be achieved on longer scales.

The rightful anger subscale could only be shortened for some of the simulated participants, shown by the median number of eight items administered. The mean value nevertheless suggested that shortening the test was sometimes possible and sometimes only three items were required. This subscale's reliability was slightly lesser than that of the complete original subscale [22]. Given that the stopping rules matched the reliability of the original subscale, less reliable evaluations always happened when all eight items were administered. This suggests that this subscale's reliability was probably slightly overestimated during the validation study. In this situation, the stopping rule acts as a safeguard against unreliable assessments. In an unfavourable scenario all the items are administered. Although this means that no time is saved

in this situation, reliability is still guaranteed to be close to that of the original scale.

Despite these positive results, our study suffers from some limitations. First, this was a simulation study. In the real world, patients could select answers that are not always perfectly coherent with their true score. In contrast, the simulation process always selects an answer which is coherent with the simulated participant's theta and this may lead to a somewhat accelerated convergence toward a reliable score. Monte-Carlo simulations are a powerful tool to optimise or accelerate certain aspects of test development, but the results obtained from simulation should always be compared with those based on real participant data. Additionally, because the reliability of the test scores in the original study was estimated in a different framework than adaptive testing, the exact proportion of test length reduction may be imprecise.

Second, we had no direct feedback from real participants. Therefore, it was difficult to assess the extent to which shorter assessments or non-exposure to inadequate or unnecessary items were viewed positively by psychiatric patients. To the best of our knowledge, few observational studies of patients' acceptance of CAT have been made [5, 24, 25], although they did suggest positive patient experiences. This should be examined explicitly in future studies. We nevertheless have no reason to believe that patients might view shorter assessments negatively as the authors have often encountered feedback during previous studies indicating that patients were tired or found questionnaires to be too long. Third, other concepts of self-stigma may be of interest to practitioners. This is the reason we will work on other adaptive tests in the future.

Conclusion

When the time required to complete a psychometric questionnaire constitutes a barrier to effective clinical evaluation, mental health professionals should have access to shorter but equally accurate tests. Efforts to modernise test engineering using CAT models make it possible to increase the comfort of testing for patients without altering data quality. In the present study, the shortened psychometric CAT demonstrated similar reliability to the initial PaSS-24 in a sample of simulated participants. Shorter assessment subscales appeared possible, demonstrating similar reliability to the original test, even though its initial scales were already very short. It is important to exercise caution in interpreting these results as they were mainly not obtained from real participant data. However, CAT is a promising opportunity to give in to the pressure to use shorter, less burdensome scales without compromising reliability and accuracy.

Correspondence

Philippe Golay
Department of Psychiatry, Lausanne University Hospital
Consultations de Chauderon
Place Chauderon 18
CH-1003 Lausanne
philippe.golay[at]chuv.ch

Ethics Statement

Approval for this study was granted by the Human Research Ethics Committee of the Canton of Vaud (protocol #2016-00768). Written informed consent was obtained from all participants and our research was carried out in accordance with the recommendations of the Human Research Ethics Committee of the Canton of Vaud and the Declaration of Helsinki.

Funding Statement

This study was financed using internal institutional funding.

Conflict of Interest Statement

No financial support and no other potential conflict of interest relevant to this article was reported.

Author Contributions

PG and DM designed the study.
PG analysed and interpreted the data.
PG, DM and CBI drafted the first version of the manuscript.
BS, SM and CBK critically revised the manuscript for important intellectual content.

Data Availability Statement

No new data were used in the research described in the article. Item parameters are available in Table 1.

References

1 Rose M, Bjorner JB, Fischer F, Anatchkova M, Gandek B, Klapp BF, et al. Computerized adaptive testing – ready for ambulatory monitoring? Psychosom Med. 2012 May;74(4):338-48.
2 Walter OB. Adaptive Tests for Measuring Anxiety and Depression. In: van der Linden WJ, Glas CA, editors. Elements of Adaptive Testing. New York (NY): Springer New York; 2010. p. 123-36.
3 Walter OB, Becker J, Bjorner JB, Fliege H, Klapp BF, Rose M. Development and evaluation of a computer adaptive test for 'Anxiety' (Anxiety-CAT). Qual Life Res. 2007;16(S1 Suppl 1):143-55.
4 Thomas ML. The value of item response theory in clinical assessment: a review. Assessment. 2011 Sep;18(3):291-307.
5 Simms LJ, Clark LA. Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). Psychol Assess. 2005 Mar;17(1):28-43.
6 Furr RM, Bacharach VR. Psychometrics: An Introduction. Thousand Oaks (CA): SAGE publications; 2013.
7 De Ayala RJ. The Theory and Practice of Item Response Theory. 2nd ed. New York (NY): The Guilford Press; 2013.
8 van der Linden WJ, Glas CA. Elements of Adaptive Testing. New York (NY): Springer New York; 2010.
9 Wainer H, Dorans NJ, Eignor D, Flaugher R, Green BF, Mislevy RJ, et al. Computerized Adaptive Testing: A primer. 2nd ed. Abingdon (OX): Routledge; 2014.
10 Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al.; PROMISE Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care. 2007 May;45(5 Suppl 1):S3-11.
11 Jordan P, Shedden-Mora MC, Löwe B. Psychometric analysis of the Generalized Anxiety Disorder scale (GAD-7) in primary care using modern item response theory. PLoS One. 2017 Aug;12(8):e0182162.
12 Thomas ML. Advances in applications of item response theory to clinical assessment. Psychol Assess. 2019 Dec;31(12):1442-55.
13 Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, et al. Development of a computerized adaptive test for depression. Arch Gen Psychiatry. 2012 Nov;69(11):1104-12.
14 Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, et al. Development of the CAT-ANX: a computerized adaptive test for anxiety. Am J Psychiatry. 2014 Feb;171(2):187-94.
15 Simms LJ, Goldberg LR, Roberts JE, Watson D, Welte J, Rotterman JH. Computerized adaptive assessment of personality disorder: introducing the CAT-PD project. J Pers Assess. 2011 Jul;93(4):380-9.
16 Forbey JD, Ben-Porath YS, Arbisi PA. The MMPI-2 computerized adaptive version (MMPI-2-CA) in a Veterans Administration medical outpatient facility. Psychol Assess. 2012 Sep;24(3):628-39.
17 Moore TM, Calkins ME, Reise SP, Gur RC, Gur RE. Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. Psychiatry Res. 2018 May;263:250-6.
18 University of Cambridge The Psychometrics Centre. Cambridge: Concerto Open Source Online Adaptive Testing Platform. Available from: https://concertoplatform.com/.
19 Corrigan PW, Kerr A, Knudsen L. The stigma of mental illness: explanatory models and methods for change. Appl Prev Psychol. 2005;11(3):179-90.
20 Corrigan PW, Michaels PJ, Vega E, Gause M, Watson AC, Rüsch N. Self-stigma of mental illness scale – short form: reliability and validity. Psychiatry Res. 2012 Aug;199(1):65-9.
21 Corrigan PW, Watson AC. The paradox of self-stigma and mental illness. Clini Psychol Sci Pract. 2002;9(1):35-53.
22 Golay P, Moga M, Devas C, Staecheli M, Poisat Y, Israël M, et al. Measuring the paradox of self-stigma: psychometric properties of a brief scale. Ann Gen Psychiatry. 2021 Jan;20(1):5.
23 Choi SW. Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. Appl Psychol Meas. 2009 Oct;33(8):644-5.
24 Anatchkova MD, Saris-Baglama RN, Kosinski M, Bjorner JB. Development and preliminary testing of a computerized adaptive assessment of chronic pain. J Pain. 2009 Sep;10(9):932-43.
25 Hart DL, Wang YC, Cook KF, Mioduski JE. A computerized adaptive test for patients with shoulder impairments produced responsive measures of function. Phys Ther. 2010 Jun;90(6):928-38.