# Computerized Adaptive Diagnosis and Testing of Mental Health Disorders

Robert D. Gibbons,[1] David J. Weiss,[2] Ellen Frank,[3] and David Kupfer[3]

[1]Center for Health Statistics and Departments of Medicine and Public Health Sciences, University of Chicago, Chicago, Illinois 60612; email: rdg@uchicago.edu

[2]Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455

[3]Department of Psychiatry and Western Psychiatric Institute and Clinic, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

Annu. Rev. Clin. Psychol. 2016. 12:83–104

First published online as a Review in Advance on November 20, 2015

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

This article's doi: 10.1146/annurev-clinpsy-021815-093634



## Keywords

computerized adaptive testing, item response theory, depression, anxiety, mania, differential item functioning, mental health measurement, mental health diagnosis, IRT, CAT

## Abstract

In this review we explore recent developments in computerized adaptive diagnostic screening and computerized adaptive testing for the presence and severity of mental health disorders such as depression, anxiety, and mania. The statistical methodology is unique in that it is based on multidimensional item response theory (severity) and random forests (diagnosis) instead of traditional mental health measurement based on classical test theory (a simple total score) or unidimensional item response theory. We show that the information contained in large item banks consisting of hundreds of symptom items can be efficiently calibrated using multidimensional item response theory, and the information contained in these large item banks can be precisely extracted using adaptive administration of a small set of items for each individual. In terms of diagnosis, computerized adaptive diagnostic screening can accurately track an hour-long face-to-face clinician diagnostic interview for major depressive disorder (as an example) in less than a minute using an average of four questions with unprecedented high sensitivity and specificity. Directions for future research and applications are discussed.

## Contents



## OVERVIEW

The importance of performing research in real-world clinical settings is widely recognized, as is the need for measurement-based care outside the bounds of clinical research. However, in busy medical and psychiatric practices and clinics, the feasibility of conducting the kind of extensive evaluations typical of clinical research is questionable. Therefore, any strategy that reduces the burden of empirically based assessment has the potential to improve outcomes through measurement-based clinical decision making.

Traditional mental health measurement is based on classical test theory, in which a patient's impairment level is estimated by a total score, which requires that the same items be administered to all respondents. These items are weighted equally, so that the question response "I am sad" is weighted of equal importance as the response "I am suicidal." In an effort to decrease patient burden, mental health instruments are often restricted to a small number of symptom items [e.g., the 9-item Patient Health Questionnaire (PHQ-9) or the 17-item Hamilton Depression Rating Scale (HAM-D)]. For a patient with a given level of depressive severity, only a few of the items will be discriminating. An alternative to administration of a full-scale or short-form assessment such as the PHQ-9 is adaptive testing, in which individuals may receive different scale items that are targeted to their specific impairment level. In adaptive testing, a person's successive item responses are used to determine a provisional estimate of his or her standing on the measured trait (for example, depression or anxiety) to be used for the selection of subsequent items; the provisional estimate is then updated after each item response. This form of testing has recently emerged in mental health research (Pilkonis et al. 2011). Procedures based on item response theory (IRT) (Embretson & Reise 2000, Hambleton & Swaminathan 1985) can be used to obtain estimates for items (for example, difficulty and discrimination) and individuals (for example, severity of depression) to more efficiently identify suitable item subsets for each individual. This approach to testing is referred to as computerized adaptive testing (CAT) and is immediately applicable to mental health measurement. For example, a depression inventory can be administered adaptively, such that an individual responds only to items that are most informative for assessing his or her

**Classical test theory:** traditional psychometric measurement based on the assumption that all items are equally difficult (severe) and have equal discrimination. Limitations include the demonstrably false assumption that all subjects are measured with the same level of certainty and that item and person characteristics cannot be separated

level of depression. The net result is that a small, optimal number of items are administered to the individual without loss (and frequently with gains) of measurement precision.

The paradigm shift is from traditional measurement, which fixes the number of items administered and allows measurement uncertainty to vary, to IRT-based CAT, which fixes measurement uncertainty and allows the number of items to vary. The results are a dramatic reduction in the number of items needed to measure mental health constructs and an increased precision of measurement. Inexpensive, efficient, and accurate screening of depression in medical and behavioral health settings is a direct application of the general theory and related methodology. For longitudinal assessments using traditional instruments, each testing session begins anew and is not informed by the results of prior testing sessions. This is not true for CAT administration, in which the next testing session can begin with the estimated severity score from the previous testing session.

Although there have been some applications of IRT-based CAT in mental health measurement (Fliege et al. 2005, Gardner et al. 2004, Pilkonis et al. 2011), this work has been based on the assumption of unidimensionality, an assumption that is generally inconsistent with the multidimensional nature of mental health constructs. Mental health questions (items) are traditionally drawn from content domains (e.g., mood, cognition, behavior), within which the items are more highly correlated than items from different content domains. This leads to a violation of the conditional independence assumption of the unidimensional IRT model, underestimation of the standard error of measurement, and greater variability in the estimated scale scores (Gibbons et al. 2007). The net result is that we overestimate the precision of measurement and prematurely conclude adaptive testing sessions. Resulting test scores are more variable, less valid, and lead to the need for larger sample sizes in clinical trials.

Diagnosis and measurement represent very different processes. Although IRT is ideal for measurement, it is not ideal for diagnostic screening where an external criterion is available [e.g., a *Diagnostic and Statistical Manual of Mental Disorders* (DSM) diagnosis of major depressive disorder (MDD)]. Decision trees (Brieman 2001, Brieman et al. 1984, Quinlan 1993) represent an attractive framework for designing adaptive predictive tests because their corresponding models can be represented as a sequence of binary decisions. Despite this intuitive appeal, decision trees have suffered from poor performance, largely as a result of variance associated with the specific algorithms used to estimate them and the limited modeling flexibility of small trees. On the other hand, models constructed of averages of hundreds of decision trees, called random forests, have received considerable attention in statistics and machine learning (Brieman 1996, Hastie et al. 2009). These models provide significant improvements in predictive performance. We refer to this general approach as computerized adaptive diagnosis (CAD).

In the following sections we describe IRT-based CAT in general, both conceptually and technically, and illustrate the application of multidimensional IRT-based CAT to problems in mental health measurement in general and to the adaptive measurement of depression, anxiety, and mania in particular. We also describe the first application of CAD to the problem of development of a diagnostic screener for MDD.

**Item response theory (IRT):** modern psychometric theory based on mathematical models that do not assume that the items are of equal difficulty and possibly not equally good at discriminating high and low levels of the latent trait of interest. Multidimensional IRT extends these ideas to the joint measurement of multiple latent variables. IRT is capable of separating characteristics of the items from characteristics of the examinee or patient

**Computerized adaptive testing (CAT):** an approach based on IRT in which an optimal set of items is selected for each individual until a previously determined level of precision of the estimate of ability or severity is obtained

**Decision tree:** a decision support tool that produces a tree-like model and associated graph. The branches of a tree are defined by the answers to each question. The next question asked depends on the sequence of answers that occurred prior to it

## CONCEPTUAL FOUNDATIONS

### Item Response Theory

Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. The difference is clarified by the following analogy, originally suggested by R. Darrell Bock. Imagine a track and field meet in which ten athletes participate in men's 110-m hurdles race and also in men's high jump. Suppose that the hurdles race is not quite conventional

in that the hurdles are not all the same height and the score is determined not only by the runner's time but also by the number of hurdles successfully cleared, i.e., not tipped over. On the other hand, the high jump is conducted in the conventional way: The crossbar is raised by, say, 2-cm increments on the uprights, and the athletes try to jump over the bar without dislodging it.

The first of these two events is like a traditionally scored objective test: Runners attempt to clear hurdles of varying heights, analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions. On the high jump, ability is measured by a scale in millimeters and centimeters at the highest scale position of the crossbar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until she can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. In IRT, ability is measured by a scale point, not a numerical count.

These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: If hurdles are arbitrarily added or removed, the number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures noncomparable. The same is true of traditional number-right scores of objective tests: Scores lose their comparability if item composition is changed.

The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2-cm positions, or if one of those positions were omitted, height cleared is unchanged, and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted, or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected. This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of educational and psychological measurement.

**Random forests:** an ensemble learning method for classification problems that constructs a large number of decision trees and uses their mode for classification

**Computerized adaptive diagnosis (CAD):** an adaptive approach to determining a binary diagnostic classification based on a decision theoretic model

**Item bank:** the total set of items (symptoms) that are initially calibrated using IRT and then selected for use in a CAT

## Computerized Adaptive Testing

Imagine a 1,000-item mathematics test with items ranging in difficulty from basic arithmetic through advanced calculus. Now consider two examinees, a fourth-grader and a graduate student in mathematics. Most questions will be uninformative for both examinees (too difficult for the first and too easy for the second). To decrease examinee burden, we could create a short test of 10 items, equally spaced along the mathematics difficulty continuum. Although this test would be quick to administer, it would provide very imprecise estimates of these two examinees' abilities because only an item or two would be appropriate for either examinee. A better approach would be to begin by administering an item of intermediate difficulty, and based on the response scored as correct or incorrect, select the next item at a level of difficulty either lower or higher. This process would continue until the uncertainty in the estimated ability is smaller than a predefined threshold. This process is called CAT. To use CAT, we must first calibrate a bank of test items using an IRT model that relates properties of the test items (e.g., their difficulty and discrimination) to the ability (or other trait) of the examinee. The paradigm shift is that rather than administering a fixed number of items that provide limited information for any given subject, we adaptively administer a small but varying number of items (from a much larger item bank) that are optimal for the subject's specific level of severity.

## The Bifactor Item Response Theory Model

Most applications of IRT are based on unidimensional models that assume that all of the association between the items is explained by a single primary latent dimension or factor (e.g., mathematical ability). However, mental health constructs are inherently multidimensional; for example, in the area of depression, items may be sampled from the mood, cognition, behavior, and somatic subdomains, which produce residual associations between items within the subdomains that are not accounted for by the primary dimension. If we attempt to fit such data to a traditional unidimensional IRT model, we will typically have to discard the majority of candidate items to achieve a reasonable fit of the model to the data. By contrast, the bifactor IRT model (Gibbons & Hedeker 1992) permits each item to tap the primary dimension of interest (e.g., depression) and one subdomain (e.g., somatic complaints), thereby accommodating the residual dependence and allowing for the retention of the majority of the items in the final model. The bifactor model was the first example of a confirmatory item factor analysis model, and Gibbons & Hedeker (1992) showed that it is computationally tractable regardless of the number of dimensions, in stark contrast to exploratory item factor analytic models. Furthermore, the estimated bifactor loadings are rotationally invariant, greatly simplifying interpretability of the model estimates (see sidebar The Bifactor Model).

**Bifactor model:** a confirmatory factor analytic model that assumes each item loads on a primary dimension and a single subdomain. Subdomains are selected in advance based on substantive criteria; however, the fit of alternative models (subdomain structures) can be compared so that the most parsimonious bifactor structure can be selected

**Item factor analysis:** a multidimensional version of IRT that permits the joint estimation of multiple latent traits, both in terms of the parameters of the items and the characteristics of the people

## Item Response Theory–Based Computerized Adaptive Testing in Mental Health Research

Although the use of CAT and IRT has been widespread in educational measurement, it has been less widely used in mental health measurement for two reasons (Gibbons et al. 2012b, 2014). First, large item banks are generally unavailable for mental health constructs. Second, mental health constructs (e.g., depression) are inherently multidimensional, and CAT has primarily been restricted to unidimensional constructs such as mathematics achievement. Application of unidimensional models to multidimensional data can result in biased trait estimates (e.g., severity), underestimates of uncertainty (Gibbons et al. 2007), and exclusion of large numbers of informative items from the bank. We have developed the underlying statistical theory and methodology necessary to apply multidimensional CAT to the measurement of depression, anxiety, and mania/hypomania symptom severity (Achtyes et al. 2015; Gibbons et al. 2012b, 2014).

# TECHNICAL FOUNDATIONS

## Overview of Item Factor Analysis

IRT-based item factor analysis makes use of all information in the original categorical responses and does not depend on pairwise indices of association such as tetrachoric or polychoric correlation coefficients. For that reason it is referred to as full-information item factor analysis. It works directly with item response models giving the probability of the observed categorical responses as a function of latent variables descriptive of the respondents and parameters descriptive of the individual items. It differs from the classical formulation in its scaling, however, because it does not assume that the response process has unit standard deviation and zero mean; rather, it assumes that the residual term has unit standard deviation and zero mean. Consider a $d$-factor solution with factor loadings $\alpha_{jv}$ for item $j$ on dimension $v$. The latent response process $y$ has zero mean and standard deviation equal to

$$\sigma_{y_j} = \sqrt{1 + \Sigma_v^d \alpha_{jv}^2}.$$

## THE BIFACTOR MODEL

In the bifactor case, the graded response model is

$$z_{jb}(\theta) = \sum_{v=1}^{d} a_{jv}\theta_v + c_{jb},$$

where only one of the $v = 2, \ldots, d$ values of $a_{jv}$ is nonzero in addition to $a_{j1}$. Assuming independence of the $\theta$, in the unrestricted case, the multidimensional model above would require a $d$-fold integral in order to compute the unconditional probability for response pattern $\mathbf{u}$, i.e.,

$$P(u = u_i) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L_i(\theta)g(\theta_1)g(\theta_2)\ldots g(\theta_d)d\theta_1 d\theta_2 \ldots d\theta_d,$$

for which numerical approximation is limited to 5 or 6 dimensions. Gibbons & Hedeker (1992) showed that for the binary response model, the bifactor restriction always results in a two-dimensional integral regardless of the number of dimensions, one for $\theta_1$ and the other for $\theta_v$, $v > 1$. The reduction formula is due to Stuart (1958), who showed that if $n$ variables follow a standardized multivariate normal distribution where the correlation $\rho_{ij} = \sum_{v=1}^{d} \alpha_{iv}\alpha_{jv}$ and $\alpha_{iv}$ is nonzero for only one $v$, then the probability that respective variables are simultaneously less than $\gamma_j$ is given by,

$$P = \prod_{v=1}^{d}\int_{-\infty}^{\infty}\left\{\prod_{j=1}^{n}\left[\Phi\left(\frac{\gamma_j - \alpha_{jv}\theta}{\sqrt{1-\alpha_{jv}^2}}\right)\right]^{u_{jv}}\right\} g(\theta)d\theta,$$

where $\gamma_j = -c_j/y_j$, $\alpha_{jv}=a_{jv}/y_j$, $y_j=(1+a_{j1}^2+a_{jv}^2)^{1/2}$, $u_{jv}=1$ denotes a nonzero loading of item $j$ on dimension $v$ (.), and $u_{jv} = 0$ otherwise. Note that for item $j$, $u_{jv} = 1$ for only one $d$. Note also that $\gamma_j$ and $\alpha_{jv}$ used by Stuart (1958) are equivalent to the item threshold and factor loading, and are related to the more traditional IRT parameterization as described above.

This result follows from the fact that if each variate is related only to a single dimension, then the $d$ dimensions are independent and the joint probability is the product of $d$ unidimensional probabilities. In this context, the result applies only to the $d - 1$ content dimensions (i.e., $v = 2, \ldots, d$). If a primary dimension exists, it will not be independent of the other $d - 1$ dimensions, since each item now loads on each of two dimensions. Gibbons & Hedeker (1992) derived the necessary two-dimensional generalization of Stuart's (1958) original result as

$$P = \int_{-\infty}^{\infty}\left\{\prod_{v=2}^{d}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n}\left(\Phi\left[\frac{\gamma_j - \alpha_{j1}\theta_1 - \alpha_{jv}\theta_v}{\sqrt{1-\alpha_{j1}^2-\alpha_{jv}^2}}\right]\right)^{u_{jv}}\right] g(\theta_v)d\theta_v\right\} g(\theta_1)d\theta_1,$$

For the graded response model, the probability of a value less than the category threshold $\gamma_{jb} = -c_{jb}/y_j$ can be obtained by substituting $\gamma_{jb}$ for $\gamma_j$ in the previous equation. Let $\delta_{ijb} = 1$ if person $i$ responds positively to item $j$ in category $b$ and $\delta_{ijb} = 0$ otherwise. The unconditional probability of a particular response pattern $\mathbf{u}_i$ is then

$$P(\mathbf{u} = \mathbf{u}_i) = \int_{-\infty}^{\infty}\left\{\prod_{v=2}^{d}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{n}\prod_{b=1}^{m_j}\left[\Phi_{jb}(\theta_1,\theta_v) - \Phi_{jb-1}(\theta_1,\theta_v)\right)^{\delta_{ijb}\cdot u_{jv}}\right] g(\theta_v)d\theta_v\right\} g(\theta_1)d\theta_1,$$

which can be approximated to any degree of practical accuracy using two-dimensional Gauss-Hermite quadrature, since for both the binary and graded bifactor response models, the dimensionality of the integral is 2 regardless of the number of subdomains (i.e., $d - 1$) that comprised the scale.

Inasmuch as the scale of the model affects the relative size of the factor loadings and thresholds, we rewrite the model for dichotomous responses in a form in which the factor loadings are replaced by factor slopes, $a_{jv}$, $\theta_v$ is one of the $d$ underlying latent variables of interest, and the threshold is absorbed in the intercept, $c_j$:

$$y_j = \sum_{v=1}^{d} a_{jv}\theta_v + c_j + \varepsilon_j.$$

To convert factor slopes into loadings, we divide by the above standard deviation and similarly convert the intercepts to thresholds:

$$\alpha_{jv} = a_{jv}/\sigma_{y_j} \text{ and } \gamma_j = -c_j/\sigma_{y_j}.$$

The threshold is the point on the latent variable where the probability of a positive response is 50%. Conversely, to convert to factor analysis units, we change the standard deviation of the residual from 1 to

$$\sigma^*_{\varepsilon_j} = \sqrt{1 - \Sigma_v^d \alpha_{jv}^2},$$

and change the scale of the slopes and intercept accordingly:

$$a_{jv} = \alpha_{jv}/\sigma^*_{\varepsilon_j} \text{ and } c_j = -\gamma_j/\sigma^*_{\varepsilon_j}.$$

For polytomous responses, the model generalizes as:

$$z_j = \sum_{v=1}^{d} a_{jv}\theta_v,$$

$$P_{jb}(\theta) = \Phi(z_j + c_{jb}) - \Phi(z_j + c_{j,b-1}),$$

where $\Phi(z_j + c_{j0}) = 0$ and $\Phi(z_j + c_{jm_j}) = 1 - \Phi(z_j + c_{j,m_j-1})$ where $\Phi$ is the unit normal integral. In the context of item factor analysis, this is the multidimensional generalization of the graded model (Samejima 1969).

## Confirmatory Item Factor Analysis

In confirmatory factor analysis, indeterminacy of rotation is resolved by assigning arbitrary fixed values to certain loadings of each factor during maximum likelihood estimation. An important example of confirmatory item factor analysis is the bifactor pattern for general and group factors, which applies to tests and scales with item content drawn from several well-defined subareas of the domain in question. To analyze these kinds of structures for dichotomously scored item responses, Gibbons & Hedeker (1992) developed full-information item bifactor analysis for binary item responses, and Gibbons extended it to the polytomous case (Gibbons et al. 2007). To illustrate, consider a set of $n$ test items for which a $d$-factor solution exists with one general factor and $d - 1$ group or method-related factors. The bifactor solution constrains each item $j$ to a nonzero loading $\alpha_{j1}$ on the primary dimension and a second loading ($\alpha_{jv}$, $v = 2, \ldots, d$) on not more than one of the $d - 1$ group factors. For four items, the bifactor pattern matrix might be

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}.$$

## PARAMETER ESTIMATION

Gibbons & Hedeker (1992) showed how parameters of the item bifactor model for binary responses can be estimated by maximum marginal likelihood using a variation of the EM algorithm described by Bock & Aitkin (1981). For the graded case, the likelihood equations are derived as follows.

Denoting the $v$th subset of the components of $\theta$ as $\theta_v^* = \begin{bmatrix} \theta_1 \\ \theta_v \end{bmatrix}$, let

$$\begin{aligned} P_i &= P(u = u_i) \\ &= \int_{\theta_1} \left\{ \prod_{v=2}^{d} \int_{\theta_v} \left[ \prod_{j=1}^{n} \prod_{b=1}^{m_j} \left( \Phi_{jb}(\theta_v^*) - \Phi_{jb-1}(\theta_v^*) \right)^{\delta_{ijb} \cdot u_{jv}} \right] g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1 \\ &= \int_{\theta_1} \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1, \end{aligned}$$

where $L_{iv}(\theta_v^*) = \prod_{j=1}^{n} \prod_{b=1}^{m_j} \left( \Phi_{jb}(\theta_v^*) - \Phi_{jb-1}(\theta_v^*) \right)^{\delta_{ijb} \cdot u_{jv}}$.

Then the log-likelihood is

$$\log L = \sum_{i=1}^{s} r_i \log P_i,$$

where $s$ denotes the number of unique response patterns, and $r_i$ the frequency of pattern $i$. As the number of items gets large, $s$ typically is the number of respondents and $r_i = 1$. Complete details of the likelihood equations and their solution are provided in Gibbons et al. 2007.

This structure, which Holzinger & Swineford (1937) termed the "bifactor" pattern, also appears in the interbattery factor analysis of Tucker (1958) and is one of the confirmatory factor analysis models considered by Jöreskog (1969). In the latter case, the model is restricted to test scores assumed to be continuously distributed. However, the bifactor pattern might also arise at the item level (Muthén 1989). Gibbons & Hedeker (1992) showed that paragraph comprehension tests, in which the primary dimension represents the targeted process skill and additional factors describe content area knowledge within paragraphs, were described well by the bifactor model. In this context, they showed that items were conditionally independent between paragraphs, but conditionally dependent within paragraphs.

The bifactor restriction leads to a major simplification of likelihood equations that (*a*) permits analysis of models with large numbers of group factors because the integration always simplifies to a two-dimensional problem, (*b*) permits conditional dependence among identified subsets of items, and (*c*) in many cases, provides more parsimonious factor solutions than an unrestricted full-information item factor analysis (see sidebar Parameter Estimation).

**Maximum marginal likelihood estimation:** a form of maximum likelihood estimation used when the number of unknown parameters increases linearly with the number of subjects. In the present case we use MMLE to estimate the parameters of the IRT model integrating over the distribution of the person parameters

## Computerized Adaptive Testing

The field of psychological testing has depended almost exclusively on conventional psychological tests since its inception about 100 years ago. In a conventional test, a set of test items is selected in advance to comprise an instrument designed to measure a particular psychological trait. All questions in that instrument are administered to every individual who takes that test.

Only a few exceptions, such as the intelligence tests based on Alfred Binet's test model and a few other individually administered tests of that type, have not used the conventional testing approach.

The Binet types of intelligence tests are adaptive tests (Weiss 1985). In an adaptive test, items are selected during the process of test administration for each individual being tested. Adaptive tests are designed to allow the test administrator to control the precision of a given measurement and to maximize the efficiency of the testing process. In the Binet test, items are classified during the process of development with respect to "mental age" levels. These levels correspond to increasing levels of item difficulty. When a test administrator administers a Binet-type test, he/she begins test administration at whatever "mental age" level the examinee appears to be functioning. Items are scored as they are administered, and when all items at a given mental age level have been administered, the test administrator determines whether additional items are needed.

If the examinee answered some of the items at a given mental age level correctly, testing continues with items of either a higher or a lower mental age level. If none or a few of the items are answered correctly, easier items are administered to that examinee. If all or most of the items at a given mental age level have been answered correctly, more difficult items are administered. Test administration continues until two mental age levels are identified: One at which the examinee answers all items incorrectly (the ceiling level) and one at which the examinee answers all items correctly (the basal level). In between the ceiling and basal level is the effective range of measurement for that individual. The result of this adaptive item-selection process is that individuals with different trait levels will be administered items at different difficulty levels.

The Binet-type tests have all the characteristics of an adaptive test, including:

1. A precalibrated bank of test items. To create an adaptive test, items must previously be administered to a group of individuals, and item difficulty and other data must be obtained on the items. An adaptive test based on IRT, for example, will use an item bank in which items are precalibrated on item difficulty, discrimination, and (if appropriate) the pseudoguessing parameter.
2. A procedure for item selection. Because items are selected based on an examinee's previous answers, items must be scored as they are administered. The next item (or item subset) to be administered is then based on how the examinee answered all previously administered items.
3. A method of scoring the test. Because the purpose of test administration is to obtain a test score for the examinee, the procedure for adaptive testing requires not only that items be scored as they are administered, but also that a test score of some type be determined at multiple points during the process of test administration.
4. A procedure for terminating the test. In contrast to a conventional test, the number of test items is not fixed in an adaptive test. Thus, in a Binet-type test, an individual may receive test items from as few as two mental age levels to as many as eight or nine, depending on how he/she performs on the test.

Research since the 1970s has shown that adaptive testing procedures are most effective when combined with IRT procedures (Kingsbury & Weiss 1980, 1983; McBride & Martin 1983). Thus, an item bank for use in adaptive testing can be calibrated according to an IRT model. The point at which a test is to be started (frequently referred to as the entry point) can be determined by taking into account individual status variables or other data about an individual (e.g., previous test scores, age, gender, clinical evaluations). Explicit procedures for estimating an entry point for an

## SEVERITY ESTIMATION

In practice, the ultimate objective is to estimate the trait level of person $i$ on the primary trait the instrument was designed to measure. For the bifactor model, the goal is to estimate the latent variable $\theta_1$ for person $i$. A good choice for this purpose (Bock & Aitkin 1981) is the expected a posteriori (EAP) value (Bayes estimate) of $\theta_1$, given the observed response vector $\mathbf{u}_i$ and levels of the other subdimensions $\theta_2 \ldots \theta_d$. The Bayesian estimate of $\theta_1$ for person $i$ is:

$$\hat{\theta}_{1i} = E(\theta_{1i}|\mathbf{u}_i, \theta_{2i} \ldots \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} \theta_{1i} \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1.$$

Similarly, the posterior variance of $\hat{\theta}_{1i}$, which may be used to express the precision of the EAP estimator, is given by

$$V(\theta_{1i}|\mathbf{u}_i, \theta_{2i} \ldots \theta_{di}) = \frac{1}{P_i} \int_{\theta_1} (\theta_{1i} - \hat{\theta}_{1i})^2 \left\{ \prod_{v=2}^{d} \int_{\theta_v} L_{iv}(\theta_v^*) g(\theta_v) d\theta_v \right\} g(\theta_1) d\theta_1.$$

These quantities can be evaluated using Gauss-Hermite quadrature as previously described.

In some applications, we are also interested in estimating a person's location on the secondary domains of interest as well. For the $v$th subdomain, the EAP estimate and its variance can be written as:

$$\hat{\theta}_{vi} = E(\theta_{vi}|\mathbf{u}_i, \theta_{1i}) = \frac{1}{P_i} \int_{\theta_v} \theta_{vi} \left\{ \int_{\theta_1} L_{iv}(\theta_v^*) g(\theta_1) d\theta_1 \right\} g(\theta_v) d\theta_v,$$

and

$$V(\theta_{vi}|\mathbf{u}_i, \theta_{1i}) = \frac{1}{P_i} \int_{\theta v} (\theta_{vi} - \hat{\theta}_{vi})^2 \left\{ \int_{\theta 1} L_{iv}(\theta_v^*) g(\theta_1) d\theta_1 \right\} g(\theta_v) d\theta_v.$$



adaptive test are available in conjunction with IRT using Bayesian statistical methods (Baker 1992, Weiss & McBride 1984).

IRT procedures for estimating an individual's trait level are applicable to the adaptive testing process. Procedures of maximum likelihood or Bayesian estimation permit estimation of trait, or in this case impairment levels, based on one or more responses made by a single individual in an adaptive test. Thus, a continuous updating of the impairment level can be accomplished after each item is administered in an adaptive test, and the next item to be administered can be based on the impairment estimate derived from all previous items administered. In addition, maximum likelihood and Bayesian estimation procedures also provide individualized standard errors of measurement (SEM) for each impairment level (see sidebar Severity Estimation).

Item selection rules derived from IRT and adaptive testing can explicitly use concepts of item information (Hambleton & Swaminathan 1985, Weiss 1985). Thus, at a given current impairment estimate, the most informative item not yet administered can be chosen for administration. When items are selected using this maximum-information item selection rule, the net effect is an extremely efficient procedure for reducing the error of measurement at each successive stage in the administration of an adaptive test (Weiss 1985). Item information describes the information contained in a given item for a specific impairment estimate. Our goal is to administer the item with maximum item information at each step in the adaptive process. Suppose there are $i = 1, 2, \ldots N$ examinees, and $j = 1, 2, \ldots n$ items. Let the probability of a response in category $h = 1, 2, \ldots m_j$ to graded response item $j$ for examinee $i$ with factor $\theta$ be denoted by $P_{ijh}(\theta)$. We call $P_{ijh}(\theta)$ a

**Posterior variance:** the uncertainty in a Bayes estimate of a person's ability or severity of illness

category probability. $P_{ijb}(\theta)$ is given by the difference between two adjacent boundaries,

$$P_{ijb}(\theta) = P(x_{ij} = b|\theta) = P^*_{ijb}(\theta) - P^*_{ijb-1}(\theta),$$

where $P^*_{ijb}(\theta)$ is the boundary probability. Under the normal ogive model, the boundary probability is given by

$$\begin{aligned} P^*_{jb}(\theta) &= \Phi(z_{jb}) \\ &= \int_{-\infty}^{z_{jb}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \end{aligned}$$

where

$$z_{jb} = a_{j1}\theta_1 + a_{j2}\theta_2 + c_{jb}.$$

When we are interested in estimating the item information function for $\theta_1$ in the presence of other subdomains, the subdomains can be integrated out of the objective function. For the purpose of CAT administration, $\theta_1$ is typically our focus; however, $\theta_2$ is also present in a bifactor model. In this case, we are interested in obtaining $I_j(\theta_1)$, which is a function only of $\theta_1$. To get $I_j(\theta_1)$, we integrate the previous bifactor item information function expression with the conditional distribution $h(\theta_2|\theta_1)$ of $\theta_2$ and obtain

$$I_j(\theta_1) = \sum_{b=1}^{m_j} \int \frac{\left[\phi(z_{jb}) - \phi(z_{jb-1})\right]^2}{\Phi(z_{jb}) - \Phi(z_{jb-1})} h(\theta_2|\theta_1) d\theta_2,$$

which provides an estimate of the information associated with $\theta_1$ averaged over the $\theta_2$ distribution.

Finally, adaptive testing procedures developed in accordance with IRT can take advantage of a number of different procedures for terminating an adaptive test. One procedure frequently applied is to reduce the individualized SEM to a prespecified level before a test is terminated (Weiss & Kingsbury 1984). An individualized SEM allows the number of test items administered to an individual to vary, but it also results in control of the subsequent level of SEM for individuals tested. Thus, for the individual who responds essentially in accordance with the IRT model, a given level of SEM will be achieved more quickly than for the individual for whom the responses are not in accordance with the IRT model, resulting in a slower reduction of the individualized SEM.

Although individually administered tests are efficient and effective, they are labor intensive, requiring a highly trained test administrator to achieve necessary levels of standardization. When adaptive testing uses IRT, however, the calculations required at each stage of item selection eliminate the possibility of using a human test administrator. Under these circumstances, the adaptive test must be administered by interactive computers, notebooks, or smart phones. Computerized adaptive testing (CAT) procedures administer items on an individual basis by presenting them on a computer screen. Responses are entered on the keyboard or by a touch screen and are immediately scored by the computer. Various algorithms for selecting items according to maximum information or other criteria are then implemented using the computational capabilities of the computer (Vale & Weiss 1984, Weiss 1985), and typically in less than one second another item is selected for administration and presented on the screen. Meanwhile, the computer continually updates the person's estimated impairment level and its SEM, again using IRT methods, and constantly monitors the appropriate termination criterion. Once the termination criterion is reached, the test is ended. Tests such as the Graduate Record Examination and the Graduate Management Admission Test have become CATs.

Research shows that adaptive tests are more efficient than conventional tests (Brown & Weiss 1977, McBride & Martin 1983). That is, in an adaptive test a given level of measurement precision can be reached much more quickly than in a test in which all examinees are administered the same items. This results from selecting items that are most informative for an individual at each

stage of test administration in the adaptive test. Typical adaptive tests result in a 50% average reduction in number of items administered, and some reductions in the range of 80% to 90% have been reported, with no decrease in measurement quality (Brown & Weiss 1977). In addition, as has been indicated, adaptive tests allow control over measurement precision. Thus, adaptive tests result in measurements that are both efficient and effective.

Although IRT was developed originally in the context of measuring ability and achievement, the family of IRT measurement models also includes numerous models that are applicable to personality instruments that are not dichotomously scored (Andrich 1978a,b, 1988; Muraki 1990; Tucker 1958). Research has demonstrated that the IRT family of models can be meaningfully applied to the measurement of attitudes and personality variables (Reise & Waller 1991) and that benefits that result from this application are similar to those observed for measuring ability and achievement. Research has also begun into improving the measurement of personality constructs using CAT (Baek 1997, Dodd et al. 1995).

The bifactor model is extremely useful for CAT of multidimensional data. The conditional dependencies produced by the subdomains can be directly incorporated in trait estimation and item information functions as shown in the previous sections, leading to improved estimates of uncertainty and elimination of premature termination of the CAT and potential bias in the estimated trait score. After each item administration, the primary ability estimate and posterior standard deviation are recomputed, and based on the estimate of theta on the primary dimension, the item with maximal information is selected as the next item to be administered. This process continues until the posterior standard deviation is less than a threshold value (e.g., 0.3). Once the primary dimension has been estimated via CAT, subdomain scores can be estimated by adding items from the subdomain that have not been previously administered, until the subdomain score is estimated with similar precision. Seo & Weiss (2015) provide and evaluate a fully multidimensional CAT algorithm for the dichotomous case of the bifactor model.

When the trait score is at a boundary (i.e., either the low or high extreme of the trait distribution), it may take a large number of items to reach the intended posterior standard deviation (SEM) convergence criterion (e.g., SEM = 0.3). In such extreme cases, we generally do not require such high levels of precision because we know that the subject either does not suffer from the condition of interest or is among the most severly impaired. A simple solution to this problem is to add a second termination condition based on item information at the current estimate of the trait score, and if there is less information than the threshold, the CAT terminates. The choice of the threshold is application specific and can be selected based on simulated CATs. A good value will affect only a small percentage of cases (e.g., <20%) and only be used in extreme (i.e., high or low) cases.

Large item banks may contain items that are too similar to be administered within a given session. These can be declared as "enemy items" and not coadministered. The idea of enemy items can be extended to the longitudinal case to ensure that the same respondent is not repeatedly administered the same items on adjacent testing sessions.

CAT will often result in a subset of the entire item bank being used exclusively, because these items have the highest loadings on primary domains and subdomains. Often the difference between the loadings of items that are selected by the CAT and those that are not is quite small and the items have similar information. To ensure that the majority of the items in the item bank are administered, we can add a probabilistic component in which a selected item is administered only if a uniform random number exceeds a threshold. Typically a threshold of 0.5 works well (for a uniform random number), but again, the exact choice can be based on simulated adaptive testing, in which the largest set of unique items is used without compromising the other characteristics of the measurement process (i.e., average number of items administered and correlation with the total bank score).

## ILLUSTRATION

The CAT-Mental Health (CAT-MH) study (Achtyes et al. 2015; Gibbons et al. 2012b, 2014) developed a bifactor-based CAT for ordinal response data (Gibbons et al. 2007) and applied it to a 1,008-item bank consisting of 452 depression, 467 anxiety, and 89 bipolar items. We review only the results for depression.

The total depression item bank consisted of 452 items. The items were organized into conceptually meaningful categories using a hierarchical approach. The hierarchy included domains (depression), subdomains (e.g., mood, cognition, behavior), and factors (e.g., within depressed mood, factors included both increased negative affect and decreased positive affect). The items were selected based on a review of more than 100 existing depression or depression-related rating scales.

Subjects for this study were male and female treatment-seeking outpatients between 18 and 80 years of age. Patients were recruited from the Western Psychiatric Institute and Clinic (WPIC) at the University of Pittsburgh, a community clinic (Dubois Regional Medical Center), and community controls.

A total of 798 subjects (WPIC) were used to calibrate the IRT model, and 816 subjects (414 WPIC and 402 Dubois) received the live CAT-Depression Inventory (CAT-DI). To study the validity of the CAT-DI, 292 consecutive subjects received a full clinician-based DSM-IV (Am. Psychiatr. Assoc. 1994) diagnostic interview (First et al. 1996) and the live CAT-DI. To examine convergent validity, data were also obtained for the HAM-D, PHQ-9, and Center for Epidemiologic Studies Depression scale (CES-D). The HAM-D was administered by a trained clinician, and the PHQ-9 and CES-D were self-reports.

Results of the calibration study revealed that the bifactor model with five subdomains (mood, cognition, behavior, somatic, and suicide) dramatically improved fit over a unidimensional IRT model (chi-square = 6,825, df = 389, $p < 0.0001$). A total of 389 items with a primary factor loading of 0.3 or greater (96% > 0.4 and 79% > 0.5) were retained in the model. Results of simulated CAT revealed that for SEM = 0.3 (approximately 5 points on a 100-point scale), an average of 12.31 items per subject (range 7 to 22) were required. The correlation between the 12-item average-length CAT and the total 389 item score was $r = 0.95$. For SEM = 0.4 (less precise), an average of 5.94 items were required (range 4 to 16), but a strong correlation with the 389-item total score ($r = 0.92$) was maintained. The average length of time required to complete the 12-item (average) CAT was 2.69 minutes in comparison with 51.66 minutes for the 389-item test.

**Figure 1** reveals the existence of two discrete distributions of depressive severity, with the lower component representing the absence of clinical depression and the higher component representing severity levels associated with clinical depression.

**Figure 2** displays the distributions of CAT-DI scores for patients with minor depression (including dysthymia), MDD, and those not meeting criteria for depression. There is a clear linear progression between CAT-DI depression severity scores and the diagnostic categories from the Structured Clinical Interview for the DSM. Statistically significant differences were found between none and minor ($p < 0.00001$), none and MDD ($p < 0.00001$), and minor and MDD ($p < 0.00001$), with corresponding effect sizes of 1.271, 1.952, and 0.724 SD units, respectively.

Convergent validity of the CAT-DI was assessed by comparing results of the CAT-DI to the PHQ-9, HAM-D, and CES-D. Correlations were $r = 0.81$ with the PHQ-9, $r = 0.75$ with the HAM-D, and $r = 0.84$ with the CES-D. In general, the distribution of scores between the diagnostic categories showed greater overlap (i.e., less diagnostic specificity), greater variability, and greater skewness for these other scales relative to the CAT-DI.

Using the 100 healthy controls as a comparator, sensitivity and specificity for predicting MDD were 0.92 and 0.88, respectively (threshold based on **Figure 1**). CAT-DI scores were significantly

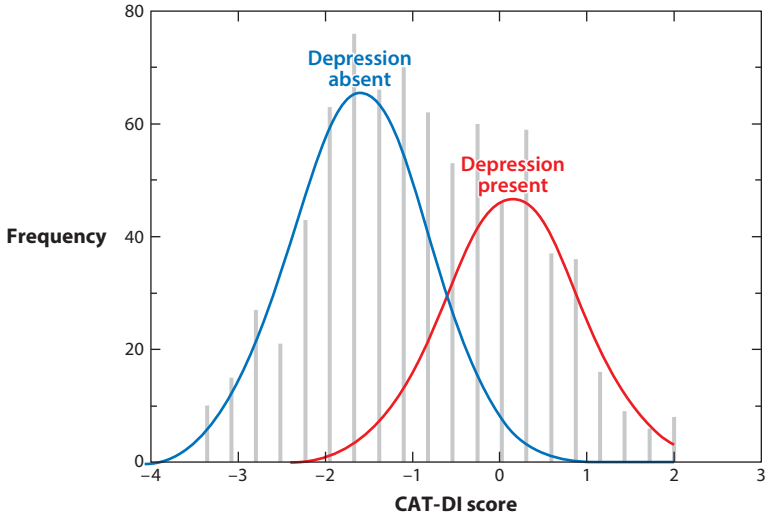


**Figure 1**

Observed and estimated frequency distributions of depressive severity using the Computerized Adaptive Testing-Depression Inventory (CAT-DI) depression scale. The lower component represents the absence of clinical depression; the higher component represents severity levels associated with clinical depression.

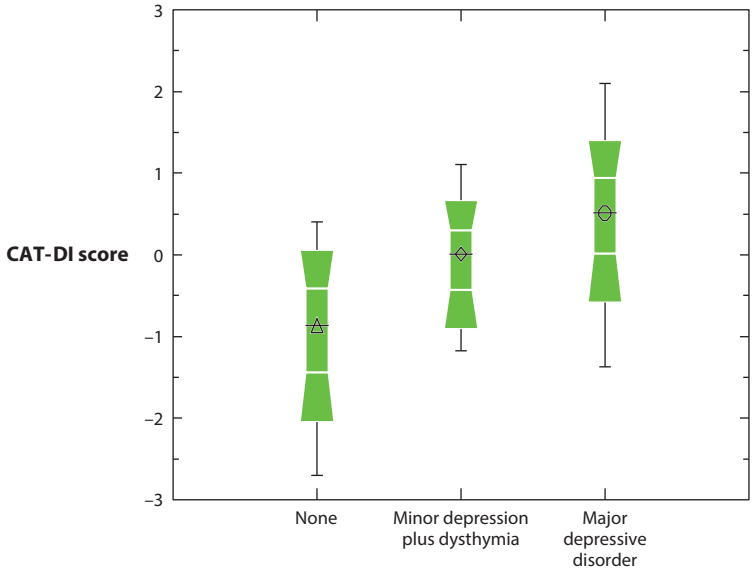


**Figure 2**

Box-and-whiskers plot for Computerized Adaptive Testing-Depression Inventory (CAT-DI) depression scores.

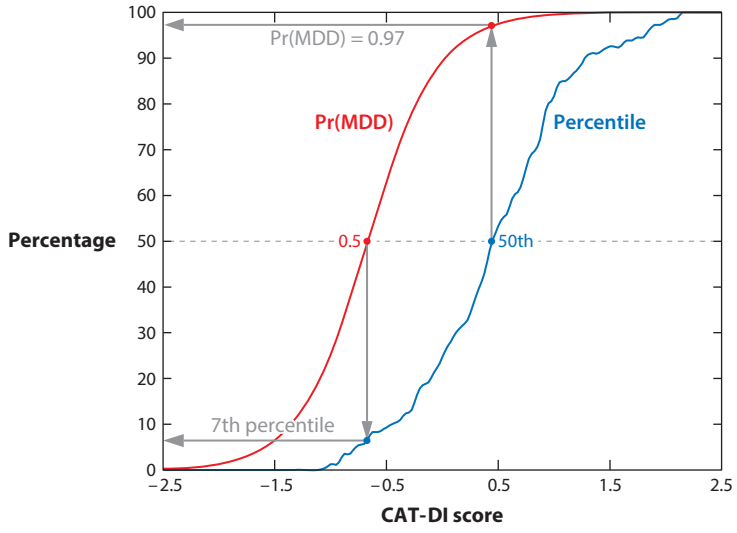


**Figure 3**

Percentile rank (Percentile) among patients with major depressive disorder (MDD) and probability (Pr) of MDD diagnosis. Abbreviation: CAT-DI, Computerized Adaptive Testing-Depression Inventory.

related to MDD diagnosis [odds ratio (OR) = 24.19, 95% CI 10.51–55.67, $p < 0.0001$]. A unit increase in CAT-DI score has an associated 24-fold increase in the probability of meeting criteria for MDD (**Figure 3**). **Figure 3** also presents the CAT-DI score percentile ranking for patients with MDD. A patient with a CAT-DI score of −0.6 has a 0.5 probability of meeting criteria for MDD but would be at the lower 7th percentile of the distribution of confirmed cases, whereas a score of 0.5 would have a 0.97 probability of MDD and would be at the 50th percentile of cases. Example adaptive testing session results are presented in **Table 1**.

Results for the CAT inventories for anxiety (CAT-ANX) (Gibbons et al. 2014) and mania (CAT-MANIA) (Achtyes et al. 2015) closely paralleled those for the CAT-DI. Using an average of 12 adaptively administered items, we found correlations of $r = 0.94$ and $r = 0.92$ for the total anxiety and mania item bank scores. For both anxiety and mania, there was a 12-fold increase in the likelihood of the corresponding DSM-5 (Am. Psychiatr. Assoc. 2013) disorder (generalized anxiety disorder or current bipolar disorder) from the low end to the high end of each scale.

## COMPUTERIZED ADAPTIVE DIAGNOSIS

The primary distinction between CAT and CAD is the use of an external criterion. In CAT there is no gold standard: The interrelationships among the items are used to define and measure the trait of interest. In CAD, the goal is to reproduce the gold standard diagnosis using an algorithm that reproduces the diagnostic classifier with a high degree of sensitivity and specificity. The gold standard may or may not be correct, but the CAD algorithm should nevertheless faithfully reproduce it. Whereas CAT is based on IRT, CAD is based on decision trees. Both methods are adaptive; however, they maximize different types of information functions.

Decision trees (Brieman 2001, Brieman et al. 1984, Quinlan 1993) represent a model in terms of a flow chart. Decisions are made by traversing the tree starting from the top node. At each node in the tree, a participant is asked to respond to a particular item. The participant progresses down the tree to the node to the left if his or her response is less than the cutoff value for the node; otherwise,

**Table 1 Item-by-item results for a subject with high severity[a]**

| Item | Response | Score | SE |
|---|---|---|---|
| 1. I felt depressed. | Most of the time | 0.474 | 0.621 |
| 2. Have you felt that life was not worth living? | Quite a bit | 0.810 | 0.551 |
| 3. Have you been in low or very low spirits? | Most of the time | 0.900 | 0.485 |
| 4. I felt gloomy. | Quite a bit | 0.917 | 0.437 |
| 5. How much have you felt that nothing was enjoyable? | Quite a bit | 0.951 | 0.424 |
| 6. How much were you distressed by blaming yourself for things? | Quite a bit | 0.973 | 0.384 |
| 7. How much were you distressed by feeling everything was an effort? | Quite a bit | 0.996 | 0.353 |
| 8. How often did you have negative feelings, such as blue mood, despair, anxiety, depression? | Often | 0.961 | 0.342 |
| 9. How much difficulty have you been having in the area of mood swings or unstable moods? | Quite a bit | 0.994 | 0.322 |
| 10. I could not get going. | Most of the time | 1.017 | 0.313 |
| 11. How much were you distressed by feelings of guilt? | Quite a bit | 1.029 | 0.302 |
| 12. I was unhappy. | Often | 1.028 | 0.299 |

[a]The subject had a score of 1.028, which corresponds to a probability of 0.995 of meeting criteria for major depressive disorder (MDD) and a percentile of 83.9% among patients with MDD. Score, 1.028; standard error (SE), 0.299.

the participant progresses to the right. The bottom node of the tree reports a classification for the participant (0 = nondepressed and 1 = depressed). Decision trees are appealing in this context because they allow the set of items presented to adapt to the responses already provided—going left at a node may result in a very different set of items being presented as compared to going right. This adaptation has the potential to considerably shorten the length of the instrument.

Despite their appeal, decision trees have frequently suffered from poor performance (Hastie et al. 2009) because algorithms used to build trees from data can exhibit sensitivity to small changes in the data sets that are provided. In contrast, ensemble models constructed of averages of hundreds of decision trees have received considerable attention in statistics and machine learning (Brieman 1996, 2001; Freund & Shapire 1996). Ensemble models provide significant improvements in predictive performance as compared to individual trees. However, averaging hundreds of trees destroys the adaptive testing structure that makes them appealing for the purposes of medical questionnaires.

In order to obtain both the advantages of individual trees and the accuracy of ensemble models, a combined approach is recommended. The first step is to fit a type of ensemble model called a random forest to the data. Random forests require minimal human intervention and have historically exhibited good performance across a wide range of domains (Brieman 2001, Hastie et al. 2009). The next step is to generate a very large artificial data set in which the items mimic the distribution of the items in the original data set. A single tree is then estimated on this artificial data set, with the intention of mimicking the output of the random forest as closely as possible while using enough data to reduce the sensitivity of the tree to small perturbations.

Trees of multiple depths (i.e., the maximum number of items administered; e.g., depth 6 and 11 items each) can be used in the analysis and compared in terms of sensitivity and specificity. For most subjects, fewer items are administered. Cross validation is performed either in an independent sample or by dividing the data into ten subgroups; in the latter, nine groups are used to build the model for each subgroup, and testing is performed on the tenth group.

## ILLUSTRATION

Gibbons et al. (2013) developed the first computerized adaptive diagnostic screening tool for depression that decreases patient and clinician burden and increases sensitivity and specificity for clinician-based DSM-IV diagnosis of MDD. A total of 656 individuals with and without minor and major depression were recruited from a psychiatric clinic, community mental health center, and through public announcements (controls without depression). The item bank consists of 88 depression scale items drawn from 73 depression measures. The focus of this study was the development of the CAD-MDD diagnostic screening tool based on a decision-theoretic approach (random forests and decision trees). An average of four items per participant was required (maximum of six items). Overall sensitivity and specificity were 0.95 and 0.87, respectively. For the PHQ-9, sensitivity was 0.70, and specificity was 0.91. As such, the CAD-MDD will identify more true positives (lower false negative rate) than the PHQ-9 while using half the number of items. Direct application of this work in primary care settings, psychiatric epidemiology, molecular genetics, and global health is inexpensive (relative to clinical assessment), efficient, and provides accurate screening of depression.

## INDEPENDENT VALIDATION STUDY

A new study was recently completed at Pine Rest Christian Mental Health Services outpatient clinics with 146 patients who received the CAT-MH and full Structured Clinical Interviews for the DSM (Achtyes et al. 2015). Sensitivity of 0.96 for MDD was found in the overall population, and specificity for differentiating patients with MDD from healthy controls was 1.00. The dimensional scales had similar ability to predict corresponding diagnoses over their range (28-fold increase for MDD and 12-fold increases for generalized anxiety disorder and bipolar disorder). The CAT for depression correlated well with the HAM-D ($r = 0.79$), PHQ-9 ($r = 0.90$), and CES-D ($r = 0.90$). Interestingly, 97% of patients indicated that the CAT-MH accurately reflected their mood, 86% preferred the computer interface to all alternatives, 97% felt comfortable taking the test, and 98% reported that they answered honestly. The entire battery of tests (depression, anxiety, mania, and MDD screener) required an average of 9.4 minutes to complete. The MDD screener alone required an average of 36 seconds to complete.

## DISCUSSION

Beyond the academic appeal of building a better and more efficient system of measurement, computerized adaptive testing of mental health constructs is important for our nation's public health. Approximately 1 in 10 primary care patients has major depressive disorder, and the presence of MDD is associated with poor health outcomes in numerous medical conditions. Rates of depression in hospitalized patients are even higher (10–20%), partially because depression increases hospitalization and rehospitalization rates (Whooley 2012). Unfortunately, clinicians often fail to identify depression in hospitalized patients. This is despite the existence of brief screening tools for depression, such as the (older) Hospital Anxiety and Depression Scale and the more widely used PHQ-9. To increase the likelihood that clinicians will perform some screening, even simpler approaches have been considered. For example, the American Heart Association recommends a two-stage screening method consisting of a two-item PHQ followed by the PHQ-9 for identifying depression in cardiovascular patients (Elderon et al. 2011). The method yields high specificity (0.91) but low sensitivity (0.52), indicating that it misses almost half of the patients with MDD.

By 2030, MDD is projected to be the number one cause of disability in developed nations and the second leading cause of disability in the world after human immunodeficiency virus and AIDS

(Mathers & Loncar 2006, Mitchell et al. 2009, Whooley 2012). Depression affects approximately 19 million Americans per year, or 10% of the adult US population (Natl. Inst. Mental Health 2013). Depression has human costs such as suicide, which ends 35,000 lives per year in the United States (Joiner 2010). Depressed people are 30 times more likely to kill themselves and 5 times more likely to abuse drugs (Hawton 1992). As discussed in a 2012 editorial in the *American Journal of Public Health* (Gibbons et al. 2012a), veterans have four times the risk of suicide relative to the general population during the first four years following military service. The ability to screen veterans for depression and suicide risk during this period of high risk and to refer them for appropriate treatment could be lifesaving. Depression is the leading cause of medical disability for people ages 14 to 44 (Stewart et al. 2003). Depressed people lose 5.6 hours of productive work every week when they are depressed (Stewart et al. 2003), and 80% of depressed people are impaired in their daily functioning (Pratt & Brody 2010). People who suffer from depression end up with six-tenths of a year less schooling, an 11% decrease in the probability of getting married, and a loss (on average) of $10,400 per year in income by age 50 (Smith & Smith 1982). The cost for the total group—over one's lifetime—is estimated at $2.1 trillion (Pratt & Brody 2010), and this does not include the increased cost of medical care that all of us must assume. Depression is a lifelong vulnerability for millions of people.

It is clear that health-care costs in general are dramatically higher for depressed patients and that integrated behavioral and physical health programs are cost effective. Recently, Bock et al. (2014) showed that in a cohort of 1,050 randomly selected multimorbid primary care patients ages 65–85 in eight German cities, mean costs were 8,144 euros in patients with depression and 3,137 euros in patients without depression. Comparison of an integrated physical and behavioral health-care system to usual care provided a cost savings of $3,363 per patient over a four-year period (Unützer et al. 2008). In a cost-effectiveness study (Pyne et al. 2003) of 12 primary care practices across 10 states, the mean incremental cost-effectiveness ratio was $15,463 per quality-adjusted life year in those practices receiving training in depression screening and treatment relative to usual care. The depression intervention compared favorably to other primary care interventions that were ultimately more expensive, such as mild hypertension ($28,552) and chronic obstructive pulmonary disease ($36,428). As a natural by-product of enhanced detection of depression, the need for quality mental health services will dramatically increase and integrated primary care and behavioral health-care programs will become commonplace.

The information obtained in only two minutes during the joint administration of the CAD-MDD and CAT-DI would take hours to obtain using traditional fixed-length tests and clinician DSM interviews. In contrast to traditional fixed tests, adaptive tests can be repeatedly administered to the same patient over time without response set bias because the questions adapt to the changing level of depressive severity. For the clinician, CAT provides a feedback loop that informs the treatment process by providing real-time outcomes measurement. For organizations, CAT/CAD provides the foundation for a performance-based behavioral health system and can detect those previously unidentified patients in primary care who are in need of behavioral health care and would otherwise be among the highest consumers of physical health-care resources. From a technological perspective, these methods can be delivered globally through the Internet and therefore do not require the patient to be in a clinic or doctor's office to be tested; rather, secure testing can be performed anywhere using any Internet-capable device (e.g., computer, tablet, smartphone). The testing results can be interfaced to an electronic medical record and/or easily maintained in clinical portals that are accessible by clinicians from any Internet-capable device.

The future direction of this body of research is immense. Screening patients in primary care for depression and other mental health disorders including risk of suicide is of enormous importance, as is monitoring their progress in terms of changes in severity during behavioral health treatment.

Of critical importance is diagnostic screening of mental health disorders such as depression, anxiety, and mania and attention-deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder in children based on adaptive self-ratings and parent ratings. A major priority should be applications of mental health CAT in military settings and among veterans who are at high risk of depression, posttraumatic stress disorder, and suicide. In genetic studies, the ability to obtain phenotypic information in large populations that can in turn be studied in terms of their genetic basis can and should be pursued. Global mental health applications should also be more rigorously understood and evaluated. Differential item functioning can be used to identify items that are good discriminators of high and low levels of depression in one language and culture but may not be effective differentiators in another. The same is true of patients identified for different indications: Somatic items are good discriminators of high and low levels of depression in a psychiatric population but may not be effective in a perinatal population or in patients presenting in an emergency department with comorbid and possibly severe physical impairments. A model of measurement based on multidimensional IRT allows for a much more sensitive understanding of these population-level differences in the measurement process and can provide rapid, efficient, and precise adaptive measurement tools that are insulated from real differences between cultures, languages, and diseases. This is the future of mental health measurement.



### SUMMARY POINTS

1. Computerized adaptive testing based on multidimensional item response theory can extract information from large item banks using a handful of optimally selected and adaptively administered symptom items.
2. Computerized adaptive diagnosis can reproduce trained clinician diagnoses in a small fraction of the time of a clinical interview (one minute versus one hour) with unprecedented high sensitivity and specificity.
3. Large item banks can be constructed that thoroughly measure a mental health construct of interest and can be administered adaptively in a small fraction of the time it would have taken to administer the bank as a fixed-length test.
4. By contrast to fixed-length short-form tests, in CAT there is no response-set bias due to repeated administration of the same items over time because CAT adapts to the changing illness severity of a patient and therefore uses different items on repeat administrations of the test.
5. Applications of this technology delivered through a cloud computing environment are widespread, ranging from screening and monitoring in integrated primary and behavioral health care, to screening entire countries in psychiatric epidemiology, to providing rapid and large-scale phenotypic information for genome-wide association studies.

## DISCLOSURE

The authors are founders of Adaptive Testing Technologies, which distributes the CAT-Mental Health. E.F. receives royalties from the American Psychological Association and Guilford Press; has equity interest in Health Rhythms and Psychiatric Assessments Inc.; serves on the Valdoxan Advisory Board; and owns the copyright for the Pittsburgh Sleep Quality Index.

## ACKNOWLEDGMENTS

This work was supported by grant R01-MH66302 from the National Institute of Mental Health.

## LITERATURE CITED

Achtyes ED, Halstead S, Smart L, Moore T, Frank E, et al. 2015. Validation of computerized adaptive testing in an outpatient non-academic setting: the VOCATIONS Trial. *Psychiatr. Serv.* 66:1091–96

Am. Psychiatr. Assoc. 1994. *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: Am. Psychiatr. Publ. 4th ed.

Am. Psychiatr. Assoc. 2013. *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: Am. Psychiatr. Publ. 5th ed.

Andrich D. 1978a. A rating formulation for ordered response categories. *Psychometrika* 43:561–71

Andrich D. 1978b. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Appl. Psychol. Meas.* 2:581–94

Andrich D. 1988. The application of an unfolding model of the IRT type to the measurement of attitude. *Appl. Psychol. Meas.* 12:33–51

Baek SG. 1997. Computerized adaptive testing using the partial credit model for attitude measurement. In *Objective Measurement: Theory into Practice*, ed. M Wilson, G Engelhard Jr, K Draney, pp. 37–52. Norwood, NJ: Ablex

Baker FB. 1992. *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker

Bock JO, Luppa M, Brettschneider C, Riedel-Heller S, Bickel H, et al. 2014. Impact of depression on health care utilization and costs among multimorbid patients—results from the MultiCare Cohort Study. *PLOS ONE* 9:e91973

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

Brieman L. 1996. Bagging predictors. *Mach. Learn.* 24:123–40

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

Brieman L, Friedman JH, Olshen R, Stone C. 1984. *Classification and Regression Trees*. Belmont CA: Wadsworth

Brown JM, Weiss DJ. 1977. *An adaptive testing strategy for achievement test batteries*. Res. Rep. 77-6. Minneapolis: Univ. Minn., Dep. Psychol., Psychom. Meth. Program, Comput. Adapt. Test. Lab.

Dodd BG, DeAyala RJ, Koch WR. 1995. Computerized adaptive testing with polytomous items. *Appl. Psychol. Meas.* 19:5–22

Elderon L, Smolderen KG, Na B, Whooley MA. 2011. Accuracy and prognostic value of American Heart Association—recommended depression screening in patients with coronary heart disease: data from the Heart and Soul Study. *Circ.: Cardiovasc. Qual. Outcomes* 4:533–40

Embretson S, Reise S. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum

First MB, Spitzer RL, Gibbon M, Williams JB. 1996. *Structured Clinical Interview for the DSM-IV Axis I Disorders Clinician Version* (*SCID-CV*). Washington, DC: Am. Psychiatr. Publ.

Fliege H, Becker J, Walter OB, Bjorner JB, Burghard F, Rose M. 2005. Development of a computer-adaptive test for depression (D-CAT). *Qual. Life Res.* 14(10):2277–91

Freund Y, Shapire R. 1996. Experiments with a new boosting algorithm, machine learning. In *Proc. 13th Int. Conf. Mach. Learn.*, pp. 148–56. Burlington, MA: Morgan Kauffman

Gardner W, Shear K, Kelleher KJ, Pajer KA, Mammen O, et al. 2004. Computerized adaptive measurement of depression: a simulation study. *BMC Psychiatry* 4:13

Gibbons RD, Bock D, Hedeker D, Weiss D, Segawa E, et al. 2007. Full-information item bifactor analysis of graded response data. *Appl. Psychol. Meas.* 31:4–19

Gibbons RD, Brown CH, Hur K. 2012a. Is the rate of suicide among veterans elevated? *Am. J. Public Health* 102:S17–19

Gibbons RD, Hedeker D. 1992. Full information item bi-factor analysis. *Psychometrika* 57:423–36

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

Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, et al. 2012b. The CAT-DI: a computerized adaptive test for depression. *Arch. Gen. Psychiatry* 69:1104–12
Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, et al. 2014. Development of the CAT-ANX: a computerized adaptive test for anxiety. *Am. J. Psychiatry* 171:187–94
Hambleton RK, Swaminathan H. 1985. *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff
Hastie T, Tibshirani R, Friedman JH. 2009. *Elements of Statistical Learning*. New York: Springer
Hawton K. 1992. Suicide and attempted suicide. In *Handbook of Affective Disorders*, ed. ES Paykel, pp. 635–50. Edinburgh: Churchill Livingstone
Holzinger KJ, Swineford F. 1937. The bi-factor method. *Psychometrika* 2:41–54
Joiner T. 2010. *Myths about Suicide*. Cambridge, MA: Harvard Univ. Press
Joreskog K. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34:183–02
Kingsbury GG, Weiss DJ. 1980. *An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests.* Res. Rep. 80-5. Minneapolis: Univ. Minn., Dep. Psychol., Psychom. Meth. Program, Comput. Adapt. Test. Lab.
Kingsbury GG, Weiss DJ. 1983. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. See Weiss 1983, pp. 257–83
Mathers CD, Loncar D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *Public Libr. Sci. Med.* 3:e442
McBride JR, Martin JR. 1983. Reliability and validity of adaptive ability tests in a military setting. See Weiss 1983, pp. 223–36
Mitchell AJ, Vaze A, Rao S. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet* 374:609–19
Muraki E. 1990. Fitting a polytomous item response model to Likert-type data. *Appl. Psychol. Meas.* 14:59–71
Muthén BO. 1989. Latent variable modeling in heterogeneous populations. *Psychometrika* 54:557–85
Natl. Inst. Mental Health. 2013. Major depressive disorder among adults. **http://www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml**
Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D. 2011. PROMIS Cooperative Group. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. *Assessment* 18:263–83
Pratt LA, Brody DJ. 2010. Depression and smoking in the U.S. household population aged 20 and over, 2005–2008. *NCHS Data Brief* 34:1–8
Pyne JM, Rost KM, Zhang M, Williams DK, Smith J, Fortney J. 2003. Cost-effectiveness of a primary care depression intervention. *J. Gen. Intern. Med.* 18:432–41
Quinlan R. 1993. *C4.5: Programs for Machine Learning*. Burlington, MA: Morgan Kaufmann
Reise SP, Waller NG. 1991. Fitting the two-parameter model to personality data. *Appl. Psychol. Meas.* 15:45–58
Samejima F. 1969, Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* 17:1–68
Seo DG, Weiss. 2015. Best design for multidimensional computerized adaptive testing with the bifactor model. *Educ. Psychol. Meas.* doi: 10.1177/0013164415575147. In press
Smith JP, Smith GC. 1982. Long-term economic costs of psychological problems during childhood. *Soc. Sci. Med.* 71:110–15
Stewart WF, Ricci JA, Chee E, Hahn SR, Morganstein D. 2003. Cost of lost productive work time among US workers with depression. *JAMA* 289:3135
Stuart A. 1958. Equally correlated variates and the multinormal integral. *J. R. Stat. Soc. Ser. B* 20:373–78
Tucker LR. 1958. An inter-battery method of factor analysis. *Psychometrika* 23:111–36
Unützer J, Katon WJ, Fan MY. 2008. Long-term cost effects of collaborative care for late-life depression. *Am. J. Manag. Care* 14:95–100
Vale CD, Weiss DJ. 1984. *A rapid item-search procedure for Bayesian adaptive testing*. Res. Rep. 77-4. Minneapolis: Univ. Minn., Dep. Psychol., Psychom. Meth. Program, Comput. Adapt. Test. Lab.
Weiss DJ, ed. 1983. *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing*. New York: Academic

Weiss DJ. 1985. Adaptive testing by computer. *J. Consult. Clin. Psychol.* 53:774–89
Weiss DJ, Kingsbury GG. 1984. Application of computerized adaptive testing to educational problems. *J. Educ. Meas.* 21:361–75
Weiss DJ, McBride JR. 1984. Bias and information of Bayesian adaptive testing. *Appl. Psychol. Meas.* 8:272–85
Whooley M. 2012. Diagnosis and treatment of depression in adults with comorbid medical conditions. *JAMA* 307:1848–57

Annual Review of
Clinical Psychology

Volume 12, 2016

# Contents

**Indexes**



**Errata**

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at http://www.annualreviews.org/errata/clinpsy