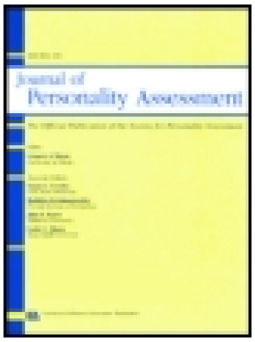

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales

Steven P. Reise & James M. Henson

To cite this article: Steven P. Reise & James M. Henson (2003) A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales, Journal of Personality Assessment, 81:2, 93-103, DOI: 10.1207/S15327752JPA8102_01

To link to this article: https://doi.org/10.1207/S15327752JPA8102_01

Published online: 10 Jun 2010.

JOURNAL OF PERSONALITY ASSESSMENT, *81*(2), 93–103

A Discussion of Modern Versus Traditional Psychometrics As Applied to Personality Assessment Scales

Steven P. Reise and James M. Henson
Department of Psychology
University of California, Los Angeles

Item response theory (IRT) methods are used by large testing firms, state agencies, and school districts to construct, analyze, and score most major aptitude, achievement, proficiency, entrance, and professional licensure exams. Personality assessment, in contrast, has not generally adopted these more powerful, modern psychometric techniques. We evaluate the possible role of IRT in the personality domain by highlighting key areas in which IRT and traditional methods differ. Although we conclude that IRT has a significant role to play in future personality measurement, there are many systemic and technical barriers to its routine application.

In large-scale cognitive assessment, which includes aptitude, achievement, proficiency, entrance, and professional licensure testing, item response theory (IRT) is the dominant psychometric paradigm for scale construction, analysis, and scoring. Although there are dozens of applications of IRT methods to personality data (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Fraley, Waller, & Brennan, 2000; Harvey & Murry, 1994; Reise & Waller, 1990; Santor, Ramsay, & Zuroff, 1994; Steinberg, 1994; von Davier & Rost, 1996), traditional classical test theory (CTT) psychometric methods continue to dominate.

The present state of affairs can be partially documented by noting that in 2000 and 2001, 20 of 39 research articles in the *Journal of Educational Measurement* and 32 of 52 in *Applied Psychological Measurement* involved IRT. In the same period, however, only 2 of 122 research articles in the *Journal of Personality Assessment* and 6 of 106 research articles in *Psychological Assessment* included IRT. Although the former pair of journals focus on innovations in psychometrics and the latter two focus on reports of scale analyses, we believe that the different content is at least partially due to the different psychometrics prevalent in the two fields.

In this article, we explore the key differences between IRT and CTT methods; however, it should be noted that this is not a tutorial on IRT. Measurement theory has too many technical complexities to summarize in a short article, and better learning tools are available elsewhere (e.g., Bond & Fox, 2001; Embretson & Reise, 2000; Thissen & Wainer, 2001). Ultimately, our chief purpose is to explore the question of whether IRT should be used as extensively in personality measurement as it is in cognitive measurement. After reviewing the basic features of CTT and IRT, we address this issue in the conclusion.

DEFINING FEATURES AND ASSUMPTIONS

Definitive treatment of strong and weak versions of true score theory or CTT can be found in Lord and Novick (1968). For our purposes, we greatly simplify. The foundation of CTT is that a respondent's observed scale score is the result of two distinct components: (a) a true score and (b) a random error component. A *true score* is defined as the expected (average) score an individual would receive if they were repeatedly administered parallel measures an infinite number of times. Simply stated, two measures are considered parallel if the true score variance is equal across both measures.

Several features of CTT are particularly relevant here. First, the true score scale is defined by a specific set of items. If an item is added or subtracted from a measure, the true score scale changes. In technical jargon, the true score scale in CTT is called *test dependent,* which means that there is a unique psychometric scale for every test. Hence, adding or removing an item from the measure results in a different psychometric scale. A second key CTT tenet is the concept of parallel measures. This concept underlies the logic of estimating a scale's *reliability* (percent of observed score variance that is due to true score variance) and the estimation of how scale precision would change if new items were added or subtracted from a measure (see Feldt & Brennan, 1989). The concept of parallel measures also plays a critical role in the comparison of respondents whom have taken different versions of a measure.

In IRT measurement models, it is assumed that a respondent has a true location on a continuous latent dimension (denoted θ or theta). However, in contrast to CTT, theta is assumed to underlie (i.e., probabilistically cause) how a person responds to an item. The objective of IRT modeling is to fit an equation that best characterizes the relationship between and the probability of endorsing an item. The equation that relates theta to the probability of endorsing an item is called an *item response function* (IRF). For example, the simplest IRF for dichotomous items is the Rasch model (Rasch, 1960), or *one-parameter logistic model* (1PLM), shown in Equation 1.[1]

$$P \mid \theta = \frac{\exp(\theta - b)}{1 + \exp(\theta - b)}\,. \qquad (1)$$

In Equation 1, theta is the respondent's position on the latent variable and *b* is the item's difficulty. In many applications of IRT, the metric of the latent variable is identified by specifying it to have a mean of zero and a standard deviation of one. In applications of the Rasch model (e.g., Equation 1), the metric of the latent variable is often identified by placing constraints on the item difficulty parameters.

Item difficulty is expressed on the same scale as the latent trait. A specific item's *difficulty* is defined as the point on the latent variable continuum where the probability of endorsing an item is .50. Typical item difficulties range from –2 to 2. Items with negative difficulties are considered "easier" and are more frequently endorsed. Even respondents with low values on the latent variable are likely to endorse such items. Items with positive difficulties are more difficult and are less frequently endorsed. Only respondents with high standing on the latent variable tend to endorse such difficult items. It is easy to show that if θ > *b,* the respondent is more likely to endorse the item, when θ < *b*, the respondent is more likely to not endorse the item, and when θ = *b,* the respondent has a .50 probability of endorsing. To illustrate this, Figure 1 displays three IRFs in which, from left to right, the item difficulties (*b*) are –1.5, 0, and 1.5, respectively. Notice that a person who has a theta level of 0 has a high probability of endorsing the first item because θ > *b,* a .50 probability of endorsing the second item because θ = *b,* but a low probability of endorsing the third item because θ < *b.*

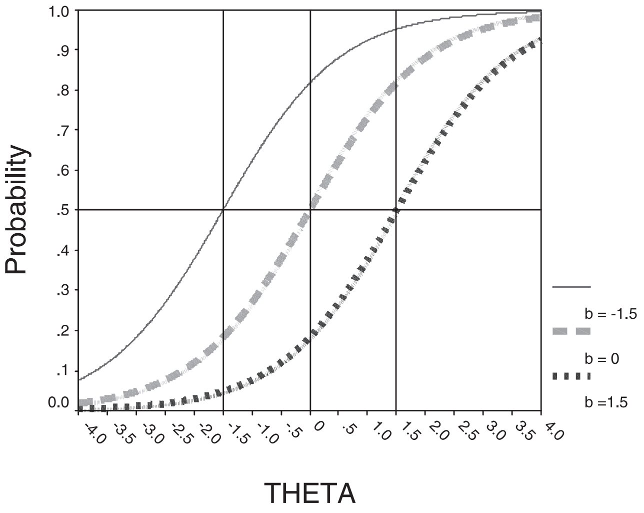

FIGURE 1 Item response functions for three items differing only in item difficulty (*b* = –1.5, 0, 1.5).

Perhaps the fundamental difference between IRT and CTT is that the latent variable scale is not dependent on a particular set of items. In other words, in IRT modeling, a respondent's true position on the latent variable scale does not depend on the specific set of items administered. A second important difference is that IRT models explicitly estimate the joint relationship between person properties (θ) and item properties (*b*) with the same model. CTT has no such feature. The consequences of this are apparent in subsequent sections.

Parametric IRT models, such as Equation 1, make strong assumptions about item response data. First, it is assumed that there is a monotonic relationship between the trait level and the probability of endorsing the item such that as the trait level increases, respondents are more likely to endorse the item. Second, for a model like Equation 1 to be valid, responses must be locally independent based on only a single common factor (θ). That is, after controlling for the common factor (i.e., the latent variable), there is no relationship among the item responses. This "unidimensionality" assumption demands that the only factor influencing response behavior is the one common variable (θ) and random error.[2]

[1] To maintain focus, we limit our discussion to parametric IRT models for dichotomous items. All concepts discussed herein could be easily generalized to IRT models for polytomous items. Moreover, we note that some authors draw a firm distinction between the measurement properties and scale construction philosophies of a Rasch or one-parameter model and other more complex IRT models (e.g., Bond & Fox, 2001). For example, an interval scale for the latent variable is only possible with a Rasch model. Also, in Rasch modeling it is common to adopt the philosophy that scale constructors should find a set of items that fit a Rasch model. This contrasts with the practice of finding the IRT model that best fits the data. The debates about the relative merits of Rasch versus more complex IRT models are important and interesting. However, to avoid being bogged down in complex psychometric arguments, we do not mention this distinction in the remainder of the article.

[2] No data set ever meets this assumption exactly—items always contain secondary factors. Therefore, researchers evaluate whether

The formal testing of model assumptions, especially the unidimensionality assumption, as well as the statistical evaluation of model-to-data goodness of fit, are major parts of applying IRT models to real data (e.g., Chernyshenko et al., 2001). For this reason, it is often said that IRT makes strong assumptions about the data (e.g., unidimensionality, model-to-data fit), whereas CTT has weak assumptions (e.g., error is independent of true score). However, this standard characterization is a bit unfair. Although formally CTT makes no claims about the dimensionality of true scores, unambiguous interpretation of scale scores as indicators of a psychological construct or dimension always requires that item responses be influenced predominantly by one and only one common factor.

Item and Scale Analysis

A typical goal of traditional scale construction is to create a fixed-length, paper-and-pencil measure that is brief enough to be parsimonious but long enough to be precise and reliable. Scale development and analysis concerns the determination of which items assess the construct most accurately and the exploration of psychometric properties for the items considered as a whole. The classical approach to addressing these issues relies on sample descriptive statistics such as the item difficulty (popularity, facility), item discrimination, and scale score reliability. In this section, we review these basic traditional psychometric indexes.

For a dichotomously scored (1,0) item, an item difficulty index is defined as the mean item response, or equivalently, as the proportion responding in the keyed direction. An item discrimination index is estimated by a correlation between item scores and total scale scores. The size of the item–test correlation reflects the degree to which an item is associated with the other items on the measure and its degree of contribution to measurement precision (i.e., item reliability). Items that correlate zero with total scale scores are interpreted to not be measuring the same construct as the other items. Dozens of alternative approaches to indexing item discrimination exist, but they all yield the same conclusions regarding which items are performing well (Marshall & Hales, 1972).

The degree to which an item set is functioning well as a whole is typically judged by an index of internal consistency reliability such as coefficient alpha (Cronbach, 1951), which is simply a function of the average interitem correlation. In CTT, a respondent's standard error of measurement (*SEM*) is inversely related to the reliability coefficient by the formula shown in Equation 2.

$$SEM = S_x\sqrt{1-r_{xx}}\,. \tag{2}$$

In Equation 2, r_{xx} is the reliability estimate and S_x is the standard deviation of total scale scores in a particular sample of respondents. A key limitation of CTT is that the reliability and *SEM* is constant for all respondents regardless of their true or observed score level. Equivalently, it is assumed that the measure is equally precise for all respondents, regardless of their standing on the construct.

The classical descriptive statistics discussed previously are not invariant across diverse samples that have different means and standard deviations on the measured variable. That is, traditional item and scale indexes are sample dependent. All else being equal, in a more heterogeneous sample (i.e., more variance in scale scores) coefficient alpha increases. The item difficulty index (mean response) changes radically depending on the average trait level of the respondent sample. Finally, the item–test correlation is influenced by the variability of scale scores in a given sample.

In IRT modeling, item analysis is similar to traditional analysis in that indexes of item discrimination and difficulty are examined but in a more powerful way. To understand this, we introduce a slightly more complicated IRT model called the *two-parameter logistic model* (2PLM) in Equation 3.

$$P\,|\,\theta = \frac{\exp(1.7a(\theta-b))}{1+\exp(1.7a(\theta-b))}. \tag{3}$$

In the 2PLM,[3] items are allowed to vary in both their difficulty (b) and discrimination (a). The item discrimination parameter (a) is proportional to the slope of the IRF and values typically range from 0.5 to 1.5. Highly discriminating items have larger slopes, and the IRF looks like a step function, whereas poorly discriminating items have smaller slopes, and the IRF looks like a flat line. To illustrate, Figure 2 displays the IRFs for two items with the same item difficulty parameter (b = 0.0) but different item discrimination parameters (a = 1.5 and 0.5). As the name implies, highly discriminating items are able to discriminate between respondents with similar levels of theta, whereas items low in discrimination are only able to discriminate between persons who are very different in their level of theta.

The item discrimination and difficulty parameters jointly determine how well an item is functioning. Specifically, instead of computing item reliability, in IRT modeling an item is judged by its item information function (IIF). Information is a psychometric concept that indicates how well an item differentiates among respondents who are at different levels of the latent variable. Scale items provide different amounts

there is a strong dominant dimension. IRT models are robust to violations of unidimensionality under this condition (Drasgow & Parsons, 1983; Tate, 2002).

[3] The 1.7 is merely a scaling factor that makes the IRF for a logistic function equal to that of a normal distribution function. In this metric, an item discrimination of a = 1.0 corresponds roughly to a factor loading of .70.

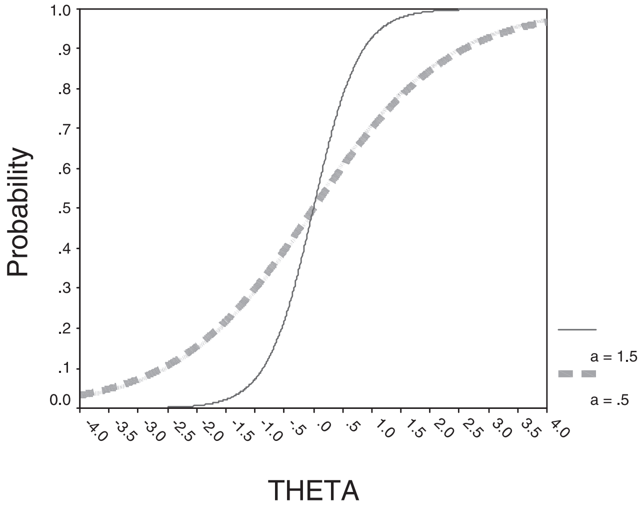

FIGURE 2 Item response functions for two items differing only it item discrimination (a = 1.5, 0.5).

of information at different ranges of the latent variable depending on the item parameters. The location on the latent variable where information is maximized is determined by the item difficulty, and the amount of information an item provides is determined by the discrimination. For example, Figure 3 displays the IIFs for the two items displayed in Figure 2 ($b = 0.0$, $a = 1.5$ and 0.5), and Figure 4 displays the IIFs for two items: one item with $b = –1$ and $a = 1.5$, and the second item with $b = 1$ and $a = 1.5$.

There are two useful features to information. The first is that it is additive across items so researchers can add together the IIFs to produce a scale information function (SIF). The SIF reveals how well a set of items is functioning as a whole. Just like a researcher may add or subtract items and recompute coefficient alpha, a researcher may also add or subtract items and recompute the SIF. The second useful feature is that information is inversely related to the *SEM* as shown in Equation 4.

$$SEM \mid \theta = \frac{1}{\sqrt{SIF \mid \theta}}. \tag{4}$$

Equation 4 demonstrates that as conditional scale information increases, the *SEM* decreases. Measures can provide different amounts of information at different levels of the latent trait. Unlike CTT, respondents will have different *SEM*s depending on where they are located on the latent variable.

Another important feature of IRT is that item parameters (e.g., *a, b*) are invariant within a linear transformation. Item parameter invariance means that their true values do not depend on the constitution of the sample. The item parameters in IRT are defined independently of sample characteristics, whereas in CTT item characteristics are based on the sample characteristics. In the 2PLM (Equation 3), b is the point on the latent variable scale at which respondents have a 50% chance of endorsing an item, and a is proportional to the slope of the IRF at the inflection point (where $\theta = b$). Neither of these indexes depends on the characteristics of a particular sample of respondents. Consider an analogous situation in ordinary linear regression where $Y = B_0 + B_1 (X)$. The (unstandardized) slope (B_1) and intercept (B_0) coefficients are the same (invariant) regardless of the mean and variance of the predictor X in a particular sample.

Item parameter invariance also means that, assuming the model is correct (assumptions are met and the model fits the data), item parameters estimated in one sample can be linearly transformed to be equal to item parameters estimated in a second sample. This can be accomplished regardless of how divergent the two sample means and variances are. The same cannot be said for classical indexes of difficulty and discrimination, although see Fan (1998) for an opposing

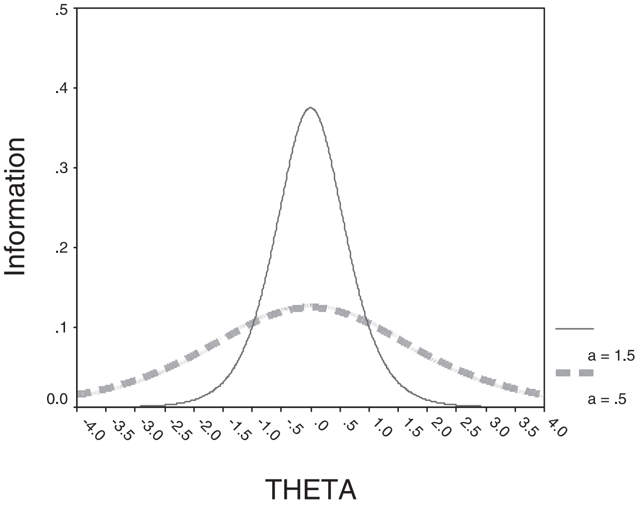

FIGURE 3 Item information curves for the two IRFs depicted in Figure 2 in which item discriminations differ (a = 1.5, 0.5).

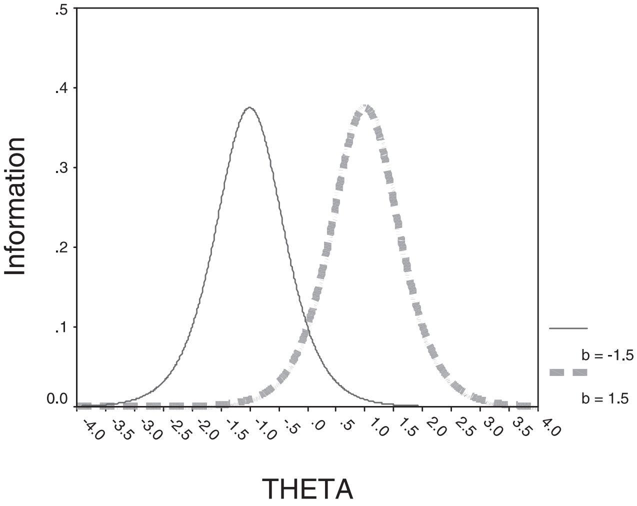

FIGURE 4 Item information curves for two items with identical discriminations (1.5) but differ in item difficulty (b = –1.5, 1.5).

point of view. It is important to note that item parameter invariance does not mean that if item parameters were estimated in two different samples, the observed results would be exactly the same; estimates of parameters are always influenced by sample characteristics, sampling error, and inefficiencies or biases in numerical estimation algorithms.

Scoring

In the majority of applications of personality assessment instruments, the summary index used to characterize individual differences is the unit weighted summed scale score (i.e., the raw score). The summed score serves as an estimate of true score and the standard error is assumed equal for all respondents. However, raw scores are rarely meaningfully interpreted. Rather, indicators of absolute standing (raw scores) are made interpretable by transformation into an index of relative standing based on norms (e.g., T scores, *z* scores). In turn, a person's (relative) standing on the construct (and their standard error) is sample dependent. How an individual's raw score is interpreted depends on who they are tested with or compared to (i.e., norms).

In IRT scaling, item responses from any subset of items with known IRFs can be used to estimate an individual's position on the latent variable continuum. The specific scoring methods, such as maximum likelihood or Bayesian estimation, are very complex and are not summarized here (see Embretson & Reise, 2000, or Thissen & Wainer, 2001). However, in contrast to traditional methods, in IRT scaling an individual's score on the latent variable is independent of the items that are administered. This type of "item-free" individual difference scaling is possible because IRT incorporates both item and person parameters into the same model. Item-free scaling is the foundation of computerized adaptive testing (CAT), as discussed in the following.

APPLICATIONS OF IRT

Previously, we noted that testing model assumptions (e.g., unidimensionality) and model-to-data fit is a major part of applying an IRT measurement model. Often the diligence of checking model assumptions seems extra cautious, especially when IRT is applied to an existing measure known to have good psychometric properties. Why is such diligence required in IRT but not in CTT? Simply put, a prime motivation for the application of IRT models is to make use of special features (e.g., linking scales across different measures, CAT, detecting item bias, optimal scaling of examinees) that are products of IRT models. For these features to work properly, the model assumptions must hold and the model must fit the data. In this section, we first briefly describe one specific application of IRT models to improve the quality and efficiency of psychological measurement, namely, CAT (Wainer, 2000). We then briefly cite two additional applications of IRT methods designed to (a) assist in the identification of biased items across subgroups of respondents and (b) compare respondents on a common scale even if they have been administered different measures.

CAT

One of the most popular applications of IRT models is CAT. The Educational Testing Service has administered over one million CATs (Gitomer, 2000, p. xiii). In CAT, a computer algorithm selects from a pool of precalibrated items (i.e., items with known IRFs) to optimize the precision of measurement for each individual. For example, a respondent is administered an item, perhaps of middle difficulty, and then he or she responds. Based on that response, the respondent's trait level and standard error are then estimated. The computer then selects another item from the pool that is optimal (i.e., provides the most psychometric information) for the respondent given the respondent's current trait level estimate.

This iterative process is repeated until either a preset number of items are administered or the standard error falls below a certain value. CAT research indicates that many existing paper-and-pencil measures could be shortened by at least 50% when administered in CAT format with no loss in measurement precision (Wainer, 2000). Waller and Reise (1989) illustrated an application of CAT in clinical decision making, and Reise and Henson (2000) illustrated the utility of CAT in normal-range personality assessment using polytomous items.

Differential Item Functioning

One question frequently researched is whether a scale is biased against or works differently across phenotypic groups, such as gender or ethnicity groups. Regardless of whether a researcher is operating under an IRT or CTT framework, for respondent scores to be comparable, items must function the same way for all respondents regardless of their group membership. That is, a valid comparison of individuals on a common scale or a valid comparison of group mean differences requires that scale items display measurement invariance (i.e., no bias for or against particular subgroups). Although there are many approaches to empirically studying whether scale items are functioning equivalently across respondent groups, IRT methods provide a particularly elegant framework for studying differential item functioning (DIF).[4]

In an IRT framework, an item exhibits DIF if the IRFs are not equivalent when estimated separately in two or more groups. In other words, if the probability of endorsing an item conditional on theta differs across subgroups, an item

[4] Other methods such as covariance structure modeling can also be used to identify DIF, but covariance models are better at studying differential scale functioning rather than item functioning.

has DIF. The study of DIF, or item bias, in CTT is severely hindered by the fact that CTT item statistics are sample dependent. In contrast, efficient and valid DIF analysis using multiple-group IRT methods is possible because of the item parameter invariance feature of IRT models. Interested readers may see Holland and Wainer (1993) for fuller treatment of this important topic; Millsap and Everson (1993) for a review of bias detection methods; Waller, Thompson, and Wenk (2000) for a provocative analysis of ethnic differences on the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943); and Reise, Smith, and Furr (2001) for a DIF analysis on a normal-range personality scale.

Comparing Individuals Who Have Taken Different Measures

In the previous paragraph, we indicted that to compare individuals on a common scale, the items must function equivalently for all respondents. Related to this problem is the issue of how to compare respondents when they have taken slightly or completely different versions of a measure. This problem occurs frequently in large-scale educational assessment (Tindal & Haladyna, 2002) in which state tests are administered yearly or quarterly and the item pool needs to be changed accordingly. Historically, comparing people who have taken different measures has been the topic of a class of statistical methods called *test-equating procedures* (Doran & Holland, 2000). These methods, although extremely valuable, suffer from well-known flaws and limitations (see Embretson & Reise, 2000).

More recent attention has turned to the topic of scale linking based on IRT methods (Choi & McCall, 2002). IRT-based linking methods are a set of procedures for insuring that peoples' scores across different measures of the same construct are transformed to the same scale and thus comparable. Of course, these methods assume that there is no DIF across subgroups as reviewed previously and that an IRT model is appropriate for the data at hand. IRT-based linking methods potentially solve two classic problems in assessment, which we phrase in question form: (a) What if some respondents do not answer all the scale items? and (b) What if different people respond to different measures, but still need to be compared?

As it is with many fields, personality researchers frequently encounter the difficult problem in which not all people respond to all items. The reasons for item nonresponse are variable and range from a lack of understanding an item to refusing to answer based on content. This creates a predicament for assessors in a CTT framework because a comparison of people requires a level playing field (i.e., an equal number of responses). One simple solution to impute missing data is to average scores across items that were answered. More complex methods of imputation are also available, but are seldom employed at the item response level. However, missing responses are not a problem for IRT scaling because a respondent's trait level can be estimated with any subset of items that have been linked to a common scale (e.g., in CAT).

A more daunting problem occurs when respondents have completed different measures, but they still need to be compared on a common scale. This problem has a variety of manifestations. First, a measure may change in content over time such as the MMPI versus MMPI–2 (see Butcher & Williams, 1992, pp. 3–11). In personality measurement, new forms, short forms, or revisions of popular scales occur frequently. Second, respondents may complete different measures of a common construct. In personality assessment, many important constructs such as depression can be assessed through one of several measures. Third, respondents may complete the same measure in different languages (Sireci, 1997). The technical complexities of linking scores onto a common scale under these circumstances are way beyond our the scope of this article (see Vale, 1986). Nevertheless, we call attention to the fact that IRT methods potentially provide elegant solutions for the comparison of scores from different measures. In turn, these linking methods promote the cross walking of research finding across investigators.

DISCUSSION

CTT methods of scale development and scoring have served personality measurement well for over 80 years. Numerous reliable, valid, and useful measures of personality constructs have been developed without the application of IRT methods. Yet it is unfortunate that whereas cognitive testing has so readily taken to IRT, the field of personality assessment has largely ignored developments and improvements in measurement theory.

Although IRT is more technically complex than CTT, this phenomenon can not be due to a lack of quantitative sophistication on the part of personality assessment psychologists. In point of fact, personality assessment researchers have historically been at the forefront of statistical and methodological innovation (e.g., Raymond Cattell, *The Scientific Analysis of Personality,* 1965). Moreover, this phenomenon can not be explained by the differential social or scientific importance of cognitive assessment relative to personality assessment. Clearly, personality assessment not only serves as the foundation of scientific soft psychology but also serves a gamut of important social functions, which range from deciding among treatment interventions, tracking clinical change, and influencing legal decisions on child custody.

At this point, the critical question is not whether IRT models are superior to CTT methods. Of course they are, in the same way that a modern CD player provides superior sound when compared to a 1960s LP player or in the same way covariance structure modeling improves on ordinary multiple regression analysis. Hattie, Jaeger, and Bond (1999) stated, "item-response theory (IRT) is an elegant and powerful model of test performance that obviates virtually all of the shortcomings of classical test theory" (p. 399). The real question is, does application of IRT result in a sufficient improvement in the quality of personality measurement to justify the added complexity? Although we provide no definitive answers, to explore this central issue, in the following we consider three questions and provide expanded commentary.

Does IRT Significantly Change the Psychometric View of a Measure?

In scale construction and analysis, the central issues are the determination of which items are functioning best as trait indicators and which items are not contributing to measurement precision. With the use of item–test correlations, item difficulties, and factor analysis, many excellent content homogeneous, internally consistent, and valid measures of personality constructs have been developed. To cite just one example, the Multidimensional Personality Questionnaire scales (Tellegen, 1982) were developed using factor analysis combined with an iterative hypothetical-deductive approach to scale construction (Tellegen & Waller, in press).[5]

Would the use of IRT item discrimination and difficulty parameters have led to better scales or different decisions about which items work and which do not? The answer is most likely "no." The reason is that all indexes of item discrimination, such as the item–test correlation, a factor loading, or the slope of an IRF, are highly related. The proportion endorsed must be highly correlated with an IRT estimate of item difficulty as well. In fact, over 50 years ago, Tucker (1946) provided transformation equations for relating IRT item parameters to classical item–test biserial correlations and proportions endorsed. Essentially, classical item descriptive statistics and IRT item parameters each use the same information. There is nothing dramatic to be learned per se by estimating an IRF versus simply computing the simple classical descriptive statistics.

However, IRT item parameters have a linear invariance property that CTT indexes do not share. In turn, this invariance property facilitates important applications as reviewed previously (e.g., DIF). There are also certain interpretational advantages to IRT item parameters. For example, the IRT *b* parameter is easier to interpret than the CTT proportion endorsed. The proportion-endorsed metric is difficult to interpret because its meaning changes across the scale: .10 versus .30 is a huge difference, whereas .50 versus .70 is not a large difference. In IRT, item difficulty is on the same scale as examinee trait level and can be defined as the amount of the construct necessary to have a .50 endorsement probability.

The most significant difference between IRT and CTT is not in the interpretation of item parameter estimates or the resulting IRFs but rather in the conceptualization of measurement error. As reviewed previously, CTT provides a single index of reliability and a standard error that is constant for all examinees. IRT, on the other hand, allows the researcher to compute an IIF and SIF and allows measurement error to change across the latent variable continuum depending on the properties of the measure. In our view, it is more realistic and valid to recognize that different measures provide different amounts of precision for respondents who are at different ranges of the latent variable.

Gray-Little, Williams, and Hancock (1997) provided a provocative IRT analysis of the Rosenberg Self-Esteem scale. This measure has been used extensively in self-esteem research and is known to provide scores with high internal consistency reliability. In Gray-Little et al.'s research, the measure also displayed high internal consistency; however, their IRT information analysis showed that the measure is rather poor at differentiating among high trait (high self-esteem) examinees. This is a critical fact to know if a researcher were planning to use this measure to study change in self-esteem or trying to distinguish between people who, on average, have high self-esteem. In other research, Reise and Henson (2000, p. 350) used information analysis to demonstrate that one item on the Revised NEO Personality Inventory Anxiety scale (Costa & McCrae, 1992) provided almost four times the information than any other item. This suggests that it would be better to administer two items like the most informative than administer the seven remaining items that make up the Anxiety scale. Other examples could be provided, but the basic point is clear. It is often more valuable to examine a measure's information/precision across the entire trait range than it is to know a single reliability coefficient.

Does IRT Make a Difference in Terms of Precision or Validity?

The traditional standard index of test performance is the unit-weighted raw score. Without going into technical details, suffice it to say that IRT provides an optimally weighted scaling of individual differences (Birnbaum, 1968). In other words, no weighting of raw item responses can result in an estimate with smaller standard error (i.e., more precision). Yet, does this superior scaling of individual differences make a practical difference? Does going to the trouble of using maximum likelihood or Bayesian methods to estimate an individual's location on a latent variable rather than summing raw scores lead to substantially different results?

The answer is "yes,“ "no," and "let's see what future research demonstrates." The answer is "no" because an individual's relative standing on a construct will show little difference when CTT methods (e.g., compute a raw score

[5] We note that the development of those scales were not informed by IRT analysis; however, they were later reanalyzed under an IRT framework (Reise & Waller, 1990).

and transform using norms) are compared to IRT scaling methods. In fact, in the 1PLM (Rasch model; Equation 1), raw scores are a sufficient statistic for estimating trait levels. In other IRT models, such as the 2PLM, latent variable scores are always highly correlated with raw scores. In our own research, we routinely encounter correlations between raw scores and trait level estimates above .98.

There is a simple reason why these high correlations occur. Although IRT provides optimal scaling, the correct weights used do not differ greatly from unit weights. For example, in the 2PLM, the optimal weight for scaling individual differences is the sum of each item response (1 or 0) multiplied by the item discrimination. In well-designed measures in which all items are functioning well, item discrimination parameters often have little variability and hence provide little more than a scaling constant to the raw score. Therefore, in terms of relative standing, there is often no great advantage to the IRT optimal scaling in comparison to raw scores. To date, we are not aware of any personality research that demonstrates that the increased precision of IRT scores leads to appreciable increases in validity coefficients (test-criterion correlations).

However, just because raw scores and trait level scores are highly correlated, this does not mean that the two scalings are equivalent or will produce similar applied results. In fact, there are well-established problems with using the raw score scale as a metric for scaling individual differences or comparing groups (Bond & Fox, 2001; Maxwell & DeLaney, 1985; Yen, 1986), investigating interactions (Embretson, 1996), and for studying change (Embretson, 1998; Fraley et al., 2000). Although IRT trait level estimates may be highly correlated with raw scores (e.g., they are a nonlinear but monotonic transformation of raw scores), the optimal scaling of individual differences with IRT can make a difference in practice and can dramatically change substantive conclusions. In the following, we highlight two studies that demonstrate this fact.

First, Embretson (1996) demonstrated that raw score scaling can result in faulty identification of interactions in an analysis of variance context. That is, using the raw score metric rather than a Rasch model IRT scaling can result in researchers missing interactions when they exist and identifying significant interactions when they do not exist. Second, Fraley et al. (2000) thoroughly investigated self-report measures of adult attachment to explore their psychometric properties. Fraley et al. found that because of scaling problems with the raw score metric caused by a clustering of within-measure item difficulties, interpretation of research on the continuity and stability of attachment is muddled. Fraley et al. suggested how IRT methods are used to improve the quality of attachment measures, which in turn will improve the validity of attachment research. Ultimately, more future research is called for investigating scaling advantages of IRT methods.

Are IRT Models Appropriate for Personality Constructs?

Before launching into this section, we pause to comment on a misconception we frequently encounter among assessment professionals and research colleagues. Namely, some researchers believe cognitive constructs are real, individual difference, psychobiological traits that cause behavior, whereas personality constructs are thought of as arbitrary, subjective, and merely summary labels of behavior. To many, it is thought that IRT methods are appropriate to use with cognitive variables but inappropriate to use with personality assessments.

We agree wholeheartedly that there are poorly thought out, redundant, intellectually flabby constructs and measures in personality assessment research (for further insightful commentary, see Block, 2002). Also, not all personality constructs are latent variables (e.g., personal concerns) and not all constructs that are commonly referred to as traits are actually "real" latent trait variables (e.g., Gough's, 1987, California Psychological Inventory Folk constructs and MMPI clinical scales are not trait measures). However, we disagree with the view that personality measurement is a qualitatively different world than cognitive measurement. In many circumstances, personality constructs are deeply embedded within psychobiological theories and are properly viewed as real traits that cause behavior in the exact same way as cognitive variables like math ability or spatial ability. Interested readers may see Tellegen (1991) for an articulate discussion of a strong view of personality traits.

With this viewpoint in mind, IRT models assume that there is a continuous latent variable or trait that influences the probability of responding to an item. Most applied IRT models are unidimensional and assume that only one common trait is influencing item performance. In turn, an individual's score is viewed as an estimate of their position on a latent trait. In essence, IRT models are essentially the same as a factor analytic model (see McDonald, 1999; Muraki & Engelhard, 1985; Wilson, Wood, & Gibbons, 1991). Factor loadings are analogous to IRT discriminations and factor thresholds are analogous to item difficulties. The main difference between IRT models and the more commonly used factor models is that the former is nonlinear, whereas the latter is linear.[6] Therefore, any personality construct/measure that can be appropriately represented by a factor analytic model can, in theory, be an appropriate context for the application of an IRT model.

However, personality theory offers a wealth of constructs that are difficult to represent as a latent variable type of con-

[6] Also, factor analytic research seldom makes use of item thresholds because analyses are conducted on a standardized or correlation matrix.

struct and therefore cannot be appropriately assessed via an IRT model. For example, the multifaceted construct is explicitly not unidimensional (Hull, Lehn, & Tedlie, 1991). Nonlinear developmental constructs such as ego development as measured by the Washington University Sentence Completion Test (Loevinger, 1993) also do not fit neatly into the IRT mold. Furthermore, emergent constructs (Bollen & Lennox, 1991; Cohen, Cohen, Teresi, Marchi, & Velez, 1990) also may not be appropriate for IRT. In a latent variable model, the latent variable is thought to cause item responses, whereas an emergent construct is simply defined by its indicators (items). Constructs such as social status, relationship quality, attractiveness, and mental health are examples of emergent constructs.

Moreover, because of the strong unidimensionality assumption, IRT models are also limited in the bandwidth of constructs that can be fit. Specifically, IRT models function best for narrower constructs (academic self-esteem) in which researchers can write a set of homogeneous items that are highly intercorrelated and thus unidimensional. Measures of broader constructs (general self-esteem) may contain item content that is too heterogeneous to achieve a good model fit when an IRT model is applied. Of course, a general construct can always be broken down into more homogenous parts and IRT applied to the subscales. Using IRT to optimally combine subscale scores into a general score is covered extensively in Thissen and Wainer (2001).

CONCLUSIONS

Clearly, IRT has many advantages over CTT that warrant the added complexity. IRT models (a) estimate both item and person parameters with the same model, (b) provide person-free item parameter estimation and item-free trait level estimation, (c) provide an optimal scaling of individual differences, and (d) facilitate important application such as CAT, linking scales, and the evaluation of DIF. It is fair to ask, if IRT is so advantageous and is so widely used in the cognitive domain, why are there not more applications in the realm of personality?

Although there are plenty of published applications of IRT to personality data, there is simply not enough research in the realm of personality assessment that communicates convincingly the relative superiority of the IRT approach. What is needed to bring the field of personality assessment up to date is research that demonstrates that the use of CTT methods can lead to incorrect substantive conclusions, whereas an IRT approach leads to more valid substantive findings. The study by Fraley et al. (2000) cited earlier is a prime example of the type of research necessary to prompt researchers to rethink their methodological practices. We also view the work of Santor et al. (1994) as a great example of how the functioning of an important clinical instrument can be better understood through the application of IRT methods.

Beyond the lack of persuasive research, there are several systemic features of personality assessment that retard change. First, the applied world of personality assessment, when compared with the world of cognitive assessment, is under little legal pressure to use better measurement practices. Educational (cognitive) assessment is under intense public scrutiny and their products are continuously being challenged in the courts. The public scrutiny and concerns over fairness and validity create an environment where the best psychometric practices must be implemented. Although personality assessment is important too, it does not receive the same ferocious treatment in the courts, by university presidents, or by special interest groups.[7]

Another issue that affects the statistical practice of personality measures relates to the nature of the domains themselves. In cognitive assessment, it is often useful to think of a domain (e.g., spelling) where the test items are a sample from this domain. In general, these tests need to be long to be reliable and because the domain is so large, researchers need a large sample of items to accurately assess the domain. In personality, many of the domains are quite restricted. For example, there are only a finite number of indicators (signs) of math self-esteem, social introversion, friendliness, or narcissism. The consequence of this is that it is difficult to create long measures for these constructs because researchers simply run out of nonredundant questions to ask. The short nature of these personality scales and lack of alternate forms often obviates the need to use CAT and linking and therefore eliminates some of the most important reasons for applying an IRT model.

Finally, instead of working with a large pool of items to measure a common construct, the world of personality assessment consists of scale tribes. Instead of trying to find a common set of indicators to measure important constructs, groups of researchers tend to stick with their preferred measure, furthering a division among researchers. There are at least three major inventories to measure the Big Five traits, at least three major measures of depression, four commonly used indicators of narcissism, and so on. In turn, many of these measures are protected by copyright and researchers are not free to change items. Moreover, once research findings begin to accumulate on a particular measure or interpretations of test scores are allowed in the courts, test companies and test authors are extremely hesitant to change their mea-

[7] Moreover, another key difference between cognitive and personality assessment is that the central applied purpose of cognitive assessment is producing precise and valid scalings of individual differences. In turn, these test scores are used to make focused decisions (accept or reject candidate). In applied personality assessment, a major focus is not scaling per se but rather test score interpretation and the prediction of wide ranging behavioral outcomes.

sure, fearing that this may invalidate previous results. All of these systemic reasons contribute to the field's reluctance to implement IRT methods to guide the analysis and creation of personality assessment scales.

In conclusion, in the introduction we stated that our chief purpose was to explore the question of whether IRT should be used as extensively in personality measurement as it is in cognitive measurement. Although the potential advantages of IRT over traditional methods are obvious and compelling, we still believe that more research is needed to demonstrate IRT's differential effectiveness in personality assessment. In particular, research needs to clarify the specific conditions under which IRT modeling is especially advantageous when compared to traditional procedures.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison-Wesley.

Block, J. (2002). *Personality as an affect-processing system.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Butcher, J. N., & Williams, C. L. (1992). *Essentials of MMPI–2 and MMPI–A interpretation.* Minneapolis: University of Minnesota Press.

Cattell, R. B. (1965). *The scientific analysis of personality.* Oxford, England: Penguin.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523–562.

Choi, S. W., & McCall, M. (2002). Linking bilingual mathematics assessments: A monolingual IRT approach. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 317–338). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14,* 183–196.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI–R) and NEO Five-Factor Inventory (NEO–FFI) professional manual.* Odessa, FL: Psychological Assessment.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Doran, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37,* 281–306.

Drasgow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201–212.

Embretson, S. E. (1998, August). *Modifiability in lifespan development: Multidimensional Rasch Model for learning and change.* Paper presented at the annual meeting of the American Psychology Association, San Francisco.

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

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357–381.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Gitomer, D. H. (2000). Forward to the second edition. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. xiii–xv). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gough, H. G. (1987). *CPI: California Psychological Inventory administrators guide.* Palo Alto, CA: Consulting Psychologists Press.

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

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers–Briggs Type Indicator: Empirical comparison of preference score versus latent-trait analyses. *Journal of Personality Assessment, 62,* 116–129.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* Minneapolis: University of Minnesota Press.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 24,* 393–446.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hull, J. G., Lehn, D. A., & Tedlie, J. C. (1991). A general approach to testing multifaceted personality constructs. *Journal of Personality and Social Psychology, 61,* 932–945.

Loevinger, J. (1993). Measurement of personality: True or false. *Psychological Inquiry, 4,* 1–16.

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

Marshall, J. C., & Hales, L. W. (1972). *Essentials of testing.* Reading, MA: Addison-Wesley.

Maxwell, S. E., & DeLaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97,* 85–93.

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

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297–334.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417–430.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmark Paedagogiske Institute.

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

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI–R Neuroticism scale. *Multivariate Behavioral Research, 36,* 83–110.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data: The parameterization of the Multidimensionality Personality Questionnaire. *Applied Psychological Measurement, 15,* 45–58.

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

Sireci, S. G. (1997). Problems and issues in linking assessment across language. *Educational Measurement: Issues and Practice, 1,* 12–29.

Steinberg, L. (1994). Context and serial order effects in personality measurement: Limits on the generality of "measuring changes the measure." *Journal of Personality and Social Psychology, 66,* 341–349.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical ade-*

quacy, and implementation (pp. 181–211). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tellegen, A. (1982). *Brief manual for the Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A. (1991). Personality traits: Issues of definition, evidence and assessment. In D. Cichetti & W. Grove (Eds.),*Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 10–35). Minneapolis: University of Minnesota Press.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), *Personality measures: Development and evaluation* (Vol. 1). Greenwich, CT: JAI.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tindal, G., & Haladyna, T. M. (2002). *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1–13.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 133–144.

von Davier, M., & Rost, J. (1996). Self monitoring—A class variable? In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 296–304). New York: Waxmann Munster.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Waller, N. G., Thompson, J., & Wenk, E. (2000). Black-white differences on the MMPI: Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales. *Psychological Methods, 5,* 125–146.

Wilson, D., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–325.

Steven P. Reise
Franz Hall
Department of Psychology
UCLA
Los Angeles, CA 90095
E-mail: Reise@psych.ucla.edu

Received November 4, 2002
Revised February 4, 2003