The Value of Item Response Theory in Clinical Assessment: A Review

Michael L. Thomas

Assessment 2011 18: 291 originally published online 19 July 2010
DOI: 10.1177/1073191110374797

Assessment
18(3) 291–307
© The Author(s) 2011
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191110374797
http://asm.sagepub.com
SAGE

The Value of Item Response Theory in Clinical Assessment: A Review

Michael L. Thomas[1]

Abstract

Item response theory (IRT) and related latent variable models represent modern psychometric theory, the successor to classical test theory in psychological assessment. Although IRT has become prevalent in the measurement of ability and achievement, its contributions to clinical domains have been less extensive. Applications of IRT to clinical assessment are reviewed to appraise its current and potential value. Benefits of IRT include comprehensive analyses and reduction of measurement error, creation of computer adaptive tests, meaningful scaling of latent variables, objective calibration and equating, evaluation of test and item bias, greater accuracy in the assessment of change due to therapeutic intervention, and evaluation of model and person fit. The theory may soon reinvent the manner in which tests are selected, developed, and scored. Although challenges remain to the widespread implementation of IRT, its application to clinical assessment holds great promise. Recommendations for research, test development, and clinical practice are provided.

Keywords

item response theory, assessment, clinical, measurement, psychometric theory, latent variable model

Among the world's least understood and ill-defined topics of study, the human psyche holds an eminent rank. Although there has been no shortage of imaginative theories to explain human thought, behavior, and emotion, scientifically based definitions, and quantifications of inherently subjective mental states have not always been common. In an effort to lay the foundations for accurate assessment and effective treatment of psychiatric disorders, researchers have turned to empirical, mathematically based models. Under the banner of psychometrics, a movement to refine the art and science of psychological measurement has been underway for more than a century. One particular branch of this movement, *item response theory* (IRT; see Embretson & Reise, 2000; Lord, 1980), has already significantly affected the development of commonly administered measures of ability and achievement (e.g., McGrew & Woodcock, 2001). The methodology and challenges for applications of IRT in clinical and personality domains have been reviewed elsewhere (Morizot, Ainsworth, & Reise, 2007; Reise & Waller, 2009). The purpose of the present review is to appraise the value of IRT in the measurement of psychopathology.

Historical and Quantitative Background

Basics of Item Response Theory

Psychometric models specify systems of mathematical relations between observed and unobserved variables. They should not be viewed as alternatives to semantic accounts of psychological events. Instead, such models serve to open scientific hypotheses to empirical examination. In the spirit of Karl Popper's (1964) promotion of risky predictions, mathematical models force researchers to test specific hypotheses. The empirical nature of scientific methodology invariably leads fields of study toward research focused on model development and evaluation. Refinement of psychological measurement through comprehensive models ought to result in better science.

Despite some early psychologists' reluctance toward modeling psychological constructs, the development of *classical test theory* (CTT; Spearman, 1904a) and *common factor theory* (Spearman, 1904b) led to the now widely held belief that traits or characteristics of an individual's psyche can be quantified—*psychometrics*. Lord and Novick's (1968) introduction of IRT, along with Rasch's (1960) treatment of probabilistic models in cognitive testing, served to increase precision in psychological measurement (see Bock, 1997). The key element in the importance of IRT is considered to be the development of models where characteristics of examinees and tests can be separated. Theorists have long

[1] Arizona State University, Tempe, AZ, USA

Corresponding Author:
Michael L. Thomas, Department of Psychology, Arizona State University, 950 S. McAllister, P.O. Box 871104, Tempe, AZ 85287, USA
Email: michael.t@asu.edu

acknowledged that something akin to a "psychometric grail" (Wright, 1980) exists in the development and understanding of measurement instruments independent of the object(s) of assessment (see Thurstone, 1928). Using a ruler to measure height, for example, should have the same meaning whether measuring an elephant or a pencil. IRT dissects the various components of a testing system in the underlying belief that any process is better understood when all relevant variables are accounted for. The humble and yet remarkably complex goal of IRT is to provide models that assign concrete values to the otherwise intangible qualities of the mind.

A collection of IRT models have been developed for this purpose; each is characterized by increasingly comprehensive systems of measurement. The majority of IRT models are *stochastic*; that is, examinees' responses (e.g., "True" vs. "False") are assumed to be probabilistic. Also, IRT is predicated on the existence of *latent variables*: constructs that cannot be directly measured, yet are inferred to exist based on peripheral associations among measurable qualities. The concept of a latent variable is well established in the minds of many psychologists, as factor analysis also assumes the existence of latent variables. Fundamentally, it must be assumed that latent variables can account for all observed covariation between test items. In other words, the observed relations between items should disappear given their association with one or more latent variables (i.e., local independence).

IRT models rest on the assumption that the probability of an examinee passing an item—where "passing" may refer to responding correctly or affirmatively—is a function of two sets of parameters: (1) their standing on the latent variable, the person parameter; and (2) the characteristics of the item, the item parameters. Similar to a familiar regression equation, the function is a theoretical proposition for how variables in a system are related. Unfortunately, the relation is not linear, and thus it is not possible to employ the typical linear regression form. However, the relation does tend to take on two lesser-known forms, the *normal ogive* (the integral or summation of the normal curve) and the *logistic ogive* (a sigmoid or "S"-shaped function known for modeling the exponential rate of natural growth followed by saturation). Examples of logistic ogives are presented in Figure 1. The *x*-axis represents a normally distributed latent variable (e.g., depression) with a mean of 0 and a standard deviation of 1 (i.e., standardized metric); however, it should be noted that nonnormal as well as nonparametric models can be accommodated within the IRT framework. The *y*-axis represents the probability of a particular response (e.g., "True"), which can range from 0 to 1. Figure 1 displays *item characteristic curves* (ICCs), graphs of the probability of passing items conditional on specific values of the latent distribution. In Figure 1, an individual with a latent variable score of 2 would have a .50 probability of answering Item 3 affirmatively, while a person with a latent variable score of 4 would have a .99 probability of answering Item 3 affirmatively. It should be noted that the probability of response is a function of both item and person parameters; however, the parameters themselves are not interdependent (beyond issues of scaling). Deviations in a person's standing on the latent variable will alter the response probability, but not item parameters.

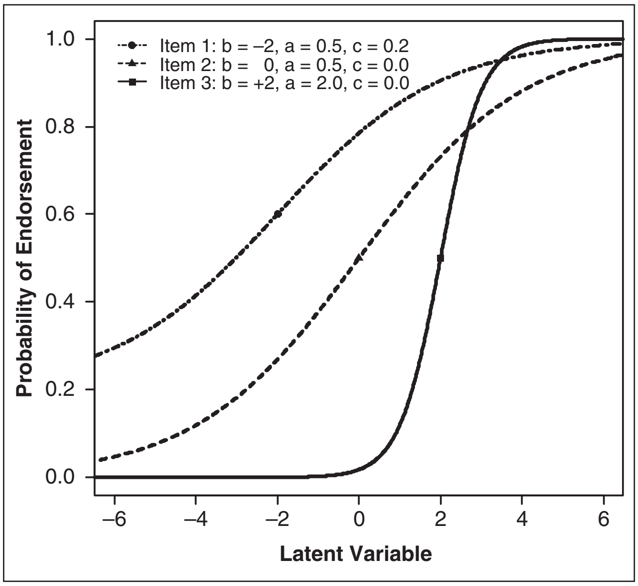

Figure 1. Item characteristic curves (ICCs) for three items with varying item difficulty (*b*), item discrimination (*a*), and item lower asymptote (*c*) parameters

The exact shape of these ICCs will vary along with the item parameters. An item's *difficulty parameter* (*b*) is the location of an ICC's point of inflection. For most IRT models, item difficulty is the value along the latent variable continuum at which an individual has a .50 probability of passing or affirming that item (this probability is altered for models that include asymptotes). From a clinical standpoint, "difficulty" can be thought of as "severity" (e.g., the severity of depression required to affirm the statement, "I've often thought of ending my own life."). In Figure 1, each item has a unique difficulty parameter; Item 1 is generally the least difficult (i.e., most likely to be endorsed) and Item 3 is generally the most difficult (i.e., least likely be endorsed). An item's *discrimination parameter* (*a*) is related to the slope of its ICC at its difficulty value (i.e., the point of inflection). Items with higher discrimination values (e.g., Item 3 in Figure 1) are more discriminating between distinct levels of the latent variable. Finally, the *lower asymptote parameter* (*c*), also known as the *pseudo-guessing parameter*, is so named to account for the fact that with some types of response formats (e.g., multiple choice tests), examinees can pass items simply by guessing. The pseudo-guessing parameter can be thought of as the probability of an examinee passing or endorsing an item when they are void of the latent variable. Item 1 in Figure 1 has a lower asymptote of .20. As can be seen, the probability of endorsement appears as though it will never dip below .20 (i.e., the ICC begins to flatten at this value).

More generally, item parameters from IRT are directly related to those from factor analysis with categorical variables (i.e., item thresholds and loadings). The models were developed in relative isolation, and hence take on superficially distinct forms (see Heinen, 1996); however, the equivalence of specific IRT models and factor analysis with categorical variables has been demonstrated (Takane & de Leeuw, 1987). Restated, under certain conditions, there is little difference between a confirmatory factor analysis and an IRT analysis. Nonetheless, IRT methodology is more mature with respect to item and scale analyses whereas confirmatory factor analysis methodology is more mature with respect to concurrent regression analyses (i.e., structural equation modeling).

Explanations of IRT are most easily understood for unidimensional scales. The three most common are the Rasch or one-parameter model, the two-parameter model, and the three-parameter model. Each successive model estimates more parameters and can be considered a more accurate representation of the data (i.e., a more comprehensive system). However, identification of the models and interpretation of their results becomes increasingly complex as parameters are added.

Item Response Theory Models

The *Rasch model* is a basic and elegant formulation of IRT. In the model, item difficulty is estimated separately for each item. A test may contain any combination of low-, medium-, and high-difficulty questions. However, the model demands that all items in a test have the same discrimination value. This implies that all items must have equivalent factor loadings and biserial correlations. The Rasch model is also sometimes referred to as the one-parameter model, as only item difficulty is allowed to vary. Creating a test composed of items with identical discrimination values is not a simple task. Despite this, the model has enjoyed a contingent of devoted theorists. This is largely because of a favorable mathematical property of Rasch models: Total scores are sufficient statistics for knowledge of latent variables. Thus, estimates of examinees' standings on latent variables based on simple, unweighted item sums are statistically sound. Extensions of the Rasch model to scales with polytomous response options (e.g., a Likert-type scale) include the *rating scale model* (Andrich, 1978) and the *partial credit model* (Masters, 1982).

The *two-parameter model* is a more general case of the Rasch model, and is most clearly aligned with common factor theory. It is often estimated using the logistic function, and thus is usually referred to as the two-parameter logistic (2PL) model. As with the Rasch model, item difficulty is estimated separately for each item. However, the two-parameter model also estimates unique item discrimination parameters. Because of this, interpretations of the two-parameter model are more ambiguous than interpretations of the Rasch model. Namely, total scores are not sufficient statistics for knowledge of latent parameters. Estimates of examinees' standings on latent variables based on simple, unweighted item sums are not statistically sound. Items must be weighted by discrimination parameters to estimate a person's standing on a latent variable. The two-parameter model does have the advantage of being more widely applicable than the Rasch model. Extensions to scales with polytomous response options include the *generalized partial credit model* (Muraki, 1992) and the *graded response model* (Samejima, 1969, 1997).

The *three-parameter model* is a more general case of the two-parameter model. As with the two-parameter model, the three-parameter model is often estimated using the logistic function and thus is typically referred to as the three-parameter logistic (3PL) model. The model adds the lower asymptote or pseudo-guessing parameter, which can be set to a constant or freely estimated for each item. As mentioned earlier, the pseudo-guessing parameter accounts for potential guessing. The model adds mathematical complexity to item parameter estimation. In addition, difficulty parameters do not have the same interpretation as they do with the Rasch and two-parameter models (specifically, the inflection point of the ICC will be greater than $p = .50$ when $c > 0$).

The models discussed so far have all rested on the assumption that a single latent variable accounts for the observed intercorrelations among items—unidimensional scales. This can be a limiting requirement. The need for multidimensional models in psychological assessment has long been recognized (e.g., Thurstone, 1947). Researchers and test developers who make incorrect assumptions of unidimensionality are either forced to remove misfitting items from tests or carry through with analyses despite the violations. This latter option is sometimes acceptable (i.e., when it does not drastically alter results). However, there are tests and items for which unidimensional models of latent variables simply do not accurately account for empirical data. In clinical assessment, for example, it can be difficult to create a depression item that is not also related to anxiety. A test developer may instead choose to model the probability of item responses based on multiple latent dimensions (see Reckase, 2009). Figure 2 presents the *item characteristic surface* for an item that is dependent on two latent variables: anxiety and depression. Notably, a third axis has been added

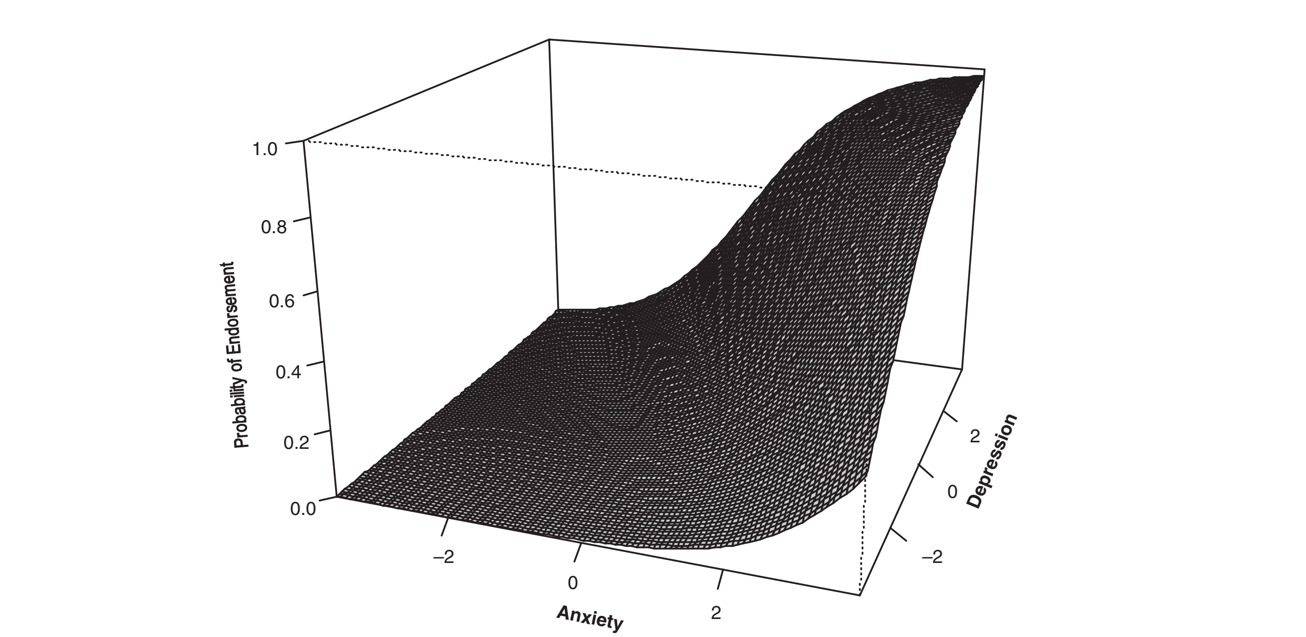

Figure 2. Item characteristic surface for a multidimensional item response model
Anxiety and depression are used as examples of two distinct latent variables that both influence the probability of item endorsement.

to the standard ICC in order to account for the second latent variable.

Nonparametric, nonmonotone, and multiple group IRT models have all emerged (see van der Linden & Hambleton, 1997). However, many of these extensions pose formidable challenges to applied researchers (e.g., complex formulas, lack of accessible software for performing the analyses, and unknown statistical properties). Technological and theoretical advances must be made before the wide array of IRT models becomes accessible.

Applications of Item Response Theory to Clinical Assessment

Model Selection

The fundamental role of an IRT equation is to model examinees' test response behavior. As previously reviewed, a number of IRT models can be selected for this purpose. In clinical assessment, Rasch models enjoy great popularity in Europe and have seen moderate use in the United States. However, because Rasch models demand identical item discrimination parameters, they often fail to fit scales developed with older technology (e.g., Tenenbaum, Furst, & Weingarten, 1985). Although Rasch models may be applicable to content-specific subscales couched within more complex frameworks (e.g., Bouman & Kok, 1987; Chang, 1996; Cole, Rabin, Smith, & Kaufman, 2004), they appear to be inappropriate for scales measuring psychological syndromes. Nevertheless, because of the beneficial properties of the model, fit of a Rasch framework should be given thorough consideration. Two-parameter models appear to be more congruent with existing clinical measures than their Rasch (one-parameter) counterparts (Reise & Waller, 1990). They have accurately reproduced observed data where Rasch models have failed (e.g., Aggen, Neale, & Kendler, 2005; Cooper & Gomez, 2008; Ferrando, 1994; Gray-Little, Williams, & Hancock, 1997). Owing to its greater flexibility and its congruence with common factor theory, the two-parameter model is more common in clinical assessment.

The three-parameter model has been applied to clinical assessment less commonly. Although the model adds flexibility to analyses, conceptualizing the impact of "pseudo-guessing" on items related to personality and psychopathology can be difficult. On clinical tests, the lower asymptote parameter has occasionally been thought of as being indicative of a response style (e.g., social desirability, true response bias, etc.; see Zumbo, Pope, Watson, & Hubley, 1997). For example, if examinees are unwilling to respond openly to an item concerning sexual practices, drug use, mental health, and so on, responses could be drawn toward more conservative options. Rouse, Finger, and Butcher (1999) fit a three-parameter model to scales from the second edition of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher et al., 2001) and found substantial correlations between estimates of lower asymptotes and indices of social desirability. Note, however, that this strategy assumes uniform response bias among examinees. Specifically, pseudo-guessing is an item parameter, not a person parameter. Therefore, the strategy cannot be used to differentiate between examinees with different response styles.

Reise and Waller (2003, 2010) expressed concern about interpreting the lower asymptote parameter as being related to response bias in the measurement of psychopathology. Specifically, the authors concluded that item extremity and nonsymmetric item ambiguity (i.e., item-level multidimensionality) likely cause both lower and upper asymptotes in clinical data—a four-parameter model. Item extremity applies to symptoms of psychopathology, such as visual hallucinations, that are not universally experienced by even the most severely disordered groups of patients (note that the reverse is true for symptoms of psychopathology with nonzero baselines). Nonsymmetric item ambiguity occurs when items take on different meanings for individuals who are high versus low on latent variables. Because of this complexity, researchers should use caution in interpretations of lower asymptotes. It would seem prudent to consider the lower asymptote as being indicative of response bias only when a researcher has substantive grounds to expect uniform misrepresentation among particular groups of examinees. For example, the assumption may be reasonable in the assessment of response bias related to cognitive ability (i.e., symptom validity testing).

Multidimensional IRT models represent an area of rapid growth in psychometric theory. Because of the models' overall complexity, however, relatively few researchers have applied multidimensional IRT in clinical domains. Yet such models have demonstrated improvements in measurement precision for both simulated and observed responses (e.g., Gardner, Kelleher, & Pajer, 2002; Gibbons et al., 2008; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007). Indeed, DeMars (2006) demonstrated that using a unidimensional model to estimate data with true multidimensional structure can lead to inaccurate parameter estimates. As such, multidimensional models would likely improve the measurement of psychiatric disorders comprising diverse and co-occurring symptoms of distress.

One of the more promising multidimensional models comes from Gibbons and Hedeker's (1992) IRT extension of the classic bifactor model. The *bifactor IRT model* is appropriate when examinees' observed responses are a function of their standing on a pervasive general latent variable as well as a series of domain-specific latent variables (for an extension to polytomous data, see Gibbons et al., 2007). Figure 3 depicts an example of a bifactor model where items are influenced both by a general internalizing (neurotic) latent variable and by a domain-specific somatization, depression, or anxiety latent variable. Bifactor models draw interest because of their ability to effectively account for residual correlations (i.e., local dependencies) in test data. In addition, the model boasts advantages with respect to predicting external outcomes in comparison with higher-order models (Chen, West, & Sousa, 2006). Early research into the applicability of the bifactor IRT model in clinical psychology has been promising (Gibbons et al., 2008; Gibbons, Rush, & Immekus, 2009; Simms, Grös, Watson, & O'Hara, 2008).

Multidimensional models have the added benefit of offering clinicians a glimpse into the underlying structure of psychological distress. Smits and De Boeck (2003), for example, used a multidimensional IRT model to identify three components contributing to the psychological experience of guilt: norm violation, worrying about what one did, and a tendency to restitute. Such explicit mathematical modeling can be used to enrich clinical descriptions of patients' symptoms. Mislevy, Levy, Kroopnick, and Rutstein (2008) note that the true value in modern psychometric theory lies in the ability to communicate increasingly complex psychological narratives. Although it cannot be denied that multidimensional models complicate clinical assessment, it seems clear that such complexity is not without purpose.

Nonparametric, nonmonotone, and multiple group IRT models have not been extensively researched in clinical domains. Some researchers have found that such models can provide good, if not better, fit for measures of personality and psychopathology (Meijer & Baneke, 2004; Roberson-Nay, Strong, Nay, Beidel, & Turner, 2007; Santor, Ramsay, & Zuroff, 1994; Stark, Chernyshenko, Drasgow, & Williams, 2006). These alternatives to traditional IRT models may become more common as the technology used to estimate their parameters becomes more available.

Reliability, Information, and Standard Error

The IRT-based concept of *information* is inversely related to standard error of measurement. Higher information equates with higher reliability, lower standard error, and more precise latent variable estimates. However, whereas the traditional CTT-based notion of reliability is assumed to be constant for all examinees, information is allowed to differ. Specifically, information as a function of the latent variable is called the *item information function*. Items are most informative at their difficulty parameter (i.e., when the probability of an examinee passing an item is .50). Intuitively, most would suspect that asking a kindergartner to solve a calculus equation would provide very little information about the child's achievement in mathematics; asking a college student to perform basic addition would be equally inappropriate. Questions that are too hard or too easy for examinees will provide little information about their ability.

The information of an entire measure is called the *test information function*. Unlike reliability, information is additive. Thus, an item's absolute contribution to a test is not

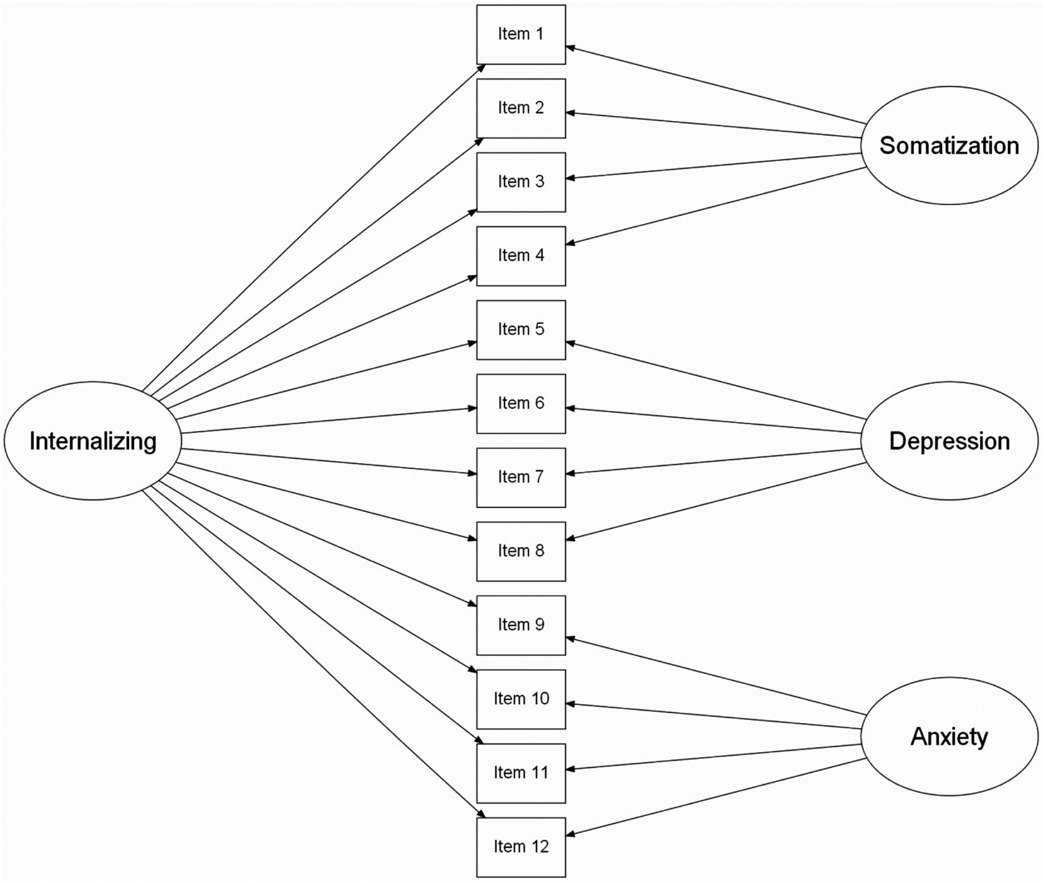

Figure 3. Bifactor model of psychopathology
Internalizing represents a general latent variable (psychological construct), whereas somatization, depression, and anxiety represent domain-specific latent variables

dependent on the group of items already contained in the test. The test information function is simply the sum of all item information functions. Figure 4 is an example of a test information function from a measure that is well-suited for use in populations with high severity of a disorder, but ill-suited for use in populations with low severity. The *standard error of measurement function*, also displayed in Figure 4, is inversely related to information.

Information functions are used to evaluate the precision of existing items (e.g., Marshall, Orlando, Jaycox, Foy, & Belzberg, 2002) and scales (e.g., Frazier, Naugle, & Haggerty, 2006). Young, Halper, Clark, and Scheftner (1992), for example, evaluated the Beck Hopelessness Scale and concluded that the test makes accurate latent variable estimates in the mid- to high spectrum, but is of little diagnostic value for individuals low on the construct. Researchers have found that the *Diagnostic and Statistical Manual of Mental Disorders*' criteria (American Psychiatric Association, 2000) for substance dependence (Langenbucher et al., 2004), depression (Aggen et al., 2005), and borderline personality (Feske, Kirisci, Tarter, & Pilkonis, 2007) all have highly peaked information functions. Indeed, researchers have found that many existing screening or diagnostic measures have information functions that peak near a cutoff on the impaired end of latent distributions (e.g., Cooke, Michie, Hart, & Hare, 1999). Such scales can only make reliable discriminations within narrow regions of latent distributions and are not appropriate for dimensional classifications of patients along entire continuums. In such instances, the scales might best be used to dichotomize individuals around these narrow regions of precision (much like passing vs. failing a driving test based on a single value of a continuous scale). Note, however, that this does not mean that the latent constructs themselves are categorical.

Diagnostic measures with peaked information functions can be viewed favorably in some circumstances. Indeed, if

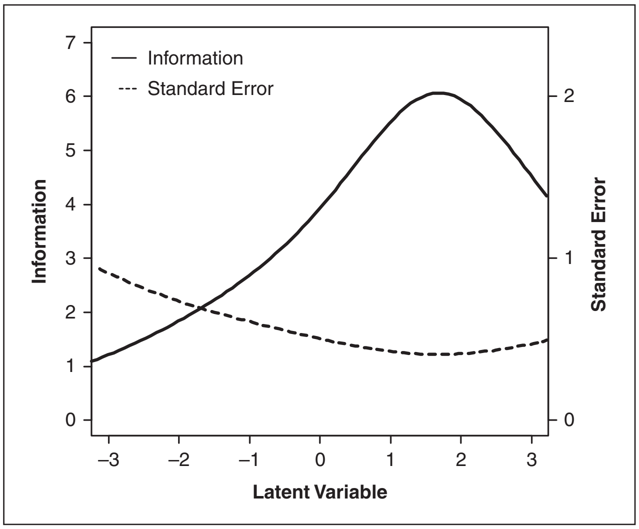

Figure 4. Test information and standard error functions

one were to develop a test used to classify examinees into impaired versus nonimpaired subgroups, a natural goal would be to maximize information at the chosen diagnostic cutoff (see Lord, 1980). Interestingly, test developers appear to have achieved this goal without the aid of IRT. This may be due, in part, to criterion-keying in test construction. Items have historically been chosen based on their point–biserial correlations with a criterion group variable (i.e., items endorsed frequently by the impaired group and infrequently by the nonimpaired group). Such items, whose difficulty parameters likely overlap near empirical criterion points (recalling that the difficulty parameter typically represents the location on a latent distribution where examinees become more, rather than less likely to endorse an item), would naturally produce peaked information functions in the impaired range of latent distributions. Thus, the use of criterion-keying in the development of psychological measures appears to be a practical, albeit somewhat imprecise, method for maximizing information at a diagnostic cutoff.

Related to this, tests can be reorganized into shorter and/or more informative versions by trimming away items that provide little information (e.g., Cooper & Gomez, 2008; Duncan-Jones, Grayson, & Moran, 1986; Grayson, 1986; Kim & Pilkonis, 1999). Test developers can choose items that will maximize information for a predetermined range of the latent variable based on need, law, or precedent. The methodology is similar to criterion-based test construction, where the criterion is a target information function (Hambleton & de Gruijter, 1983). *Mastery* or *screening tests*, for example, can be designed to provide peak information at a chosen threshold used to classify cases from noncases. Kessler et al. (2002) used item information to create a screening measure for psychological distress by choosing items that maximized information in the 90th to 99th percentile of the latent distribution.

Perhaps, the most opportunistic use of information functions comes through *computer adaptive tests* (CATs; e.g., Walter et al., 2007). CATs generate immediate estimates of examinees' standings on latent variables and choose subsequent items in a manner that will maximize information. This normally involves administering a slightly more difficult (severe) item when an examinee answers affirmatively and administering a slightly less difficult item when an examinee answers nonaffirmatively. To appreciate the strategy, consider a hypothetical situation where a clinical graduate student comes across his or her first depressed patient. After initial introductions, the student asks the patient, "Are you feeling hopeless?" to which the patient replies, "Yes, I am." And then, given the student's naïve understanding of depression, he follows up by asking the next question on his list, "And have you been feeling a little blue lately?" It does not take a veteran clinician to realize that a person who is hopeless is almost certainly "feeling a little blue lately." The question contributes almost no useful information to the assessment. In this instance, the student could have asked a more efficient follow-up question, for example, "Have you thought of ending your own life?"

CATs function like efficient clinical interviewers. They make ongoing estimates of examinees' standings on latent variables, and choose to administer items that will provide the greatest amount of information. Doing so can drastically reduce testing time and burden. For example, Waller and Reise (1989) examined the value of adaptive testing for the Absorption scale of the Multidimensional Personality Questionnaire and were able to reach accurate estimates of latent variables using on average only 25% of the original items. Kamakura and Balasubramanian (1989) did the same for the socialization subscale of the California Psychological Inventory using on average only 33% of the original items. Reise and Henson (2000) reduced item administration for the revised Neuroticism–Extroversion–Openness Inventory by half. Unfortunately, the aforementioned studies were all simulated CATs. That is, live adaptation of tests did not occur. In one of the few studies of live administration within a clinical domain, Simms and Clark's (2005) CAT for the Schedule for Nonadaptive and Adaptive Personality yielded greater psychometric information per administered item than did a traditional paper and pencil version of the measure. In addition, study participants preferred the CAT version of the measure primarily because of its reduced testing time.

Practical implementation of CATs in the assessment of psychopathology is not without barriers (see Revicki & Sloan, 2007); however, the cost- and time-effective nature of such testing will remain attractive to clinicians, insurance providers, and, perhaps most important, patients. The National Institutes of Health is currently funding an ambitious attempt to bring computerized adaptive testing to clinical assessment: the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS is a multisite collaborative research effort to standardize a large bank of physical and psychoemotional health items (Cella et al., 2007). Although the PROMIS system currently invites researchers to use developmental CATs in collaborative research efforts, the scales are not yet available for clinical practice at the time of this review.

Scaling and Equating

IRT's explicit measurement models facilitate meaningful scaling of item and person parameters. CTT-based total scores neither directly quantify psychiatric disorders nor are they directly related to the behavioral, cognitive, and emotional symptoms of distress. Nonetheless, a common assumption is that endorsement of more symptoms equates with higher likelihood of a disorder. That is, total scores are assumed to maintain ordinal properties with respect to latent variables. As such, percentile rankings are said to be the only "permissible statistic" that can be used to meaningfully summarize total scores (Stevens, 1946). In contrast, it can be argued that IRT approaches interval scaling of the relations between item parameters and person parameters (see Reise & Haviland, 2005); increases in the latent variable equate with additive or linear increases in the log-odds of item endorsement.

To understand why this is important, consider Juan's and Erica's hypothetical scores on a measure of depression (assuming a normally distributed dimensional variable). Within a CTT framework, the relation between Juan's and Erica's scores is typically described in reference to a normative population (or to each other). For example, because Juan's 99th percentile total score is higher than Erica's 7th percentile total score, we might conclude that Juan is more depressed. However, it is difficult to generalize beyond this comparison. In IRT, the difference between Juan's and Erica's latent variables can additionally be interpreted with respect to the behavioral, cognitive, and emotional symptoms related to the construct under investigation. Odds and/or probabilities of symptom endorsement can be provided because logistic and normal ogive models link latent variables with the probability of affirming various items. For example, we might conclude that Juan's standardized latent variable score of +2.50 equates with a 99% chance he is feeling sad, a 90% chance he is feeling hopeless, and a 30% chance he is contemplating suicide. On the other hand, Erica's standardized latent variable score of –0.50 equates with a 20% chance she is feeling sad, a 5% chance she is feeling hopeless, and a 1% chance she is contemplating suicide. We could even determine how much of a decrease in the latent variable is required to reduce Juan's suicidal ideation to 5%. Such meaningful descriptions of examinees' scores are not only useful in the diagnosis of psychiatric disorders, but also in the development of symptom-oriented treatment plans and the management of risk.

In research situations, simultaneous scaling of person and item parameters contributes to understanding of the relations between latent variables and items. For example, analyses of the Psychopathy Checklist–Revised revealed that individuals with a "lack of empathy" are more psychopathic than those with a "need for stimulation" (Cooke et al., 1999; Cooke & Michie, 1997). That is, endorsement of the interpersonal and affective features of psychopathy indicates greater severity of the disorder than does endorsement of the impulsive and antisocial features. Activities of daily living related to mobility, bathing, dressing, and eating are more easily impaired (i.e., have lower difficulty/severity values) than activities related to communication, bowel continence, and orientation (Teresi, Cross, & Golden, 1989). Analyses of the Beck Depression Inventory reveal that endorsement of items related to suicide requires greater severity of depression than does endorsement of items related to crying (Gibbons, Clarke, VonAmmon Cavanaugh, & Davis, 1985). Although such descriptive results can be approximated using CTT technology, IRT offers greater precision in scaling relations among variables. A mathematical relation linking the probability of suicidality given severity of depression, for example, is established in an IRT model.

IRT-based latent variables are meant to estimate constructs, not total scores. Therefore, two or more tests of the same construct can be calibrated on the same scale of measurement—a task made comparatively more difficult within CTT because of the nonindependence of person and item parameters. Test and item equating within IRT is simplified by only requiring that item parameters and person parameters for two or more tests of the same construct be calibrated on the same scale of measurement. By equating distinct tests of depression (e.g., Carmody et al., 2006; Orlando, Sherbourne, & Thissen, 2000) and general health (e.g., Martin et al., 2007), researchers have shown how to simplify the diagnostic process and improve accuracy. Such gains occur because the information provided by two or more distinct measures of the same construct are additive. Thus, instead of having two related but psychometrically distinct measures, IRT allows clinicians to combine information into a single estimate.

Item and Test Bias

Item parameters derived from distinct populations cannot always be calibrated on the same scale of measurement. This occurs when item parameters are population dependent, opening the door for item bias. Nonbiased items are those for which the probabilities of examinees from different populations affirming the items are equal when their standings on latent variables are equal. That is, examinees' memberships in specific populations should not alter their response probabilities. Items that do not maintain the above property are said to display *differential item functioning* (DIF); items that do are said to display *measurement invariance*. Examinations of DIF within an IRT framework are more precise due to separation of item and person parameters. Legitimate differences in group abilities can be accurately distinguished from test bias.

Examinations of DIF have become common in multicultural settings (e.g., Hui, Drasgow, & Chang, 1983; Pedraza et al., 2009). Hulin, Drasgow, and Komocar (1982), for example, found DIF between items in English and Spanish versions of a job satisfaction measure because of discrepancies in translation. Carle, Millsap, and Cole (2008), on the other hand, found that difference between boys' and girls' scores on a measure of children's depression could be attributed to legitimate group differences rather than DIF. IRT has the additional advantage of examining in great detail potential bias in the predication of outcomes (e.g., Leung & Drasgow, 1986). The identification of DIF allows researchers to account for discrepancies in their data and test developers to remove poor items from their scales. Moreover, reduction of measurement bias is an explicit ethical principle that psychologists subscribe to in their code of conduct (9.05; American Psychological Association, 2002). As such, clinical scientists and practitioners should employ all available methods for detecting and avoiding such bias in their work.

Longitudinal Research

Bias can also be found in the context of longitudinal within-group comparisons. Researchers who study change (e.g., because of time or treatment) must also be cautious of DIF (Horn & McArdle, 1992). Changes in latent variable estimates ought to be caused by true changes in examinees' standing on latent variables rather than changes in item parameters (i.e., lack of metric invariance) or changes in the structure of constructs relating to items (i.e., lack of configural invariance). IRT and confirmatory factor analysis are both valuable in the investigation of measurement invariance. However, confirmatory factor analysis may be better suited for the assessment of invariance in latent structure and IRT may be better suited for the assessment of invariance in item parameters (Meade, Lautenschlager, & Hecht, 2005). As mentioned earlier, estimates in IRT are directly related to those in factor analysis, but IRT presents a more explicit parameterization.

Millsap (2010) presents the methodology for testing measurement invariance in longitudinal data with IRT. As with between-group comparisons of DIF, within-group evaluations of longitudinal measurement invariance involve estimating item parameters separately and then testing for equivalence between measurement occasions. For example, Long, Harring, Brekke, Test, and Greenberg (2007) demonstrated the longitudinal construct validity of a screening measure for psychological distress by showing invariance for item parameters across repeated measurements. Meade et al. (2005) demonstrated the use of longitudinal IRT with a job satisfaction survey, finding DIF with respect to item difficulty (severity) across measurement occasions.

Model and Person Fit

One does not typically discuss model fit in the context of CTT. Indeed, CTT does not specify the relations between true scores and latent variables. Yet the study of reliability in CTT does involve rarely tested assumptions about true scores (i.e., strict parallelism, parallelism, or tau-equivalence). McDonald (1999) has demonstrated how such assumptions can be thought of as special cases of the Spearman single-factor model, an assumption that all items measure a common attribute. Thus, the fit of some underlying model should be evaluated in CTT, even though it is not standard practice. Researchers employing IRT models, on the other hand, commonly test model assumptions using global and local fit indices. As already mentioned, researchers have questioned whether unidimensional, parametric, and monotone IRT models are appropriate for clinical constructs. In IRT, all such assumptions can be examined with empirical data. If the assumptions are inconsistent with observed responses, quantitative comparisons should reveal the error (e.g., Stark et al., 2006).

An unavoidable limitation of all assessment instruments is the potential for erroneous diagnostic outcomes. However, although most psychologists are willing to accept that true scores or latent variable estimates are accurate only within the limits of standard error, a more concerning situation arises when the relations between item parameters and person parameters become systematically distorted for particular examinees. That is, there are some individuals for whom the IRT model simply will not fit. Analyses of *person fit* serve to identify examinees for whom the response model does not accurately predict their performance. Specifically, the strategy is like an analysis of DIF at the person level. Take, for example, a measure that fits a Rasch model. The likelihood of an examinee endorsing high-difficulty items while failing to endorse low-difficulty items is highly improbable given the model's structure. Deviant item response patterns suggest that the test is not accurately estimating examinees' latent variables. Possible explanations for such patterns include fatigue, poor effort, and cheating/misrepresentation.

Drasgow, Levine, and Williams (1985) developed a z-score index for maximum likelihood estimates (lz) that can be used to determine how deviant a response pattern is in comparison with an assumed normal distribution of response patterns. This is possible because the likelihood of latent variable estimates may differ even when the actual parameter estimates do not. The use of *person characteristic curves*, an alternative approach to examining person fit, can be used in conjunction with *lz* to explore the causes of aberrant responses (Nering & Meijer, 1998). The accuracy of such person fit statistics with actual data is mixed. Reise and Waller (1993) applied the *lz* person fit statistic to items from a personality measure and concluded that it does have potential for identifying aberrant response styles, but was too unreliable to provide meaningful moderation of other variables. Zickar and Drasgow (1996) used person-fit algorithms to assess misrepresentation on personality tests and found limited success. Ferrando and Chico (2001) compared IRT person-fit analyses of misrepresentation to traditional measures of misrepresentation (i.e., response bias scales) and found the IRT person-fit approach to be less accurate.

Perhaps the limitation with the approach is that it can only be used to identify sequentially improbable response options. If an examinee responds to a measure in an unusually exaggerated manner, but endorses items in the correct sequence, the person-fit statistics will not identify the response style as aberrant. In the extreme case, the response pattern of an examinee who endorses 59 of 60 items from the Infrequency scale (*F*) on the MMPI-2 would not be considered aberrant even though the *T*-score for such a pattern would literally be off the page. Nonetheless, whether indicative of misrepresentation or not, aberrant responding does lead to poor classifications of examinees (Frost & Orban, 1990; Hendrawan, Glas, & Meijer, 2005).

Recent work on the use of multidimensional IRT models that account for poor person fit (e.g., Bolt & Johnson, 2009) may help to nullify the detrimental effects of response bias. Such models offer to produce estimates of examinees' standings on latent variables that are less influenced by subjective responding by incorporating the effects of response styles into the overall probability of item endorsement. These analyses seem particularly applicable to the assessment of personality and psychopathology, where symptoms of distress can be interpreted uniquely by individual examinees. In addition, when used to assess deliberate response bias (i.e., malingering), multidimensional IRT models can be used to account for the effects of purposeful distortion on a continuous rather than discrete basis. Consider a forensic neuropsychological examination, for example, where it is often paradoxically found that patients with mild traumatic brain injuries report greater impairment than patients with severe traumatic brain injuries (e.g., Thomas & Youngjohn, 2009). Researchers have hypothesized that these aberrant response profiles are due to patients' conscious and/or unconscious attempts to gain compensation through litigation. This suggests that a multidimensional IRT model underlies the data. The probability of symptom endorsement is influenced not only by severity of injury but also by desire for compensation. A comprehensive IRT model could be used to tease apart these separate causes for item endorsement, thereby preserving measurement of injury severity.

Use of Item Response Theory in Clinical Assessment

Current Use

In many respects, the use of IRT in clinical assessment has yet to mature. Many of the articles cited in this article are didactic in nature; they have yet to directly influence the day-to-day practice of clinical psychologists. Most clinicians are not familiar with IRT terminology, few tests appear to have been meaningfully altered because of the theory, and coverage of the topic remains relatively sparse in assessment textbooks (e.g., Kaplan & Saccuzzo, 2008). Most discouraging, few IRT-based instruments are available for use in clinical settings. Yet this bleak picture ignores widespread application of the technology within research communities. A glance through any of the prominent research journals in psychological assessment (e.g., *Assessment*, *Psychological Assessment*, *Journal of Personality Assessment*, etc.) will likely reveal several articles employing the methodology. Moreover, if one considers the impact of the broader family of latent variable models, it becomes clear that modern psychometric theory is thriving in the field. Clinical psychologists might even be surprised to learn that IRT underlies some of their more commonly used diagnostic instruments. On the Woodcock–Johnson III (McGrew & Woodcock, 2001), for example, users can request estimates of examinees' latent abilities based on the Rasch model (i.e., the *W* value). In domains of personality and psychopathology as well, IRT concepts can be found tucked away in technical manuals (e.g., Personality Assessment Inventory; Morey, 1991).

Unfortunately, existing applications within clinical psychology have largely been restricted to self-report measures of psychological distress and psychoeducational measures related to disorders of learning and attention. It is surprising that IRT has had relatively little impact in the realm of clinical neuropsychology. The clear connection between the assessment of cognitive impairment (i.e., clinical neuropsychology) and cognitive ability/achievement (i.e., educational psychology) should make integration relatively seamless. Yet only recently have neuropsychologists begun to use IRT in their work (e.g., La Femina, Senese, Grossi, & Venuti, 2009; Pedraza et al., 2009; Schultz-Larsen, Kreiner, & Lomholt, 2007). This seems unfortunate, as great potential lies in the application of IRT to clinical neuropsychology. In particular, CATs offer to reduce the lengthy burden of testing for both patients and neuropsychologists—a pervasive concern in the field. Also, IRT's precision with respect to the estimation of latent variables would be of great value to neuropsychologists' growing interest in preclinical and/or subtle cognitive impairments (e.g., mild cognitive impairment, mild traumatic brain injury, etc.).

It is notable that the National Institutes of Health is currently funding a major attempt to standardize patient-reported outcomes using IRT (PROMIS). In general, the evidence reviewed in this article suggests that clinical psychologists are moving toward modern measurement of psychopathology. Yet a schism currently exists between researchers and practitioners; assessment specialists and treatment specialists. The hefty investment that clinical psychologists must endure to comprehend IRT has made the technology less appealing than previous innovations in clinical assessment (e.g., projective or performance-based testing). Because of this, IRT specialists have operated in somewhat of a vacuum. Collaboration between psychometricians and clinical psychologists is essential for propagation of IRT in the assessment of psychopathology. To foster these relationships, clinicians should work toward improving their level of sophistication in measurement theory. Programs awarding doctoral-level degrees in psychology—particularly those focused on the scientist-practitioner model—might give serious consideration to bolstering training in this respect. IRT is but one extension of an increasingly sophisticated field of psychometric theory. Continuous effort must be directed toward incorporating such advancements into the assessment of psychopathology, if clinicians are to maintain their expertise in applied psychological measurement.

Test Selection

It is likely that IRT's implications for the selection of clinical measures will soon have a major impact on the field. The theory provides a degree of precision in test analysis that simply has not existed in the realm of CTT. For example, we know from IRT that tests are not equally reliable for all intervals of latent variable distributions. Single-value summaries of test reliability (e.g., Cronbach's alpha) fail to capture meaningful nuances in test accuracy. Clinical psychologists can use IRT to better distinguish between measures in both research and practice settings. That is, by reviewing published accounts of test information (e.g., Thomas & Locke, in press), psychologists can make informed decisions about the appropriate use of clinical measures within specific populations. Test developers can facilitate the process by providing this information in technical manuals.

CATs represent a clear advantage over traditional instruments. The ability to decrease exam time while increasing the accuracy of latent variable estimates cannot be readily dismissed. Skipped items, unique item sets, and amalgams of items from various scales can all be effectively accounted for within the IRT-based CAT framework. Unfortunately, the current availability of CATs in clinical psychology is very limited. This can be expected to change. Interest in CATs is growing, and not just in psychology. In educational testing, for example, CATs have become common alternatives to paper-based test formats (e.g., Educational Testing Service, 2010). In addition, some of the limitations that hinder the use of CATs in other settings—such as the chronic need for novel test items—are of less concern in clinical assessment. When available, clinical psychologists should give strong consideration to the use of CATs in their research and practice.

IRT makes possible the ability to pool information from distinct measures of psychopathology. As previously reviewed, items related to the same underlying construct can all be combined—even when originating from unique measures—to improve the overall accuracy of latent variable estimates. One can imagine that research focused on equating popular measures of psychopathology would be of great interest to the field. For example, a researcher might choose to collect data on depression-related MMPI-2 items, depression-related PROMIS items, and items from the Beck Depression Inventory in order to calibrate all on the same scale of measurement. By doing so, psychologists could then use these published linking equations to combine information from all equated tests. This would not only improve diagnostic accuracy within particular assessment and research settings, but would also facilitate more general comparisons of published research (i.e., meta-analyses).

Model Development

Common factor theory revolutionized psychologists' ability to assess constructs in personality and psychopathology. Traditional use of this technology has been described as "blind" data reduction focused on the identification of key latent variables (Mulaik, 2010). Unfortunately, many test developers have learned that the internal structures of self-report inventories often do not coincide with theoretical categories of psychiatric nosology (i.e., the "neo-Kraepelinian" nomenclature). In other words, simple unidimensional models do not fit observed data. Because of this, test developers have increasingly focused on modeling unidimensional symptom clusters instead of broader syndromes of psychological distress (see Smith & Combs, 2010). This appears to have led to opposing views as to the appropriate strategy for developing clinical instruments: a modern view advocating the inclusion only of items shown to fit unidimensional factor models, and a traditional view advocating the inclusion of all items shown to predict criterion-related validity variables. That is, choosing to maximize scale homogeneity versus external validity.

Indeed, the conflict has led to heated exchanges over revisions to popular clinical measures. A special issue of the *Journal of Personality Assessment* (October 2006, Vol. 87, Issue 2) was dedicated to debate over a recent version of the MMPI (i.e., the MMPI-2 Restructured Form; Ben-Porath & Tellegen, 2008). Currently, little effort has been made toward incorporating unidimensional constructs into comprehensive models of psychiatric disorders.

IRT, on the other hand, is not seen as a data reduction tool, but as a technology for modeling observed data (de Ayala, 2009). Indeed, the growth of nonparametric and multidimensional IRT models is a direct reflection of theorists' desire to alter models to fit data, not data to fit models (although the Rasch tradition represents a notable exception). This flexibility in IRT has allowed for great emphasis on mathematical modeling of psychological constructs (e.g., Embretson, 1984, 2010). Accordingly, sophisticated IRT-related models of disorders are beginning to emerge (see Batchelder, 2010). IRT offers to free test developers from the restrictions imposed by the quest for unidimensional scales; in essence, a movement toward sophisticated models of psychopathology. For example, the growth of Markov chain Monte Carlo techniques in the estimation of latent variable models has allowed researchers to consider increasingly intricate structures of psychopathology (e.g., Reise & Waller, 2010). Such techniques can be used to model multidimensional, discrete, and/or multilevel latent constructs, rater effects, missing data, covariates, and so forth (Levy, 2009; Patz & Junker, 1999).

Strauss and Smith (2009) reviewed construct validity in psychological measurement, concluding that the use of, ". . . multifaceted, complex constructs as predictors or criteria in validity or theory studies is difficult to defend" (p. 15). Multidimensional constructs should not be portrayed as unidimensional variables. As masons construct buildings from mortar and brick, so too must researchers construct models from foundational materials. IRT can provide the crucial tools for psychologists who desire to study complex models of psychopathology. Consider existing clinical theories of psychological distress. Until recently, psychologists have lacked the ability to translate complex theories into mathematical measurement models. However, there now exist numerous psychometric models suitable for studying detailed theories of psychiatric disorders (see DiBello, Roussos, & Stout, 2007). Researchers and practitioners should be encouraged that psychometric models can now mimic substantive models and should raise their expectations for what can be accomplished though the process of clinical measurement. No longer must simple models dominate the field. Although there is good reason to rely on traditional and well-researched measurement instruments, revolutionary tools for evaluating psychiatric populations based on advanced clinical models may soon usher in a new era of assessment.

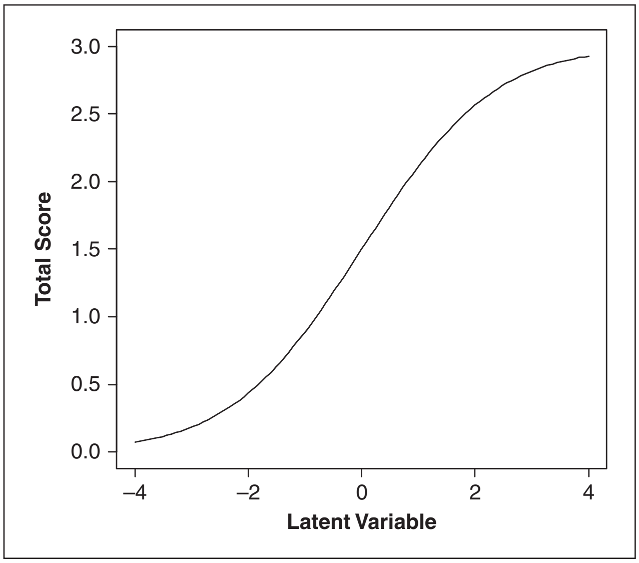

Figure 5. Test characteristic curve representing the relation between examinees' status on a latent variable (*x*-axis) and their expected scores on a test (*y*-axis)

Test Scoring

CTT, common factor theory, and IRT all have dissimilar traditions with respect to estimating patients' latent distress. In CTT, the focus is on sums of items: total scores. In common factor models, latent factor scores can be estimated, but person parameters are not explicit in the equations. It is only when discrete item responses are modeled to be the result of latent response probabilities that estimation of person parameters becomes necessary. In fact, a major limitation of IRT during its early history was the complexity that arises from simultaneous estimation of person and item parameters. For this reason, IRT has developed considerable sophistication in the estimation of examinees' latent variable scores.

This sophistication in IRT will allow clinical psychologists to replace total scores with more accurate IRT-based latent variable estimates. Figure 5 demonstrates the relation between values of a unidimensional latent variable and total scores through the use of a *test characteristic curve*: predicted total scores plotted against a range of the latent construct. As in the example, predicted total scores are generally related to the latent variable by a monotonically increasing function (Baker, 2001). However, the association is nonlinear, with total scores demonstrating increasingly imprecise relations to the latent variable near the extremes of distributions (Dumenci & Achenbach, 2008); that is, high and low total scores tend to be inaccurate. Mathematically, it can be demonstrated that total scores are inefficient estimates of latent variables (McDonald, 1999). IRT-based weighted item sums provide more information than unweighted item sums. Both are unbiased estimates of patients' latent distress, but total scores cannot be more reliable or more valid than latent variable estimates when models accurately reflect observed data.

Beyond this, model-derived latent variable estimates facilitate the many benefits accompanying IRT (e.g., CATs, meaningful scaling, etc.). It would seem unnecessary to revert to simplistic total score models when IRT lies beneath these sophisticated analyses of clinical measures. In addition, advanced estimation of patients' standings on latent variables (e.g., Bayesian methods) readily allow for the use of covariates and prior knowledge. If, for example, a clinician suspects a low base of schizophrenia in an outpatient clinic, a parameter (or prior distribution) reflecting this belief can be used to influence test outcomes.

Limitations

Although IRT holds great promise for clinical assessment, there are clear limitations to the methodology. Researchers and test developers have abundant didactic applications of IRT to guide their work. However, the mathematical bases underlying the methodology may be difficult for some researchers, test developers, and clinicians to grasp. Compounding this problem, most applications of the models cannot currently be conducted with commonly used "point-and-click" statistical programs (e.g., *SPSS*). Yet this does not prevent applied psychologists from gaining functional understandings of IRT. Most have performed heuristic item analyses for years without the aid of sophisticated technology. In addition, some less-common statistical modeling programs (e.g., *Mplus*, *Bilog*, *R*) have improved their accessibility to the theory. As the field integrates IRT terminology with clinical practice (e.g., replacing or converging the concept of reliability with information), a more general integration will likely occur. Statisticians and methodologists can facilitate the process by providing accessible explanations of the theory.

There is a natural assumption that IRT will improve test validity by improving test reliability. However, almost no studies reviewed in this article have directly addressed the overall construct validity (see Strauss & Smith, 2009) of psychological measures. IRT models, as are commonly employed, focus on the internal structure of psychological measures. Although internal structure is sometimes considered to be an index of measurement validity, many clinicians will demand to see relations to external diagnostic variables before accepting the validity of new instruments and test construction techniques. Indeed, some have questioned whether modern refinements do more harm than good with respect to the validity of psychological measures (e.g., Caldwell, 2006; Fava, Ruini, & Rafanelli, 2004; Gordon, 2006). This is an empirical matter that must be addressed. Although mathematics and simulation work suggests a clear advantage to IRT methodology, confirmation of this is necessary to convey the practical advantages of IRT in clinical measurement.

Conclusion

In 1884, Sir Francis Galton optimistically wrote that ". . . the powers of man are finite, and if finite they are not too large for measurement" (1971, p. 4). This review has addressed IRT's contribution toward confirming Galton's belief in clinical assessment. The theory is only beginning to affect the day-to-day activities of clinical practitioners; however, it has significantly altered clinical scientists' understanding of existing tests, psychiatric disorders, and measurement processes. Tools for improving analyses of measurement error, computer adaptive testing, scaling of latent variables, calibration and equating, evaluation of test and item bias, assessment of change due to therapeutic intervention, and evaluation of model and person fit have seen growing use in the field. In this respect, the theory has already demonstrated great value. In addition, IRT has the potential to drastically alter test selection, model development, and scoring, all of which can improve accuracy in clinical psychology. The evolution of measurement is an ongoing process. There is, and always has been, a lag between theory and practice. Much work remains before the value of IRT in clinical assessment comes to be fruition. In the interim, the outlook is promising.

Acknowledgments

I thank Richard Lanyon, Roy Levy, Roger Millsap, Manuel Barrera, and the anonymous reviewers for their insightful comments on previous versions of this article.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author received no financial support for the research and/or authorship of this article.

References

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35*, 475-487.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text revision). Washington, DC: Author.

American Psychological Association. (2002). *American Psychological Association ethical principles of psychologists and code of conduct*. Retrieved from http://www.apa.org/ethics/code/index.aspx

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

Baker, F. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71-93). Washington, DC: American Psychological Association.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21-33.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335-352.

Bouman, T. K., & Kok, A. R. (1987). Homogeneity of Beck's depression inventory (BDI): Applying Rasch analysis in conceptual exploration. *Acta Psychiatrica Scandinavica, 76*, 568-573.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration, scoring, and interpretation* (Rev. ed.). Minneapolis: University of Minnesota Press.

Caldwell, A. B. (2006). Maximal measurement or meaningful measurement: The interpretive challenges of the MMPI-2 Restructured Clinical (RC) scales. *Journal of Personality Assessment, 87*, 193-201.

Carle, A. C., Millsap, R. E., & Cole, D. A. (2008). Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent variable models. *Educational and Psychological Measurement, 68*, 281-303.

Carmody, T. J., Rush, A. J., Bernstein, I. H., Brannan, S., Husain, M. M., & Trivedi, M. H. (2006). Making clinicians lives easier: Guidance on use of the QIDS self-report in place of the MADRS. *Journal of Affective Disorders, 95*, 115-118.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*, S3-S11.

Chang, C. (1996). Finding two dimensions in MMPI-2 depression. *Structural Equation Modeling, 3*, 41-49.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189-225.

Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman, A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychological Assessment, 16*, 360-372.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist–Revised. *Psychological Assessment, 9*, 3-14.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist–Revised (PCL:SV): An item response theory analysis. *Psychological Assessment, 11*, 3-13.

Cooper, A., & Gomez, R. (2008). The development of a short form of the Sensitivity to Punishment and Sensitivity to Reward Questionnaire. *Journal of Individual Differences, 29*, 90-104.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*, 145-168.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26: Psychometrics* (pp. 979-1030). Amsterdam, Netherlands: Elsevier North-Holland.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20*, 55-62.

Duncan-Jones, P., Grayson, D. A., & Moran, P. A. (1986). The utility of latent trait models in psychiatric epidemiology. *Psychological Medicine, 16*, 391-405.

Educational Testing Service. (2010). *About the GRE general test*. Retrieved from http://www.ets.org/gre/general/about

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.

Embretson, S. E. (2010). *Measuring psychological constructs*. Washington, DC: American Psychological Association.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fava, G. A., Ruini, C., & Rafanelli, C. (2004). Psychometric theory is an obstacle to the progress of clinical research. *Psychotherapy and Psychosomatics, 73*, 145-148.

Ferrando, P. J. (1994). Fitting item response models to the EPI-A Impulsivity subscale. *Educational and Psychological Measurement, 54*, 118-127.

Ferrando, P. J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*, 997-1012.

Feske, U., Kirisci, L., Tarter, R. E., & Pilkonis, P. A. (2007). An application of item response theory to the DSM-III-R criteria for borderline personality disorder. *Journal of Personality Disorders, 21*, 418-433.

Frazier, T. W., Naugle, R. I., & Haggerty, K. A. (2006). Psychometric adequacy and comparability of the short and full forms of the Personality Assessment Inventory. *Psychological Assessment, 18*, 324-333.

Frost, A. G., & Orban, J. A. (1990). An examination of an appropriateness index and its effect on validity coefficients. *Journal of Business and Psychology, 5*, 23-36.

Galton, F. (1971). The measurement of character. In L. D. Goodstein & R. I. Lanyon (Eds.), *Readings in personality assessment* (pp. 4-10). New York, NY: Wiley. (Reprinted from *Fortnightly Review* (of London), 1884, 42, 179-185)

Gardner, W., Kelleher, K. J., & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care, 40*, 812-823.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., Clarke, D. C., VonAmmon Cavanaugh, S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19*, 43-55.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423-436.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research, 43*, 401-410.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J.,Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Gordon, R. M. (2006). False assumptions about psychopathology, hysteria and the MMPI-2 Restructured Clinical scales. *Psychological Reports, 98*, 870-872.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem scale. *Personality and Social Psychology Bulletin, 23*, 443-451.

Grayson, D. A. (1986). Latent trait analysis of the Eysenck Personality Questionnaire. *Journal of Psychiatric Research, 20*, 217-235.

Hambleton, R. K., & de Gruijter, D. N. (1983). Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement, 20*, 355-367.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Thousand Oaks, CA: SAGE.

Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29*, 26-44.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Hui, C. H., Drasgow, F., & Chang, B. (1983). Analysis of the modernity scale: An item response theory approach. *Journal of Cross-Cultural Psychology, 14*, 259-278.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818-825.

Kamakura, W. A., & Balasubramanian, S. K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment, 53*, 502-519.

Kaplan, R. M., & Saccuzzo, D. P. (2008). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. T., . . . Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in nonspecific psychological distress. *Psychological Medicine, 32*, 959-976.

Kim, Y., & Pilkonis, P. A. (1999). Selecting the most informative items in the IIP scales for personality disorders: An application of item response theory. *Journal of Personality Disorders, 13*, 157-174.

La Femina, F., Senese, V. P., Grossi, D., & Venuti, P. (2009). A battery for the assessment of visuo-spatial abilities involved in drawing tasks. *Clinical Neuropsychologist, 23*, 691-714.

Langenbucher, J. W., Labouvie, E., Martin, C. S., Sanjuan, P. M., Bavly, L., Kirisci, L., & Chung, T. (2004). An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in *DSM-IV*. *Journal of Abnormal Psychology, 113*, 72-80.

Leung, K., & Drasgow, F. (1986). Relation between self-esteem and delinquent behavior in three ethnic groups: An application of item response theory. *Journal of Cross-Cultural Psychology, 17*, 151-167.

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics, 2009*, 1-18. doi:10.1155/2009/537139

Long, J. D., Harring, J. R., Brekke, J. S., Test, M. A., & Greenberg, J. (2007). Longitudinal construct validity of Brief Symptom Inventory subscales in schizophrenia. *Psychological Assessment, 19*, 298-308.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Marshall, G. N., Orlando, M., Jaycox, L. H., Foy, D. W., & Belzberg, H. (2002). Development and validation of a modified version of the Peritraumatic Dissociative Experiences Questionnaire. *Psychological Assessment, 14*, 123-134.

Martin, M., Kosinski, M., Bjorner, J. B., Ware, J. E., Jr., MacLean, R., & Li, T. (2007). Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 Physical Function scale. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16*, 647-660.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III technical manual*. Itasca, IL: Riverside.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing, 5*, 279-300.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*, 5-9.

Mislevy, R. J., Levy, R., Kroopnick, M., & Rutstein, D. (2008). Evidentiary foundations of mixture item response theory models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 149-175). Charlotte, NC: Information Age.

Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407-423). New York, NY: Guilford Press.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function to the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53-69.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment, 12*, 354-359.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Pedraza, O., Graff-Radford, N. R., Smith, G. E., Ivnik, R. J., Willis, F. B., Petersen, R. C., & Lucas, J. (2009). Differential item functioning of the Boston Naming Test in cognitively normal African American and Caucasian older adults. *Journal of the International Neuropsychological Society, 15*, 758-768.

Popper, K. (1964). Scientific theory and falsifiability. In J. A. Mourant & E. H. Freund (Eds.), *Problems of philosophy: A book of readings* (pp. 541-547). New York, NY: Macmillan.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*, 228-238.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7*, 347-364.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*, 164-184.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 25-46.

Reise, S. P., & Waller, N. G. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 147-173). Washington, DC: American Psychological Association.

Revicki, D. A., & Sloan, J. (2007). Practical and philosophical issues surrounding a national item bank: If we build it will they come? *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16*, 167-174.

Roberson-Nay, R., Strong, D. R., Nay, W. T., Beidel, D. C., & Turner, S. M. (2007). Development of an abbreviated Social Phobia and Anxiety Inventory (SPAI) using item response theory: The SPAI-23. *Psychological Assessment, 19*, 133-145.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6*, 255-270.

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007). Mini-mental status examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology, 60*, 268-279.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17*, 28-43.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety, 25*, E34-E46. doi:10.1002/da.20432.

Spearman, C. (1904a). The proof and measurement of association between two things*. American Journal of Psychology, 15*, 72-101.

Spearman, C. (1904b). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Smith, G. T., & Combs, J. (2010). Issues of construct validity in psychological diagnoses. In R. F. Krueger & E. Simonsen (Eds.), *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (pp.205-222). New York, NY: Guilford Press.

Smits, D. J. M., & De Boeck, P. (2003). A componential IRT model for guilt. *Multivariate Behavioral Research, 38*, 161-188.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1-25.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Tenenbaum, G., Furst, D., & Weingarten, G. (1985). A statistical reevaluation of the STAI anxiety questionnaire. *Journal of Clinical Psychology, 41*, 239-244.

Teresi, J. A., Cross, P. S., & Golden, R. R. (1989). Some applications of latent trait analysis to the measurement of ADL. *Journals of Gerontology, 44*, S196-S204.

Thomas, M. L., & Locke, D.E.C. (in press). Psychometric properties of the MMPI-2-RF Somatic Complaints scale. *Psychological Assessment*.

Thomas, M. L., & Youngjohn, J. R. (2009). Let's not get hysterical: Comparing the MMPI-2 Validity, Clinical, and RC scales in TBI litigants tested for effort. *Clinical Neuropsychologist, 23*, 1067-1084.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago, IL: University of Chicago Press.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology, 57*, 1051-1058.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for anxiety (anxiety-CAT). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 16*, 143-155.

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Wright, B. (1980). Foreword. In G. Rasch, *Probabilistic models for some intelligence and attainment tests* (Expand ed.). Chicago, IL: University of Chicago Press.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83-105.

Young, M. A., Halper, I. S., Clark, D. C., & Scheftner, W. A. (1992). An item-response theory evaluation of the Beck Hopelessness Scale. *Cognitive Therapy and Research, 16*, 579-587.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement, 57*, 963-969.