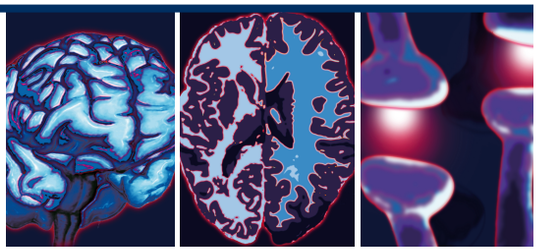
REVIEW



Psychometric considerations of depression symptom rating scales

Shawn M McClintock[1,2], Charlotte Haley[3] & Ira H Bernstein*

Practice points

- Depression is a condition that is inferred from its symptoms and signs rather than a condition that can be observed directly such as a fracture, elevated body temperature or blood pressure.
- As such, depression can be viewed as a latent variable, inferred variable or construct rather than an observable variable.
- Latent variables have been commonly studied in psychometrics using what is known as classical test theory.
- More recently, various forms of modern test theory, in particular item response theory (IRT), have proven particularly useful in articulating features of tests used to infer depression.
- An IRT model developed by Samejima is particularly useful with polytomous (multipoint) scales, such as Likert scales, as opposed to scales that use binary responses.
- One of the several useful aspects of IRT is that it can be used to equate scales (i.e., provide a formal basis for saying that a score of X on test A corresponds to a score of Y on test B).
- This paper contrasts various scales used to infer depression with regard to their internal properties such as reliability and dimensionality.
- A key point is that reliability and related statistics are a function of the sample in which they are inferred (i.e., there is no such thing as 'the reliability of test X').
- Although it is important to ensure that the various scales measure depression in a meaningful sense, their results are more similar than different because of their similar item content. Because of this and space limitations, we will not be concerned with this issue of validity.

[1]Division of Psychology, Department of Psychiatry, The University of Texas Southwestern Medical Center, Dallas, TX 75390-8898, USA
[2]Department of Psychiatry, Columbia University/New York State Psychiatric Institute, 1051 Riverside Drive, New York, NY 10032, USA
[3]Department of Clinical Sciences, Duke-NUS Graduate Medical School-Singapore, 8 College Road, 169857, Singapore
*Author for correspondence: Department of Clinical Sciences, The University of Texas Southwestern Medical Center, Florence Bioinformation Center, Suite E506, 5323 Harry Hines Blvd, Dallas, TX 75390-9066, USA; Tel.: +1 214 648 9543; Fax: +1 214 648 3934; ira.bernstein@utsouthwestern.edu

part of future Medicine fsg

Practice points cont.

- One important distinction among the scales is whether depression is inferred from self-report or by clinical observation.
- The essential point is that a clinician has a choice of a number of useful, albeit imperfect scales to infer depression that will produce similar results.

SUMMARY A major difficulty associated with the diagnosis of major depressive disorder (MDD) is that it cannot be unequivocally tied to observable bodily functions in the sense that a condition like hypertension can be described in terms of elevated blood pressure. Thus, MDD exemplifies what is called a 'latent variable', 'inferred variable' or 'construct'. Creation of scales to measure MDD symptomatology in adult populations is a popular line of research. The traditional approaches to constructing and evaluating scales are known as classical test theory, but modern test theory – more specifically, item response theory – offers some advantages. One useful feature is the ability to equate a score of X on scale A with a score of Y on scale B. Both classical test theory and item response theory are examined in this review. Perhaps the main point is that scales for depression, whether based upon self reporting or clinician rating, use fairly similar item content so that they are more similar than different. Thus, the purpose of this paper is to review the general considerations involved in constructing depression scales, which applies to the many other latent variables studied in psychiatry, and to compare important factors regarding the utility of scales for research and clinical practice.

As perhaps in most of the biological and physical sciences, terms in medicine such as 'systolic blood pressure' and 'bodily temperature' can be defined in terms of a single, observable event. This does not necessarily mean 'unequivocal', since two health service providers with different acuities may derive different blood pressure readings from a given patient. The more critical point is that in much of psychiatry, similar to most of the social and behavioral sciences, terms such as depression are of necessity what are called 'latent variables', 'inferred variables' or 'constructs'. That is, the magnitude of a patient's depression is an inference based upon several signs and symptoms such as sadness or loss of interest. Such symptoms are not as directly observable as, for example, systolic blood pressure. The various symptoms of depression are given formal recognition in the DSM-IV-TR of the American Psychiatric Association [1], and the International Classification of Diseases, tenth edition, clinical modification (ICD-10-CM) of the WHO [2].

The fact that depressive symptomatology has to be some form of an aggregate has given rise to a fairly large number of different depression measurement scales, and it is the purpose of this paper to compare and contrast some of the more commonly used scales. To do so involves both a traditional approach to the analysis of scales (psychometrics) known as classical test theory (CTT) and a newer set known as modern test theory, specifically item response theory (IRT). Note that although it is common to use the term 'modern', these methods go back well over 50 years. What makes them 'modern' is their application in areas like psychiatry and the more recent availability of computers to implement these models [3]. Because responses to items on depression scales, similar to most other scales used in psychiatry, typically have several categories (e.g., Likert or other rating scales), and are thus polytomous rather than binary (dichotomous), a specific IRT model formalized by Samejima [4] is of particular use and will be discussed further in this paper. Andrich also developed a model for use with polytomous items [5]. Those interested in learning more about Andrich's use of the Rasch model may refer to De Ayala [6].

In the section that follows, the more widely used depression scales will be introduced. We will then describe the CTT and IRT criteria that can be used to evaluate them and summarize the major results of using these criteria. To anticipate, however, the fact that these scales tend to use similar items indicates that the results of administering two or more to the same individual will likely provide results that are more similar than they are different. However, research has

suggested that even minimal differences in the constitutional item content between depression measures can produce dissimilar measurement conclusions/outcomes, despite a high correlation between the scales, at times resulting in measuring slightly different constructs [7,8]. As such, it is important to consider the various properties of the scales as well as the populations for which they are intended. This article will focus on scales used in measuring major depressive disorder (MDD) symptomatology in adult populations.

Depression scales in adult populations

At the present time, there are no generally accepted biological markers of MDD, so depression symptom severity rating scales are examined to characterize the presenting depressive symptoms (see **Table 1**). There are a plethora of rating scales available that vary in terms of dimensionality and item content, recording method, use for general or specific populations and cost. Depression rating measures can include only those items unique to depression, which would produce unidimensionality, or they can include items that capture related constructs (e.g., anxiety), which would produce multidimensionality [9]. For example, both the Inventory of Depressive Symptomatology (IDS) [10] and the Hamilton Rating Scale for Depression (HAM-D) [11] are multidimensional scales as they measure depression and anxiety constructs [12]. On the other hand, the Quick Inventory of Depressive Symptomatology (QIDS) [13] and the Patient Health Questionnaire-9 (PHQ-9) [14] are unidimensional scales as they only measure the construct of depression. There are typically two reporting methods for depression rating scales: self-report or clinician-rated. The former involves the patient completing a depression rating scale by checking the presence or absence of symptoms, and the latter entails a semistructured interview by a clinician to capture the presenting symptomatological information. Some depression scales are available in both self-report and clinician-rated formats such as the IDS and QIDS, whereas others are only clinician-rated, such as the HAM-D, or self-report, such as the PHQ-9 or the Beck Depression Inventory-II (BDI-II) [15]. Fava *et al.* found that these different formats, either self-report or clinician-rated, of depression rating scales can provide different information [16]. Many depression rating scales can be used with most general populations with psychiatric illness, although some were developed for use with special populations. For example, the Edinburgh Postnatal Depression Scale (EPDS) [17] measures depressive symptoms in women who are pregnant or have given birth, and the Geriatric Depression Scale (GDS) [18] assesses depressive symptoms in elderly adults (aged 55 years and older). Lastly, some depression rating scales must be purchased for a specified fee (e.g., BDI-II), and others can be used at no cost (e.g., QIDS and HAM-D). Given the many depression symptom rating scales available for use in clinical and research practices, it is prudent to select a measure with excellent psychometric properties, and importantly, one that meets the needs of the respective environment.

Types of scales

Self-report versus clinician-rated

One of the more obvious differences among the scales is that some are designed for the patient to respond directly, whereas others have a clinician mediate the patient's responses. Both the self-report and clinician-rated methods provide valuable information to inform the quality and quantity of presenting depressive symptoms, and each have respective advantages and disadvantages. For example, there is an economic advantage to the self-report method in terms of time relative to the clinician-rated method.

One less obvious difference between the methods is the frame of reference that is likely to be used. For example, the self report of a patient's sadness is likely to reflect that person's comparison of his/her current versus usual feeling state, so the comparison is what is known as an ipsative (self-measuring) comparison. By contrast, a clinician is likely to compare how that person's sadness compares with other patients in a similar context, thus forming a normative comparison. An important difference between self-report and clinician-rated scales is that the patient is far less practiced at responding to a scale than the clinician. Nonetheless, there are studies that used the same general scales, such as the QIDS, which are published in both forms and have reasonably high, albeit imperfect, correlations. This is particularly true when measurement is made late in antidepressant treatment since the patient has had practice and gained experience at self-reporting depressive symptoms.

Methods of choosing items

For most depression severity measures, items are typically chosen to reflect the constructor's

Table 1. Psychometric properties of commonly used depression symptom rating scales.

Instrument name	Scale versions	Dimensionality	Number of items	Rating metric	Symptom domain content	Coefficient α (context)
Beck Depression Inventory-II (BDI-II)	Self-report	Multidimensional	21	Four-point scale (0, 1, 2, 3)	Cognitive, behavioral, affective, somatic	0.92 (outpatients) [54] 0.94 (primary care patients) [55] 0.90 (African–American general medical patients) [56]
Center for Epidemiological Study – Depression Scale (CES-D)	Self-report	Multidimensional	20	Five-point scale (0, 1, 2, 3, 4)	Depressive affect, somatic symptoms, positive affect, interpersonal relations	0.85 (community samples) [57] 0.90 (psychiatric samples) [57] Similar in racially diverse studies [58]
Edinburgh Postnatal Depression Scale (EPDS)	Self-report	Unidimensional	10	Four-point scale (0, 1, 2, 3)	Sadness, anxiety, suicide, self-blame, coping, joy	0.87 (community postpartum sample with major or minor depression) [17]
Hamilton Rating Scale for Depression (HAM-D)	Clinician-rated	Multidimensional	There are multiple versions of the HAM-D. For example, the 17-item version was the published first, but there are reports of six-, 21-, 24- and 28-item versions, among others	Three-point scale (0, 1, 2) and five-point scale (0, 1, 2, 3, 4)	Affective, somatization–anxiety, cognitive, suicide, insomnia, weight/appetite change, libido, sadness, psychomotor, obsessive–compulsive, paranoia, self-critical	0.76 (community sample) [59] 0.92 (psychiatric outpatients/community sample) [60] 0.83–0.89 (outpatients with major depressive disorder) [13,52] Note: higher internal consistency with structured interview versus unstructured [61]
Geriatric Depression Scale (GDS)	Self-report	Multidimensional	Original version: 30 items Short version: 15 items	Two-point scale (0, 1)	Lowered affect, somatic concern, cognitive complaints, functional impairments, feelings of discrimination, lack of future orientation, decreased self-esteem	0.94 (30-item; depressed inpatient elderly and nondepressed elderly community sample) [18] 0.80 (15-item; elderly primary care sample) [62] 0.88 (30-item); 0.79 (15-item; nursing home patients) [63]
Inventory of Depressive Symptomatology (IDS)	Clinician-rated, self-report	Multidimensional	30	Four-point scale (0, 1, 2, 3)	Insomnia, sad mood, appetite/weight change, concentration, outlook, suicidal ideation, involvement, energy/fatigability, psychomotor function, anxiety, mood reactivity, mood quality, anhedonia, libido, self-criticalness	0.94 (depressed and nondepressed outpatients) [19] 0.90–0.92 (depressed outpatients) [13,64]
Montgomery–Asberg Depression Rating Scale (MADRS)	Clinician-rated	Multidimensional	10	Seven-point scale (0, 1, 2, 3, 4, 5, 6)	Sadness, tension, insomnia, decreased appetite, concentration, lassitude, anhedonia, pessimism, suicide	0.89 (elderly depressed and nondepressed clinical and community sample) [37]
Patient Health Questionnaire-9 (PHQ-9)	Self-report	Unidimensional	9	Four-point scale (0, 1, 2, 3)	Interest, appetite change, sleep change, sad mood, suicide, concentration, self-esteem, energy, self-critical, psychomotor disturbance	0.83 (depression in primary care patient sample) [65]

Table 1. Psychometric properties of commonly used depression symptom rating scales (cont.).

Instrument name	Scale versions	Dimensionality	Number of items	Rating metric	Symptom domain content	Coefficient α (context)
Quick Inventory of Depressive Symptomatology (QIDS)	Clinician-rated, self-report	Unidimensional	16	Four-point scale (0, 1, 2, 3)	Insomnia, sad mood, appetite/weight change, concentration, outlook, suicidal ideation, involvement, energy/fatigability, psychomotor function	0.86–0.87 (depressed outpatients) [13,52] 0.85–0.87 (elderly depressed and nondepressed clinical and community sample) [37]
Zung Self-Rating Depression Scale (ZSDS)	Self-report	Multidimensional	20	Four-point scale (1, 2, 3, 4)	Depression pervasive effect, physiological disturbances, other disturbances, psychomotor activities	0.79 (community sample) [66]

experience with depressed patients. For instance, Hamilton developed the HAM-D based on his clinical observations of patients who were treated for mood disorders on an inpatient unit [11]. During that time, it was believed that endogenous depression was more severe than exogenous forms, thus HAM-D items were differentially weighted (e.g., sad mood can get a maximum score of four whereas lack of insight can get a maximum score of two). An interesting development was Rush *et al.*'s use [19] of the DSM-IV-TR [1] in the construction of the IDS and QIDS, thus creating depression symptom inventories that reflect psychiatric diagnostic criteria. Using the DSM-IV-TR criteria as a platform to create a depression symptom inventory allows the creator(s) to develop a measure that is more generalizable to depressed populations as a whole, rather than creating a unique inventory that is only applicable to a limited depressed cohort.

Important test properties

Important test properties include item level properties, scale-level properties and dimensionality. These properties are defined differently depending on the analytic method used, such as CTT or IRT, but they essentially have the same underlying concepts.

■ Item-level properties

Two main item properties include difficulty and discrimination, the former term being a carry-over from the origins of psychometric theory in the assessment of skills. In this area, a difficult item, as its name implies, is one that a subject is unlikely to answer in the keyed direction. Translating this to the measurement of depression (with a certain linguistic difficulty), a 'difficult' item is a symptom that a patient is unlikely to endorse (e.g., paranoia, which is an item on the 24-item HAM-D) and an 'easy' item is one that a patient is likely to acknowledge (e.g., insomnia). By contrast, discrimination in both settings refers to how highly related the item (symptom in the case of depression) is to the trait (disease of depression).

■ Scale-level properties

Scale-level properties include the mean and standard deviation of the item set, but more importantly from a psychometric standpoint, the internal consistency reliability. The general definition of the latter is the ratio of true variance (actual variance in depression) to total

variance (variation in observed test scores) and it provides the relations among the items and their number. One vital point is that this is specific to a given sample – the more variable the sample, the higher the internal consistency reliability is likely to be. As a result, it is absolutely incorrect to refer to the internal consistency reliability of a test as a single number, although one can refer to its internal consistency reliability in a given context (e.g., a typical inpatient population). For an example of how reliability can vary across settings, being higher where variability is greater, using the Montreal Test of Cognitive Ability (MoCA) as an example, see Bernstein *et al.* [20].

In contrast to internal consistency reliability, there is also temporal stability reliability, which measures how highly related scale scores are at different points in time. Importantly, the two meanings are essentially unrelated, as a measure can be of high or low internal consistency independent of whether it is a trait-like or state-like measure.

■ Dimensionality

A third issue is whether the items measure one trait (unidimensional) or multiple traits (multidimensional). Ideally, a scale would be unidimensional so that its items would measure a single 'thing'. For example, if a scale measured both depression and anxiety and was to relate to some other event, it would be difficult to determine if the depression or anxiety were responsible. However, an important consideration here is what the other traits are. If the other trait were to reflect something substantive, such as anxiety, the situation would be as noted. However, suppose some items were to be phrased so that an affirmative response were to denote depression but others phrased so that a negative response were to denote depression, it is quite possible that this methodological difference would produce multidimensionality as admitting depression is not the same thing as denying its absence. For example, Carlson *et al.* found that item phrasing resulted in poorer psychometric properties for the Center for Epidemiologic Studies – Depression (CES-D) rating scale [21], and Dunbar *et al.* showed similar findings on a measure of global self-esteem [22]. In both studies, the negative-phrased items tended to form their own factor, leading to a multidimensional scale rather than a unidimensional scale. Nonetheless, there are useful reasons to incorporate both types of items, so this outcome is not necessarily undesirable. Similarly, if a scale (probably one that is clinician-rated) were to combine items that are essentially symptoms of depression (asking the person if they feel sad) with signs (does the person appear sad), the resulting multidimensionality would also not be problematic. Indeed, this may even enhance relations with external criteria, and thus there are benefits to multidimensional scales.

CTT

CTT is largely, but not completely, based upon observable rather than latent variables (for a comprehensive review of this topic, see Nunnally and Bernstein [23]). A person's estimated score is the obtained test score. For example, if a student completed a ten-item set of long division problems and correctly answered eight items, his/her estimated ability would be eight. However, this raw score of eight could be transformed in a variety of ways, such as into percentile ranks or z-scores. Underlying this is the recognition that a person could then be given a different set of ten items and perhaps obtain a different score. Indeed, it is this possible variation across different but equivalent forms of the test, relative to variation among different people, which forms the basis for the concept of internal consistency reliability. This measure reflects the number of items in the scale and their average intercorrelation. A common standard by which to judge how high this number should be is 0.8 for classifying individuals and 0.7 for research applications (group differences) [23].

The reliability coefficient is a major CTT test property and is usually expressed in terms of Cronbach's coefficient α. The computation of Cronbach's α reflects the basic theoretical definition of internal consistency as the ratio of true variance to total variance. Several measures to inform the psychometric characteristics of scales can be obtained from the reliability coefficient such as the standard error of measurement and confidence intervals.

One of the major advantages of CTT is that most of the relevant indices involve commonly studied concepts, even for those who have not formally studied psychometrics. Perhaps the most common measures of difficulty and discrimination are the item means and item-total correlation, the latter representing the correlation between the score on a given item and the total test score subtracting out the item in question, which is done to correct for the overlap produced by the presence of the item in the total score. The

item standard deviation or variance is also typically reported, although this is redundant with respect to the mean for binary items.

Scale dimensionality is typically inferred from several criteria produced from a principal component analysis. Perhaps the most common criteria is the Kaiser–Guttman rule [24–26], which identifies the dimensionality as the number of component eigenvalues greater than 1 (since 'λ' is a common abbreviation for an eigenvalue, it is also known as the 'λ > 1' rule). Eigenvalues are obtained from a rather technical process, but all that one needs know is that the first eigenvalue accounts for the most variance among the measures, the second accounts for the most variance that is left over after adjusting for the first, and so forth. The successive eigenvalues form what is known as the scree.

Although the Kaiser–Guttman rule is the default in many statistical software packages, more recent work stresses alternatives. An important alternative is parallel analysis [27–30], which is a process that involves generating the scree from data derived from random normal deviates (in practice, several sets of results are generated and averaged). Although parallel analysis is commonly associated with CTT analysis, it can be conducted using IRT modeling [23]. The dimensionality is the number of eigenvalues generated from the real data that exceed the randomly generated counterparts. Thus, if a scale is unidimensional, its first eigenvalue exceeds the first eigenvalue of the randomly generated data, but its successive eigenvalues are all smaller than their counterparts. **Figure 1** illustrates a unidimensional outcome comparing scree from the ten components obtained from a version of the Montgomery–Asberg Depression Rating Scale (MADRS) compared with a randomly generated scree. One of several alternatives to parallel analysis to fit a one-factor model is the use of confirmatory factor analysis, which has been used in the development of depression severity measures [31,32]. A model that meets this criterion is certainly unidimensional. However, models can misfit such as a model that meets the parallel analysis criterion for incidental reasons, as discussed by Carlson *et al.* [21]. For example, one common problem that can result in a confirmatory factor analysis model misfit is the non-normality of depressive symptomatology.

Limitations of CTT

The most commonly raised limitations of CTT are [23]:

- It is difficult to discriminate across groups (e.g., to evaluate gender differences);
- It is difficult to equate scores on different tests (i.e., have meaningful bases for saying that scores on different tests are equivalent);
- Item and scale statistics apply only to those taking the specific test or the population from which they were sampled;
- It is difficult to generalize to other groups or other forms of the test;
- Properties of the test are confounded with properties of people taking the test;
- Each item contributes equally to the test score;
- It is difficult to estimate how likely it is that a person will respond correctly to any given item;
- Items are assumed to be measured on the same interval scale;
- It is difficult to compare a person's score on different tests;
- Traditional equating procedures imply normal distributions, which are typically inapplicable, as most distributions do not meet criteria for normalcy;
- Errors of measurement are generally assumed to be equal throughout the continuum. Thus, while it is possible that two tests have the same reliabilities in a given sample, it is also possible that one test is more reliable among people scoring lower on the test and another is more reliable among people scoring higher on the test;
- People have to be presented with all of the items, as opposed to adaptive testing, where a limited subset are presented.

Advantages of CTT

One main advantage of CTT is that the relations of items with total test score is monotonic (the higher the response to a given item, the more seriously the person is depressed) rather than any specific form of relationship. Thus, CTT statistics can be conducted with small cohorts. The advantage of monotonicity is also applicable to Rasch models. For example, focusing on the construct of depression, when the data satisfy the assumptions of the Rasch model, the total raw score on a depression measure would be a sufficient statistic for the level of depression. For example, Cole *et al.* found monotonicity in

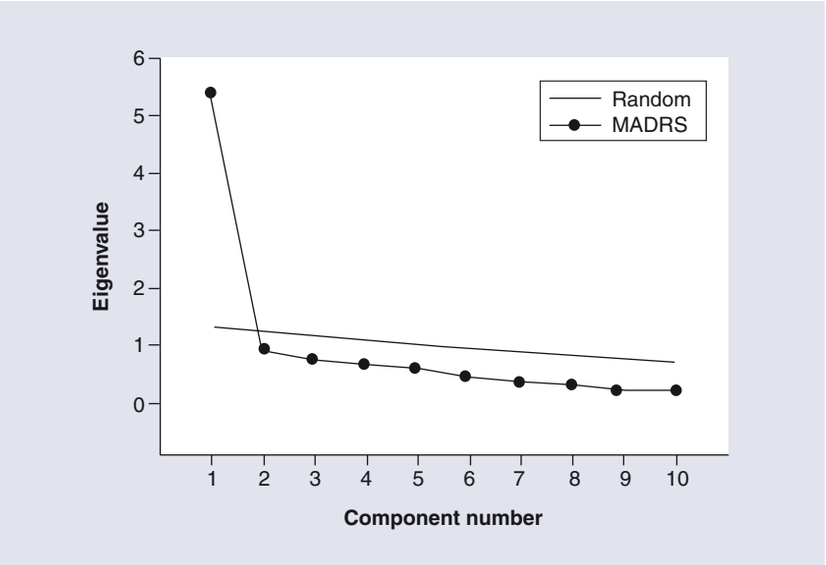


Figure 1. Scree plot example of unidimensionality. The comparison of a scree of the ten items of the MADRS with a randomly generated scree.
MADRS: Montgomery–Asberg Depression Rating Scale.

a short form of the CES-D scale using Rasch modeling [33].

IRT

Although there are many specific IRT models, two postulates are most common: unidimensionality, as noted above; and local independence, which means that the probabilities of answering two items correctly are independent once overall trait magnitude (depression in the present context) is controlled. As a result, trait magnitude is the only thing responsible for the tendency of items to correlate. This leads to two deductions: performance on scales can be predicted by a set of factors generically denoted 'Θ' (but referring specifically to depression in this article); and the relation between performance on a given item and Θ (depression) can be described by the item characteristic curve, item response function or item trace. Θ is normally scaled in z-score units. In other words, 50% of the sample falls at or below Θ = 0; 84% of the sample falls at or below Θ = +1; and so on. This does not depend upon which sample the item is given to or on the other items comprising the test. There are several widely used statistical software programs for IRT including Mplus (Muthén and Muthén, LA, USA), Multilog (Scientific Software International, IL, USA), Parscale (Scientific Software International) and RUMM (RUMM Laboratory, WA, USA). Orlando *et al.* developed a useful program for test equating [34]. For a general introduction to IRT, see De Ayala [6].

The item trace for a given item is most commonly expressed as a particular S-shaped curve known as a logistic ogive, although some earlier IRT models used very similar but mathematically less convenient cumulative normal curves. The logistic ogive function (see **Figure 2**) is described by two parameters: slope, usually symbolized as 'a', representing discrimination, and location, symbolized as 'b', representing difficulty. When estimated, standard errors are usually provided, allowing for tests of the null hypothesis that each is zero.

The necessity of two-parameter models

There are models that assert that all items have a common slope (i.e., equally discriminating). These include the Rasch and one-parameter logistic (which are different approaches to some authors and the same to others) [35]. These models can be quite attractive conceptually; for example, the person score is a sufficient estimate of depression. However, we will not consider them here. The reason is that all depression scales basically use symptoms and occasionally signs. Using a one-parameter model forces what we feel is the totally unacceptable assumption that all symptoms are equally discriminating; for example, difficulty sleeping is as diagnostic as sad mood or suicidal ideation [3]. For examples of the range of slopes found with common depression scales that have been subject to two-parameter analysis, see Bernstein *et al.* [36] and Doraiswamy *et al.* [37].

The Samejima model

The Samejima model for polytomous items asserts that an item with k alternative responses (e.g., the four response choices per item on the QIDS) provides k – 1 parallel functions (see **Figure 3**). As in the dichotomous case, the difficulty is symbolized as 'a'. The successive locations are symbolized 'b_0' and 'b_1' (generically 'b_i'). Two more important and related concepts are item and test information functions. Item information functions describe how sensitive that item is to slight changes in Θ over the dimension, and the test information function does the same for the score on the test as a whole. As implied earlier, this is similar to the CTT concept of internal consistency reliability, but it is a function of Θ rather than a constant (refer to **Figure 2**).

One of the most useful features of IRT is the ability to test differences in parameter estimates between groups or conditions. One way to do so is by measuring the fit provided by a model, which is distributed as a form of χ^2, commonly abbreviated G^2. The technique is to allow the estimate(s) to vary freely between the two groups in a more general model and then constrain the estimate(s) to equality in a second model in which all other parameters are allowed to be the same. This latter model is said to be a nested form of the first model. If the G^2 is nonsignificant in the more general model and the sample size is sufficiently great (although unfortunately, this is not well defined), one can test the difference in G^2 for significance based upon comparing the number of parameter estimates. A nonsignificant difference implies that the parameter estimate(s) are equal. As a rule of thumb, it is common to see group differences in location (difficulty), which are equivalent to mean differences in CTT scores, but it is less common to see differences in discrimination. Thus, one may see two groups differ in how sad they are, but typically not see a difference in how strongly sadness relates to overall depression in the groups. Group differences in any of these parameters are commonly referred to as differential item functioning (DIF). Note that there are useful methods for evaluating DIF within a CTT context (see Nunnally and Bernstein [23]).

Test equating is handled relatively easily if it can be assumed that the group taking test A is equivalent to the group taking test B (as when the same individuals are given both tests). Observed scores on each are linked to values of Θ from which they can be linked to each other. In other words, if a raw score of five on test A implies $\Theta = 0$ and a raw score of seven on test B also implies $\Theta = 0$, the two raw scores are considered equated (in practice, equatings are often approximate). For three recent applications of test equating to depression measures, see Carmody *et al.* [38], Bernstein *et al.* [39] and Bernstein *et al.* [36].

■ Components of IRT analysis

An IRT analysis [23,40] of a depression scale usually consists of some or all of the following components: parameter estimates (usually slope and intercepts); test information function (a plot of the ability of the test to detect differences in depression as a function of depression magnitude); analysis of DIF relevant to groups of interest; and equating different tests or forms of the same test (indicating which scores describe the same latent trait magnitude; e.g., showing that

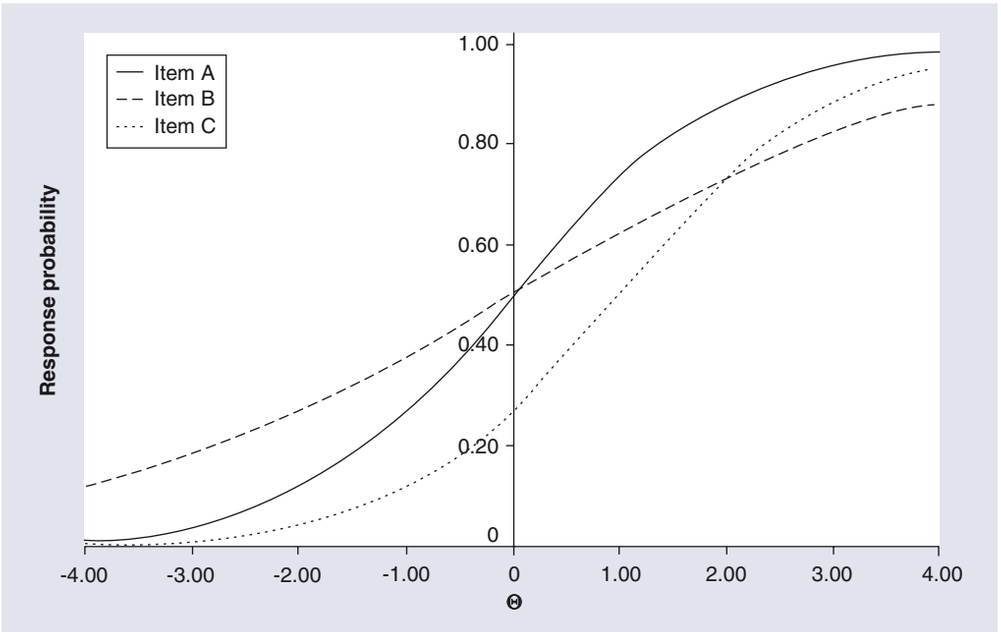


Figure 2. Operating characteristic curves. Three operating characteristic curves as would be used in the relatively simple case of a binary response to three questions like 'I am sad'. Note that the functions for items A and B differ in slope, the item response theory measure of discrimination. Conversely, items A and C differ in location, the item response theory measure of difficulty.

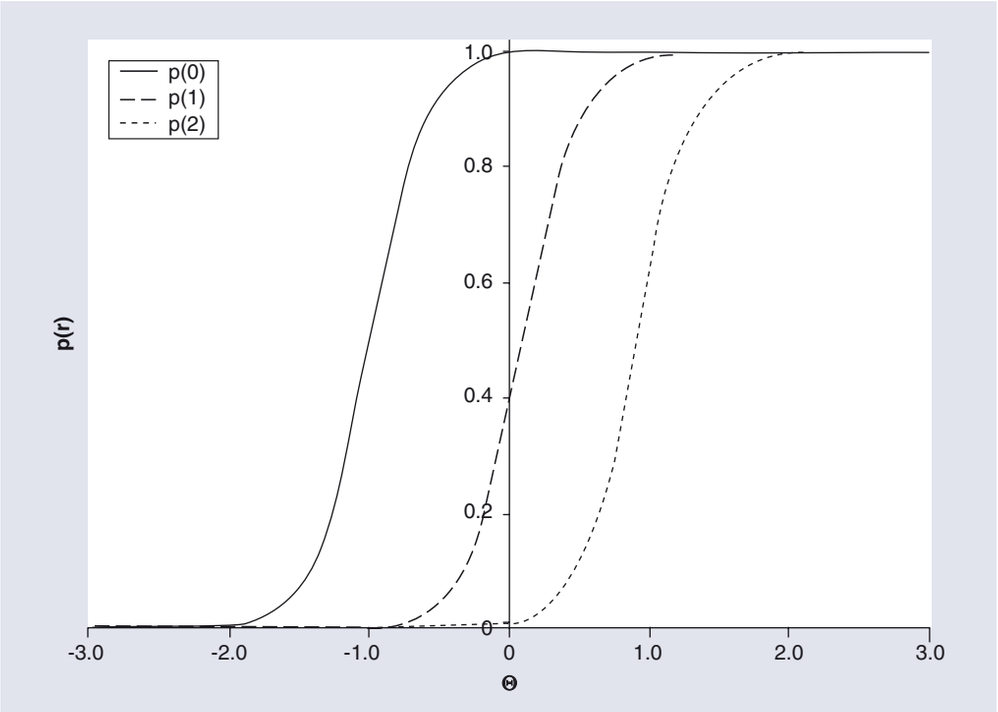


Figure 3. Polytomous functions based on the Samejima model. Polytomous functions for item two (sad mood) of the Quick Inventory of Depressive Symptomatology (QIDS). The leftmost function (p[0]) dichotomizes responses into zero versus one, two or three; the middle function (p[1]) dichotomizes responses into zero or one versus two or three; and the rightmost function (p[2]) dichotomizes responses into zero, one or two versus three (assuming a 0–3 scale). Note, for example, that if a person falls approximately one standard deviation below the mean on sad mood (Θ), then that individual has approximately a 0.5 probability of choosing normal response 0 and a 0.5 probability of choosing one of the pathological categories in the sample. This is said to be at threshold for that item. Note that the probability of answering in the pathological direction is lower in a more normal sample, but the threshold difference between categories and between items in Θ units tends to remain the same.

a score of X on test A and a score of Y on test B both correspond to a depression z-score of -0.5 in the sample). This is usually accompanied by some analysis designed to demonstrate the scale's unidimensionality, such as a factor analysis.

■ Advantages & disadvantages of IRT

Since the concepts of IRT are less familiar than those of CTT, it is important to stress what one gains from the use of IRT. The main points are: the ability to compare groups explicitly with respect to item characteristics (difficulty and discrimination); the ability to evaluate what, in effect, is reliability at different points along the continuum (of depression in this case) instead of as a single composite measure; and the ability to equate tests in a formal manner. Conversely, perhaps the major disadvantage is that concepts like item trace intercept and slope are less familiar in the present context than item means and item/total correlations. Indeed, because the comparable CTT- and IRT-based constructs are usually strongly related, it may not be necessary to present both. Perhaps the main point is that CTT and IRT are complementary approaches; one should not think of IRT as replacing CTT.

Applications of CTT & IRT to depression measure development

Both CTT and IRT analytic techniques have respective advantages and disadvantages, and when combined, their convergence can result in the development of psychometrically informed depression measures. Examples of measures that have benefited from CTT and IRT analyses include, but are not limited to, the CES-D,

HAM-D, IDS and QIDS. The CES-D includes 20 items to assess depression in the general population [41]. In the original study of its development, CTT statistics found it to have a high coefficient α and coefficient of stability, as well consisting of four factors [41]. Despite having four factors, the author recommended using only one composite score. Later CTT analysis suggested that internal consistency was improved when two items were removed [42], and a recent IRT analysis suggested that the four reversed-scored items of the CES-D be altered to further improve the scale's internal consistency and unidimensionality [21]. Thus, CTT and IRT together helped to determine the dimensionality of the CES-D, alter item content and lessen the item content.

As the HAM-D was one of the first depression measures to be widely employed in clinical research [11], it has been substantially analyzed for its psychometric properties [43–45]. Through various CTT and IRT analyses, the HAM-D has been found to have multiple factors that resulted in the development of shorter forms, including a six-item version that measures a unidimensional construct of depression across depressive subtypes [46], and a six-item subscale that measures anxiety and somatization [47,48]. Interestingly, for the anxiety and somatization factor, no cut-off score was established to inform the point at which a person would be identified as having depression and anxiety, although in a recent investigation, IRT analyses resulted in the identification of such a cut-off score [12]. The culmination of the many psychometric analyses has led to a discussion regarding the different cut-off scores to define the presence and absence of depression [49] and whether the HAM-D provides more advantages or disadvantages in clinical research [50]. Since its introduction by Hamilton in 1960 [11], CTT and IRT statistics have informed different versions of the HAM-D, each containing various items corresponding to uni- and multi-dimensional versions, and unique cut-off scores to determine depression presence and severity, as well as absence.

Relative to the CES-D and the HAM-D, the IDS is a recently developed depression symptom measure that originally contained 28 items [51], but two additional items that measure atypical features of depression were added so that it would measure the depressive items as outlined by the DSM-IV [19] that are specific to the atypical features depressive subtype. An analysis of the IDS found it to be multidimensional with three specific factors (general depression, anxiety and atypical symptoms), suggesting that a different measure could be developed comprised of items regarding only the construct of depression [10], which resulted in the development of the QIDS [13]. This development of the QIDS from the IDS exemplifies the utility of CTT analyses. For example, the two atypical symptom items were added to the IDS to increase construct validity, but they did not contribute to the overall internal consistency, and in turn resulted in a new factor score that was independent of the overall general construct of depression. Thus, test developers require both theory and research in order to know what items to include in a depression measure, but they must perform psychometric analyses to inform, revise, and confirm the depression measure. Additional studies of the IDS and QIDS using IRT methods identified cut-off scores for depression severity [13,52] and select factor scores [12], equated test scores with other depression measures to allow for score conversions [52] and validated its use in select populations such as the elderly [37] and adolescents [53].

The CES-D, HAM-D, IDS and QIDS examples highlight how CTT and IRT analytic methods have helped depression measure developers to: select, retain and eliminate items to improve internal consistency and create uni- or multi-dimensional measures; identify optimal threshold scores for the detection of depression or other factors (e.g., HAM-D somatization/anxiety factor); and equate scores between measures to enhance test interpretation ability (e.g., a QIDS – Self-Report [QIDS-SR16] score of ten is equivalent to a 17-item HAM-D score of 13).

Conclusion & future perspective

As there is no current codified biologic marker of MDD, the field of psychiatry is dependent upon psychometrically sound instruments that measure depressive symptoms. These instruments help to establish the presence or absence of depressive symptoms and, importantly, the severity of the symptoms. Moreover, they can help to establish consistency between clinicians, and if there are identical clinician- and self-rated instruments, can do the same between clinicians and patients. The caveat to using a depression symptom rating instrument is that it should be constructed and evaluated based on CTT, IRT or both. It would be imprudent to use a depression symptom instrument that was not designed or evaluated with

those methods, as there is no established validity or reliability. Other important measurement topics to focus on when designing and assessing psychometrics of scales include reliability of a measure over time and sensitivity to measuring change. In conclusion, a depression scale should be selected based not on its popularity, but rather on its psychometric properties, while taking into consideration any practical limitations such as cost or time that may exist.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:

- ▪ of interest
- ▪▪ of considerable interest

1 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition, Text Revision)*. American Psychiatric Association, Washington, DC, USA (2000).

2 WHO. *International Classification of Diseases, Tenth Revision, Clinical Modification.* WHO Press, Switzerland (2011).

3 Streiner DL. Measure for measure: new developments in measurement and item response theory. *Can. J. Psychiatry* 55(3), 180–186 (2010).

▪▪ **Comprehensive review of item response theory.**

4 Samejima F. Graded response model. In: *Handbook of Modern Item Response Theory.* Van Linden W, Hambleton RK (Eds). Springer-Verlag, NY, USA, 85–100 (1997).

5 Andrich D. A rating formulation for ordered response categories. *Psychometrika* 43, 561–573 (1978).

6 De Ayala RJ. *The Theory and Practice of Item Response Theory.* The Guilford Press, NY, USA (2009).

7 Snaith P. What do depression rating scales measure? *Br. J. Psychiatry* 163(3), 293–298 (1993).

▪ **Provides constructive critique regarding the development of depression rating measures.**

8 Cameron IM, Crawford JR, Lawton K, Reid IC. Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br. J. Gen. Pract.* 58(546), 32–36 (2008).

9 Gullion C, Rush A. Toward a generalizable model of symptoms in major depressive disorder. *Biol. Psychiatry* 44(10), 959–972 (1998).

▪ **Comprehensive analysis of depression rating scale item content and the construction of factor scores and dimensionality.**

10 Rush AJ, Carmody T, Reimitz PE. The Inventory of Depressive Symptomatology (IDS): Clinician (IDS-C) and Self-Report (IDS-SR) ratings of depressive symptoms. *Int. J. Methods Psychiatr. Res.* 9, 45–59 (2001).

11 Hamilton M. A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* 23, 56–62 (1960).

12 Mcclintock SM, Husain MM, Bernstein IH *et al.* Assessing anxious features in depressed outpatients. *Int. J. Methods Psychiatr. Res.* (2011) (In Press).

13 Rush A, Trivedi MH, Ibrahim HM *et al.* The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol. Psychiatry* 54(5), 573–583 (2003).

14 Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. *J. Gen. Intern. Med.* 16(9), 606–613 (2001).

15 Beck AT, Steer RA. *Beck Depression Inventory.* The Psychological Corporation, TX, USA (1993).

16 Fava GA, Kellner R, Lisansky J, Park S, Perini GI, Zielezny M. Rating depression in normals and depressives: observer versus self-rating scales. *J. Affect. Disord.* 11(1), 29–33 (1986).

▪ **Describes the acquistion of unique information relative to the use of clinician-rated or self-report depression rating measures.**

17 Cox J, Holden J, Sagovsky R. Detection of postnatal depression. Development of the 10-item Edinburgh Postnatal Depression Scale. *Br. J. Psychiatry* 150(6), 782–786 (1987).

18 Yesavage JA, Brink TL, Rose TL *et al.* Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* 17(1), 37–49 (1982).

19 Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. *Psychol. Med.* 26(3), 477–486 (1996).

20 Bernstein I, Lacritz L, Barlow C, Weiner M, Defina L. Psychometric evaluation of the Montreal Cognitive Assessment (MoCA) in three diverse samples. *Clin. Neuropsychol.* 25(1), 119–126 (2011).

21 Carlson M, Wilcox R, Chou C-P *et al.* Psychometric properties of reverse-scored items on the CES-D in a sample of ethnically diverse older adults. *Psychol. Assess.* 23(2), 558–562 (2011).

22 Dunbar M, Ford G, Hunt K, Der G. Question wording effects in the assessment of global self-esteem. *Eur. J. Psychol. Assess.* 16(1), 13–19 (2000).

23 Nunnally JC, Bernstein IH. *Psychometric Theory.* McGraw-Hill, NY, USA (1994).

▪▪ **Classic textbook on psychometric theory and application.**

24 Guttman L. Some necessary conditions for common-factor analysis. *Psychometrika* 19(2), 149–161 (1954).

25 Kaiser H. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20(1), 141–151 (1960).

26 Kaiser H. A second generation little jiffy. *Psychometrika* 35(4), 401–415 (1970).

27 Horn JL. An empirical comparison of various methods for estimating common factor scores. *Educ. Psychol. Meas.* 25(2), 313–322 (1965).

28 Humphreys LG, Ilgen DR. Note on a criterion for the number of common factors. *Educ. Psychol. Meas.* 29(3), 571–578 (1969).

29 Humphreys LG, Montanelli RG. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivar. Behav. Res.* 10(2), 193–205 (1975).

30 Montanelli R, Humphreys L. Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: a Monte Carlo study. *Psychometrika* 41(3), 341–348 (1976).

31 Cole DA. Utility of confirmatory factor analysis in test validation research. *J. Consult. Clin. Psych.* 55(4), 584–594 (1987).

32 Morley S, Williams AC, Black S. A confirmatory factor analysis of the Beck Depression Inventory in chronic pain. *Pain* 99(1–2), 289–298 (2002).

33 Cole JC, Rabin AS, Smith TL, Kaufman AS. Development and validation of a Rasch-derived CES-D short form. *Psychol. Assess.* 16(4), 360–372 (2004).

34 Orlando M, Sherbourne CD, Thissen D. Summed-score linking using item response theory: application to depression measurement. *Psychol. Assess.* 12(3), 354–359 (2000).

35 Wright BD, Mok M. Understanding Rasch measurement: Rasch models overview. *J. Appl. Meas.* 1(1), 83–106 (2000).

36 Bernstein IH, Rush AJ, Stegman D, Macleod L, Witte B, Trivedi MH. A comparison of the QIDS-C16, QIDS-SR16, and MADRS in an adult outpatient clinical sample. *CNS Spectrums* 15, 458–468 (2010).

37 Doraiswamy PM, Bernstein IH, Rush AJ *et al.* Diagnostic utility of the Quick Inventory of Depressive Symptomatology (QIDS-C16 and QIDS-SR16) in the elderly. *Acta Psychiatr. Scand.* 122(3), 226–234 (2010).

38 Carmody TJ, Rush AJ, Bernstein I *et al.* The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur. Neuropsychopharmacol.* 16(8), 601–611 (2006).

39 Bernstein IH, Wendt B, Nasr SJ, Rush AJ. Screening for major depression in private practice. *J. Psychiatr. Pract.* 15(2), 87–94 (2009).

40 Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use (3rd Edition).* Oxford University Press, UK (2003).

41 Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psych. Meas.* 1(3), 385–401 (1977).

42 Zimmerman M, Coryell W. Screening for major depressive disorder in the community: a comparison of measures. *Psychol. Assess.* 6(1), 71–74 (1994).

43 Ruhe HG, Dekker JJ, Peen J, Holman R, Jonghe FD. Clinical use of the Hamilton Depression Rating Scale: is increased efficiency possible? A *post hoc* comparison of Hamilton Depression Rating Scale, Maier and Bech Subscales, Clinical Global Impression, and Symptom Checklist-90 scores. *Compr. Psychiat.* 46(6), 417–427 (2005).

44 Maier W, Philipp M, Heuser I, Schlegel S, Buller R, Wetzel H. Improving depression severity assessment – I. Reliability, internal validity and sensitivity to change of three observer depression scales. *J. Psychiatr. Res.* 22(1), 3–12 (1988).

45 Bech P, Allerup P, Gram LF *et al.* The Hamilton Depression Scale. *Acta Psychiatr. Scand.* 63(3), 290–299 (1981).

46 O'Sullivan RL, Fava M, Agustin C, Baer L, Rosenbaum JF. Sensitivity of the six-item Hamilton Depression Rating Scale. *Acta Psychiatr. Scand.* 95, 379–384 (1997).

47 Cleary P, Guy W. Factor analysis of the Hamilton Depression Scale. *Drugs Exp. Clin. Res.* 1, 115–120 (1977).

48 Overall JE, Rhoades HM. Use of the Hamilton Rating Scale for classification of depressive disorders. *Compr. Psychiat.* 23, 370–376 (1982).

49 Zimmerman M, Posternak MA, Chelminski I. Is the cutoff to define remission on the Hamilton Rating Scale for Depression too high? *J. Nerv. Ment. Dis.* 193(3), 170–175 (2005).

50 Zimmerman M, Posternak MA, Chelminski I. Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? *J. Clin. Psychopharmacol.* 25(2), 105–110 (2005).

51 Rush A, Giles DE, Schlesser MA, Fulton CL *et al.* The Inventory for Depressive Symptomatology (IDS): preliminary findings. *Psychiatry Res.* 18(1), 65–87 (1986).

52 Rush AJ, Bernstein IH, Trivedi MH *et al.* An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression. A sequenced treatment alternatives to relieve depression trial report. *Biol. Psychiatry* 59(6), 493–501 (2006).

53 Bernstein IH, Rush AJ, Trivedi MH *et al.* Psychometric properties of the Quick Inventory of Depressive Symptomatology in adolescents. *Int. J. Methods Psychiatr. Res.* 19(4), 185–194 (2010).

54 Beck AT, Steer RA, Brown GK. *BDI-II Beck Depression Inventory: Manual.* The Psychological Corporation, TX, USA (1996).

55 Arnau RC, Meagher MW, Norris MP, Bramson R. Empirical articles – psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychol.* 20(2), 112–119 (2001).

56 Grothe KB, Dutton GR, Jones GN, Bodenlos J, Ancona M, Brantley PJ. Validation of the Beck Depression Inventory-II in a low-income African American sample of medical outpatients. *Psychol. Assess.* 17(1), 110–114 (2005).

57 Radloff LS. The CES-D scale. *Appl. Psych. Meas.* 1(3), 385–401 (1977).

58 Roberts RE. Reliability of the CES-D scale in different ethnic contexts. *Psychiatry Res.* 2(2), 125–134 (1980).

59 Rehm LP, O'Hara MW. Item characteristics of the Hamilton Rating Scale for Depression. *J. Psychiatr. Res.* 19(1), 31–41 (1985).

60 Reynolds WM, Kobak KA. Reliability and validity of the hamilton depression inventory: a paper-and-pencil version of the Hamilton Depression Rating Scale clinical interview. *Psychol. Assess.* 7(4), 472–483 (1995).

61 Potts MK, Daniels M, Burnam MA, Wells KB. A structured interview version of the Hamilton Depression Rating Scale: evidence of reliability and versatility of administration. *J. Psychiatr. Res.* 24(4), 335–350 (1990).

62 D'Ath P, Katona P, Mullan E, Evans S. Screening, detection and management of depression in elderly primary care attenders. I: The acceptability and performance of the 15 item Geriatric Depression Scale (GDS15) and the development of short versions. *Fam. Pract.* 11(3), 260–266 (1994).

63 Jongenelis K, Pot AM, Eisses AM *et al.* Diagnostic accuracy of the original 30-item and shortened versions of the Geriatric Depression Scale in nursing home patients. *Int. J. Geriatric Psychiatry* 20(11), 1067–1074 (2005).

64 Trivedi MH, Rush AJ, Ibrahim HM *et al.* The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychol. Med.* 34(1), 73–82 (2004).

65 Cameron IM, Crawford JR, Lawton K *et al.* Assessing the validity of the PHQ-9, HADS, BDI-II and OIDS-SR-sub-1-sub-6 in measuring severity of depression in a UK sample of primary care patients with a diagnosis of depression: study protocol. *Prim. Care* 13(2), 67–71 (2008).

66 Knight RG, Waal-Manning HJ, Spears GF. Some norms and reliability data for the State–Trait Anxiety Inventory and the Zung Self-Rating Depression scale. *Br. J. Clin. Psychol.* 22, 245–249 (1983).