Current Psychiatry Reports (2019) 21: 67
https://doi.org/10.1007/s11920-019-1053-9

PSYCHIATRY IN THE DIGITAL AGE (J SHORE, SECTION EDITOR)

Without Wasting a Word: Extreme Improvements in Efficiency and Accuracy Using Computerized Adaptive Testing for Mental Health Disorders (CAT-MH)

Robert D. Gibbons[1] · Frank V. deGruy[2]

Published online: 1 July 2019

Abstract

Purpose of Review We review recent literature on the adaptive assessment of complex mental health disorders and provide a detailed comparison of classical test theory and adaptive testing based on multidimensional item response theory.

Recent Findings Adaptive tests for a wide variety of mental health traits (e.g., depression, anxiety, mania, substance misuse, suicidality) are now available in a cloud-based environment. These tests have been validated in a variety of settings against lengthy structured clinical interviews with excellent results and even higher reliability than fixed-length tests. Applications include screening and assessments in emergency departments, psychiatric and primary care clinics, student health clinics, perinatal medicine clinics, child welfare settings, and the judicial system.

Summary The future of mental health measurement will be based on automated screening and assessments. Adaptive tests will provide increased precision of measurement and decreased burden of measurement. Integration into the electronic health record is important and now easily accomplished.

Keywords Mental health measurement · Depression · Suicidality · Multidimensional item response theory · Computerized adaptive testing · Bifactor model

Introduction

With global access to the internet and the development of modern server technology, we can now provide national and even international screening and assessment of complex mental health constructs. Gone are the days when only a chosen few patients received lengthy face-to-face clinical interviews administered by skilled clinicians. Gone also are the days when we relied on traditional short-form assessments in which patients are screened for psychiatric conditions such as major depressive disorder using as few as 2 questions, or measured for the severity of depression or anxiety using a fixed set of 10 or fewer questions. We no longer need rely on classical test theory, in which all patients must be asked the same questions regardless of their severity of illness, in which all questions are weighted equally regardless of the severity, and then summed into a total score. The future of mental health measurement depends on a new, better strategy: questions are drawn from exhaustively large "item banks" that have been calibrated using modern psychometric methods; items are adaptively selected and administered so that they quickly converge on illness severity. Moreover, modern item response theory (IRT) can accommodate multidimensional constructs such as mental health disorders, unlike traditional educational measurement where constructs such as mathematical ability are essentially unidimensional and measured using simple unidimensional IRT. In this review, we will discuss the development and evolution of new mental health measurement systems for both screening and severity assessment that (1) preserve the multidimensionality of complex mental health constructs; (2) reduce patient burden; (3) completely eliminate clinician burden; and (4) at the same time maximize the precision of measurement. This is nothing less than a fundamental paradigm

This article is part of the Topical Collection on *Psychiatry in the Digital Age*

✉ Robert D. Gibbons
rdg@uchicago.edu

[1] Blum-Riese Professor of Biostatistics, University of Chicago, Chicago, IL, USA

[2] Department of Family Medicine, University of Colorado, Aurora, CO, USA

upgrade—from traditional fixed-length mental health tests with unvarying items but variable precision of measurement, to adaptive tests that fix the precision of measurement by allowing the items to vary, both in number and content. No longer do we need to sacrifice the precision of measurement for the speed of measurement. Through adaptive testing based on multidimensional item response theory (MIRT), we can specify a level of precision, then administer a small, statistically optimal set of items targeted to each patient's underlying level of severity at that particular time.

What Is Measurement?

What exactly is measurement? A traditional definition is "Measurement is the process of obtaining the magnitude of a quantity relative to an agreed-upon standard." This definition does not work for our purposes here because it presumes that there is an agreed-upon standard. While there may be in the physical sciences, there is not in the social and behavioral sciences, where we attempt to measure latent constructs such as depression or suicidality. We cannot "spike" a person with a known amount of depression and then compare different depression tests to determine which if any adequately reproduce the patient's level of severity. Rather, we draw statistical inferences about the magnitude of the latent variable through a statistically thoughtful examination of its manifestations, namely the symptoms that a patient has and the severity level of those symptoms. Some symptoms may be better discriminators than others in terms of the severity of the underlying latent variable. Some symptoms may be informative at the low end of the scale (e.g., Do you feel sad?) while others may be more informative at the higher end of the scale (e.g., Do you feel that others would be better off if you were dead?). The traditional (yet statistically naïve) approach of simply summing the individual item scores to derive a measure of the underlying latent variable as a "total score" assumes that all items are equally discriminating and that all items are equally severe. We can do much better.

Classical Test Theory

Most mental health measurement is based on subjective judgment and classical test theory (CTT). CTT is an application of the 1809 Gaussian theory of errors [1] to the measurement of individual differences. Originating in the work of Charles Spearman in 1907 [2], the classical theory was first applied to scores from cognitive tests in which item responses were scored "right" or "wrong." The test score of a respondent was the number of right responses. Later the theory was extended to any multiple-item psychological test in which items can be meaningfully scored in a consistent direction. Classical test theory assumes that the test score obtained by counting "right" responses is an additive linear model consisting of two random components—one due to the individual differences in the population of respondents and the other due to error, defined as the item-by-respondent interaction. In practice, CTT instruments (e.g., the PHQ-9) are characterized by a simple score (the sum of the 9-item responses) without regard to measures of either of the two variance components. The degree of uncertainty in the resulting test score is unknown, and as noted previously, the items are treated equally despite the possibility of large differences in severity. Moreover, changing the number of items (e.g., eliminating the suicide item) renders the results of the two different tests non-comparable.

Item Response Theory

In contrast to CTT, IRT takes a model-based approach to measurement. Its origins date back to the pioneering work of Lawley and Lord in the 1940s and 1950s [3•,4•]. In IRT, the observed responses (e.g., symptom severity ratings) arise from underlying variation in a latent variable of interest (e.g., depression) which is discretized with a threshold (e.g., meets/does not meet criteria). The corresponding probability of a specified response to an item (or rating of a symptom) is a function of the underlying severity of the illness and characteristics of the items, both of which can be estimated statistically from the response patterns. The item parameters typically include a parameter that describes the prevalence with which the specified response occurs (similar to the 50% lethal dose or LD50 in bioassay) and a second parameter which describes how well the item discriminates between patients of low and high levels of severity, similar to a slope in a regression model or a factor loading in a factor analysis model. Readers familiar with the concept of probit analysis will see that IRT is a form of probit analysis where the dosage metameter is unobserved and the outcome is a vector of discrete responses (one for each symptom item) instead of a single binary variable (e.g., dead or alive).

Multidimensional Item Response Theory

For complex traits such as depression, anxiety, or suicidality in which items are sampled from subdomains such as cognition, mood, and somatization, the assumption that the item responses are independent conditional on a single latent variable is no longer valid. The net result is that we underestimate the true uncertainty in the resulting test score, unnecessarily discard items due to poor item fit, and produce test scores with greater empirical variability (between subject) than desired [5••]. Multidimensional item response theory (MIRT) [6] obviates these problems by allowing us to directly incorporate this multidimensionality into test score estimation, resulting in separate estimates of each of the different subdomains. In our case, however, a more parsimonious solution is the bifactor restriction [7••,8] in which each item loads on the primary dimension of interest (e.g., depression) and a single subdomain from which that item was drawn (e.g., mood disorder). This allows us to describe the underlying disorder as a single-valued index that directly incorporates the conditional dependencies produced by the sampling of items within subdomains. Gibbons and coworkers [9••] first applied the bifactor model to the problem of measuring depression based on discrete item-response data in 2012.

Computerized Adaptive Testing

CAT makes use of the property of scaled measurement inherent in IRT, such that different subjects can respond to different items, yet still be similarly measured in terms of the latent attribute of interest. CAT requires that a large bank of items (potentially hundreds) be previously calibrated using an IRT or MIRT model (e.g., a bifactor model) so that those items that are good discriminators of high and low levels of the characteristic of interest can be identified and ordered in terms of their estimated severity on the latent variable metric of interest (e.g., depression). After each item is administered, an estimate of the patient's severity is made along with its uncertainty. Based on that severity estimate, the next most informative item is administered based on a statistical selection criterion. The process continues until a predefined uncertainty threshold (e.g., 5 points on a 100-point scale) is met. CAT has been extended to incorporate the inherent multidimensionality of mental health constructs [5••, 9••, 10], which is greatly facilitated by the bifactor model. To illustrate how this works, let us use a simple unidimensional example. Imagine that we wish to use CAT to test students' mathematical ability, and we have two examinees: a fourth grader, and a mathematics graduate student. CAT would begin by administering an algebra item, selected from somewhere midway along the continuum of difficulty. When the 4th grader got it wrong, CAT would select an easier item, but when the graduate student got it right would select a more difficult item. How far to move and exactly which item to administer next is the statistical key to the problem, which is more difficult for multidimensional constructs such as depression, anxiety, or mania than it is for essentially unidimensional constructs like mathematics.

Advantages of IRT Over CTT

There are several limitations of CTT that are overcome using IRT and MIRT, and many of the advantages of IRT and MIRT are further improved through combination with CAT. In the following explanation, we refer to IRT in general to describe both unidimensional and multidimensional models.

1. CTT provides no estimate of uncertainty in the estimated score, whereas IRT provides an estimate of both the scale score (e.g., severity of depression) and the uncertainty in that estimate. Adding CAT allows us to *ante hoc* specify a level of acceptable uncertainty, so that all subjects are measured to the same level of uncertainty regardless of their level of impairment. This is not true for fixed-length tests where uncertainty or precision will vary both between individuals and within individuals repeatedly measured over time.
2. CTT assumes that the items are a random sample of items from the population of items and that each item reflects the same level of severity of the underlying latent variable (e.g., depression) and each item has the same ability to discriminate high and low levels of the latent variable (e.g., depression). All of these assumptions are patently false. IRT jointly estimates characteristics of the subjects (i.e., a severity score and a corresponding uncertainty estimate) and characteristics of the items (i.e., prevalence of the item in the population, such as severity level and the item's discrimination of high and low levels of the underlying construct). CAT adds to this by adaptively selecting the most informative item for each subject based on an estimate of her severity level at any stage in the testing session. Since the person and item characteristics are measured on the same scale, we can select the most appropriate item for a given individual on a given measurement occasion. This is in direct contrast to traditional CTT instruments used in mental health measurement which must always administer the same set of items regardless of the severity level of the patient. Even if we have strong evidence to believe that the patient's depression is severe based on prior testing or a suicide attempt, CTT-based tests are unaltered whereas IRT-based CAT can use this information to further refine the testing session and tailor it (in a personalized medicine-like approach) to each patient.
3. For repeat assessments, CTT-based tests administer the same items over and over again, leading to response set bias. In other words, after repeated administration of the same items in the same order, responses are remembered and tend to be the same, even when severity might be changing. This bias is completely eliminated in IRT-based CAT where different items are presented in different testing sessions. Despite the use of different items, test-retest reliability is actually higher for IRT-based CAT (CAT-depression inventory (CAT-DI) $r = 0.92$ versus the PHQ-9 $r = 0.84$) [11]. These advantages are highlighted by recent applications of IRT-based CAT in short-term assessments (e.g., every 30 min) used to evaluate the

effects of fast acting drugs like ketamine (study at Columbia University, Michael Grunebaum and J. John Mann PIs) or for the positioning of electrodes in deep brain stimulation for treatment-refractory depression [12].

4. IRT methods permit the item content of tests to be updated during use without compromising the long-term comparability of the estimated scores. This allows items to be added to or deleted from the item bank as we learn more about the disorder of interest, while still maintaining comparability and interpretability between the scale scores before and after the change in the item bank. This is not possible with CTT—an added or deleted item renders scores noncomparable.
5. Perhaps the most important difference between CTT and IRT-based CAT is the ability of the latter to simultaneously reduce test burden and increase precision of measurement. The CAT-DI [9••] using an average of 12 adaptively administered items was able to reproduce the entire 389 item bank scores almost exactly ($r = 0.95$). Traditional CTT-based tests sacrifice the precision of measurement for the speed of measurement, but no such sacrifice is made with CAT, so long as items are drawn from large item banks. The distinct advantage of the CAT-DI is that it is based on multidimensional IRT so that the majority of items in the original item bank are maintained in the final item bank. As an example, the CAT-DI was based on an original item bank of 452 items, of which 389 of those items were retained in the final item bank based on strong loadings on the primary depressive severity dimension. Similar approaches to IRT-based CAT based on unidimensional IRT also began with large item banks (300 items) but fewer than 30 items remained after calibration due to the demonstrably false assumption of unidimensionality (e.g., NIH–PROMIS [13]). The majority of items administered via IRT-based CAT are targeted to that patient's level of severity and as a consequence precision of measurement is dramatically increased. Having an estimate of precision for any given test score enables this measurement process to continue until the desired level of precision is obtained. And remember that the desired level of precision changes according to the purpose: desired precision might be quite different for an RCT of a novel new antidepressant medication where we might want higher precision (less uncertainty but longer adaptive tests) than for a national psychiatric epidemiologic survey where we can accept lower precision but need more rapid assessments.
6. Some CTT measures of depression such as the HAM-D [14] have the additional problem of items with variable numbers of response categories. The HAM-D, for example, has items with 5-category responses, and items with binary responses. It is unlikely that Dr. Max Hamilton intended to count 5-category items as being 2.5 times more important than binary items, yet this is exactly the consequence of having items with different response formats. This is not a problem for IRT where differences in numbers of categories are absorbed in the calibration.

Table 1 displays examples of CTT (PHQ-9) and MIRT-based CAT (CAT-MH) for the screening (CAD-MDD) and measurement (CAT-DI) of depression at intake and two follow-up interviews. The table illustrates the reduction of items for adaptive screening (4 for CAD-MDD versus 9 for PHQ-9) and the reduction in items for CAT based on repeat assessments (13 items for week 1 versus 9 items for week 2), and the added information regarding uncertainty in the severity score for the CAT-DI. The CAT-DI also provides the estimated probability of an MDD diagnosis and the percentile among SCID diagnosed cases of MDD with that score. Finally, Table 1 illustrates how the CAT-MH adapts to the changing severity level of the patient and changes both the number and specific questions asked, required to achieve the same level of precision.

Diagnosis Versus Severity Measurement

It is important to note that diagnosis and scaled measurement (i.e., severity) of dimensional constructs are quite different things. The simple thresholding of a continuous measure to produce a diagnosis is inherently inefficient. Unlike measurement which should be based on IRT, diagnostic screening based on the mapping of a set of symptoms onto a "gold standard" diagnosis represents a statistical prediction problem that can be solved using machine learning techniques. For diagnosis, we want to assess symptoms with a severity level at the point at which the diagnosis shifts from negative to positive. A computerized adaptive diagnostic screener for major depressive disorder (CAD-MDD) has been developed [15] and shown to have sensitivity of 0.95 and specificity of 0.87 against an hour-long SCID DSM-5 diagnostic interview, when using an average of only 4 symptom items (max = 6), administered in less than a minute.

While patient self-reports of symptoms can be validly and reliably assessed using IRT-based CAT, clinicians use a very different observational paradigm based on clinical experience gleaned over years of repeated clinical exposure. Nevertheless, data indicate that the use of IRT-based measurement from patient self-reports is generally in close agreement with clinician ratings. For example, the correlation between the CAT-DI and the clinician rated HAM-D was found to be $r = 0.75$ in a general psychiatric patient sample [9••]. In addition, the CAD-MDD which is based on a machine learning algorithm [16] is indeed quite different from the process by which a clinician arrives at a diagnosis, yet produces diagnoses with excellent agreement with clinical experience in a tiny fraction (less than one-hundredth) of the time. The advantage of these new adaptive tools is that they focus clinical attention

Table 1 Example baseline screening and weeks 1 and 2 assessments for PHQ-9 versus CAT-MH

PHQ-9	Baseline	Week 1		Week 2	
	Responses	Responses	Uncertainty	Responses	Uncertainty
Little interest or pleasure in doing things	2	3	?	2	?
Feeling down depressed or hopeless	2	2	?	1	?
Trouble falling/staying sleep, Sleeping too much	3	2	?	1	?
Feeling tired or having little energy	3	3	?	2	?
Poor appetite or overeating	3	3	?	3	?
Feeling bad about yourself or that you are a failure or have let your family down	3	3	?	2	?
Trouble concentrating on things such as reading the newspaper or watching television	2	2	?	2	?
Moving or speaking slowly that other people could have noticed. Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual	3	3	?	2	?
Thoughts that you would be better off dead or of hurting yourself	3	2	?	2	?
MDD screen	Positive				
Confidence	?				
Total score		23		17	
Uncertainty		?		?	
0 = not at all, 1 = several days, 2 = more than half the days, 3 = nearly every day					
CAT-MH (CAD-MDD baseline, CAT-DI Weeks 1 and 2)	Baseline Responses	Week 1 Responses	Uncertainty	Week 2 Responses	Uncertainty
Baseline					
How much did any feelings of depression bother you?	Often				
How much have you felt nothing was interesting or fun?	Extremely				
I felt sad	Quite a bit				
How much of the time have you felt downhearted and blue*?*	Most of the time				
MDD Screen	Positive				
Confidence	99.30%				
Week 1					
I felt depressed		Most of the time	10.4		
Have you felt that life was not worth living?		Extremely	9.3		
How often did you feel hopeless?		Often	7.8		
How much of the time have you felt downhearted and blue?		Most of the time	7.1		
Has feeling depressed interfered with what you usually do?		Most of the time	6.2		
Have you found yourself wishing you were dead and away from it all?		Much more than usual	6.2		
How much of the time have you felt so down in the dumps that nothing could cheer you up?		All of the time	6.0		
I had difficulty sleeping		Quite a bit	5.8		
How much have you felt discouraged?		Extremely	5.6		
I felt like I was at the end of my rope		Most of the time	5.3		
Have you been feeling sluggish?		Quite a bit	5.2		

Table 1 (continued)

How much have you been disappointed in yourself?	Quite a bit	5.1		
How much were you distressed by trouble concentrating?	Quite a bit	4.8		
Severity Score	82.9	4.8		
Category	Severe			
Probability of MDD	99.80%			
Percentile among patients with positive SCID DSM-5 MDD	90.4th			
Week 2				
Have you been in low or very low spirits?			Some of the time	10.2
I had difficulty sleeping			Somewhat	8.7
Has feeling depressed interfered with what you usually do?			Some of the time	8.2
To what degree has fatigue caused you distress?			Somewhat	6.9
How often did you feel sad?			Some of the time	6.6
How much were you distressed by blaming yourself for things?			Moderately	6.2
How much have you felt withdrawn from others?			Moderately	5.7
How much have you been disappointed in yourself?			Moderately	5.3
I had difficulty concentrating			Moderately	5.0
Severity Score			56.3	5.0
Category			Mild	
Probability of MDD			90.20%	
Percentile among patients with positive SCID DSM-5 MDD			26.4th	

on those in greatest need while sparing the clinician time spent screening large numbers of patients, the majority of which are not in need of their care.

What Can We Test?

MIRT-based CATs are now available for the dimensional measurement in adults of depression, anxiety, mania/hypomania, PTSD, substance abuse, psychosis, and suicidality, in English and Spanish. In perinatal patients, we can measure depression, anxiety, and mania/hypomania. For youth (ages 7–17), we have instruments to measure depression, anxiety, mania/hypomania, ADHD, conduct disorder, oppositional defiant disorder, and suicidality rated by both the child and parent [5••, 9••, 10, 17, 18•,19•, 20, 21].

Example Applications

There are numerous applications of this new technology.

- Large-scale screening and monitoring of depression and anxiety in integrated primary care and behavioral health-care settings is a natural application of this work and is already in practice in a number of major institutions.
- Insurers can now monitor the progress of patients through treatment without the patient needing to be in the clinic for testing.
- High frequency monitoring, even hourly (or less) in response to fast acting new pharmacologic treatments (e.g., ketamine and deep brain stimulation) are possible because the same items are not repeatedly administered to the same patient. The ability to monitor patients daily for changes in depressive severity is now feasible.
- Patients seeking treatment can be screened by telephone so that waiting lists can be prioritized based on need rather than waiting time.
- Large-scale screening of adolescents in schools for depression and suicide risk is now possible as a first step in prevention efforts using these tests, and now tests are available for lower school students and their parents as well.
- In pharmaceutical studies, adaptive testing provides a method for easily identifying the most severely ill patients for enrollment as well as providing outcome measures with increased precision for identifying pharmacologic treatment effects. A recent study at Columbia University has shown that the adaptive depression and suicidality tests outperformed clinician ratings of depression and

suicidality in a randomized clinical trial of ketamine versus midazolam.

- Large-scale molecular genetic studies (e.g., genome-wide association studies) can use rapid adaptive testing to provide mental health phenotypes needed to better understand the genetic basis for psychiatric disorders and make advances in personalized medicine.
- CAT is also uniquely suited to detect falsification of responses that may be made to either give the impression of an illness (e.g., in a jail) or to mask the presence of an illness (e.g., in the military or among athletes). As an example, MIRT-based CAT is in use in the Cook County Bond Court to identify potential inmates who are in need of mental health treatment.
- Large-scale University student screening and monitoring programs can now be implemented and students that are in need of treatment can be immediately identified and treated with internet-based cognitive behavior therapy, or more intensive treatment, depending on their severity scores. The entire entering class at UCLA (6000 students) has recently been screened and assessed for depression, anxiety, and suicidality in this way. Those with mild to moderate severity are immediately enrolled in internet-based cognitive behavior therapy (iCBT) and peer support counseling.
- Finally, CAT has been used for state-wide assessment of mental health disorders in the state of Tennessee child welfare system where over 300 case-workers have been trained in its use.

Integration With the Electronic Health Record

As of 2015, 87% of ambulatory care practices used an electronic health record (EHR) [22]. Simultaneously, use of the internet and web-enabled digital devices has grown enormously; currently 88% of US adults use the internet and 77% use smartphones [23, 24]. The simultaneous expansion of health information technology (IT) and internet use provides a unique opportunity for health-care systems to use novel methods to integrate patient reported outcomes (PROs) into clinical care.

A new strategy for integrating PROs into clinical care is by using patient portals. Patient portals are secure websites which give patients access to their health information via a web connection [25] and allow for secure messaging between providers and patients. Patient portals which are linked to an EHR, also known as *tethered*, allow for seamless data flow from the health-care system to patients. Patients who use portals have fewer no-show appointments [26] and higher levels of satisfaction and engagement [27–29] with only slightly increased physician workloads [30]. Studies have demonstrated that patient portals increase communication between patients and providers [30]; however, to date, tethered patient portals have not been used proactively by health-care systems to systematically collect patient PRO data. Using patient portals to collect PROs proactively could be highly efficient and actionable, especially when data are triaged and pushed to clinical teams as necessary. Portal-based outreach by the health-care system could improve population health management, increase patient engagement, improve clinical workflow by off-loading work, and improve quality of care.

To date, very limited research exists in this domain, and no randomized controlled trials have examined whether patient portals can be used to provide population-level PRO measurement. In order to motivate health-care systems to invest in the health IT infrastructure to integrate PROs via portals, it is essential to conduct a randomized controlled trial (RCT) to understand whether portals can extend clinical care to new populations, and in essence, evaluate the return on investment. To assess whether patient portals increase the capture of PROs, it is essential to test a PRO, like depression symptomatology, that is clinically important, highly prevalent, measurable, treatable, and has a long-standing history of being under assessed and managed. For decades, MDD has had validated brief questionnaires, and more recently computerized adaptive tests, which could be used to diagnose and assess the severity of MDD. However, these tests are not used routinely in clinical practice, resulting in only half of patients being diagnosed, and only half of patients with MDD being adequately treated in primary care.

Recently, an AHRQ-funded study (Dr. Neda Laiteerapong PI) is under way to test this hypothesis. To date, a MIRT-based CAT for the measurement of depression [9••] and a machine learning computerized adaptive diagnostic screener for MDD [15] have been fully integrated into the Epic EHR at the University of Chicago, and clinical workflows designed around their integration. The system enables depression screening and measurement in the clinic using computers and remote screening and assessment via the patient portal, with results immediately displayed in the patient's medical record. This randomized controlled study will evaluate whether portal invitations to complete depression screening increase the rate of screening compared with screening for depression during routine clinic visits and whether evaluation of depression symptoms via the portal increases major depression remission rates compared with usual care.

Implications for Clinical Practice

So, how can this new technology be used to improve clinical practice? To begin, both diagnostic screeners and dimensional severity measures for depression, anxiety and mania can be administered either in the clinic or in a patient's home prior to their primary care visit. Results of these tests can then be integrated into the electronic health record so that the primary care doctor can be alerted to the possibility of a psychiatric issue for discussion, observation, treatment, or referral, as indicated. For all clinicians, the dimensional scores can be administered either at clinic visits or remotely at any interval of time (e.g., weekly) so that the effectiveness of treatment can be monitored and mental health care adjusted sensitively and frequently. Similarly, mental health screening can be initiated in the emergency department where rates of undiagnosed and untreated depression are likely to be high. These patients have been shown to be high users of physical health services including ED and inpatient services, so their identification may lead to decreased health care costs and improved outcomes if successfully referred and treated [31•]. In other high-risk populations such as perinatal clinics, repeated assessments during the course of pregnancy and post-partum can lead to early detection of the onset of perinatal depression and subsequent better outcomes. Finally, these tests can form the basis for a stepped care approach whereby less severely ill patients can be treated with inexpensive yet efficacious mobile psychotherapy apps (e.g., iCBT), reserving psychotherapy and/or pharmacotherapy for those with moderate to severe depression. In our recent ED study [31•], 22% of patients screened positive for depression but only 7% were in the moderate to severe range. The majority of these screen-positive subjects (15%) could likely be treated using far less expensive and less clinician-intensive interventions. Remarkably, those with MDD (adaptively assessed using the CAD-MDD) had a 61% increased rate of ED visits and a 49% increased rate of hospitalizations. Across the range of the CAT-DI depressive severity scale there was a 250% increase in both ED visits and hospitalizations. These depression assessments, so useful at focusing the visit, were completed in the ED in an average of 2 min.

With new US Preventive Services Task Force recommendations for widespread screening for depression in primary care, perinatal care and across the age span [32•], the advantages of scaled measurement become all the more relevant.

In the world of primary care, innovation around practice redesign is at a fever pitch: integration of behavioral health into primary care, advanced self-management strategies for chronic diseases, incorporation of community resources into personal health plans, and dealing with the social determinants of health are all changes that will benefit from this technology. One of the key advancements is in the area of patient engagement—self-management in dealing with the common mental, emotional, and behavioral problems in primary care. IRT-based CAT offers us an unprecedented platform for shared patient participation in the management of mental health disorders, substance misuse and suicidality by virtue of its ease and accuracy of use by patients at times other than clinic visits, and protection against testing bias; offering massive improvements in the care of these conditions in primary care. We simply cannot use traditional CTT-based instruments in this way. It is inevitable that we will adopt these new methods of measurement in primary care.

Conclusions

Just as model-based measurement replaced classical test theory in educational measurement, the same transition is inevitable in measurement in the social and behavioral sciences in general, and in mental health in particular. While the heavy lift from unidimensional to multidimensional constructs leads to statistical complexity in estimation and development, this statistical hurdle has been cleared, and the practical use of these measures is no more complex than their more limited CTT counterparts. Clinical workflow and seamless integration with the electronic health record is now feasible for routine implementation in our health-care systems. The application of MIRT-based CAT in mental health is not limited to clinical care, but goes beyond to screening and assessments in our schools, child welfare systems, and criminal justice system. Indeed, there are applications that we have not yet considered. Mental health is fundamental to the quality of our lives and to the world's public health; and we are now in possession of tools that make the assessment of mental health much, much easier, and more accurate.

Acknowledgments Robert Gibbons is a founder of Adaptive Testing Technologies which distributes the CAT-MH™ suite of tests. The terms of this arrangement have been reviewed and approved by the University of Chicago in accordance with its conflict of interest policies.

Funding Information This work was supported by grants R01-MH-66302, R01-MH-100155, 2 R01-MH-100155-06, U01-MH-104311, and 1R44MH118780 from the National Institute of Mental Health.

Compliance with Ethical Standards

Conflict of Interest Frank V. deGruy declares no potential conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Gauss CF. Theoria motus corporum coelestium in sectionibus conicis solem ambien-tium. Perthes et Besser, Hamburg. Werke, 1809; 7: 1–280. Translated by C. H. Davis as Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections. Little, Brown, Boston, 1857. Reprinted by Dover, New York, 1963.
2. Spearman C. Demonstration of formulae for true measurement of correlation. Am J Psychol. 1907;18(2):161–9.

3.• DN L. On problems connected with item selection and test construction. Proc Roy Soc Edinb. 1943;61:273–87. **This is a foundational paper on item response theory.**

4.• Lord, F. (1952). A Theory of Test Scores (Psychometric Monograph No. 7). Richmond, VA: Psychometric Corporation. Retrieved from www.psychometrika.org/journal/online/MN07.pdf. **This is a foundational paper on item response theory.**

5.•• Gibbons RD Computerized adaptive diagnosis and testing of mental health disorders. Annu Rev Clin Psychol. 2016;12:83–104. **This paper describes the development of MIRT-based CAT, providing both statistical and clinical details.**

6. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika. 1981;46:443–59.

7.•• Gibbons RD, Hedeker D. Full-information item bi-factor analysis. Psychometrika. 1992;57:423–36. **This paper developed the bifactor IRT model.**

8. Gibbons RD, Bock RD, Hedeker D, Weiss D, Segawa E, et al. Full-information item bi-factor analysis of graded response data. Appl Psychol Meas. 2007;31(2007):4–19.

9.•• Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, et al. The CAT-DI: a computerized adaptive test for depression. Arch Gen Psychiatry. 2012;69:1104–12. **This is the first paper to describe the development of an MIRT-based CAT for mental health measurement.**

10. Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, et al. Development of the CAT-ANX: a computerized adaptive test for anxiety. Am J Psychiatr. 2014;171:187–94.

11. Beiser D, Vu M, Gibbons RD. Test-retest reliability of a computerized adaptive depression test. Psychiatr Serv. 2016;67:1039–41.

12. Sani S, Busnello J, Kochanski R, Cohen Y, Gibbons RD. High frequency measurement of depressive severity in a patient treated for severe treatment resistant depression with deep brain stimulation. Transl Psychiatry. 2017;7:e1207.

13. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, et al. PROMIS cooperative group. Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): depression, anxiety, and anger. Assessment. 2011;18:263–83.

14. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56–62.

15. Gibbons RD, Hooker G, Finkelman MD, Weiss DJ, Pilkonis PA, et al. The CAD-MDD: a computerized adaptive diagnostic screening tool for depression. J Clin Psychiatry. 2013;74:669–74.

16. Brieman L. Random forests. Mach Learn. 2001;45:5–32.

17. Achtyes ED, Halstead S, Smart L, Moore T, Frank E, et al. Validation of computerized adaptive testing in an outpatient nonacademic setting. Psychiatr Serv. 2015;66:1091–6.

18.• Kim JJ, Silver RK, Elue R, Adams MG, La Porte LM, et al. The experience of depression, anxiety and mania among perinatal women. Arch Womens Ment Health. 2017;19:94–100. **This paper introduced MIRT-based CAT for the measurement of perinatal depression, anxiety and mania.**

19.• Gibbons RD, Kupfer D, Frank E, Moore T, Boudreaux E. Development of a computerized adaptive suicide scale. J Clin Psychiatry. 2017;78:1376–82. **This paper introduced the first adaptive suicide scale.**

20. Gibbons RD, Alegria M, Cai L, Herrera L, Markle SL, et al. Successful validation of the CAT-MH scales in a sample of Latin American migrants in the United States and Spain: Psychological Assessments. Published on-line ahead of print 30(10), 1267–76.

21. www.adaptivetestingtechnologies.com. August 22, 2017

22. Heisey-Grove D, Patel VONC. Data brief: any, certified, and basic: quantifying physician EHR adoption through 2014. In: The Office of the National Coordinator for Health Information Technology, editor. ; 2015. p. 1–10.

23. Mobile Fact Sheet. 2017. (Accessed August 22, 2017, at http://www.pewinternet.org/fact-sheet/mobile/.)

24. Internet/Broadband Fact Sheet. 2017. (Accessed August 22, 2017, at http://www.pewinternet.org/fact-sheet/internet-broadband/.)

25. Byrne JM, Elliott S, Firek A. Initial experience with patient-clinician secure messaging at a VA medical center. JAMIA. 2009;16:267–70.

26. Nijland N, van Gemert-Pijnen JE, Kelders SM, Brandenburg BJ, Seydel ER. Factors influencing the use of a web-based application for supporting the self-care of patients with type 2 diabetes: a longitudinal study. J Med Internet Res. 2011;13:e71.

27. Nazi KM, Hogan TP, McInnes DK, Woods SS, Graham G. Evaluating patient access to electronic health records: results from a survey of veterans. Med Care. 2013;51:S52–6.

28. Ketterer T, West DW, Sanders VP, Hossain J, Kondo MC, et al. Correlates of patient portal enrollment and activation in primary care pediatrics. Acad Pediatr. 2013;13:264–71.

29. Lam R, Lin VS, Senelick WS, Tran HP, Moore AA, et al. Older adult consumers' attitudes and preferences on electronic patient-physician messaging. Am J Manag Care. 2013;19:eSP7–11.

30. Neuner J, Fedders M, Caravella M, Bradford L, Schapira M. Meaningful use and the patient portal: patient enrollment, use, and satisfaction with patient portals at a later-adopting center. Am J Med Qual. 2015;30:105–13.

31.• Beiser DJ, Ward CE, Vu M, Laiteerapong N, Gibbons RD. Depression in emergency department patients and association with healthcare utilization. Acad Emerg Med. Published on-line 18 March 2019 https://doi.org/10.1111/acem.13726. **This paper introduced the use of CAT for emergency medicine**.

32.• Siu AL, the US Preventive Services Task Force. Screening for depression in adults: US Preventive Services Task Force recommendations statement. JAMA. 2016;315(4):380–7. **This paper mandated depression screening in primary care.**

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.