Item Response Theory and Clinical Measurement

Steven P. Reise[1] and Niels G. Waller[2]

[1]Department of Psychology, University of California, Los Angeles 90095; email: reise@psych.ucla.edu

[2]Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455-0344; email: nwaller@umn.edu

Annu. Rev. Clin. Psychol. 2009. 5:27–48

First published online as a Review in Advance on October 31, 2008

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

This article's doi: 10.1146/annurev.clinpsy.032408.153553

1548-5943/09/0427-0027$20.00

Key Words

computerized adaptive testing, differential item functioning, linking scales, scale information curve, latent trait metric, quasi-trait

Abstract

In this review, we examine studies that use item response theory (IRT) to explore the psychometric properties of clinical measures. Next, we consider how IRT has been used in clinical research for: scale linking, computerized adaptive testing, and differential item functioning analysis. Finally, we consider the scale properties of IRT trait scores. We conclude that there are notable differences between cognitive and clinical measures that have relevance for IRT modeling. Future research should be directed toward a better understanding of the metric of the latent trait and the psychological processes that lead to individual differences in item response behaviors.

Contents

INTRODUCTION

Item response theory (IRT; Embretson & Reise 2000) is a set of psychometric models for developing and refining psychological measures, administering scales, and scaling individual differences. Over the past several decades, these models have profoundly changed the administration and scoring of large-scale aptitude tests, state-wide achievement tests, and professional licensure exams. More recently, IRT has also been adopted by health outcomes researchers (Orlando-Edelen & Reeve 2007). Although there has been no shortage of scholars championing the potential of IRT in these domains (e.g., Reise et al. 2005), its use in personality and psychopathology measurement has lagged behind that of other areas. A perusal of *Journal of Abnormal Psychology*, *Journal of Personality Assessment*, and *Psychological Assessment* reveals that application of IRT in clinical assessment remains the exception and not the rule.

Nevertheless, there are signs that clinical researchers are beginning to use IRT with greater frequency. One sign is that the literature has moved beyond the stage populated with didactic articles addressing "Why use IRT modeling?" (Hays et al. 2000), or "Why IRT is superior to traditional methods" (Reise & Henson 2003), to a more mature stage that includes dozens of informative applications. This welcome development has revealed many compelling advantages of IRT over traditional psychometric practices. However, it has also uncovered many conceptual and practical challenges. In this review, we highlight some of these challenges and demonstrate that the translation of IRT from the cognitive to the clinical domain is less straightforward than suggested in many articles and texts.

IRT: item response theory

Important Item Response Theory Concepts

To model item responses to a clinical instrument, a researcher must first assume that the item covariation is caused by a continuous latent variable (common factor). IRT is thus an "effects" indicator model (Bollen & Lennox 1991) in the sense that the latent trait presumably "causes" the item response variance. Although this may appear to be a trivial observation, it is a critical recognition for two reasons. First, it sets limits on the type of constructs that can be appropriately modeled by IRT. Second, it has ramifications for how latent variables should be validated. An IRT model earns validity when it

helps elucidate how and why changes on a latent variable cause changes in item responding (for a similar view, see Borsboom 2005). Notice that this view downplays the traditional role of establishing a network of external correlations with other scales and instead emphasizes the study of response processes and the meaning of latent traits.

A primary objective of IRT is to estimate the parameters of a mathematical function, typically a logistic function, that "models" the relation between the latent trait and the item responses. For binary items, this function is referred to as an item response curve (IRC). For polytomous items, the relation between the latent trait and a category option is called a category response curve (CRC). A commonly applied IRT model for binary items is the two-parameter logistic (2PL) model. It is called a "two"-parameter model because items are allowed to differ in two ways—discrimination and location (severity). A popular polytomous IRT model is the graded-response model (Samejima 1969). This model can be viewed as a generalization of the 2PL model.

Throughout this review, we emphasize that IRCs and CRCs—and all indices derived from these functions (e.g., information curves)—must be interpreted in reference to a well-defined metric. Of course, this same principal holds true in the common factor model, although it is seldom recognized because popular software routinely standardizes factor loadings.

After estimating the parameters of an IRT model, researchers can investigate the fidelity by which items measure a latent trait. In IRT, the concept of score fidelity or reliability is replaced by the concepts of item and test information (Samejima 1977). If the item responses are locally independent (Embretson & Reise 2000, p. 231), then an item information curve can be summed across items to produce a scale information curve (SIC). This latter function is important because the standard error of a (maximum likelihood) trait estimate is inversely related to the square root of the SIC. Estimating an SIC and exploring how the standard error of measurement changes as a function of trait level is perhaps the most widely cited rationale for using IRT in clinical measurement.

This is but one example of how IRT and traditional psychometrics differ in their evaluation of items and scales. A second example lies in the scaling of individual differences. As noted in many sources (e.g., Embretson & Reise 2000, pp. 158–186), from an IRT perspective, individuals are assumed to have a "true" position on a latent continuum that can be identified independently of a particular item pool (under appropriate restrictions). This property is a critical feature that underlies the logic of many IRT applications such as computerized adaptive testing (CAT) and scale linking.

Overview

The remainder of this review is divided into three major sections. First, we examine studies that have used IRT to explore the psychometric properties of clinical measures. We then review three important applications of IRT that have relevance for clinical research: linking (Kolen & Brennan 1995), CAT (Wainer et al. 2000), and differential item functioning (DIF) analysis (Holland & Wainer 1993). Finally, we examine the implications of scoring individuals using IRT models and call attention to heretofore neglected challenges in applying IRT models to clinical scales.

THE PSYCHOMETRIC ANALYSIS OF CLINICAL MEASURES

Scale Analysis

In many clinical studies, IRT has been used to explore the psychometric properties of existing clinical instruments. These studies often appear with titles of the following form: "An IRT Analysis of ____" (fill in your favorite scale or construct name), and contain statements such as "IRT was used to assess the strength of the relationship between the items and the constructs of interest and the information available across the latent construct" (Hill et al. 2007, p. S39).

Item response curve (IRC): a probabilistic function that relates item responses to an underlying latent trait

CRC: category response curves

Item information curve: a function displaying the amount of psychometric information across the range of the latent trait

Scale information curve (SIC): the sum of the item information curves for a scale

Computerized adaptive testing (CAT): an IRT method for constructing and administering individually tailored tests

Scale linking: a method of placing items from distinct measures onto a common metric

Differential item functioning (DIF): an IRT technique for studying item and test bias by identifying items that have different psychometric characteristics in two or more groups

Examples of this type of research can be found in the study of attachment (Fraley et al. 2000), anxiety disorders (Krueger & Finger 2001), schizophrenia (Bell et al. 1994, Santor et al. 2007), alcohol problems (Krueger et al. 2004, Saha et al. 2006), interpersonal problems (Kim & Pilkonis 1999), depression (Aggen et al. 2005, Stansbury et al. 2006), distress (Ferrando 2001, Kessler et al. 2002), social inhibition (Emons et al. 2007), physical functioning (Hays et al. 2007), obsessive-compulsive disorder (Uher et al. 2008), and quality of life (e.g., Hill et al. 2007, Uttaro & Lehman 1999).

A review of this literature reveals several striking contrasts between the IRT modeling of clinical and cognitive tests. For instance, in clinical assessment, item discrimination parameters (slopes of IRCs) are often surprisingly high (i.e., greater than 2.5; logistic metric), suggesting that the measured construct is conceptually narrow. As an example, in a recent evaluation of the criteria for major depression in the *Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised* (DSM-III-R; American Psychiatric Association 1987), Aggen et al. (2005) report item discrimination estimates of 3.21 and 2.59 (logistic metric) for the items "depressed" and "disinterest," respectively. Hays et al. (2007) report item discriminations ranging from 1.88 to 4.24 on a physical functioning measure, and Chan et al. (2004) report item discriminations of 4.43 ("could not shake the blues") and 4.14 ("felt sad") on a popular depression measure.

Unusual results have also been reported for item threshold parameters in clinical applications of polytomous IRT models. For instance, threshold parameters (*a*) are frequently clustered within a limited range of the latent trait as opposed to being spread out across the trait range and (*b*) often have extreme values (i.e., >3 and <-3). For an example of this latter finding, Gomez et al. (2005) analyzed a four-option reward responsiveness scale and found that threshold estimates for the first category ranged from –7 to –4 and that estimates for the second category ranged from –3 to –2. These results imply that even people who are two standard deviations below the mean on the latent trait (in this case, reward responsiveness) are likely to respond in the highest response categories (i.e., categories 3 and 4). For one item, "it would excite me to win a contest" the third threshold was –0.15 indicating that even people below the mean on this positive trait are likely to respond in the highest category (category 4).

Reports of high discriminations, a restricted range of thresholds, and extreme threshold values are by no means limited to a few studies in this literature. Rather, such results appear to be normative. To further explore this conclusion, consider a counter example. Uttaro & Lehman (1999) applied IRT to a multifaceted seven-option measure of quality of life. What is provocative in this study is that all item discrimination estimates were in a reasonable range (0.97 to 1.89; logistic metric), and the six item thresholds were evenly spread across the trait range from around –2.5 to 2.5.

This study differs from the typical IRT study in clinical assessment in two consequential ways. First, many IRT applications used scales that assess conceptually narrow constructs and consequently contain homogeneous item content (headache impact, alcohol problems, depression). In such cases, high item discriminations are expected due to high item intercorrelations. In contrast, Uttaro & Lehman (1999) analyzed a multifaceted scale having three to six items measuring each of seven distinct aspects of quality of life (e.g., leisure, social, family). This content heterogeneity reflects a broader construct and thus the scale had lower item intercorrelations and smaller item discriminations. A second distinguishing feature of the Uttaro & Lehman (1999) study was that scores on their quality-of-life measure were normally distributed. In the clinical literature we reviewed, scale scores tended to be positively skewed and threshold parameters clustered in the high (pathological) trait range.

We believe that differences between the Uttaro & Lehman (1999) and other studies are noteworthy because they highlight potential differences between cognitive and clinical

constructs and indicate subtle problems of applying IRT to this latter domain. For instance, many clinical instruments have peaked scale information curves with maximum information in the pathological trait range. The following quotes illustrate the ubiquity of the problem: " . . . items are most informative at the higher end" (Emmons et al. 2006, p. 27); "items discriminated well at the more severe end of the depression latent trait" (Sharp et al. 2006, p. 379); "The CY-BOCS discriminated better at the severe end of the spectrum" (Uher et al. 2008, p. 979); " . . . the DSM-III-R MD criteria are insensitive to discriminating at low levels of risk in a population-based sample" (Aggen et al. 2005, p. 481). These findings suggest that clinical instruments possess a limited range of item location parameters and provide measurement precision in only a narrow portion of the trait continuum. Similar findings are reported for instruments that use polytomous response formats (which is surprising since this format should spread the item information over the trait continuum).

Prior to offering a tentative explanation of these results, we remind readers that an SIC, like all other IRT indices, must be interpreted relative to the metric of a calibration sample. Some of the aforementioned studies used clinical samples and thus the mean of the latent trait reflects the mean of those specific populations. Other studies have used community samples and thus the mean of the latent trait ostensibly reflects a general-population norm. In still other studies, combined community and clinical samples were used, yielding a mean on the latent trait that is difficult to interpret.

Regardless of the sample used, many studies reported peaked information for elevated trait scores. A case in point is the recent Krueger et al. (2004) investigation of 110 dichotomous indicators of alcohol problems in an adult male community sample. Similar to other studies in this area, Kreuger et al. found that the majority of their item location parameters ranged between 1.0 and 2.7. This suggests that the items provide maximum information for individuals who are at least one standard deviation higher than the mean of this community sample. Moreover, for individuals who are at two standard deviations above the mean, scale information was approximately 140 (see Krueger et al. 2004, Figure 3). At this level, the standard error of a trait estimate is approximately 0.08. On the other side of the scale, information was near 10 for individuals who are at two standard deviations below the mean. This value implies a standard error of 0.30. Considered together, these results suggest that standard errors are four times larger for low-trait individuals relative to high-trait individuals. What is remarkable about these results is that the relatively larger standard error for low-trait individuals required administration of 110 items (i.e., the entire item set) to achieve.

What do such findings imply about the nature of psychopathology and health outcomes measurement? We believe that the observed peaked information functions are not due to deficient scale construction (e.g., poor sampling of the content domain), a product of using dichotomous (in contrast to polytomous) items, or use of a community as opposed to a clinical sample. Rather, we believe that the peaked information function for many clinical scales reflects the quasi-trait status of many psychopathology constructs. By the term "quasi-trait," we mean that the trait is unipolar (relevant only in one direction) and that variation at the low end of the scale is less informative in both a substantive as well as a psychometric sense. For example, the low end of depression is not happiness but rather the lack of depression; the low end of narcissism is not self-hatred but rather an absence of self-absorption; the low end of physical problems is not athleticism but rather an absence of mobility concerns.

The existence of quasi-traits, and their associated peaked (in the severe trait range) information curves, is consequential for many IRT applications. For example, many studies that find a peaked information curve conclude, "Items need to be written to provide information in the low end of the continuum." If a construct is really a quasi-trait, such attempts might be wasted effort. Second, and

Quasi-trait: a unipolar construct in which one end of the scale represents severity and the other pole represents its absence (depression versus not depressed). This is in contrast to a bipolar construct, where both ends of the scale represent meaningful variation (depression versus happiness)

related to the first problem, when working with a quasi-trait it will be difficult to find items that have information spread across the trait range. In turn, this presents unique challenges to computerized adaptive testing (reviewed below). Additionally, studying change scores on quasi-traits may be especially problematic due to the markedly different precision for individuals at different trait ranges. Fraley et al. (2000) illustrated this point in an IRT analysis of adult attachment measures.

Exploring the Continuity Hypothesis of Psychiatric Syndromes by Item Response Theory

In this section, we review studies that have used IRT to investigate the underlying continuity of psychiatric symptoms. We begin by discussing a study by Aggen et al. (2005). These authors used IRT to study the scaling properties of the DSM-III diagnostic criteria for major depression (see also Sharp et al. 2006 for an IRT study of the scalability of depression symptoms in children). Using factor analysis and IRT, the authors argued that the DSM-III-R criteria for major depression form a "reasonably coherent unidimensional scale" (Aggen et al. 2005, p. 475). They also discuss several advantages of using IRT-based scoring when forming clinical profiles. They conclude that "item response models that treat symptoms as ordered indicators of risk rather than as counts towards a diagnostic threshold more fully exploit the information available in symptom endorsement data patterns" (p. 475).

In a related vein, Krueger et al. (2004) used IRT to determine whether alcohol problems could be scaled onto a common dimension (see also Saha et al. 2006 for a similar work). These authors note that clinicians have long debated the typological or dimensional status of alcohol-related behaviors. To address this issue, the authors fit an IRT model to 110 alcohol use/abuse indicators in a large community sample of men. Overall, their findings supported the hypothesis that alcohol-related behaviors can be scaled along a single continuum. Specifically, item location parameters aligned along a continuum from "intoxication and heavier use . . . through abuse . . . through persistence of problems for longer periods of time . . . to dependence . . . to very serious medical and psychological complications . . . " (Krueger et al. 2004, p. 116).

Exploring Alternative Item Responses Models

Although they are applied almost universally in clinical applications, not all researchers agree that the 2PL (dichotomous) or graded-response (polytomous) models are appropriate vehicles for characterizing item response behavior in this domain. To address this concern, in this section we review lines of research that suggest alternative models that may be more appropriate for clinical data, namely (*a*) a 4-parameter model, (*b*) an ideal-point model, and (*c*) nonparametric modeling. Finally, we describe a literature that challenges the response process at the level of the individual. This research addresses the important notion of person-fit.

4-Parameter models. In two of our recent papers (Reise & Waller 2003, Waller & Reise in press), we have questioned the adequacy of the 2PL model for representing psychopathology items. Specifically, using a large clinical sample, we examined "empirical" IRCs from several factor scales of the Minnesota Multiphasic Personality Inventory-Adolescent assessment. Empirical IRCs show item endorsement rates plotted as a function of corrected total scores. Careful examination of these plots revealed that many items violate a basic feature of the 2PL. Namely, the plots showed that many items did not have lower asymptotes of 0.00 and upper asymptotes of 1.00, as required by the 2PL. A 4PL model including upper and lower asymptote parameters would be needed to accommodate such findings.

Although the majority of items conformed to the 2PL, for several items the response probability for low-scoring individuals was greater than zero. Conversely, the empirical IRCs indicated that for many items, the observed

response proportion did not asymptote at 1.00 at the high end of the scale. Specifically, for individuals who scored in the highest trait range, the observed response proportions achieved maximum values between 0.50 and 0.80. These findings suggest that even in a group of individuals who score in the severe range of a latent trait, the probability that an individual will manifest a particular symptom is less than—sometimes substantially less than—100%. Such findings could not be identified by standard IRT models, and we were forced to program a Bayesian 4PL model to investigate these issues. An important finding from this research is that fitting a 2PL model to 4PL data can lead researchers to form an overly positive evaluation of a scale's measurement precision.

Ideal-point models. An implicit (and thus, rarely considered) assumption of factor analysis, traditional item analysis, and standard IRT models is that the data conform to a dominance-response process (Coombs 1964). This model implies that increasing trait levels are reflected in higher item endorsement probabilities.

The aforementioned research by Reise and Waller maintains a dominance view but suggests limits on the process. Stark et al. (2006) go one step further and ask whether the dominance model is universally valid for personality and other noncognitive items (see also Chernyshenko et al. 2007). These researchers suggest that an ideal-point model may be appropriate in these domains. In an ideal-point model, an individual is more likely to endorse an item when their trait level is near the item's location and less likely to endorse an item as their trait level deviates from the item's location. This feature of the model produces a bell-shaped IRC.

Stark et al. (2006) fit both dominance- and ideal-point models to scales from a well-known personality inventory (see also Weekers & Meijer 2008). Results indicated that several items operated more consistently with an ideal-point representation than a dominance representation. Moreover, they found that failure to properly model the items (i.e., applying the 2PL instead of the ideal-point model) resulted in meaningful changes in the rank ordering of individuals. In turn, these authors suggest that this could have serious implications if a scale, incorrectly analyzed, were to be used for hiring decisions (where assessment focuses on the rank order of applicants).

Nonparametric IRT. Over the past 10 years, there has been increasing interest in nonparametric IRT analyses (Meijer & Baneke 2004). In this context, nonparametric IRT modeling refers mainly to two related types of analysis. One technique uses graphical plots for intensive exploration of response category functioning. This approach was illustrated by Santor and colleagues (Santor et al. 1994, Santor & Ramsey 1998), who used graphical techniques to identify ordering problems with the response categories on a popular depression measure. A second type is known as Mokken scaling (Sijtsma & Molenaar 2002). This method also uses graphical and statistical analyses to test more fundamental psychometric features of the data, such as the monotonicity of the item response function and scalability of the total test scores.

Importantly, in nonparametric IRT analysis, the existence of a quantitative latent variable is not postulated (analyses are conducted with raw scores). Moreover, as the name implies, no formal parametric model (e.g., the 2PL or graded-response model) is used to characterize item responses. For example, rather than fitting a 2PL model, a nonparametric researcher would begin by graphing, for each item, a function relating item endorsement proportions to total test scores. This "empirical" IRC is free to take on any shape. Thus far, nonparametric analyses have been used widely (*a*) to overcome limitations of classical test theory item statistics, (*b*) to justify ordering individuals on the basis of raw scores, (*c*) to justify the application of a parametric IRT model, and (*d*) as a tool for the detailed analysis of response category functioning.

A excellent example of this latter use is reported in Santor et al. (2007). These authors

CTT: classical test theory

used empirical CRCs to study the category options of the Positive and Negative Syndrome Scale (Kay et al. 1987). Findings suggested that the empirical response curves for many response categories overlapped (which represents a violation of standard IRT models). For example, on several items, a response of "2" did not indicate a higher trait level than a response of "1." Substantively, this suggests that trained clinicians cannot reliably rate differences between the two categories on this measure, perhaps because the construct is ill-defined at that end (e.g., quasi-trait).

Person-Fit: What Does It Mean?

The analytic methods described above suggest alternatives to the standard parametric item response curves. The methods described in this section evaluate the interpretability of individual trait scores under an assumed IRT model. Because IRT uses mathematical models to relate trait levels to item responses, it is possible to evaluate model fit at the global, item, or person levels. Indeed, this idea is represented in a sizable number of person-fit indices (scalability, appropriateness) that assess the degree to which an individual's item response pattern is consistent with an IRT model (Meijer & Sijtsma 2001). This idea recognizes that some item response patterns cannot be meaningfully interpreted.

Reise & Flannery (1996) reviewed the potential of person-fit assessment in personality and psychopathology research and found that this approach has been relatively unsuccessful at identifying dissimulation. Ferrando & Chico (2006, p. 1009) also concluded, "the person-fit indices based on the results given by the 2PLM fail to detect deliberate dissimulation to any practical degree." In our opinion, these results are not surprising because faked protocols are mostly characterized by a mean shift in item responses, not response inconsistency. However, the use of person-fit indices as a tool for identifying model-inconsistent item response patterns has been more successful as revealed in the following research.

In a study aimed at identifying the psychological causes of misfitting item response patterns, Meijer et al. (2008) recently conducted an extensive study of children who responded to a popular measure of self-concept. After identifying several children with poor person-fit, and replicating the results with repeated assessment, interviews were conducted with the children and their teachers to identify the causes of misfit. The authors concluded, "For some children in the sample, item scores did not adequately reflect their trait level. Based on teacher interviews, this was found to be due most likely to a less developed self-concept and/or problems understanding the meaning of the questions" (Meijer et al. 2008, p. 227).

An important take-home message of the Meijer et al. (2008) study is that individuals who display poor person-fit are not necessarily merely generating random or otherwise faulty responses. Indeed, there may be interesting psychological reasons why an individual's response pattern is inconsistent with a model (see also Reise & Waller 1993, Waller & Reise 1992). In short, researchers should not assume that all individuals can be meaningfully scaled on a construct.

ITEM RESPONSE THEORY APPLICATIONS: LINKING, COMPUTERIZED ADAPTIVE TESTING, AND DIFFERENTIAL ITEM FUNCTIONING ASSESSMENT

Herein we review research that has used IRT to achieve specific analytic goals that are not well handled by classical test theory (CTT). We begin by citing research that has used IRT-based linking (Kolen & Brennan 1995) to place items from distinct measures onto a common scale. We then review how IRT has changed the way scales are administered and scored via CAT (Wainer et al. 2000). Next, we call attention to the large literature that has used IRT to study scale equivalence across sociodemographic groups by studying DIF (Teresi & Fleishman 2007).

Item Response Theory and Linking

During the past decade, several researchers have used IRT "linking" procedures to place personality or psychopathology items from separate measures onto a common scale. Steinberg (2001, 2008), for example, has used linking techniques to study context effects (whether an item's ordinal position in an inventory affects responses), and the psychometric properties of an item as a function of the number of response alternatives. For a second example, Walton et al. (2008) used a common persons linking design to study whether items from normal- and abnormal-range personality scales could be placed on the same scale. Study results were used to argue for the continuity of personality and psychopathology traits.

Linking methods have also been used to place different types of measures onto a common scale. For example, Uher et al. (2008) studied the comparability of self, parent, and interview measures of obsessive-compulsive disorder in children, and Chan et al. (2004) used linking to investigate mode effects (phone versus mail) on a depression measure. Although the results of these studies are provocative, we would counsel researchers to proceed cautiously before applying linking in contexts where multidimensionality may result from method factors. The validity of a linking study hinges critically on the degree to which to all model assumptions are satisfied.

Finally, in a notable example of how IRT linking can be used to bring order to a construct domain, McHorney & Cohen (2000) report that there are over 75 self-report instruments to measure the construct of physical "functional status." To address this Tower of Babel, these authors used a common (anchor) item linking design to place items from different instruments onto a common scale. This effort produced a well-calibrated item pool with information at trait levels ranging from –1.5 (wash hands, use a spoon, take medications) to 1.7 (scrub floor on hands and knees, paint walls, walk 2 miles).

Computerized Adaptive Testing

A chief motivation behind the expansion of IRT in large-scale cognitive testing is the recognition that IRT can facilitate the development of an item bank (see above cited linking studies) and efficient computerized adaptive tests (Cook et al. 2005). For instance, in cognitive testing, CAT has been shown to significantly shorten assessments without sacrificing score fidelity (Wainer et al. 2000). Moreover, through building item banks, researchers can "standardize" measurement within fields that are characterized by multiple competing measures. This latter point was successfully illustrated by the Patient Reported Outcomes Measurement Information System (PROMIS; Cella et al. 2007) initiative. To our knowledge, this project is the most ambitious application of IRT outside of cognitive testing. The primary goal of PROMIS is to create item banks, and subsequent CATs, to better standardize the measurement of physical (e.g., fatigue), mental (e.g., depression), and social (e.g., social role participation) health outcomes.

Item bank: a comprehensive pool of items that measure a common trait

Given the attractive advantages of computerized testing, it is not surprising that clinical researchers have vigorously pursued the feasibility of CAT. For instance, Ware et al. (2000) successfully used item banking and CAT to measure headache impact (see also Bjorner et al. 2003 for similar findings). In this research, linking methods were used to develop a comprehensive item bank from four major headache impact scales and new items suggested from prior research. Using a real-data simulation (a strategy in which a CAT is simulated using data collected from a paper-and-pencil administration), the researchers reported that they could recapture full-length scores on the original four instruments with impressive accuracy by adaptively administering five items or less.

Another interesting application of CAT in clinical assessment was reported by Fliege et al. (2005). After intense item content and psychometric analyses, these researchers culled 64 items from a larger pool of 144 descriptors to form a CAT depression item-bank. Using

real-data simulations, these authors suggested that fewer than six items were needed to achieve standard errors below 0.32 for most respondents. This study is particularly notable in that it is one of the few that were able to create an item pool with item locations spread across the trait range. However, note that they accomplish this desiderata by treating depression as a bipolar continuum marked by happiness items (e.g., optimistic) on one end and depression items on the other. This scale feature is noteworthy given our previous comments on quasi-traits, and the findings in Stansbury et al. (2006) that positively worded items on a well-known depression measure needed to be eliminated to achieve adequate fit to an IRT model.

In the above studies, CAT was evaluated using real-data simulations. Unfortunately, few clinical studies have implemented a CAT in real time. A notable exception is the recent study by Simms & Clark (2005). These researchers used live testing to conduct a validation study of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark 1993). They reported that "... computerized administration had little effect on descriptive statistics, rank ordering of scores, reliability, and concurrent validity..." (Simms & Clark 2005, p. 28). Although CAT trait estimates were less precise than those from the full 375-item instrument, by using CAT the researchers were able to administer 36% fewer items and cut testing time in half.

Interestingly, although the SNAP has a well-defined higher-order structure, Simms & Clark (2005) investigated CAT efficiency by adaptively administering SNAP items one scale at a time. For many psychopathology inventories, this approach will be inefficient because it fails to take advantage of high scale correlations. In such cases, a better approach may be to use multidimensional CAT (Gardner et al. 2002, Wang et al. 2004, Wang & Chen 2004). In this procedure, each item response provides information for multiple traits.

As an illustration of multidimensional CAT in clinical assessment, Gardner et al. (2002) reported simulated adaptive results for a 35-item parent-report, four-factor measure of psychosocial problems. By using this powerful technology, an average of only 11.6 of 35 questions was administered to each subject. Although impressive, this result hides the even more remarkable finding that for 50% of the sample, accurate trait estimation was achieved by administering five items or less. In a separate study, Wang et al. (2004) concluded that multidimensional CAT can achieve an item savings of around 50% relative to the one-scale-at-a-time method. They also note that the degree of items savings depends on the correlation among subscales; higher correlations yield more item savings.

The Bifactor Model and Multidimensional Computerized Adaptive Testing

Introduced in the 1930s (Holzinger & Swineford 1937) and then ignored for several decades, the bifactor model has reemerged as a powerful technique for modeling multidimensionality in scales and items (Gibbons & Hedeker 1992). In this model, items are free to discriminate on both a general dimension and on item group factors. Research by Simms et al. (2007) and Krueger et al. (2007) argues strongly that a bifactor perspective is appropriate for representing certain psychological domains such as mood and anxiety disorders. Moreover, Reise et al. (2007) demonstrated the appropriateness of bifactor IRT for multifaceted health outcomes data. Finally, the previously cited Uttaro & Lehman (1999) study found that the bifactor model provided a superior fit relative to a unidimensional model when characterizing quality-of-life indicators.

To the degree that item banks are consistent with the bifactor model, one of the more intriguing developments in this area is Gibbons et al.'s (2008) bifactor approach to IRT modeling and CAT administration. To illustrate these ideas, this research team developed a 616-item pool to measure mood and anxiety disorders. Unsurprisingly, the authors found that different aspects of these disorders

were highly correlated and thus all items loaded on a general factor. Each item also loaded on one of four "content" dimensions (mood, panic, obsessive-compulsive, and social phobia). Using a real-data simulation and an actual CAT administration, Gibbons et al. (2008) reported that item administration could be reduced by 95% if researchers were only interested in measuring the general trait. On the other hand, to measure both the general factor and the content dimensions, the number of CAT-administered items was much higher. For instance, in two distinct samples, averages of 98 and 88 items were needed to accurately assess the four subscales from their item pool.

These results illustrate that it requires more items to precisely measure subscales in the bifactor approach relative to either the one-scale-at-a-time or multidimensional (correlated traits) approaches to CAT. The reason for this is subtle. In the bifactor model, due to the general factor, item discrimination parameters on the group factors are relatively lower than what their values would be if the items were analyzed as a separate scale. To the degree that the item group factor item discrimination parameters are lower, more items are needed to achieve precise estimates of individual-trait levels.

More than 20 years ago, Weiss (1985, p. 774) concluded that "adaptive tests can decrease testing by about 50% while resulting in more precise measurements in comparison to conventional tests." Our literature review leads us to conclude that the efficiency produced by CAT is even more compelling in clinical assessment. Although many signs suggest that the future of CAT in clinical assessment is bright, we also see signs of reduced visibility ahead.

One obstacle concerns the putative efficiency of CAT with polytomous items. For instance, Reise & Henson (2000) explored CAT on the eight-item NEO subscales (Costa & McCrae 1992) and reported that the scales could be reduced by half with no appreciable loss in measurement precision. However, after inspecting the administered items, the authors found that most people received essentially the same item sequence. In other words, the most discriminating item was administered, followed by the second most discriminating item, and so on. The authors attributed this finding to the polytomous item response format, which for the NEO subscales, tends to spread the item information out across a broad trait range. Because researchers are not reporting item administration statistics, it is difficult to know whether this phenomenon is common (see Thissen et al. 2007 for related concerns).

Differential Item Functioning

Fair measurement requires that test scores have the same meaning across all relevant examinee groups. To test this requirement, with the emergence of IRT and other latent variable models, there has been an explosion of item and test bias studies in the clinical domain. This research has eschewed the term "bias" and now parades under the banner of DIF analysis. In nontechnical terms, an item displays DIF when the IRCs (or CRCs) differ for distinct respondent groups. The presence of many items that display DIF may compromise the ability to scale distinct groups onto a common metric.

Recent examples of DIF in clinical assessment include studies that have examined the equivalence of a Spanish and English versions of a posttraumatic stress disorder measure (Orlando & Marshall 2002), the equivalence of health care ratings across Hispanic and non-Hispanic populations (Morales et al. 2000), and the invariance of personality and psychopathology measures across black and white populations (Sheppard et al. 2006, Waller et al. 2000). This literature is difficult to review due to its sheer volume, the diversity of DIF detection methods used, the variable quality of measures, and the diverse populations studied. Nevertheless, within the clinical literature, McHorney & Fleishman (2006) uncovered three consistent findings. Namely, DIF research suggests that, after controlling for trait-level differences, (*a*) women are more likely to report emotional distress, pain, fatigue, and other markers of negative affects, (*b*) older individuals are more likely to paint a Pollyannic self-portrait, and

(*c*) Hispanics, as compared to other ethnic groups, are more likely to endorse the extreme values on Likert scales.

We do not dispute McHorney & Fleishman's (2006) observations or their substantive explanations. Nevertheless, we believe it is important to realize that the presence of item-level DIF does not necessarily lead to bias at the level of scale scores. Although researchers have found DIF on many instruments, these differences may not detrimentally affect the ability to scale and compare individuals. Rather, for many studies an appropriate conclusion would run as follows: "The two versions were not fully equivalent on an item-by-item basis, in that 6 of the 17 items displayed uniform DIF. No bias was observed, however, at the level of the composite . . . " (Orlando & Marshall 2002, p. 50). Such "default" conclusions do not devalue DIF investigations. Rather, they suggest that is it better to test and find DIF than to ignore a potential problem.

In closing this section, we focus on a major criticism of DIF analysis. Namely, the interpretation of most DIF findings has been posthoc. For instance, writing within a special issue on DIF assessment in health outcomes, McHorney & Fleishman (2006) noted that the " . . . crucial goal of DIF research is to explain the occurrence of DIF in terms of meaningful psychological constructs. None of the DIF articles in this supplement fully explored or hypothesized explanations of the observed DIF." In our mind, such observations are troubling; researchers must not ignore the "psycho[logy]" in psychometrics. DIF studies must become theoretically motivated in order to make a larger contribution to our substantive understanding of psychological measurement.

To move the field in this direction, researchers may wish to consider individual groupings that are based on psychological rather than sociological characteristics. Cohen & Bolt (2005), endorse this idea and argue that researchers are looking for qualitative differences in trait manifestations (i.e., DIF) in the wrong places. Many DIF researchers have studied blatant groupings (men versus woman; white versus black, clinical versus community) rather than latent groupings of greater psychological interest (e.g., psychotic depressives versus nonpsychotic depressives). Cohen & Bolt (2005) further argue that psychological differences do not sort easily by demographic grouping. Instead, they recommend a mixture-modeling IRT approach prior to DIF investigation. McHorney & Fleishman (2006, p. S208) concur with this view and state that a mixture approach "could provide insight into the psychological processes that lead to DIF by sociodemographic groups."

THE ROLE OF METRIC IN CLINICAL MEASUREMENT

Heretofore, we have discussed how IRT has been used to solve many longstanding problems in clinical measurement, such as linking items across scales, the optimal construction of short forms via CAT, and the assessment of DIF. In this section, we turn our sights on more theoretical concerns and focus on the metric of IRT trait scores. It is our belief that the so-called metric question is one of the most important and neglected topics in clinical assessment, and that our ability to draw meaningful inferences from test scores is directly proportionate to our understanding of the metric of our scales. To justify this claim and to situate our discussion into a broader context, we review three milestones in the history of psychological measurement that have framed contemporary debates on the meaning of test scores (Blanton & Jaccard 2006; Borsboom 2006; Borsboom & Mellenbergh 2004; Embretson 1994; Michell 1997, 2004). The three milestones are (*a*) the assembly of the Ferguson Committee (Ferguson et al. 1940; Michell 1999, pp. 143–161), (*b*) the development of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley 1940), and (*c*) the development of IRT (Ferguson 1942, Finney 1944, Lawley 1943). We begin with a discussion of the Ferguson Committee.

In 1932, the British Association for the Advancement of Science convened an assembly

of 19 prominent physicists and psychologists to determine whether psychological attributes were quantitative. Collectively, these individuals became known as the Ferguson Committee (Ferguson et al. 1940). The physicists on this committee held that quantitative variables must have nonarbitrary units and be amenable to the operations of measurement, as outlined by Campbell (1920; see Michell 1990 for a contemporary account of these axioms). They also held that psychological variables were not quantitative. These concerns were echoed by many others (Campbell 1933, Johnson 1936) who believed that psychological measurement was little more than pseudomathematics (see Michell 1999).

After eight years of debate, the members of the Ferguson Committee were unable to reach a consensus. Nevertheless, their critique of the sone scale (loudness magnitude) led S. S. Stevens to propose a radical redefinition of measurement as being "the assignment of numbers to objects or events according to a rule" (Stevens 1946, p. 667). Stevens, as is well known, proposed a typology of scale types (nominal, ordinal, interval, and ratio) and a set of proscriptions on meaningful data transformations and permissible statistics (Davison & Sharma 1988, 1990; Hand 1996; Lord 1953; Maxwell & Delaney 1985; Stine 1989). His work also led to the development of more advanced theories of measurement (Krantz et al. 1971, Luce & Tukey 1964, Luce et al. 1990, Suppes et al. 1989). Unfortunately, this later work has gone virtually unnoticed by applied psychologists and clinicians (Borsboom 2006, Cliff 1992). We hope that this situation will change now that accessible introductions to this material and its relevance for understanding psychological constructs are available (Borsboom 2005; Michell 1990, 1999, 2000).

Ironically, at the same time that the Fergusen committee was debating the possibility of psychological measurement, two Minnesota clinicians—Starke R. Hathaway (a psychologist) and J. Charnley McKinley (a psychiatrist)—paid no attention to this theoretical dispute and developed one of the most popular psychological tests of all time: the MMPI (Hathaway & McKinley 1940). Treatment of the history and merits of the MMPI, with its iconoclastic method of scale construction, are available elsewhere. Here, we merely wish to point out that the MMPI authors made no attempt to construct unidimensional scales that would yield interval-level measurement (a similar critique can be made about other popular clinical tests, such as the Thematic Apperception Test and Rorschach).

Rather, original MMPI scales were constructed to classify examinees into distinct diagnostic taxa, and each scale was designed to tap the richness of a clinical syndrome in all of its psychometric messiness. Although the test was often interpreted as if scores reflected individual differences on a latent dimension, during scale development no attention was paid to issues such as trait scalability, score homogeneity, factor pureness, or the metric of the scale scores. Instead, users developed configural scoring rules (Meehl 1950) to better capture the dynamics of clinical assessment (Meehl 1945). This highly nuanced, or unnecessarily complex (Jackson 1971), way to measure clinical constructs makes it virtually impossible to interpret MMPI scores (from the clinical scales) from the perspective of classic measurement theory (Campbell 1920, Michell 1990). We leave it to others to judge whether this is a strength or a liability of the instrument. In either case, these issues do raise the possibility that clinical syndromes may not represent quantitative variables as traditionally defined (De Boeck et al. 2005). They also provide a context for better understanding why the application of IRT models to the MMPI, and to similar instruments, can present unique challenges (Childs et al. 2000, Waller 1999, Waller et al. 2000).

The early 1940s also witnessed the development of the first IRT models (Ferguson 1942, Finney 1944, Lawley 1943). A history of these models (Bock 1997, Goldstein & Wood 1989, Wright 1997) and a discussion of how they differ from CTT (Embretson 1996, Embretson & Reise 2000, Hambleton & Jones 1993) are available elsewhere. In this section, we ask whether

IRT truly represents a theory or model of item responding and we consider the implications of our answer for understanding the metric of IRT trait scores. As a point of comparison, we note that "*classical test theory is a tautology rather than a model or a theory . . . the model must hold with respect to any given set of data*" (Lord & Novick 1968, p. 48; emphasis in original). In IRT, however, single items or entire item pools can fail to fit a latent trait model (Embretson & Reise 2000, pp. 226–246).

Items that fit a Rasch model (Rasch 1960, 1977) and thus satisfy the axioms of conjoint measurement theory (Luce & Tukey 1964) are presumed to yield interval-level trait scores (Bond & Fox 2001, Brogden 1977, Fischer 1995, Perline et al. 1979). Some authors (Harwell & Gatti 2001, Kirisci et al. 2006, Mungas & Reed 2000) have made a similar suggestion (i.e., that trait scores are interval level) when items responses fit other IRT models.

Concerning the Rasch model, we believe that the jury is still out with regard to interval-level measurement (Kyngdon 2008; Michell 1993, 2004, 2008; Wood 1978). Concerning other IRT models, we believe that the question was settled more than 30 years ago (Lord 1975, 1980). IRT models that include varying discrimination parameters yield ordinal-level trait scores (Mislevy 1987). Lord (1975) made this point in an article that has been almost forgotten. Lord showed that if a data set fits a 2PL or 3PL IRT model, then it also fits an infinite number of other logistic IRT models. Moreover, trait scores from the alternative models need not be linearly related. Surprisingly, Lord also suggested that "the ability scale θ . . . may have undesirable properties" (1975, p. 210) and that "the θ scale seems to be inadequate for many tests" (1975, p. 216). In a later publication, Lord (1980) noted,

> [t]he ability scale θ is the scale on which all item response functions have a particular mathematical form $P_i(\theta)$. This is a specified form chosen by the psychometrician . . . [o]nce we have found the scale θ on which all item response curves are (say) logistic, it is often thought that this scale has unique virtues. This conclusion is incorrect . . . (p. 84).

> [Moreover w]e cannot draw any useful conclusions from the shape of a single information function unless we assert that the ability scale we are using is unique except for a linear transformation. Most important, *we cannot know at what ability level the test or test score discriminates best, unless we have an ability scale that is not subject to challenge* (pp. 87–88; emphasis in original).

In most cases, finding a theoretically justifiable trait scale will be a daunting task. Many IRT programs attempt to solve this issue by assuming that the score distribution is Gaussian (see Thomas 1982 for a discussion of why normal distributions do not signify interval-level measurement). This assumption, however, may not be reasonable when measuring psychopathology constructs, and the development of less-restrictive IRT models is an area of active research (Woods 2006, Woods & Thissen 2006).

It bears repeating that Lord (1975) showed that if a data set fits one IRT model, it also fits many other logistic models. Related to this idea, Goldstein (1980; see also Goldstein & Wood 1989) and Garcia-Perez (1999) have shown that "logistic IRT models can fit a set of data generated by IRFs other than logistic functions just as well as they fit logistic data, even though the response processes and parameter spaces involved in each case are substantially different" (Garcia-Perez 1999, p. 74). A disturbing aspect of this finding is that these alternative models can yield trait scores with different rank orders. Stevens cautioned readers that "[w]hen only the rank-order of data is known, we should proceed cautiously with our statistics, and especially with the conclusions we draw from them" (1946, p. 679). Unfortunately, researchers have less guidance on how to proceed when different models yield trait scores that are not monotonically related.

Goldstein & Wood (1989) have noted that most IRT models were "developed with a

stunning disregard for psychological theory which might provide theoretically sound IRFs as replacements for logistic functions" (1989, p. 76). This fact, combined with the epistemic conclusions of Roberts & Pashler (2000) that absolute model fit cannot lift the veil on nature (and elucidate her hidden secrets), should alert IRT researchers to the importance of considering alternative models. It is noteworthy that the structural equation modeling literature on alternative models is vast (MacCallum et al. 1993, Meehl & Waller 2002), whereas a comparable IRT literature is almost nonexistent.

Using Item Response Theory Trait Estimates in Lieu of Sum Scores

From a practical standpoint, many researchers want to know the statistical consequences of using IRT trait estimates in lieu of simple sum-scores. Several papers have addressed this question (Dumenci & Achenbach 2008, Fan 1998, Ferrando & Chico 2007, Lawson 1991, Lu et al. 2005, MacDonald & Paunonen 2002). Many of these papers have used a variant of the following design: (*a*) collect (or simulate) items responses from a large population of examinees, (*b*) create pairs of random samples with equal numbers of subjects, (*c*) estimate item and ability parameters in each sample using CTT and IRT, and (*d*) compare the ability estimates across the two models and the item parameter estimates across the two samples. A well-replicated finding from these studies is that trait estimates from IRT and CTT correlate approximately 0.90 or higher.

The putative exchangeability of IRT and CTT trait scores has led some researchers to question the relative benefits of IRT (Barrett 2008, Fan 1998, MacDonald & Paunonen 2002). In our opinion, this pessimistic conclusion is unwarranted. Pearson correlations are relatively insensitive (within limits) to monotonic transformations of variables, and thus the above findings are neither surprising nor particularly relevant to the question of which model provides the better metric for scaling. An advantage of IRT scoring is that the trait estimates are relatively more spread out at the distribution tails. This desideratum should be of particular interest to clinicians. It is also important to realize, when interpreting the aforementioned results, that Pearson correlations computed on large samples may hide important subsample differences. The following two examples illustrate this point.

Table 1 Hypothetical scores from two scoring methods

	Method 1	Method 2
Fred	500	250
Ron	550	550
Ben	575	575
George	600	600
Alan	625	800

Imagine that the scores in **Table 1** represent two methods for scoring a depression scale. Scores from the general populations have a mean of 500 and a standard deviation of 100. In what sense, if any, are these scores exchangeable? For the five individuals in **Table 1** the two sets of scores correlate 0.96. Nevertheless, they paint very different pictures about the clinical status of Fred and Alan. This is one example of the relative insensitivity of Pearson correlations to monotonic scale transformations.

A second limitation of assessing score exchangeability with Pearson correlations was discussed by Waller & Reise (2008) in their evaluation of the 4 PL IRT model. In a large sample of MMPI data (from a unidimensional factor scale), these authors compared trait estimates from both the 3PL and 4PL models. Interestingly, whereas the trait estimates from these models produced a Spearman correlation of 0.99, the Spearman correlation for high-scoring subjects (i.e., those with 4PL trait estimates over 1.00) was only 0.45. In other words, the rank order for subjects in the clinically significant range was demonstrably different across the two models.

Research that has moved beyond simple correlations has also shown that the use of CTT or IRT trait estimates can generate important

differences in statistical models. For instance, the two scoring schemes have been shown to produce starkly different conclusions in the analysis of change (Fraley et al. 2000, Seltzer et al. 1994), the presence of statistical interactions (Embretson 1996, Kang & Waller 2005), and the magnitude of biometric parameter estimates (Berg et al. 2007). Interestingly, they have not been shown to produce large differences in linear regression models (Lu et al. 2005). When contemplating such comparisons, it is well to keep in mind the distinction between IRT trait scores and trait score estimates (Hoijtink & Boomsma 1996). IRT trait estimates are not error free.

CONCLUSION

Numerous IRT developments originated in the context of large-scale cognitive testing. In that domain, IRT serves chiefly to solve practical problems in test assembly, analysis, and administration. Over the past two decades, IRT methods have slowly emigrated to other areas of psychology. The early period of this transition was marked by enthusiastic didactic articles and empirical studies that cautiously explored the applicability of this new set of psychometric tools to personality, psychopathology, and health outcomes assessment. Presently, the applicability of IRT to typical performance measures is no longer challenged, and IRT applications are beginning to appear in clinical assessment.

In this review, we cited numerous examples that have applied IRT to develop, analyze, administer, and score clinical outcome measures. We also highlighted creative applications of IRT that addressed important questions ranging from the continuity of traits, the nature of the response process, and qualitative differences in trait manifestation between sociodemographic groups. We believe that the field has benefited tremendously from these applications. However, we also believe that the field may benefit from a deeper appreciation of the differences between cognitive and clinical constructs.

Large-scale cognitive testing is characterized by large samples that represent the relevant population, reasonably normal score distributions, well-articulated content domains, potentially infinite item pools, broadband constructs (e.g., quantitative ability), constructs where both ends of the continuum are interpretable (i.e., no knowledge versus mastery knowledge), and moderate correlations among tests of different abilities due to the omnipresent *g* factor.

On the other hand, the clinical measures we examined are characterized by relatively small samples of poorly defined mixtures of patient groups of convenience, highly skewed score distributions, poorly articulated content domains, constructs with a limited number of potential indicators, narrow band constructs (e.g., fatigue), quasi-traits, and high correlations among scales measuring different traits (e.g., due to the omnipresent negative affectivity dimension).

In our view, these differences merit greater attention as they potentially influence all aspects of IRT modeling, from dimensionality assessment to parameter estimation and interpretation and on to applications such as scale linking, the development of item pools, the administration of adaptive tests, and DIF assessment. In closing, we emphasize that the most important future research need lies in better understanding of the latent trait. This will involve a better understanding of the latent trait metric and the psychological processes that lead an individual to endorse an item.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENT

This research was supported by funds provided by PROMIS[1] (PI: David Cella, PhD, U01AR52177).

LITERATURE CITED

Aggen SH, Neale MC, Kendler KS. 2005. DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychol. Med.* 35:475–87

American Psychiatric Association. 1987. *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: Am. Psychiatr. Press. 3rd ed., rev.

Barrett P. 2008. The consequence of sustaining a pathology: scientific stagnation—a commentary on the target article "Is psychometrics a pathological science?" by Joel Michell. *Measurement* 6:78–123

Bell RC, Low LH, Jackson HJ, Dudgeon PL, Copolov DL, Singh BS. 1994. Latent trait modelling of symptoms of schizophrenia. *Psychol. Med.* 24:335–45

Berg S, Glas C, Boomsma D. 2007. Variance decomposition using an IRT measurement model. *Behav. Genet.* 37:604–16

Bjorner JB, Kosinski M, Ware JE. 2003. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to Headache Impact Test (HIT)^{TM}. *Qual. Life Res.* 12:913–33

Blanton H, Jaccard J. 2006. Arbitrary metrics in psychology. *Am. Psychol.* 61:27–41

Bock RD. 1997. A brief history of item response theory. *Educ. Meas. Issues Pract.* 16:21–33

Bollen K, Lennox R. 1991. Conventional wisdom on measurement: a structural equations perspective. *Psychol. Bull.* 110:305–14

Bond T, Fox C. 2001. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Erlbaum

Borsboom D. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. London: Cambridge Univ. Press

Borsboom D. 2006. The attack of the psychometricians. *Psychometrika* 71:425–40

Borsboom D, Mellenbergh GJ. 2004. Why psychometrics is not pathological: a comment on Michell. *Theory Psychol.* 14:105–20

Brogden HE. 1977. The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika* 42:631–34

Campbell NR. 1920. *Physics: The Elements*. London: Cambridge Univ. Press

Campbell NR. 1933. The measurement of visual sensations. *Proc. Phys. Soc.* 45:565–71

Cella D, Yount S, Rothrock N, Gershon R, Cook K, et al. 2007. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Med. Care* 45(5 Suppl. 1):S3–11

Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. 2004. The interview mode effect on the center for epidemiological studies depression (CES-D) scale. *Med. Care* 42:281–89

Chernyshenko OS, Stark S, Drasgow F, Roberts BW. 2007. Constructing personality scales under the assumptions of an ideal-pint response process: toward measuring the flexibility of personality measures. *Psychol. Assess.* 19:88–106

[1]The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, PhD, Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztein, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the Web site at **www.nihpromis.org** for additional information on the PROMIS cooperative group.

Clark LA. 1993. *Manual for the Schedule for Nonadaptive and Adaptive Personality (SNAP)*. Minneapolis: Univ. Minn. Press
Childs R, Dahlstrom GW, Kemp SM, Panter AT. 2000. Item response theory in personality assessment: a demonstration using the MMPI-2 Depression scale. *Assessment* 7:37–54
Cliff N. 1992. Abstract measurement theory and the revolution that never happened. *Psychol. Sci.* 3:186–90
Cohen AS, Bolt DM. 2005. A mixture model analysis of differential item functioning. *J. Educ. Meas.* 42:133–48
Cook KF, Roddey TS, O'Malley K. 2005. Dynamic assessment of health outcomes: time to let the CAT out of the bag? *Health. Serv. Res.* 40:1694–711
Coombs CH. 1964. *A Theory of Data*. New York: Wiley
Costa PT Jr, McCrae RR. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL: Psychol. Assess. Resourc.
Davison ML, Sharma AR. 1988. Parametric statistics and levels of measurement. *Psychol. Bull.* 104:137–44
Davison ML, Sharma AR. 1990. Parametric statistics and levels of measurement: factorial designs and multiple regression. *Psychol. Bull.* 107:394–400
De Boeck P, Wilson M, Acton SG. 2005. A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychol. Rev.* 112:129–58
Dumenci L, Achenbach TM. 2008. Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychol. Assess.* 20:55–62
Embretson SE. 1994. Comparing changes between groups. Some perplexities arising from psychometrics. In *Modern Theories of Measurement: Problems and Issues*, ed. D Laveault, BD Zumbo, ME Gessareli, MW Boss, pp. 213–48. Ottawa, ON: Univ. Ottawa
Embretson SE. 1996. Item response theory models and spurious interaction effects in factorial ANOVA designs. *Appl. Psychol. Meas.* 20:201–12
Embretson SE, Reise SP. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum
Emons WHM, Meijer RR, Denollet J. 2007. Negative affectivity and social inhibition in cardiovascular disease: evaluating type-D personality and its assessment using item response theory. *J. Psychosom. Res.* 63:27–39
Fan X. 1998. Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educ. Psychol. Meas.* 58:357–81
Ferguson A, Myers CS, Bartlett RJ, Banister H, Bartlett FC, et al. 1940. Quantitative estimates of sensory events: final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Adv. Sci.* 1:331–49
Ferguson GA. 1942. Item selection by the constant process. *Psychometrika* 7:19–29
Ferrando PJ. 2001. The measurement of neuroticism using MMQ, MPI, EPI, and EPQ items: a psychometric analysis based on item response theory. *Personal. Individ. Differ.* 30:641–56
Ferrando PJ, Chico E. 2001. Detecting dissimulation in personality test scores: a comparison between person-fit indices and detection scales. *Educ. Psychol. Meas.* 61:997–1012
Ferrando PJ, Chico E. 2007. The external validity of scores based on the two parameter logistic model: some comparisons between IRT and CTT. *Psicológica* 28:237–57
Finney DJ. 1944. The application of probit analysis to the results of mental tests. *Psychometrika* 8:31–39
Fischer G. 1995. Derivations of the Rasch model. In *Rasch Models: Foundations, Recent Developments, and Applications*, ed. G Fischer, I Molenaar, pp. 15–38. New York: Springer-Verlag
Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. 2005. Development of a computer-adaptive test for depression (D-CAT). *Qual. Life Res.* 14:2277–91
Fraley RC, Waller NG, Brennan KA. 2000. An item response theory analysis of self-report measures of adult attachment. *J. Personal. Soc. Psychol.* 78:350–65
Garcia-Perez MA. 1999. Fitting logistic IRT models: small wonder. *Spanish J. Psychol.* 2:74–94
Gardner W, Kelleher KJ, Pajer KA. 2002. Multidimensional adaptive testing for mental health problems in primary care. *Med. Care* 40:812–23
Gibbons RD, Grochocinski VJ, Weiss DJ, Bhaumik DK, Kupfer DJ, et al. 2008. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychol. Serv.* 59:361–68
Gibbons RD, Hedeker D. 1992. Full-information item bifactor analysis. *Psychometrika* 57:423–36
Goldstein H. 1980. Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *Br. J. Math. Stat. Psychol.* 33:234–46

Goldstein H, Wood R. 1989. Five decades of item response modelling. *Br. J. Math. Stat. Psychol.* 42:139–67
Gomez R, Cooper A, Gomez A. 2005. An item response theory analysis of the Carver and White 1994 BIS/BAS scales. *Personal. Individ. Differ.* 39:1093–103
Hambleton RK, Jones RW. 1993. Comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.* 12:38–47
Hand DJ. 1996. Statistics and the theory of measurement, with discussion. *J. R. Stat. Soc. A* 159:445–92
Harwell MR, Gatti GG. 2001. Rescaling ordinal data to interval data in educational research. *Rev. Educ. Res.* 71:105–31
Hathaway SR, McKinley JC. 1940. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *J. Psychol.* 10:249–54
Hays RD, Liu H, Spritzer K, Cella D. 2007. Item response theory analyses of physical functioning items in the medical outcomes study. *Med. Care* 45(5 Suppl. 1):S32–38
Hays RD, Morales LS, Reise SP. 2000. Item response theory and health outcomes measurement in the 21st century. *Med. Care* 38(9 Suppl.):1128–42
Hill CD, Edwards MC, Thissen D, Langer MM, Wirth RJ, et al. 2007. Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. 45(5 Suppl. 1):S39–47
Hoijtink H, Boomsma A. 1996. Statistical inference based on latent ability estimates. *Psychometrika* 61:313–30
Holland PW, Wainer H. 1993. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum
Holzinger KJ, Swineford F. 1937. The bifactor method. *Psychometrika* 2:41–54
Jackson DN. 1971. The dynamics of structured personality tests. *Psychol. Rev.* 78:229–48
Johnson HM. 1936. Pseudo mathematics in the mental and social sciences. *Am. J. Psychol.* 48:342–51
Kang SM, Waller NG. 2005. Moderated multiple regression, spurious interaction effects, and IRT. *Appl. Psychol. Meas.* 29:87–105
Kay SR, Fiszbein A, Opler LA. 1987. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr. Bull.* 2:261–76
Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, et al. 2002. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychol. Med.* 32:959–76
Kim Y, Pilkonis PA. 1999. Selecting the most informative items in the IIP scales for personality disorders. *J. Personal. Disord.* 13:157–74
Kirisci L, Tarter RE, Vanyukov M, Martin C, Mezzich A, Brown S. 2006. Application of item response theory to quantify substance use disorder severity. *Addict. Behav.* 31:1035–49
Kolen MJ, Brennan RL. 1995. *Test Equating: Methods and Practices*. New York: Springer
Krantz DH, Luce RD, Suppes P, Tversky A. 1971. *Foundations of Measurement: Vol. 1. Additive and Polynomial Representations*. New York: Academic
Krueger RF, Finger MS. 2001. Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychol. Assess.* 13:140–51
Krueger RF, Nichol PE, Hicks BM, Markon KE, Patrick CJ, et al. 2004. Using latent trait modeling to conceptualize an alcohol problems continuum. *Psychol. Assess.* 16:107–19
Krueger RF, Markon KE, Patrick CJ, Benning SD, Kramer MD. 2007. Linking antisocial behavior, substance use, and personality: an integrative quantitative model of the adult externalizing spectrum. *J. Abnorm. Psychol.* 116:645–66
Kyngdon A. 2008. The Rasch model from the perspective of the representational theory of measurement. *Theor. Psychol.* 18:89–109
Lawley DN. 1943. On problems connected with item selection and test construction. *Proc. R. Soc. Edinburgh* 61:273–87
Lawson S. 1991. One parameter latent trait measurement: Do the results justify the effort? In *Advances in Educational Research: Substantive Findings, Methodological Developments*, ed. B Thompson, Vol. 1, pp. 159–68. Greenwich, CT: JAI
Lord FM. 1953. On the statistical treatment of football numbers. *Am. Psychol.* 8:750–51
Lord FM. 1975. The "ability" scale in item characteristic curve theory. *Psychometrika* 40:205–17
Lord FM. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum

Lord FM, Novick MR. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley
Lu IR, Thomas DR, Zumbo BD. 2005. Embedding IRT in structural equation models: a comparison with regression based on IRT scores. *Struct. Equat. Model.* 12:263–77
Luce RD, Tukey JW. 1964. Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psychol.* 1:1–27
Luce RD, Krantz DH, Suppes P, Tversky A. 1990. *Foundations of Measurement: Vol. 3. Representation, Axiomatization, and Invariance*. San Diego, CA: Academic
MacCallum RC, Wegener DT, Uchino BN, Fabrigar LR. 1993. The problem of equivalent models in applications of covariance structure analysis. *Psychol. Bull.* 114:185–99
MacDonald P, Paunonen SV. 2002. A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educ. Psychol. Meas.* 62:921–43
Maxwell SE, Delaney HD. 1985. Measurement and statistics: an examination of construct validity. *Psychol. Bull.* 97:85–93
McHorney CA, Cohen AS. 2000. Equating health status measures with item response theory: illustration with functional status items. *Med. Care* 38:11–45
McHorney CA, Fleishman JA. 2006. Assessing and understanding measurement equivalence in health outcomes measures: issues for further quantitative and qualitative inquiry. *Med. Care* 44(Suppl. 3):S205–10
Meehl PE. 1945. The dynamics of "structured" personality tests. *J. Clin. Psychol.* 1:296–303
Meehl PE. 1950. Configural scoring. *J. Consult. Psychol.* 14:165–71
Meehl PE, Waller NG. 2002. The path analysis controversy: a new statistical approach to strong appraisal of verisimilitude. *Psychol. Methods* 7:283–300
Meijer RR, Baneke JJ. 2004. Analyzing psychopathology items: a case for nonparametric item response modeling. *Psychol. Methods* 9:354–68
Meijer RR, Egberink IJL, Emons WHM, Sijtsma K. 2008. Detection and validation of unscalable item score patterns using item response theory: an illustration with Harter's self-perception profile for children. *J. Personal. Assess.* 90:227–38
Meijer RR, Sijtsma K. 2001. Methodology review: evaluating person fit. *Appl. Psychol. Meas.* 25:107–35
Michell J. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale, NJ: Erlbaum
Michell J. 1993. Numbers, ratios, and structural relations. *Australas. J. Philos.* 71:325–32
Michell J. 1997. Quantitative science and the definition of measurement in psychology. *Br. J. Psychol.* 88:355–83
Michell J. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. London: Cambridge Univ. Press
Michell J. 2000. Normal science, pathological science and psychometrics. *Theor. Psychol.* 10:639–67
Michell J. 2004. Item response models, pathological science and the shape of error: reply to Borsboom and Mellenbergh. *Theor. Psychol.* 10:121–29
Michell J. 2008. Conjoint measurement and the Rasch paradox: a response to Kyngdon. *Theor. Psychol.* 18:119–24
Mislevy RJ. 1987. Recent developments in item response theory with applications for teacher certification. *Rev. Educ. Res.* 14:239–75
Morales LS, Reise SP, Hays RD. 2000. Evaluating the equivalence of health care ratings by whites and Hispanics. *Med. Care* 38:517–27
Mungas D, Reed BR. 2000. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statist. Med.* 19:1631–44
Orlando M, Marshall GN. 2002. Differential item functioning in a Spanish translation of the PTSD Checklist: detection and evaluation of impact. *Psychol. Assess.* 14:50–59
Orlando-Edelen M, Reeve BB. 2007. Applying item response theory IRT modeling to questionnaire development, evaluation, and refinement. *Qual. Life Res.* 15:5–18
Perline R, Wright BD, Wainer H. 1979. The Rasch model as additive conjoint measurement. *Appl. Psychol. Meas.* 3:237–56
Rasch G. 1960/1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: Univ. Chicago Press

Rasch G. 1977. On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Dan. Yearbook Philos.* 14:58–93
Reise SP, Ainsworth AT, Haviland MG. 2005. Item response theory: fundamentals, applications, and promise in psychological research. *Curr. Dir. Psychol. Sci.* 14:95–101
Reise SP, Flannery WP. 1996. Assessing person-fit on measures of typical performance. *Appl. Meas. Educ.* 9:9–26
Reise SP, Henson JM. 2000. Computerization and adaptive administration of the NEO PI-R. *Assessment* 4:347–63
Reise SP, Henson JM. 2003. A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *J. Personal. Assess.* 81:93–103
Reise SP, Morizot J, Hays RD. 2007. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual. Life Res.* 16:19–31
Reise SP, Waller NG. 1993. Traitedness and the assessment of response pattern scalability. *J. Personal. Soc. Psychol.* 65:143–151
Reise SP, Waller NG. 2003. How many IRT parameters does it take to model psychopathology items? *Psychol. Meth.* 8:164–84
Roberts S, Pashler H. 2000. How persuasive is good fit? A comment on theory testing. *Psychol. Rev.* 107:358–67
Saha TD, Chou SP, Grant BF. 2006. Toward an alcohol use disorder continuum using item response theory: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychol. Med.* 36:931–41
Samejima F. 1969. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monog.* 34(Suppl.):100–13
Samejima F. 1977. A use of the information function in tailored testing. *Appl. Psychol. Meas.* 1:233–47
Santor DA, Ascher-Svanum H, Lindenmayer JP, Obenchain RL. 2007. Item response analysis of the positive and negative syndrome scale. *BMC Psychiatry* 7:66
Santor DA, Ramsey JO. 1998. Progress in the technology of measurement: applications of item response models. *Psychol. Assess.* 10:345–59
Santor DA, Ramsey JG, Zuroff DC. 1994. Nonparametric item analysis of the Beck Depression Inventory. Evaluating gender item bias and response option weights. *Psychol. Assess.* 6:255–70
Seltzer MH, Frank KA, Bryk AS. 1994. The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric. *Educ. Eval. Policy Anal.* 16:41–49
Sharp C, Goodyer IM, Croudace TJ. 2006. The short mood and feelings questionnaire SMFQ: a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7- through 11-year-old children. *J. Abnorm. Child Psychol.* 34:379–91
Sheppard R, Han K, Colarelli SM, Dai G, King D. 2006. Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment* 13:442–53
Sijtsma K, Molenaar IW. 2002. *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage
Simms LJ, Clark LA. 2005. Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychol. Assess.* 17:28–43
Simms LJ, Gros DF, Watson D, O'Hara MW. 2007. Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depress. Anxiety* 10:1–13
Stansbury JP, Ried LD, Velozo CA. 2006. Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) Scale. *J. Personal. Assess.* 86:10–22
Stark S, Chernyshenko OS, Drasgow F, Williams BA. 2006. Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *J. Appl. Psychol.* 91:25–39
Steinberg LS. 2001. The consequences of pairing questions: context effect in personality measurement. *J. Personal. Soc. Psychol.* 81:332–42
Steinberg LS. 2008. How strongly can one disagree? Investigating the consequences of the number of Likert-scale points using item response theory. Manuscr. under review
Stevens SS. 1946. On the theory of scales of measurement. *Science* 103:677–80
Stine WW. 1989. Meaningful inference: the role of measurement in statistics. *Psychol. Bull.* 105:147–55

Suppes P, Krantz DH, Luce RD, Tversky A. 1989. *Foundations of Measurement: Vol. 2. Geometrical, Threshold, and Probabilistic Representations*. San Diego, CA: Academic
Teresi JA, Fleishman JA. 2007. Differential item functioning and health assessment. *Qual. Life Res.* 16:33–42
Thissen D, Reeve BB, Bjorner JB, Chang C. 2007. Methodological issues for building item banks and computerized adaptive tests. *Qual. Life Res.* 16(Suppl. 1):109–19
Thomas H. 1982. IQ, interval scales, and normal distributions. *Psychol. Bull.* 91:198–202
Uher R, Heyman I, Turner CM, Shafran R. 2008. Self-, parent-report and interview measures of obsessive-compulsive disorder in children and adolescents. *J. Anxiety Disord.* 22:979–90
Uttaro T, Lehman A. 1999. Graded response modeling of the quality of life interview. *Eval. Prog. Plan.* 22:41–52
Wainer H, Dorans NJ, Eignor D, Flaugher R, Green BF, et al. 2000. *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Erlbaum. 2nd ed.
Waller NG, Reise SP. 1992. Genetic and environmental influences on item response pattern scalability. *Behav. Genet.* 22:135–52
Waller NG. 1999. Searching for structure in the MMPI. In *The New Rules of Measurement: What Every Psychologist and Educator Should Know*, ed. S Embretson, S Hershberger, pp. 185–217. Mahwah, NJ: Erlbaum
Waller NG, Reise SP. 2008. Measuring psychopathology with nonstandard IRT models: fitting the four parameter model to the MMPI. In *New Directions in Psychological Measurement with Model-Based Approaches*, ed. S Embretson, JS Roberts. Washington, DC: Am. Psychol. Assoc. In press
Waller NG, Thompson JS, Wenk E. 2000. Black-white differences on the MMPI. Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: an illustration with the MMPI. *Psychol. Meth.* 5:125–46
Walton KE, Roberts BW, Krueger RF, Blonigen DM, Hicks BM. 2008. Capturing abnormal personality with normal personality inventories: an item response theory approach. *J. Personal.* In press
Wang WC, Chen PH. 2004. Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Appl. Psychol. Meas.* 28:295–316
Wang WC, Chen PH, Cheng YY. 2004. Improving measurement precision of test batteries using multidimensional item response models. *Psychol. Meth.* 9:116–36
Ware JE, Bjorner JB, Kosinski M. 2000. Practical implications of item response theory and computerized adaptive testing. *Med. Care* 38(Suppl. II):73–82
Weekers AM, Meijer RR. 2008. Scaling response processes on personality items using unfolding and dominance models. *Eur. J. Psychol. Assess.* 24:65–77
Weiss DJ. 1985. Adaptive testing by computer. *J. Consult. Clin. Psychol.* 53:774–89
Wood R. 1978. Fitting the Rasch model: a heady tale. *Br. J. Math. Stat. Psychol.* 31:27–32
Woods CM. 2006. Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychol. Meth.* 11:253–70
Woods CM, Thissen D. 2006. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika* 71:281–301
Wright B. 1997. A history of social science measurement. *Educ. Meas. Issues Pract.* 16:36–52

Annual Review of
Clinical Psychology
Volume 5, 2009

Contents

Indexes

Errata

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