HHS Public Access

Author manuscript

Psychol Assess. Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

Psychol Assess. 2019 December ; 31(12): 1442–1455. doi:10.1037/pas0000597.

Advances in Applications of Item Response Theory to Clinical Assessment

Michael L. Thomas[1]

[1]Department of Psychiatry, University of California San Diego, La Jolla, CA, United States

Abstract

Item response theory (IRT) is moving to the forefront of methodologies used to develop, evaluate, and score clinical measures. Funding agencies and test developers are routinely supporting IRT work, and the theory has become closely tied to technological advances within the field. As a result, familiarity with IRT has grown increasingly relevant to mental health research and practice. But to what end? This paper reviews advances in applications of IRT to clinical measurement in an effort to identify tangible improvements that can be attributed to the methodology. Although IRT shares similarities with classical test theory and factor analysis, the approach has certain practical benefits, but also limitations, when applied to measurement challenges. Major opportunities include the use of computerized adaptive tests to prevent conditional measurement error, multidimensional models to prevent misinterpretation of scores, and analyses of differential item functioning to prevent bias. Whereas these methods and technologies were once only discussed as future possibilities, they are now accessible due to recent support of IRT focused clinical research. Despite this, much work still remains in widely disseminating methods and technologies from IRT into mental health research and practice. Clinicians have been reluctant to fully embrace the approach, especially in terms of prospective test development and adaptive item administration. Widespread use of IRT technologies will require continued cooperation among psychometricians, clinicians, and other stakeholders. There are also many opportunities to expand the methodology, especially with respect to integrating modern measurement theory with models from personality and cognitive psychology, as well as neuroscience.

Keywords

item response theory; clinical psychology; assessment; psychometrics; applied measurement

Item response theory (IRT) is a modern psychometric methodology that has changed the rules for developing, evaluating, and scoring psychological tests (Embretson, 1996). The methodology has broad applications in clinical assessment (Reise & Waller, 2009), and is garnering increased attention from funding agencies and test publishers (Cella et al., 2007; Pearson, 2017). Moreover, as IRT is becoming closely tied to technological developments in the field—especially computerized adaptive testing—familiarity with the theory is increasingly relevant to professional practice. The purpose of this review is to describe how

Correspondence concerning this article should be addressed to Michael L. Thomas, Ph.D.; University of California, San Diego; Department of Psychiatry; 9500 Gilman Drive MC: 0738; La Jolla, CA 92093-0738. mlthomas@ucsd.edu.

advances in IRT are being used to improve psychological assessment in clinical contexts; it is distinguished from previous reviews by focusing not on the potential of IRT methods and technologies, but on tangible advances within the field. Specifically, the review seeks to answer the questions: Are we using IRT, how are we using it, and what are the next steps?

Item Response Theory

Models

A basic description of IRT is provided for readers who are unfamiliar with the approach. The literature should be consulted for broad coverage of methodology (e.g., Embretson & Reise, 2000), technical concerns (e.g., Baker & Kim, 2004), historical context (Bock, 1997), and comparisons with other psychometric frameworks (Hambleton & Jones, 1993).

IRT is one of several measurement theories that provides a framework for relating observed item and test scores to the latent variables that are directly relevant to assessment and diagnosis. This is accomplished via measurement models. IRT cannot be characterized by any single model, but rather dozens that are designed to meet the requirements of distinct theories and data (de Ayala, 2009). These models can be classified by three characteristics: (1) the number and type of person parameters; (2) the number and type of item parameters; and (3) the mathematical function relating person and item parameters to observed data.

Person parameters quantify individual differences in the latent abilities or traits measured by a test. IRT models are analogous to those from factor analysis in that latent ability can be unidimensional or multidimensional (Reckase, 2009). Item parameters can include difficulty, discrimination, and guessing. Item difficulty, as the name suggests, quantifies how difficult it is to respond to an item correctly. Models that have been designed to be fitted to ordinal categorical items, such as scales with polytomous symptom ratings, typically refer to threshold parameters between response options. Item discrimination quantifies the strength of the relationship between latent ability and the item response. Guessing quantifies the likelihood that an examinee will respond correctly, or affirmatively, in the absence of the ability, or the inclination, to do so.

IRT models are routinely fitted to measures of achievement, ability, personality, and clinical symptoms, among other constructs. Unfortunately, because most of the IRT models commonly used in clinical psychology were designed by researchers who were primarily interested in educational assessment, certain terminology can lead to confusion. For example, the term ‘difficulty’ seems misapplied when models are fitted to measures of clinical symptoms. Because of this, the difficulty parameter is sometimes referred to as the severity parameter in clinical assessment, or simply as the item’s location.

Most IRT models are part of the generalized linear model framework (Skrondal & Rabe-Hesketh, 2004), and use link functions to relate a linear term based on the person and item parameters to discrete response data. That is, a mathematical function must be used to allow predictions from a linear model to map onto nonlinear response terms (e.g., log odds of accuracy). The two most commonly used link functions are the logit—the inverse of the cumulative distribution function for the logistic distribution—and the probit—the inverse of

the cumulative distribution function for the normal distribution (see Madsen & Thyregod, 2010). The choice between these functions has little impact on model predictions, as a scaling constant can be used to achieve nearly equivalent metrics (Camilli, 1994). Model functions can be extended to account for ordered-categorical and nominal response data (Bock, 1972; Samejima, 1969), as well as many other functional forms (van der Linden & Hambleton, 1997).

There is a distinction in IRT between Rasch and non-Rasch type models (de Ayala, 2009). Whereas Rasch type models require discrimination parameters to be equal across items, or rather, omit the parameter, non-Rasch type models allow item discrimination to vary. Rasch type models include the classic Rasch or 1-parameter model for dichotomous item responses (e.g., correct vs. incorrect or true vs. false), and both the partial credit model and the rating scale model for ordered categorical item responses. Non-Rasch type models include the 2-parameter model (i.e., difficulty and discrimination parameters) and the 3-parameter model (i.e., difficulty, discrimination, and guessing parameters) for dichotomous item responses, and the generalized partial credit model and graded response model for ordered categorical item responses.

An example of an item response function (IRF) for a unidimensional 2-parameter logistic IRT model is shown in Figure 1. The *x*-axis represents the latent person parameter, which is typically scaled to have a mean of 0.0 and a standard deviation of 1.0. The Greek letter theta (θ) is used to denote the latent ability or trait. The *y*-axis represents the probability of responding correctly or affirmatively. The IRF, the solid line, indicates the response probability for an examinee with a particular parameter value. The IRF represents an item with a difficulty value, Greek letter beta (β), of 0.5 and a discrimination value, Greek letter alpha (α), of 1.2. Notably, the slope of the IRF varies over ability, reflecting the nonlinear relationship between changes in ability and the probability of a correct item response.

Estimation and Software

IRT person and item parameters are latent, and must therefore be estimated from observed data. Estimation of latent item parameters—when person parameters are known—or estimation of latent person parameters—when item parameters are known—is relatively simple, and is typically based on maximum likelihood approaches. However, because both person and item parameters are typically unknown, estimation can be challenging in practice, and typically requires advanced software, strong parametric assumptions about the distributions of model parameters, and copious data. Marginal maximum likelihood has become the preferred approach (Baker & Kim, 2004), but Bayesian methods are also common (Fox, 2010).

There is considerable discussion about the sample size needed to estimate model parameters. Recommendations range from hundreds to thousands of participants (e.g., Jiang, Wang, & Weiss, 2016; Reise & Yu, 1990; Stone & Yumoto, 2004). Sample size requirements are influenced by several factors, including the number of parameters estimated, the distributions of these parameters, and the estimation approach. Moreover, what is considered acceptable, or not, may vary by context. For example, the consequences of poorly estimated parameters are likely to be more serious in high-stakes clinical assessment (e.g., forensic

testing) when compared to assessments done for clinical research. Ultimately, determination of appropriate sample size should be based on factors that include both statistical and applied concerns. Generally, however, fitting Rasch type models to data in samples of fewer than 100 examinees, 2- and 3-parameter models to data in samples of fewer than 200 examinees, and multidimensional models to data in samples of fewer than 400 examinees is viewed negatively.

Several computer programs can be used for IRT. IRTPRO (Cai, Thissen, & du Toit, 2011) and ACER ConQuest (Adams, Wu, & Wilson, 2015) are two of the more popular commercial programs (for a detailed example describing the use of IRTPRO see Toland, 2015), and several freely available packages within the R software environment, including ltm (Rizopoulos, 2006) and mirt (Chalmers, 2012), offer similar functionality. Even programs not specifically used for IRT, such as Mplus (Muthén & Muthén, 1998–2017), can estimate parameters in IRT metrics.

Other Psychometric Frameworks

IRT is often compared to other psychometric modeling frameworks for didactic purposes. These comparisons often generate controversy. This is because it is difficult to draw clear distinctions between theories, and thus strengths and weaknesses are invariably overstated. Below, distinctions are highlighted while acknowledging the considerable overlap.

Classical test theory.—The relative strengths and limitations of classical test theory (CTT) have been widely discussed in the literature (Hambleton & Jones, 1993; McDonald, 1999; Nunnally & Bernstein, 1994). CTT models are commonly distinguished from IRT models based on strength of assumptions; whereas CTT models typically rely on weak assumptions, IRT models tend to rely on strong assumptions (DeMars, 2010). Novick (1966, p. 2) describes weak models as those which, "…make no specific assumptions concerning the functional form of observed score, true score [latent person parameter], or error score distributions." Weak assumptions have the benefit of making a model widely applicable, but can also limit its usefulness. For example, unlike IRT, CTT item and person parameters are sample dependent (Lord, 1980). As a result, CTT estimates of item difficulty (i.e., proportion correct) and item discrimination (i.e., item-total score correlation), as well as estimates of reliability, are dependent on the sample from which they were obtained (Hambleton & Jones, 1993). CTT models sometimes embody strong assumptions (Brennan & Lee, 1999), at which point the distinction between CTT and IRT becomes less clear (cf. McDonald, 1999).

IRT models are considered strong because they make restrictive assumptions, such as the number of latent abilities or traits measured, the form of the association (or function) relating item responses to latent ability, and the conditional associations among errors (e.g., Bejar, 1983). These assumptions must be met to trust the conclusions derived from an IRT analysis. Indeed, one reason psychometricians have developed so many IRT models is that considerable flexibility is needed for applied researchers to find just the right model, and assumptions, necessary for a particular test.

Factor analysis.—Common factor theory gives rise to two of the more popular psychometric modeling approaches used in clinical psychology: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Mulaik, 2010). Although a fundamental equivalence between some IRT and factor analysis models has been recognized by psychometricians for many years (Kamata & Bauer, 2008; Takane & Deleeuw, 1987), the theories are still sometimes incorrectly viewed as being entirely distinct.

Unlike IRT models, which are nearly always designed to be used with ordinal category response data, EFA and CFA models originally focused on continuous data. However, models that relate factors to discrete data (e.g., Christoffersson, 1975) were developed soon after Jöreskog's (1969) seminal work in CFA, and long before IRT became popular in clinical assessment. As Kamata and Bauer (2008) explain, for some IRT models, item discrimination parameters can be directly transformed into factor loadings, and vice versa, as can item difficulty and item intercept parameters. Thus, distinctions between IRT and factor analysis tend to be mostly practical. For example, whereas methods commonly used to estimate parameters in IRT are relatively more effective when a test comprises a large number of items, methods used to estimate parameters in CFA are more readily extended to structural equation models.

Are We Using IRT?

As the previous discussion suggests, IRT is a sophisticated but also complex methodology. It is important to ask whether clinical psychologists are using the approach, especially in comparison to techniques from factor analysis and CTT. To help answer this question, abstracts published within the last 10 years in the top 50 clinical psychology journals were searched for key words, phrases, and acronyms related to CTT, factor analysis, and IRT (see Supplemental Material). The results, reported in Figure 2, suggest that factor analysis is by far the most commonly reported psychometric methodology, followed by CTT and then IRT. Interestingly, it also appears that the gap between the use of IRT and CTT has narrowed in recent years. It is important to keep in mind that IRT, and to a lesser extent factor analysis, are considered novel, and thus authors may be more likely to highlight these methodologies in abstracts in comparison to CTT. Nonetheless, it does appear that IRT can now be considered one of the "big three" psychometric methodologies used in clinical psychology.

IRT abstracts were further evaluated to identify key words, phrases and acronyms that would suggest the use of Rasch versus non-Rasch type models. The results indicate that 37% of clinical psychology IRT-abstracts mentioned Rasch type models and 22% mentioned non-Rasch. There are several explanations for the more common use of Rasch type models. One is that estimation and parameter interpretation is comparatively simpler for Rasch type models. For this reason, Rasch type models might be preferred in clinical research and practice. Also, whereas 38% of IRT-abstracts that mentioned non-Rasch type models also mentioned Rasch type models, the reverse was true of only 22% of IRT-abstracts. This suggests that clinical researchers who fit Rasch type models to data may be less likely to evaluate and compare alternative frameworks.

How Are We Using IRT?

Computerized Adaptive Testing

Standard error of measurement (SEM) reflects uncertainty in estimates of latent true scores, and is framed as the amount of variation of an examinee's score on a series of measurements (Cronbach, 1960). Reliability refers to the consistency of scores. SEM and reliability play many roles in clinical measurement, including as components of power calculations for treatment studies, and in determining reliable change for psychotherapeutic outcomes. Moreover, standards of professional practice require that estimates of reliability and SEM be reported, and considered, when selecting tests (AERA, APA, & NCME, 2014).

Although methods for estimating SEM vary, the most commonly referenced expression derives SEM from reliability: SEM equals the standard deviation of the score multiplied by the square root of one minus its reliability (Haynes, Smith, & Hunsley, 2011). Valid interpretation of SEM produced by this formula is based on an assumption that values are invariant over scores within a population of examinees. However, due to the fact that most psychological measures employ categorical response options (e.g., correct vs. incorrect or ordered ratings), measurement error often varies systematically (Embretson, 1996; McDonald, 1999).

The *Standards for Educational and Psychological Testing* state that, "When possible and appropriate, conditional standard errors of measurement should be reported at several score levels…" (Standard 2.14; AERA et al., 2014). Using a single SEM value can give misleading conclusions regarding the precision or imprecision of score estimates. To the extent that SEM is unexpectedly larger for some individuals, or populations, researchers can expect relative attenuation of effect size and loss of statistical power, and clinicians can expect poorer diagnostic sensitivity and specificity (Thomas et al., 2017). Moreover, conditional standard errors can lead to erroneous conclusions when interpreting intra- and inter-individual deficits.

In IRT, measurement error is quantified as standard error of the estimated ability parameters (SE_θ) rather than SEM. The distinction is that SE_θ refers to the amount of estimation error in person parameters (i.e., the latent ability score). And, unlike most applications of SEM, there is no assumption that SE_θ is constant over scores. Figure 1 shows the SE_θ function for the hypothetical IRF discussed earlier. Notably, SE_θ is a non-linear, "U"-shaped function of ability. Estimation error is conditional, and typically becomes worse in the tails of the ability distribution. SE_θ is tied to a mathematical quantity known as information: the change (or variance) in the observed response data given a change in latent ability (Lord, 1980). Information drops, and thus SE_θ rises, when the probability of an observed item response approaches the accuracy floor (i.e., 0% or guessing) or ceiling (100%). Typically, information, rather than SE_θ values have been reported in IRT papers; however, clinicians often prefer to report and interpret SE_θ values because they are conceptually similar to SEM statistics from CTT.

The information, and thus SE_θ, function for an entire test is based on the sum of all item information functions. A typical goal of an IRT analysis is to estimate item parameters in

order to determine the SE_θ function of a test. An example is provided by Fayers et al. (2005) who analyzed item data from the Mini-Mental State Examination (MMSE) obtained in a large sample of older Norwegian adults with and without cognitive impairment. Test response and SE_θ functions for the MMSE based on the results of Fayers et al. are shown in Figure 3. The solid line is the test response function and the dashed line is the SE_θ function. The test response function provides the examinee's expected total score on the MMSE based on their latent ability. As can be seen, while the range of expected total scores for examinees with below average ability is large, the range of expected total scores for examinees with above average ability is small. This is a "ceiling effect" that is also associated with relatively larger SE_θ for examinees with above average ability. That is, examines with cognitive impairment are measured more accurately than examinees without cognitive impairment. Fayers' et al. results suggest that while the MMSE may be well suited as a screening measure, its use as an outcome measure is suspicious, especially when used to assess initially high functioning individuals.

Simulated data help demonstrate the risks of conditional standard error. Hypothetical samples of examinees were drawn from normal distributions assuming either below or above average baseline ability. Random latent change scores were added to the second observation in order to simulate decline. Change in MMSE total scores based on this simulation are shown in Figure 4. Although the amount of latent change was equal between groups, observed change in the MMSE scores was larger in the below average ability group compared to the above average ability group (cf. Mungas & Reed, 2000).

The problem of conditional error becomes exaggerated to the extent that a test comprises fewer, discretely scored items with a restricted range of difficulty. Researchers have demonstrated these issues for measures of cognition (Gavett & Horwitz, 2012; Pedraza, Sachs, Ferman, Rush, & Lucas, 2011), personality (Ranger & Ortner, 2011; Spence, Owens, & Goodyer, 2012), and psychiatric symptoms (Olino et al., 2012). Indeed, over the last 20 years the literature has become saturated with studies demonstrating that conditional measurement error is not the exception, but rather the rule in clinical measurement.

Psychometricians and test developers have known for many years that reliability can be improved by adaptively tailoring items to the ability or trait levels of individual test takers. Binet and Simon (1905), for example, pioneered an approach that is still used in cognitive testing where examinees are administered items of increasing difficulty until they can no longer respond correctly. Unfortunately, adaptive testing was not fully realized until many years later.

Today, IRT is providing the methodological framework needed to implement adaptive testing (Linden & Glas, 2010), and inexpensive, fast personal computers, tablets, and smartphones are providing the medium of delivery. Conditional measurement error occurs when tests comprise items that are generally too hard (too severe) or too easy (too mild) for at least some examinees. Therefore, adaptive tests seek to administer items that have just the right amount of difficulty for each test taker. What constitutes "just right" depends on the particular model used, but is ultimately determined by item information. Adaptive tests administer items that contribute the greatest information to estimates of latent ability. A

primary challenge to adaptive testing is that optimal item selection is dependent on knowing the examinee's latent ability; the very quantity targeted for measurement. In practice, computerized adaptive testing is an iterative process that involves administering items, provisionally estimating latent ability, determining whether SE_θ meets a minimum value needed to end testing, and if SE_θ is too high, administering additional items that are tailored to the examinee's current ability estimate.

There are two primary benefits of computerized adaptive testing. First, test tolerability, speed, and efficiency can be improved because examinees are only administered items that are matched to their ability or trait level. Second, because the minimum level of SE_θ needed to end testing is under the examiner's control, SE_θ is, in theory, no longer conditionally related to ability. However, in practice, the benefits of adaptive testing are often not fully realized due to complexities of implementing the technology. Most notably, adaptive testing requires a large pool of candidate items: the item bank. In order to maintain the potential to administer items that are "just right" for each examinee within a population, test developers must ensure that item banks comprise a wide range of difficulty. Moreover, item parameters typically need to be pre-calibrated in large, diverse samples of many hundreds or thousands of participants.

Fortunately, the last decade has seen major advances in funding and research efforts needed to develop and maintain computerized adaptive tests for clinical assessment. The most notable example is the National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System (PROMIS; Northwestern University, 2017). PROMIS includes self-report measures for a broad array of functions, symptoms, behaviors, and feelings that were developed by a large network of institutions and collaborators (Gershon, Rothrock, Hanrahan, Bass, & Cella, 2010). PROMIS items and instruments are available in three forms. First, items and their parameter estimates have been reported in the literature, and are available online, allowing researchers to create their own adaptive tests. Second, PROMIS investigators have developed short forms—small subsets of items chosen for their optimal parameters—which tend to compare favorably to many existing psychometric tools (e.g., Olino et al., 2013). Finally, and usually for a fee, items can be administered adaptively using a variety of computerized and online technologies. PROMIS measures are not just being used in research, but also in clinical practice (Howell et al., 2015), and were recommended for use in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013).

But, do PROMIS measures effectively combat the problem of conditional measurement error? Several studies have compared adaptive PROMIS measures to non-adaptive (short form) PROMIS scales as well as to other measures. Khanna et al. (2011), for example, evaluated SE_θ values produced by fixed versus adaptive PROMIS scales, and found that whereas both administration types tended to provide consistently low SE_θ for estimates within the average to below average range of ability, only the adaptive measures continued to produce acceptably low SE_θ in the above average range. Varni et al. (2014), on the other hand, examined several fixed and adaptive pediatric PROMIS scales—including measures of depression and anxiety—and found that adaptive tests provided relatively little benefit over short forms. On aggregate, while in no instance can the SE_θ functions produced by adaptive

PROMIS measures be described as constant, PROMIS measures do appear to produce less exaggerated patterns of conditional error.

PROMIS measures may be the most salient, but not the only examples of adaptive testing in clinical assessment. Gibbons and colleagues (Gibbons et al., 2012, 2014) have developed computerized adaptive measures of depression and anxiety based on a multidimensional IRT bifactor model (see below). Moore et al. (2017) developed a computerized adaptive version of the Schizotypal Personality Questionnaire. Forbey and colleagues (Forbey, Ben-Porath, & Arbisi, 2012) developed and validated a computerized adaptive version of the Minnesota Multiphasic Personality Inventory–2 (MMPI-2-CA). Simms et al. (2011) developed an adaptive measure of traits that are consistent with personality disorders from the DSM-5.

Interestingly, the majority of computerized adaptive tests developed specifically for clinical psychology have been directed towards the assessment of clinical symptoms or constructs related to abnormal personality. This is in contrast to educational assessment where adaptive tests are typically used to assess ability or achievement. Two salient examples of computerized adaptive tests developed for cognitive assessment come from the NIH Toolbox (Weintraub et al., 2013), which includes two adaptive measures of language—the NIH-TB Oral Reading Recognition Test and the NIH-TB Picture Vocabulary Test (Gershon et al., 2014).

Beyond these ‘live’ tests, several simulation studies have demonstrated the potential benefits of adaptive testing in clinical assessment (e.g., Reise & Henson, 2000; Sunderland et al., 2017). There are also several computer programs available for simulation research (Choi, Podrabsky, & McKinney, 2012; Han, 2012; Magis & Barrada, 2017), as well as both commercial and open-source software for hosting computerized adaptive tests (International Association for Computerized Adaptive Testing, 2017; Scalise & Allen, 2015) making the technology accessible.

Despite the increased availability of computerized adaptive testing in the field, evidence suggests that relatively few clinical researchers are actually evaluating or using the approach. Only 4% of IRT abstracts mentioned terms related to computerized adaptive testing. Although this likely ignores studies that used adaptive tests, such as PROMIS measures, but chose not to mention this the abstract, it does appear that the outright focus on adaptive testing in clinical research and practice is rare. This is perhaps not surprising given that the field has traditionally relied heavily on clinical interviewing and paper-and-pencil tests. Rabin et al. (2014), for example, surveyed neuropsychologists in the United States and Canada and found that only 6% of the instruments being used were computerized. Moreover, nearly 60% of respondents reported ‘never’ or ‘only rarely’ using computerized tests. These results are particularly striking given that neuropsychologists are some of the heaviest users of psychological tests in clinical assessment.

Why has adaptive testing technology been slow to catch on? Perhaps clinical psychologists do not want to abandon existing tools, are concerned about threats to clients’ privacy, or simply are unaware or even unimpressed with the benefits of the technology. Whatever the

cause, the gap between what computerized adaptive tests can do to improve clinical assessment, and what they are actually doing, is still large.

Multidimensional Models

Measurement research in psychology has a strong concern with preventing erroneous interpretations of scores. That is, psychologists are not just interested in how observed scores relate to true scores, but also in understanding their psychological determinants and functional significance (Strauss & Smith, 2009). This is because item and test responses—according to common interpretations—are behaviors that are determined by latent cognitive and affective processes. A key question in psychometric research, therefore, is the latent dimensionality of scores. That is, how many abilities account for individual differences in response data.

Interpretations and evaluations of test scores are often based on an assumption of unidimensionality. Unidimensional measures are those for which patterns of association between items can be explained by only one individual difference variable. Spearman (1904) observed that systematic correlations between variables can provide evidence that scores are determined by a common ability. However, unidimensionality is rarely if ever unambiguously supported by data, and the dimensionality of scores can change as a function of other item and test scores included in an analysis (Haynes et al., 2011).

Strauss and Smith (2009) describe challenges to establishing the validity of test scores that arise from multiple psychological constructs. Comparing individuals on the putative abilities or traits thought to determine scores can become clouded. For example, because measures of depression often include a subset of items that relate to somatic concerns, scores can become artificially elevated in non-depressed populations with medical illnesses (Kathol, Mutgi, Williams, Clamon, & Noyes, 1990). Moreover, when correlating observed scores with variables in order to establish their validity, it may be unclear what abilities or traits drove the association.

Several concerns must be considered when determining the dimensionality of scores. IRT models require item responses to be locally (or conditionally) independent (de Ayala, 2009); that is, items are assumed to be uncorrelated after variance explained by the model has been removed. Violations of the local independence assumption can lead to biased parameter estimates and inflated estimates of reliability (DeMars, 2006). Because many test developers and users prefer unidimensional scores, a common use of IRT methodology is to identify and remove items that violate local independence. This approach is reasonable when psychological theories support the measurement of a unidimensional trait; however, the removal of items from scales purely due to psychometric concerns can be problematic. Artificially imposing unidimensionality onto measures can limit a researcher’s or clinician’s ability to address specific hypotheses and develop detailed theories that have real-world value (Nichols, 2006).

Multidimensional IRT models (Reckase, 2009) provide an alternative for researchers and clinicians who wish to embrace multidimensionality as a tool that can be leveraged towards conducting more informative assessments. There are generally two classes of models used

for this purpose: between-item and within-item multidimensional models. Between-item models are those for which responses to individual items are determined by just a single latent variable (i.e., one discrimination parameter per item), but where multiple abilities are specified. Within-item multidimensional models are those for which responses to individual items are determined by more than one latent variable (i.e., multiple discrimination parameters per item). Figure 5 demonstrates this distinction. The terms between-item and within-item are analogous to simple versus complex factor structure respectively.

Between-item models are often preferred because their parameters are generally simpler to interpret and estimate. Michel et al. (2016) for example developed a computerized adaptive measure of quality of life for patients with schizophrenia that assesses domains including self-esteem and resilience. The dimensions are correlated, but items relate to just one dimension.

Unfortunately, between-item models are not always plausible, and may not capture the full complexity of psychological theories. Within-item models are more flexible, in that items can discriminate individual differences in multiple latent person parameters, but this flexibility comes at a price. As in factor analysis, within-item multidimensional IRT models can suffer from rotational indeterminacy; that is, there are infinite combinations of ability and discrimination parameters that would all provide equivalent fit to data. To resolve indeterminacy, researchers must rotate parameters so that the pattern of item discrimination either matches psychological theory, or moves towards a structure that approximates a between-item model (i.e., simple structure). Although exploratory rotations are common in factor analysis, the approach can be antithetical to IRT, which, traditionally, has been more confirmatory than exploratory in nature. That is, one typically turns to IRT in the final stages of item selection and parameter estimation, or to evaluate items for an existing measure. IRT measurement models are often assumed to be known *a priori*, and violations of model assumptions are often dealt with by changing the items, not the measurement model. Because of this, within-item multidimensional IRT models have been rare in clinical assessment, with one exception: the confirmatory bifactor model.

Confirmatory bifactor models were proposed by Holzinger and Swineford (1937) and later extended to dichotomous and polytomous item response data in IRT by Gibbons and colleagues (Gibbons et al., 2007; Gibbons & Hedeker, 1992). The bifactor model has a within-item multidimensional structure where all items discriminate individual differences in one general ability or trait as well as one additional domain specific dimension. All latent person parameters are constrained to be uncorrelated, and thus individual difference variance is cleanly partitioned into orthogonal general versus specific components.

There have been several uses of IRT bifactor modeling in clinical assessment. One is to identify variance due to nuisance dimensions—that is, abilities or traits that are not the primary target of measurement. For example, Rubright, Nandakumar, and Karlawish (2016) fitted an IRT bifactor model to MMSE items and found clusters of items related to specific domains such as orientation and language, independent of general cognition. Common variance due to these factors led to violations of local independence when data were assumed unidimensional. Gibbons and colleagues' (Gibbons et al., 2012, 2014)

computerized adaptive measures of depression and anxiety are both derived from IRT bifactor models. The measures separate general variance in depression and anxiety from specific domains such as mood and somatic symptoms. Specific domains account for variance that is not explained by general depression and anxiety.

Another use of the IRT bifactor model is to improve the interpretability of scale scores across several correlated measures. For example, Thomas (2012) fitted an IRT bifactor model to the Brief Symptom Inventory (BSI; Derogatis, 1993)—a screening measure of general psychological distress that comprises scales such as depression, nervous tension, and psychoticism. Scores on the BSI are highly correlated, and patients often display patterns of consistently elevated scores across scales, making the assessment of specific areas of distress challenging. The bifactor model cleanly separated variance into general versus specific components, thereby improving the measure's interpretability, as well as the sensitivity and specificity of diagnoses (Thomas, 2012).

Multidimensional measures cannot always be reliably interpreted as such. Reise, Moore, and Haviland (2010) show how bifactor models can be used to quantify the reliability of distinct sources of variance. As an example, the authors examined the Observer Alexithymia Scale (OAS; Haviland, Warren, & Riggs, 2000), a measure of difficulties recognizing, processing, and regulating emotions with five subscales. Reise et al. found that while the scale's general factor accounted for 57% of common variance, domain specific factors accounted for less than 5% of variance each. The authors concluded that, "…despite the empirical fact that the data are multidimensional, scores derived from the OAS primarily reflect a single common source, alexithymia" (Reise et al., 2010, p. 556). Thus, when variance of a multidimensional measure is vastly dominated by a general factor, reliable interpretations of scores may be restricted to the overall level and not pattern of scale elevations.

The bifactor model is just one of many multidimensional IRT models that can be fitted to clinical measures. Multidimensional models can be fitted to dichotomous, ordered categorical, and nominal item data (Reckase, 2009), across, 1-, 2-, and 3-parameter frameworks. These models also serve as a bridge to mathematical modeling approaches used elsewhere in psychology, especially those designed to improve the construct validity of scores (Embretson, 2010), and are frequently touted as the gateway to newer, better clinical tools (Gibbons et al., 2012). When viewed from the generalized linear model framework (Skrondal & Rabe-Hesketh, 2004), it becomes clear that both unidimensional and multidimensional IRT models are simply instances of a much broader and more flexible measurement framework. Thus, the bounds on what approaches should, or should not, be considered IRT soon become blurred in the field of multidimensional modeling. The field can likely expect much more integrated modeling work in clinical measurement over the coming decades (see below).

What evidence is there that multidimensional models are currently being used in clinical assessment? IRT abstracts were searched for key words that would suggest the use of unidimensional versus multidimensional IRT models (see Supplemental Material). Whereas 21% of abstracts mentioned terms related to unidimensional models, 17% mentioned terms related to multidimensional models. Although this undoubtedly underestimates the number

of papers that fitted unidimensional models to data—as unidimensionality is often considered a default—it is notable that nearly a fifth of IRT papers in the field mentioned a multidimensional approach.

The advent of multidimensional measurement models in clinical assessment could have dramatic, if not yet fully understood ramifications. Clinical test batteries often comprise multiple, highly correlated measures of overlapping constructs. Moreover, many psychiatric and neurological disorders have similar clinical presentations, making it difficult to tie specific symptoms to specific causes. Because of this, clinicians are often trained to interpret outcomes based on configurations or patterns of scores. Use of multidimensional models could support, and in some instances, replace, clinician-based interpretations of score patterns. Multidimensional IRT models fitted to items across an entire battery of measures could replace a convoluted pattern of observed scores with a cleaner, more interpretable pattern of latent scores. Moreover, with the aid of computerized adaptive testing, automated algorithms could be used to ensure each of the latent constructs targeted is measured with optimal precision.

In neuropsychological assessment, for example, tests are designed to measure unique cognitive abilities (e.g., naming and episodic memory), and it is the pattern of deficits observed across measures that would lead a clinician to suspect a certain etiology (e.g., Alzheimer's vs. vascular based dementia). Some domains are vulnerable to a wide array of brain injuries or neurological diseases (e.g., measures of attention), while others are specific (e.g., measures of aphasia). Also, there is a distinction between "don't hold" tests—which are expected to be impacted by clinical disorders—and "hold" tests—which are typically not (e.g., measures of vocabulary and general knowledge). Thus, neuropsychologists must integrate large amounts of data, and flexibly adapt test batteries in real time in order to ensure that all psychometric considerations are addressed. Could multidimensional IRT be used to aid neuropsychologists and other clinicians in this complex work? It is possible that newer methods, which combine multidimensional IRT with other, complex modeling frameworks such as mathematical cognitive psychology and Bayesian networks, might one day be used for this purpose.

Differential Item Functioning and Person Misfit

Analyses of differential item functioning (DIF) are meant to identify items that produce biased scores for certain groups of individuals defined by factors such as gender and race (Millsap, 2011). DIF exists when the probability that examinees from separate populations have unequal probabilities of responding to an item correctly (or affirmative), even when they have the same value of the underlying ability (or trait) targeted for measurement. More practically, DIF implies that item parameters vary by group. There are two types of DIF: uniform DIF, where item difficulty parameters are unequal, and nonuniform DIF, where item discrimination parameters can also be unequal. Uniform DIF implies that items are symmetrically harder for individuals from a certain population. Nonuniform DIF, on the other hand, need not systematically advantage or disadvantage any group, but does produce invalid scores for certain individuals. A related approach based on confirmatory factor

analysis tests for bias in the form of placing invariance constraints on item parameters (Millsap, 2011).

Examples of DIF analyses abound in the clinical assessment literature. Teresi et al. (2009), for example, performed a DIF analysis on the PROMIS depression bank items comparing groups defined by gender. The authors found that the items “I felt hopeless” and “I felt sad” both demonstrated nonuniform DIF. The item “I felt like crying”, on the other hand, demonstrated uniform DIF, where females were expected to receive systematically higher scores, and thus positively biased estimates of depression. Weinstock, Strong, Uebelacker, and Miller (2009) performed a DIF analysis of Major Depressive Episode symptom criteria from the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000). The authors compared individuals with a history of Bipolar I and Bipolar II diagnoses, and found uniform DIF for the suicide criterion. Specifically, individuals with Bipolar I diagnoses were more likely to report suicidal ideation relative to individuals with Bipolar II diagnoses, even when they had the same underlying level of depression.

Ideally, DIF should be identified during test development in order to remove offending items (Pedraza & Mungas, 2008). However, latent variable models can be used to adjust for DIF in the process of scoring measures. For example, Sacco, Casado, and Unick (2011) conducted a DIF analysis of the Interpersonal Support Evaluation List (ISEL-12; Cohen, Mermelstein, Kamarck, & Hoberman, 1985), a measure of perceived social support. Focusing on race, the authors found DIF on 10 of the instrument’s 12 items. Moreover, while ISEL-12 total scores indicated significant differences between Hispanics and Whites, and between Blacks and Whites, but not Asians and Whites, when data were scored using a latent variable model that accounted for DIF, group differences reversed, with a difference found only between Asians and Whites.

An alternative approach to identifying invalid scores is to focus on examinees instead of items. In particular, the likelihood of an examinee’s observed response data, given an assumed model and a known set of item parameters, can be used to identify aberrant response profiles with person fit statistics (Meijer & Sijtsma, 2001). Person fit statistics identify patterns of item responses that are odd or unexpected given the properties of items (e.g., an examinee who denies sadness but endorses suicidal ideation and hopelessness). Conijn, Emons, De Jong, and Sijtsma (2015) calculated person fit statistics for the Outcome Questionnaire–45 (OQ-45; Lambert, Gregersen, & Burlingame, 2004), a measure of psychotherapy outcomes, and found that misfit was associated with severity of distress. Patients with psychotic, somatoform, and substance-related disorders, in particular, were more likely to show misfit. Thomas and Lanyon (2012) examined item data from the Psychological Screening Inventory-2 (Lanyon, 2010) in a sample of forensic examinees charged with criminal offenses and a comparison sample of non-forensic controls. Analyses indicated that 61% of the forensic examinees but only 5% of controls had misfitting item response profiles. In both studies, results suggested that certain characteristics of individuals (e.g., severity of illness and motivation to deceive) might lead to invalid scores.

Despite concerns, the practical consequences of DIF and person misfit have not been well studied in the clinical literature. Technical reports tend to suggest that the overall impact of DIF and person misfit on the estimates, reliability, and rank orderings of scores tends to be relatively minor (Lee & Zhang, 2010; Meijer & Sijtsma, 2001). However, little is known about the impact of DIF and misfit on practical issues that tend to concern clinicians (e.g., using scores to inform diagnoses and the measurement of change). Moreover, group statistics may mask the practical impact that biased and invalid response data might have on individual patients. For example, while DIF associated with group differences based on race may not greatly impact the reliability of scores, it may nonetheless lead to serious consequences for individual patients, such as overpathologizing and denial of insurance claims. These issues require further study, with a focus on questions that are relevant to clinical decision making, before the added value of DIF and person misfit can be properly weighed. Moreover, ways to integrate DIF and person misfit statistics into assessment technologies and clinical interpretations need further study.

Notably, DIF terms were highly mentioned in recently published IRT abstracts. Indeed, 22% of IRT abstracts in top clinical psychology journal mentioned DIF analyses, suggesting that clinicians consider the assessment of bias to be one of the more important and accessible methods from IRT. Person fit analyses, on the other hand, were quite rare, with only 2% of abstracts mentioning related terms.

Next Steps?

What Characteristics Will Define the Next Generation of IRT Models Used in Clinical Assessment (And Who Will Develop Them)?

There are already a dizzying number of IRT models. Despite this, several authors have argued for expansion. These approaches are based on the common belief that current models are too simple, and therefore mischaracterize psychological processes driving response behavior, or miss important, latent information that is embedded within the profile of item responses.

In personality assessment, Drasgow, Chernyshenko, and Stark (2010) argued that ideal point models should be preferred over dominance models. When fitted to ordered categorical item data, dominance models assume that the examinee will respond with ever increasing (monotonic) probability, or levels of agreement, along with increases in the latent trait assessed. Ideal point models, on the other hand, assume that examinees are likely to endorse items near their ideal point, but not items representing lower or higher locations on the latent trait continuum. Drasgow et al. (2010) give the example of the item, “I enjoy chatting quietly with a friend at a café” for measuring extraversion. The authors argue that while extreme introverts might deny the item because they are concerned about talking in public, extreme extraverts might deny the item because they enjoy talking in more exciting locations.

Much of the work on ideal point models is focused on personality research in healthy populations. Moreover, although there are now several studies that could be used to inform clinical work (Cho, Drasgow, & Cao, 2015), it is not clear that clinical measures and items are ideally suited for this approach. Personality inventories that are common in clinical

assessment, such as the Minnesota Multiphasic Personality Inventory (MMPI) and Personality Assessment Inventory (PAI) combine items that relate to both personality and symptoms of psychopathology. Ideal point models are likely not uniformly appropriate for the wide range of items that appear on such scales. Moreover, while Drasgow et al. (2010) clearly articulated that ideal point models are only appropriate for non-cognitive measures, Reise (2010) further emphasized that such models are also not appropriate for personality items that do not explicitly concern attitudes. Thus, whether and how ideal point models can be integrated into the scoring and psychometric evaluation of complex clinical measures is as yet unclear.

In the realm of cognitive assessment, especially in relation to clinical neuropsychology, several authors have argued for an approach that merges mathematical cognitive models with psychometric models in the assessment of clinical constructs (Batchelder, 2010; Brown, Thomas, & Patt, 2017). For example, authors have merged ideas from IRT with multinomial processing tree models (Batchelder, 2010), diffusion process models (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011), and signal detection theory models (Thomas et al., 2017), among others. Many of these represent extensions of the multidimensional modeling work described earlier. The approach allows for the measurement of individual difference variables within the context of strong, experimentally-validated, and even brain-based neurocognitive theories (e.g., McKenna, Brown, Drummond, Turner, & Mano, 2013).

However, the cognitive modeling literature is replete with ongoing debates about which model is “best”, and the popularity of certain modeling frameworks tends to wax and wane over time. Would clinicians accept a test that produces estimates of latent scores based on mathematical cognitive models that are unlikely to endure? Brown et al. (2017) considered this point in detail, and argued that it is natural for theories and measurement frameworks to change, and improve over time. Nonetheless, there are several practical challenges to a shifting scoring framework that is based on ever-changing measurement theories. Whether or not clinicians would embrace this approach, likely depends on practical benefits. If cognitive-psychometric models can be shown to improve diagnostic sensitivity and specificity, clarity in the interpretation of treatment outcomes, or otherwise improve the day-to-day research and practice activities of clinical psychologists, the ephemeral nature of certain tests and scoring frameworks could be forgiven. For now, demonstrations of the clinical benefits of cognitive-psychometric modeling are lacking.

Another future modeling direction for clinical assessment comes from mixture IRT models. Mixture IRT models assume that there are both quantitative and qualitative differences between examinees (see de Ayala & Santiago, 2017). Specifically, standard IRT models—which assume that examinees can be compared along a latent ability or trait continuum—are augmented to include a finite number of latent classes. Person estimates include both the examinee’s θ score (or scores), and the probability that the examinee belongs to each of the latent classes assumed. Item parameters are allowed to vary between groups, which is similar to the DIF approach described earlier, except that the grouping variable is assumed to be latent rather than observed. Mixture IRT models may have clinical utility in identify important clinical subpopulations. Finch and Pierson (2011), for example, fitted an IRT mixture model to items from a survey of risky youth behavior, and found four distinct

classes that varied both in terms of the types of risky behaviors engaged (e.g., risky sexual behavior). Moreover, because item parameters were allowed to vary by group, the authors could identify subsets of items that were particularly useful for identifying certain subclasses of at-risk youths.

There is little doubt that advanced psychometric models require sophisticated software and additional technical knowledge for users. Conceivably, most of the complex work needed to develop and implement these models and tools could be accomplished by statistical and technical experts, with applied researchers and clinicians reaping the benefits. But, does this divorce clinicians from the development of the very tests they use? Historically, statistically-minded clinicians have been at the forefront of test development in clinical psychology (e.g., Hathaway & McKinley, 1940). However, as fields of psychology become increasingly specialized, there is a risk that clinicians might lose the interest, or the time, to train in the types of advanced mathematical modeling techniques being used to develop the next generation of measures.

The implementation of computerized adaptive testing may provide a template for overcoming this problem. Development of computerized adaptive tests often brings together experts from several fields, including clinicians, psychometricians, programming experts, and experts in implementation science. Similarly, collaborative science could be used to promote the development and use of advanced mathematical measurement models. In neuropsychological assessment, for instance, experts in cognitive neuroscience, mathematical cognitive modeling, psychometrics, and clinical neuropsychology might combine their efforts to develop tools that are reliable, valid, and informed by modern brain and cognitive theory.

Can IRT be Extended to Small-Scale Research?

Estimation of item parameters for most IRT models requires large sample size. Because of this, much of the published clinical work in IRT focuses on established measures that are already routinely used in research and practice. Investigators often perform IRT analyses as a secondary use of data, and when they are able to cobble together large datasets from several archival studies. This leads to a restrictive focus of IRT. Almost no studies have been published in which an investigator uses IRT to evaluate measures or paradigms used for small-scale research. And, while there are many large, national datasets that can support IRT work, there is also a recent trend to focus on basic, experimental paradigms.

The National Institute of Mental Health (NIMH) has embarked on an era of experimental therapeutics in which the immediate goal is to identify malleable biological and psychological targets that can one day be used to design effective interventions. As part of this, the number of non-standardized testing paradigms developed to study disordered cognitive and affective processes has grown exponentially, as have in-kind discoveries but also challenges. Many experimental measures are designed for the purposes of a single study, and even those tasks that have been developed for widespread use are often validated using relatively modest samples (Ragland et al., 2012). This is a problem, because non-standardized tests share the same statistical liabilities as standardized tests, and thus failure

to rigorously evaluate and update these tools can lead to underpowered and irreproducible research.

To meet the demands of cutting-edge mental health research, psychometric modeling approaches are needed that can be applied to novel testing paradigms using only limited data. Newer statistical techniques might be used to circumvent the large sample requirements of IRT. The use of explanatory IRT models (de Boeck & Wilson, 2004), for example, can reduce estimation complexity by assuming that item parameters can be predicted by experimental task factors, such as the number of word syllables and presentation time for items used in a memory experiment. Moreover, methods for estimating mixed-effects IRT models allow certain parameters to be treated as random, thus alleviating the need to precisely quantify all parameters in a model.

Out with the Old, in with the New?

Parenthetically, many clinicians acknowledge that test selection can be driven by habit, convention, or costs, rather than ongoing research on reliability and validity. Moreover, relying on the same measure for many decades can be wise, especially given that clinicians tend to develop advanced, non-standard knowledge of how to interpret and use test scores that is only gained after years of practice. Because of this, clinical researchers and practitioners often have little incentive to use newer measures and measurement technologies. Revicki and Sloan (2007) considered these and other issues when addressing whether relevant stakeholders would be interested in using the PROMIS item bank. The authors noted that instrument developers, researchers, clinicians, the Food and Drug Administration, the pharmaceutical industry, patients, and study participants all have a stake in the development and use of modern psychometric tools. The authors concluded that it was, "… uncertain whether there is sufficient interest among key stakeholders for IRT-based measures", and that acceptance would require "demonstrated success" and "continued meetings and discussions" among stakeholders.

Ten years later, it is difficult to say, for certain, whether IRT-based measures and technologies such as PROMIS have been accepted by clinicians and other relevant stakeholders. At the very least, where IRT was only rarely seen in clinical publications just a few decades ago, it is now difficult to read any major psychological assessment journal and not find at least one article that uses IRT, or a measure developed or refined using the theory. Moreover, a tacit endorsement of IRT methods by the DSM-5 (American Psychiatric Association, 2013), as well as continued interest by the NIMH and private test developers suggests that the methodology has expanded beyond the purview of psychometricians. Nonetheless, additional efforts are needed to demonstrate the value of IRT in clinical assessment and to explain how advances can improve the day-to-day practice and research efforts of psychologists.

There appear to be two primary barriers to the adoption of IRT methods and technologies in clinical assessment. First, much of the published work claiming to show the superiority of IRT over related methods from CTT focuses on technical issues that primarily capture the attention of statisticians and psychometricians. Given that CTT is more familiar, and that test administration and scoring based on simple paper-and-pencil forms and sums of item scores

is decidedly less complex, the onus is on proponents of IRT to demonstrate that the methodology directly impacts issues that are important to clinicians such as diagnostic accuracy, and power and effect size in clinical studies. Moreover, it is likely not enough to show that IRT-based technologies such as computerized adaptive tests are superior to non-adaptive measures, but also that their benefits outweigh any added costs of the approach. Second, even if clinicians are open to the approach, IRT may seem too complicated. In reality, many of the rewards of IRT can be enjoyed without any advanced knowledge of the theory. For those clinicians who are only interested in using IRT-based technology, for example, it is not unreasonable to begin administering and interpreting scores based on computerized adaptive tests, such as the NIH PROMIS instruments, even with little technical knowledge of how they are produced. Computerized adaptive tests usually require no advanced input from the test administrator, and produce scores in metrics that are familiar to clinicians (e.g., Z-scores, T-scores, and other scaled scores).

Summary

In clinical and education assessment, IRT seems to have captured the current zeitgeist of progress. Although IRT is not entirely distinct from prior psychometric frameworks, it does, nonetheless, offer certain practical benefits. Clinical psychologists are using these benefits to more effectively combat error in clinical measurement. Work remains in developing the next generation of psychometric models and technologies, and in implementing these in the broad arenas of clinical research and practice. Nonetheless, the rapid acceptance and expansion of IRT over the last decade suggest that the methodology has become a mainstay of clinical assessment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported, in part, by the National Institute of Mental Health of the National Institutes of Health under award number K23 MH102420. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Adams RJ, Wu ML, & Wilson MR (2015). ACER ConQuest: Generalised Item Response Modelling Software (Version 4). Camberwell, Victoria: Australian Council for Educational Research.

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (text rev. 4th ed.). Washington: Author.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, D.C.: American Psychiatric Association.

Baker FB, & Kim SH (2004). Item response theory: Parameter estimation techniques (2nd ed. ed.). New York: Dekker.

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

Bejar II (1983). Introduction to item response models and their assumptions In Hambleton RK (Ed.), Applications of item response theory (pp. 1–23). Vancouver: Educational Research Institute of British Columbia.

Binet A, & Simon T (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel anormaux [New methods for the diagnosis of levels of intellectual abnormality]. Annee Psychologique, 11, 191–244.

Bock RD (1972). Estimating item parameters and latent ability when responses are scored in 2 or more nominal categories. Psychometrika, 37(1), 29–&.

Bock RD (1997). A brief history of item response theory. Educational Measurement: Issues and Practice, 16, 21–33.

Brennan RL, & Lee WC (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. Educational and Psychological Measurement, 59(1), 5–24.

Brown GG, Thomas ML, & Patt V (2017). Parametric model measurement: Reframing traditional measurement ideas in neuropsychological practice and research. The Clinical Neuropsychologist, 1–26.

Cai L, Thissen D, & du Toit SHC (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli G (1994). Origin of the scaling constant D=1.7 in item response theory. Journal of Educational and Behavioral Statistics, 19(3), 293–295.

Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, … Rose M (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap cooperative group during its first two years. Medical Care, 45(5 Suppl 1), S3–S11.

Chalmers RP (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29.

Cho S, Drasgow F, & Cao MY (2015). An investigation of emotional intelligence measures using item response theory. Psychological Assessment, 27(4), 1241–1252. [PubMed: 25961137]

Choi SW, Podrabsky T, & McKinney N (2012). Firestar-D: Computerized adaptive testing simulation program for dichotomous item response theory models. Applied Psychological Measurement, 36(1), 67–68.

Christoffersson A (1975). Factor-analysis of dichotomized variables. Psychometrika, 40(1), 5–32.

Cohen S, Mermelstein R, Kamarck T, & Hoberman HM (1985). Measuring the functional components of social support In Sarason IG & Sarason BR (Eds.), Social support: Theory, research, and applications (pp. 73–94). The Hague, Holland: Martinus Nijhoff.

Conijn JM, Emons WHM, De Jong K, & Sijtsma K (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire-45. Assessment, 22(4), 513–524. [PubMed: 25520211]

Cronbach LJ (1960). Essentials of psychological testing. New York: Harper & Brothers.

de Ayala RJ (2009). The theory and practice of item response theory. New York: Guilford Press.

de Ayala RJ, & Santiago SY (2017). An introduction to mixture item response theory models. Journal of School Psychology, 60, 25–40. [PubMed: 28164797]

de Boeck P, & Wilson M (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York, NY: Springer.

DeMars C (2010). Item response theory. Oxford; New York: Oxford University Press.

DeMars CE (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement, 43(2), 145–168.

Derogatis LR (1993). BSI, Brief Symptom Inventory: Administration, scoring, and procedures manual (4th ed.). Minneapolis, MN: National Computer Systems.

Drasgow F, Chernyshenko OS, & Stark S (2010). 75 Years After Likert: Thurstone Was Right! Industrial and Organizational Psychology-Perspectives on Science and Practice, 3(4), 465–476.

Embretson SE (1996). The new rules of measurement. Psychological Assessment, 8(4), 341–349.

Embretson SE (2010). Cognitive design systems: A structural modeling approach applied to developing a spatial ability test In Embretson SE (Ed.), Measuring psychological constructs:

Advances in model-based approaches. (pp. 71–93). Washington: American Psychological Association.

Embretson SE, & Reise SP (2000). Item response theory for psychologists. Mahwah, N.J.: L. Erlbaum Associates.

Fayers PM, Hjermstad MJ, Ranhoff AH, Kaasa S, Skogstad L, Klepstad P, & Loge JH (2005). Which Mini-Mental State Exam items can be used to screen for delirium and cognitive impairment? Journal of Pain and Symptom Management, 30(1), 41–50. [PubMed: 16043006]

Finch WH, & Pierson EE (2011). A mixture IRT analysis of risky youth behavior. Frontiers in Psychology, 2.

Forbey JD, Ben-Porath YS, & Arbisi PA (2012). The MMPI-2 Computerized Adaptive Version (MMPI-2-CA) in a Veterans Administration medical outpatient facility. Psychological Assessment, 24(3), 628–639. [PubMed: 22149324]

Fox JP (2010). Bayesian item response modeling: Theory and applications. Bayesian Item Response Modeling: Theory and Applications, 1–313.

Gavett BE, & Horwitz JE (2012). Immediate list recall as a measure of short-term episodic memory: Insights from the serial position effect and item response theory. Archives of Clinical Neuropsychology, 27(2), 125–135. [PubMed: 22138320]

Gershon RC, Cook KF, Mungas D, Manly JJ, Slotkin J, Beaumont JL, & Weintraub S (2014). Language measures of the NIH Toolbox Cognition Battery. Journal of the International Neuropsychological Society, 20(6), 642–651. [PubMed: 24960128]

Gershon RC, Rothrock N, Hanrahan R, Bass M, & Cella D (2010). The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. Journal of Applied Measurement, 11(3), 304–314. [PubMed: 20847477]

Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, … Stover A (2007). Full-information item bifactor analysis of graded response data. Applied Psychological Measurement, 31(1), 4–19.

Gibbons RD, & Hedeker DR (1992). Full-information item bifactor analysis. Psychometrika, 57(3), 423–436.

Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, & Kupfer DJ (2012). Development of a Computerized Adaptive Test for Depression. Archives of General Psychiatry, 69(11), 1104–1112. [PubMed: 23117634]

Gibbons RD, Weiss DJ, Pilkonis PA, Frank E, Moore T, Kim JB, & Kupfer DJ (2014). Development of the CAT-ANX: A Computerized Adaptive Test for Anxiety. American Journal of Psychiatry, 171(2), 187–194. [PubMed: 23929270]

Hambleton RK, & Jones RW (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12, 38–47.

Han KT (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. Applied Psychological Measurement, 36(1), 64–66.

Hathaway SR, & McKinley JC (1940). A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. Journal of Psychology, 10(2), 249–254.

Haviland MG, Warren WL, & Riggs ML (2000). An observer scale to measure alexithymia. Psychosomatics, 41(5), 385–392. [PubMed: 11015624]

Haynes SN, Smith G, & Hunsley JD (2011). Scientific foundations of clinical assessment. New York: Routledge.

Holzinger KJ, & Swineford F (1937). The bi-factor method. Psychometrika, 2(1), 41–54.

Howell D, Molloy S, Wilkinson K, Green E, Orchard K, Wang K, & Liberty J (2015). Patient-reported outcomes in routine cancer clinical practice: A scoping review of use, impact on health outcomes, and implementation factors. Annals of Oncology, 26(9), 1846–1858. [PubMed: 25888610]

International Association for Computerized Adaptive Testing. (2017). CAT Software. Retrieved from http://iacat.org/content/cat-software

Jiang SY, Wang C, & Weiss DJ (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. Frontiers in Psychology, 7, 1–10. [PubMed: 26858668]

Joreskog KG (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2p1), 183–&.

Kamata A, & Bauer DJ (2008). A note on the relation between factor analytic and item response theory models. Structural Equation Modeling, 15, 136–153.

Kathol RG, Mutgi A, Williams J, Clamon G, & Noyes R Jr. (1990). Diagnosis of major depression in cancer patients according to four sets of criteria. American Journal of Psychiatry, 147(8), 1021–1024. [PubMed: 2375435]

Khanna D, Krishnan E, Dewitt EM, Khanna PP, Spiegel B, & Hays RD (2011). The future of measuring patient-reported outcomes in rheumatology. Arthritis Care & Research, 63, S486–S490. [PubMed: 22588770]

Lambert MJ, Gregersen AT, & Burlingame GM (2004). The Outcome Questionnaire In Maruish M (Ed.), The use of psychological tests for treatment planning and outcome assessment (3rd ed., pp. 191–234). Malwah, NJ: Lawrence Erlbaum.

Lanyon RI (2010). Psychological Screening Inventory-2: Technical manual. Port Huron, MI: Sigma Assessment Systems.

Lee Y-H, & Zhang J (2010). Differential item functioning: Its consequences. Princeton, New Jersey: Educational Testing Service.

Linden W. J. v. d., & Glas CAW (2010). Elements of adaptive testing. New York: Springer.

Lord FM (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: L. Erlbaum Associates.

Madsen H, & Thyregod P (2010). Introduction to general and generalized linear models. Boca Raton: CRC Press.

Magis D, & Barrada JR (2017). Computerized adaptive testing with R: Recent updates of the package catR. Journal of Statistical Software, 76(Cn1), 1–19.

McDonald RP (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McKenna BS, Brown GG, Drummond SPA, Turner TH, & Mano QR (2013). Linking mathematical modeling with human neuroimaging to segregate verbal working memory maintenance processes from stimulus encoding. Neuropsychology, 27(2), 243–255. [PubMed: 23527652]

Meijer RR, & Sijtsma K (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25(2), 107–135.

Michel P, Baumstarck K, Lancon C, Ghattas B, Loundou A, Auquier P, & Boyer L (2016). Modernizing quality of life assessment: development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia. Quality of Life Research, 25, 111–111. [PubMed: 26198665]

Millsap RE (2011). Statistical approaches to measurement invariance. New York: Routledge/Taylor & Francis Group.

Moore TM, Calkins ME, Reise SP, Port AM, Jackson CT, Ruparel K, & Gur RE (2017). Development and public release of a computerized adaptive (CAT) version of the schizotypal personality questionnaire. submitted.

Mulaik SA (2010). Foundations of factor analysis (2nd ed.). Boca Raton, FL: CRC Press.

Mungas D, & Reed BR (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. Statistics in Medicine, 19(11–12), 1631–1644. [PubMed: 10844724]

Muthén LK, & Muthén BO (1998–2017). Mplus User's Guide. Eighth Edition Los Angeles, CA: Muthén & Muthén.

Nichols DS (2006). The trials of separating bath water from baby: A review and critique of the MMPI-2 restructured clinical scales. Journal of Personality Assessment, 87(3), 358–358.

Northwestern University. (2017). Patient-Reported Outcomes Measurement Information System. Retrieved from http://www.healthmeasures.net/explore-measurement-systems/promis

Novick MR (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3(1), 1–18.

Nunnally JC, & Bernstein IH (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Olino TM, Yu L, Klein DN, Rohde P, Seeley JR, Pilkonis PA, & Lewinsohn PM (2012). Measuring depression using item response theory: An examination of three measures of depressive symptomatology. International Journal of Methods in Psychiatric Research, 21(1), 76–85. [PubMed: 22290656]

Olino TM, Yu L, McMakin DL, Forbes EE, Seeley JR, Lewinsohn PM, & Pilkonis PA (2013). Comparisons across depression assessment instruments in adolescence and young adulthood: An item response theory study using two linking methods. Journal of Abnormal Child Psychology, 41(8), 1267–1277. [PubMed: 23686132]

Pearson. (2017). Computer-based and Computerized Adaptive Testing. Retrieved from http://www.pearsonassessments.com/research/researchpub/researchlist.topic.computer_based_and_computerized_adaptive_testing.html

Pedraza O, & Mungas D (2008). Measurement in cross-cultural neuropsychology. Neuropsychology Review, 18(3), 184–193. [PubMed: 18814034]

Pedraza O, Sachs BC, Ferman TJ, Rush BK, & Lucas JA (2011). Difficulty and discrimination parameters of Boston Naming Test items in a consecutive clinical series. Archives of Clinical Neuropsychology, 26(5), 434–444. [PubMed: 21593059]

Rabin LA, Spadaccini AT, Brodale DL, Grant KS, Elbulok-Charcape MM, & Barr WB (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. Professional Psychology-Research and Practice, 45(5), 368–377.

Ragland JD, Ranganath C, Barch DM, Gold JM, Haley B, MacDonald AW, … Carter CS (2012). Relational and Item-Specific Encoding (RISE): Task development and psychometric characteristics. Schizophrenia Bulletin, 38(1), 114–124. [PubMed: 22124089]

Ranger J, & Ortner TM (2011). Assessing personality traits through response latencies using item response theory. Educational and Psychological Measurement, 71(2), 389–406.

Reckase MD (2009). Multidimensional item response theory. New York: Springer.

Reise SP (2010). Thurstone Might Have Been Right About Attitudes, but Drasgow, Chernyshenko, and Stark Fail to Make the Case for Personality. Industrial and Organizational Psychology-Perspectives on Science and Practice, 3(4), 485–488.

Reise SP, & Henson JM (2000). Computerization and adaptive administration of the NEO PI-R. Assessment, 7(4), 347–364. [PubMed: 11151961]

Reise SP, Moore TM, & Haviland MG (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of Personality Assessment, 92(6), 544–559. [PubMed: 20954056]

Reise SP, & Waller NG (2009). Item response theory and clinical measurement. Annual Review of Clinical Psychology, 5, 27–48.

Reise SP, & Yu JY (1990). Parameter recovery in the graded response model using Multilog. Journal of Educational Measurement, 27(2), 133–144.

Revicki DA, & Sloan J (2007). Practical and philosophical issues surrounding a national item bank: If we build it will they come? Qual Life Res, 16 Suppl 1, 167–174. [PubMed: 17468940]

Rizopoulos D (2006). ltm: An R package for latent variable modeling and item response theory analyses. Journal of Statistical Software, 17(5), 1–25.

Rubright JD, Nandakumar R, & Karlawish J (2016). Identifying an appropriate measurement modeling approach for the Mini-Mental State Examination. Psychological Assessment, 28(2), 125–133. [PubMed: 26029945]

Sacco P, Casado BL, & Unick GJ (2011). Differential item functioning across race in aging research: An example using a social support measure. Clinical Gerontologist, 34(1), 57–70.

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika, 34(4p2), 1–&.

Scalise K, & Allen DD (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. British Journal of Mathematical & Statistical Psychology, 68(3), 478–496. [PubMed: 26061260]

Simms LJ, Goldberg LR, Roberts JE, Watson D, Welte J, & Rotterman JH (2011). Computerized adaptive assessment of personality disorder: introducing the CAT-PD project. Journal of Personality Assessment, 93(4), 380–389. [PubMed: 22804677]

Skrondal A, & Rabe-Hesketh S (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Boca Raton, FL: Chapman and Hall/CRC.

Spearman C (1904). “General intelligence “ objectively determined and measured. American Journal of Psychology, 15, 201–292.

Spence R, Owens M, & Goodyer I (2012). Item response theory and validity of the NEO-FFI in adolescents. Personality and Individual Differences, 53(6), 801–807. [PubMed: 23049153]

Stone M, & Yumoto F (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. Journal of Applied Measurement, 5(1), 48–61. [PubMed: 14757991]

Strauss ME, & Smith GT (2009). Construct validity: Advances in theory and methodology. Annual Review of Clinical Psychology, 5, 1–25.

Sunderland M, Slade T, Krueger RF, Markon KE, Patrick CJ, & Kramer MD (2017). Efficiently measuring dimensions of the externalizing spectrum model: Development of the Externalizing Spectrum Inventory-Computerized Adaptive Test (ESI-CAT). Psychological Assessment, 29(7), 868–880. [PubMed: 27841446]

Takane Y, & Deleeuw J (1987). On the relationship between item response theory and factor-analysis of discretized variables. Psychometrika, 52(3), 393–408.

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, … Cella D (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly, 51(2), 148–180. [PubMed: 20336180]

Thomas ML (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. Psychological Assessment, 24(1), 101–113. [PubMed: 21767026]

Thomas ML, & Lanyon RI (2012). A Bayesian item response theory approach to symptom validity detection: Evaluating Psychological Screening Inventory-2 response profile likelihoods. Psychology Injury and Law, 5(3–4), 221–234.

Thomas ML, Patt VM, Bismark A, Sprock J, Tarasenko M, Light GA, & Brown GG (2017). Evidence of systematic attenuation in the measurement of cognitive deficits in schizophrenia. Journal of Abnormal Psychology, 126(3), 312–324. [PubMed: 28277736]

Toland MD (2015). Practical Guide to Conducting an Item Response Theory Analysis (vol 34, pg 120, 2014). Journal of Early Adolescence, 35(3), Np2–Np2.

van der Linden WJ, & Hambleton RK (1997). Handbook of modern item response theory. New York: Springer.

van der Maas HLJ, Molenaar D, Maris G, Kievit RA, & Borsboom D (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. Psychological Review, 118(2), 339–356. [PubMed: 21401290]

Varni JW, Magnus B, Stucky BD, Liu Y, Quinn H, Thissen D, … DeWalt DA (2014). Psychometric properties of the PROMIS (R) pediatric scales: Precision, stability, and comparison of different scoring and administration options. Qual Life Res, 23(4), 1233–1243. [PubMed: 24085345]

Weinstock LM, Strong D, Uebelacker LA, & Miller IW (2009). Differential item functioning of DSM-IV depressive symptoms in individuals with a history of mania versus those without: An item response theory analysis. Bipolar Disorders, 11(3), 289–297. [PubMed: 19419386]

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, … Gershon RC (2013). Cognition assessment using the NIH Toolbox. Neurology, 80, S54–S64. [PubMed: 23479546]

Public Significance Statement

Item response theory (IRT) is a methodology used to develop, evaluate, and score psychological tests. This paper reviews the literature for recent advancements in the use of IRT as applied to clinical research and practice. IRT has contributed to major advancements in computerized adaptive testing, multidimensional modeling, and the identification of bias, but there are still many opportunities to further develop applicatios of the theory.

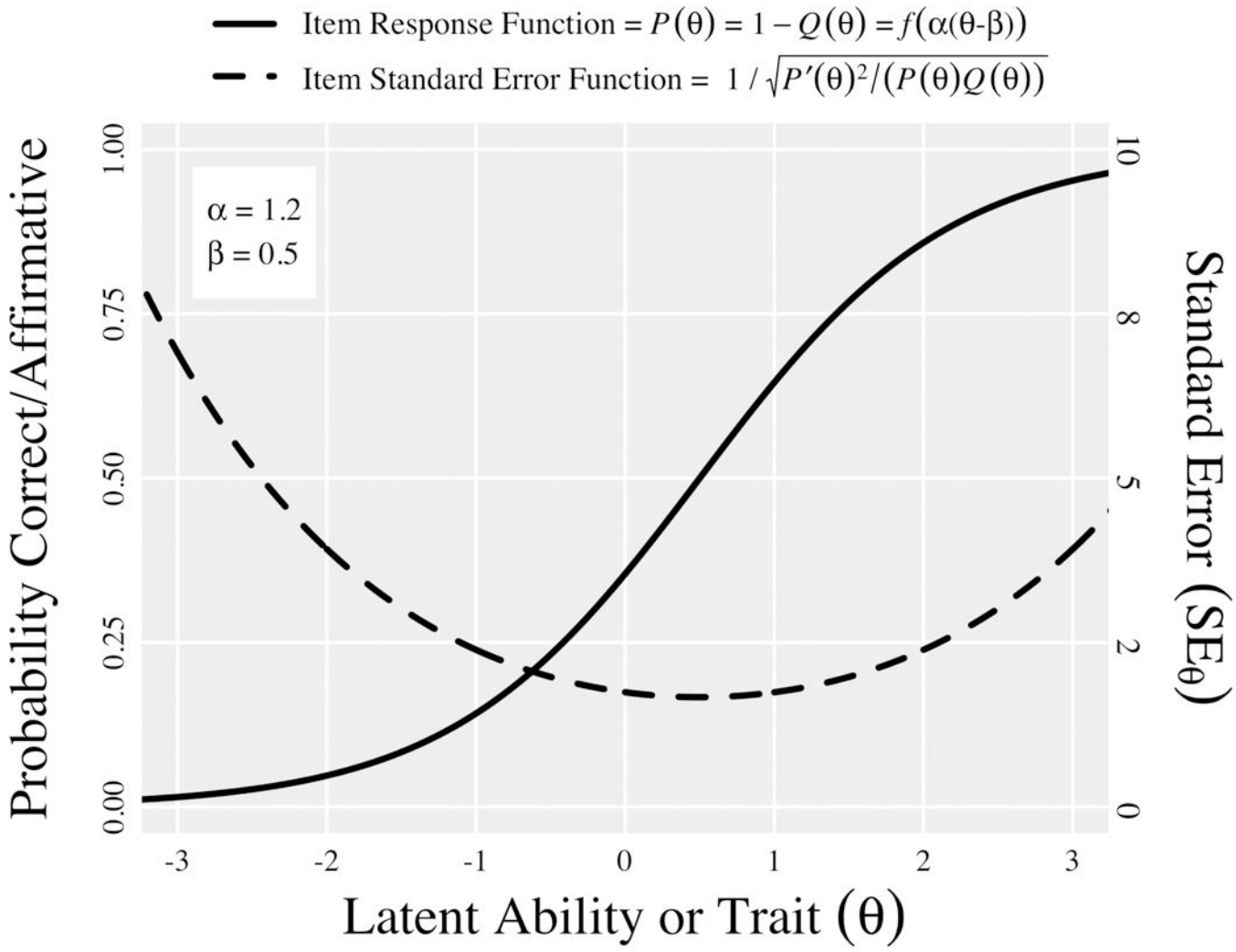

Figure 1.

Item response and standard error of estimate functions for a hypothetical item scored using the two-parameter logistic item response model. The function *f* is the logistic function: 1/1+exp(−x). The prime symbol indicates the derivative of function *f*.

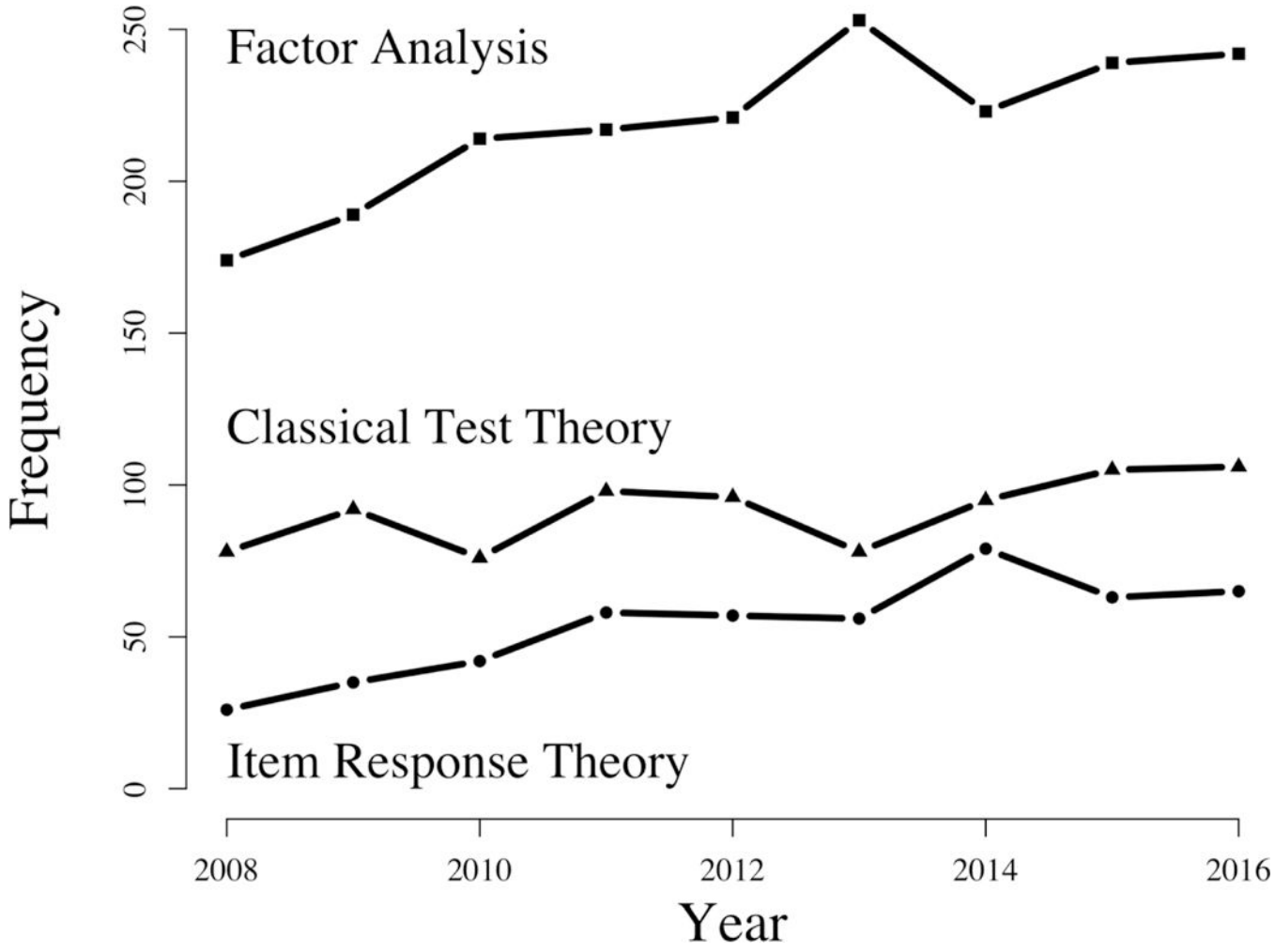

Figure 2.
Frequency of key words, phrases, and acronyms relevant to item response theory, classical test theory, and factor analysis found of Abstracts from the top 50 journals in clinical psychology journals between 2008 and 2016.

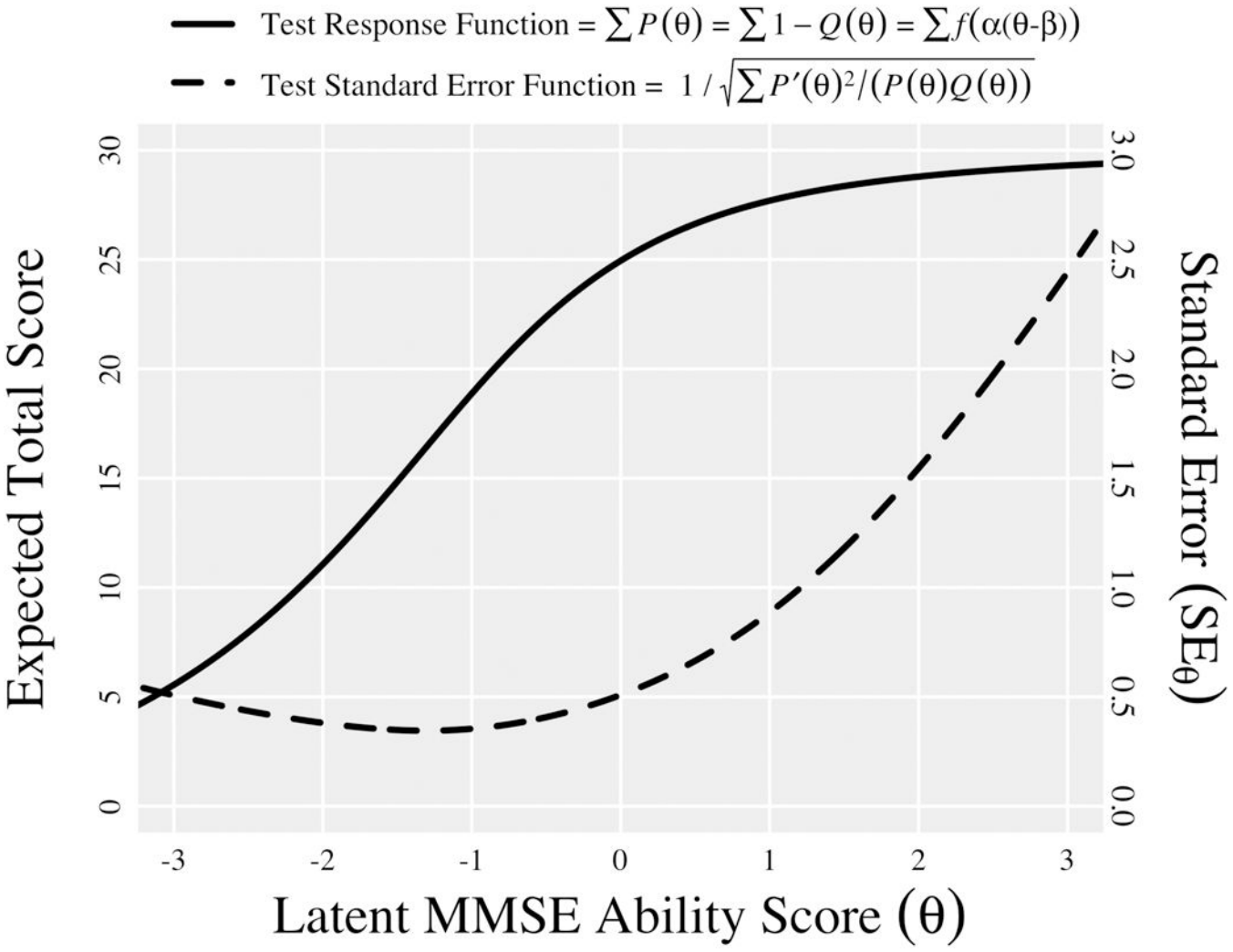

Figure 3.

Test response and standard error of estimate functions for Mini-Mental State Examination (MMSE) item parameters reported in Table 3 of Fayers et al. (2005). The function *f* is the logistic function: 1/1+exp(−x). The prime symbol indicates the derivative of function *f*. Summations are over items. Fayers et al. (2005) only report parameters for 20 items; therefore, the test response function was rescaled to have a new maximum value of 30 so that the MMSE range would match what is commonly reported in the literature.

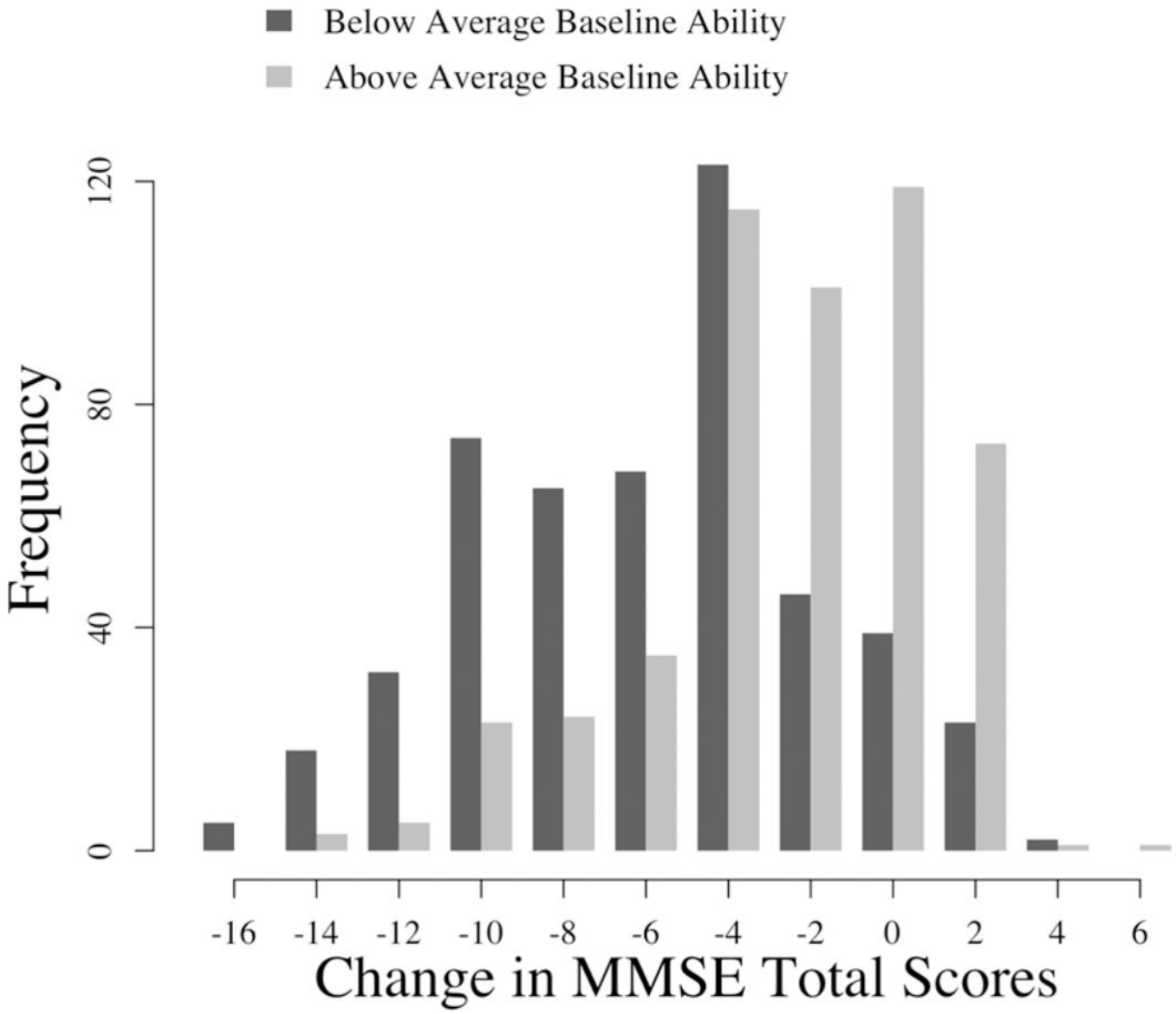

Figure 4.

Results of simulated change in Mini-Mental State Examination (MMSE) total scores. Hypothetical samples of examinees were drawn from normal distributions assuming either below average baseline ability, $N \sim (-1.5, 1.0)$, or above average baseline ability, $N \sim (1.5, 1.0)$. Random latent change scores, $N \sim (-1.0, 0.1)$, were added to the second observation in order to simulate decline in ability. Samples of 500 cases were drawn for each group. Change in MMSE total scores based on simulated item response data are shown. As can be seen, observed declines were relatively larger in the below average baseline sample, despite both groups experiencing the same average amount of true decline.

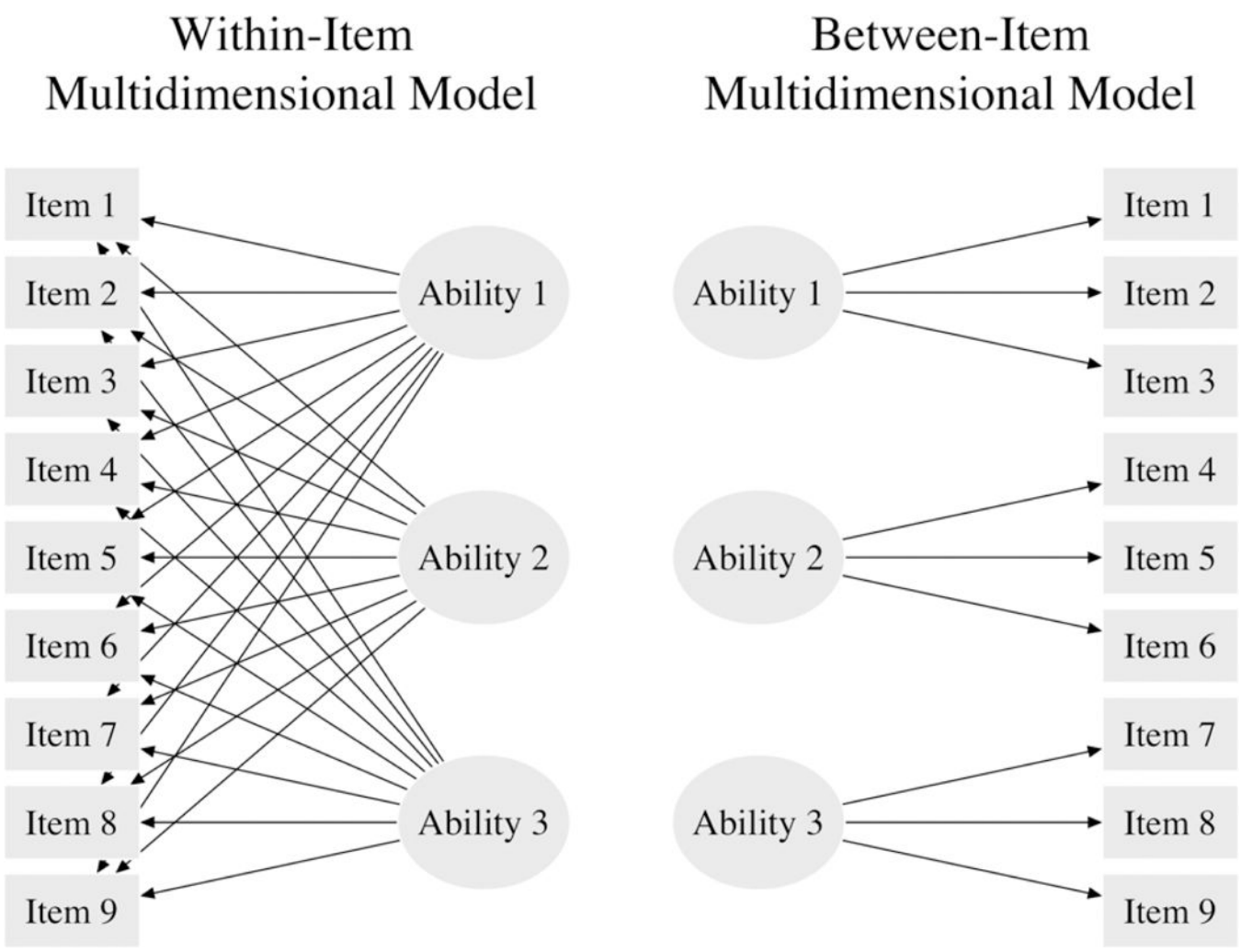

Figure 5.
Within-item multidimensional model versus between-item multidimensional model. Ovals are latent abilities and rectangles are observed item scores.