MODERN PSYCHOMETRIC THEORY IN CLINICAL ASSESSMENT

by

Michael Lee Thomas

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

AUGUST 2011

MODERN PSYCHOMETRIC THEORY IN CLINICAL ASSESSMENT

by

Michael Lee Thomas

has been approved

May 2010

Graduate Supervisory Committee:

Richard Lanyon, Chair
Manuel Barrera
Roy Levy
Roger Millsap

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

Item response theory (IRT) and related latent variable models represent modern psychometric theory, the successor to classical test theory in psychological assessment. While IRT has become prevalent in the assessment of ability and achievement, it has not been widely embraced by clinical psychologists. This appears due, in part, to psychometrists' use of unidimensional models despite evidence that psychiatric disorders are inherently multidimensional. The construct validity of unidimensional and multidimensional latent variable models was compared to evaluate the utility of modern psychometric theory in clinical assessment. Archival data consisting of 688 outpatients’ presenting concerns, psychiatric diagnoses, and item level responses to the Brief Symptom Inventory (BSI) were extracted from files at a university mental health clinic. Confirmatory factor analyses revealed that models with oblique factors and/or item cross-loadings better represented the internal structure of the BSI in comparison to a strictly unidimensional model. The models were generally equivalent in their ability to account for variance in criterion-related validity variables; however, bifactor models demonstrated superior validity in differentiating between mood and anxiety disorder diagnoses. Multidimensional IRT analyses showed that the orthogonal bifactor model partitioned distinct, clinically relevant sources of item variance. Similar results were also achieved through multivariate prediction with an oblique simple structure model. Receiver operating characteristic curves confirmed improved sensitivity and specificity through multidimensional models of psychopathology. Clinical researchers are encouraged to consider these and other comprehensive models of psychological distress.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER

Page

CHAPTER

Page

APPENDIX

APPENDIX

Page

APPENDIX

Page

APPENDIX

Page

Chapter 1

Modern Psychometric Theory Improve in Clinical Assessment

Overview

Among the world's least understood and ill-defined topics of study, the human psyche holds an eminent rank. While there has been no shortage of imaginative theories to explain human thought, behavior, and emotion, scientifically based definitions and quantifications of inherently subjective mental states have been less common. In an effort to lay the foundations for accurate assessment and effective treatment of psychiatric disorders, researchers have turned to empirical, mathematically-based theories of mental health. Under the banner of psychometrics, a movement to refine the art and science of psychological measurement has been underway for more than a century. One particular branch of this movement, *item response theory* (IRT), has revolutionized the theory behind measurement of ability and achievement (see Embretson & Hershberger, 1999; Embretson & Reise, 2000) and significantly impacted the development of commonly administered tests (e.g., McGrew & Woodcock, 2001).

Curiously, psychologists interested in clinical domains have been slow to embrace modern psychometrics. Some assert that psychiatric disorders are simply too complex to be captured by quantitative measurement models based on reduced components (e.g., Fava, Ruini, & Rafanelli, 2004; Gordon, 2006). Unfortunately, little overt evidence has emerged to assuage these concerns. Despite great advancements in psychometric theory, clear demonstrations of diagnostic benefits have been slow to come. The purpose of this project is to explore modern psychometric theory's impact on test validity in the realm of clinical assessment. First, IRT is explained in comparison and in contrast to related theories of measurement. Second, applications of IRT in clinical assessment are reviewed. Third, an analysis of clinically relevant data is conducted in an attempt to elucidate the value of modern psychometric theory in the diagnosis and conceptualization of psychiatric disorders.

The Evolution of Measurement

Psychometric models specify systems of mathematical relations between observed and unobserved variables. They should not be viewed as alternatives to semantic accounts of psychological phenomena. Instead, such models serve to open scientific hypotheses to empirical examination. In the spirit of Karl Popper's (1964) promotion of risky predictions, mathematical models force researchers to test specific hypotheses. The empirical nature of scientific methodology invariably leads fields of study towards research focused on model development and evaluation. Refinement of psychological measurement through the use of comprehensive models ought to result in better science. In physics, for example, the laws of thermodynamics predict how changes in temperature, pressure, and volume will affect a physical system. Models are used to predict the consequences of changing one variable (e.g., temperature) on another variable in the system (e.g., pressure). The discovery of such comprehensive models is the objective of many psychologists (as conveyed by the well-known colloquial term "physics envy").

The attempt to characterize psychological events with mathematical models is by no means a modern endeavor. Cognitive psychologists have made such attempts since the inception of the discipline—consider the well-known Weber–Fechner law first described in the 1800s. Nevertheless, there has long remained a degree of reluctance towards modeling psychological constructs void of physical anchors (i.e., hypothetical variables without physical antecedents or consequences). Wilhelm Wundt, commonly regarded as the founder of experimental psychology, felt that advanced mental processes were too complex to be studied experimentally (Hergenhahn, 2001). In the United States, William James's pragmatic approach towards psychology, along with the radical behaviorists' movement of the early 20th century, further entrenched this skeptical view of assessing cognition.

Louis Thurstone's pioneering series of publications in the 1920s were among the first empirical attempts to measure complex thought (for a review see Thurstone, 1959). Thurstone (1928) took his law of comparative judgment—based on the work of Weber and Fechner—and adapted it to create an equidistant interval scale of measurement for attitudes (called the *Thurstone scale*). In 1904 Charles Spearman wrote two seminal papers that led to the development of *common factor theory* (further developed by Thurstone) and *classical test theory* (CTT), the 20th century's dominant model in psychological measurement

(Spearman, 1904a, 1904b). Both theories assume that traits or characteristics of an individual's psyche can be meaningfully quantified—*psychometrics*. Well-known extensions of CTT provided by Guttman (1945) and Cronbach (1951), among others, allowed for widespread standards in the measurement of individual differences. Cronbach, Gleser, Nanda, and Rajaratnam (1972) later generalized the concepts of CTT in what came to be known as *generalizability* or *G theory*, a rich statistical framework for investigations of reliability.

CTT, as traditionally defined, appears to have reached a peak in conceptual and mathematical development during the mid 20th century (e.g., Gulliksen, 1950). And, as is natural in the progression of any science, an emergent class of measurement models including common factor theory, IRT, and latent class analysis among others came to be known as more powerful branches of psychometric theory. Oddly, while many psychologists followed the development of common factor theory, they failed to embrace the development of IRT, creating an artificial split between the two theories. The divide seems artificial because there is no obvious reason why CTT, common factor theory, and IRT should remain separate; CTT, for example, is simply a less explicit model nested within the common factor theory or IRT framework (McDonald, 1999). Indeed, Lord and Novick (1968) introduced IRT in connection with CTT, not as a distinct theory. McDonald (1999) argued that all branches of psychometric theory are mathematically and conceptually unified: a difficult fact to escape once the formulas are examined in detail.

CTT, also sometimes referred to as *classical true-score theory*, postulates that an examinee's observed score (*X*) on a test or measure of any psychological construct is determined by his or her true score (*T*) measured with some amount of error (*E*):

$$X = T + E. \tag{1}$$

An examinee's true score is commonly thought of as their expected score on an item or sum of items (i.e., the test). The CTT model has been described as being necessarily true rather than theoretical or falsifiable (Lord & Novick, 1968). That is, *Equation 1* is difficult to disprove. While necessarily true models are superficially attractive, philosophical tenets suggest that they are "weak" from a scientific perspective (Popper, 1964). To examine the meaningful qualities of a test (e.g., measurement error) a number of basic assumptions must be added to the model. Most importantly, it must be assumed that error scores are independent of true scores (i.e., $r_{TE} = 0$) and that error scores among items are independent (i.e., $r_{EE} = 0$). Assumptions must also be made about the nature of true scores (i.e., strictly parallel, parallel, tau-equivalent, or congeneric), but are rarely verified in practice. No direct relation between the psychological construct being assessed (e.g., depression) and a person's true score is postulated in the model. While relevant qualitative explanations are commonly offered for the meaning of a true score, the model itself remains ambiguous in this regard.

Some have argued that a potentially dangerous belief has emerged in which statistical analyses based on CTT are thought to equate with mathematical modeling (Thurstone, 1959, pp. 9). The two should not be confused. Statistical analyses are used to describe samples, infer population parameters, and test hypotheses. Accordingly, a reliability coefficient derived from the CTT model is a descriptive statistic. A mathematical theory is more than a tool to describe samples or to confirm hypotheses: it is a hypothesis within itself. Theories comprise rules and predictions to which observations should conform. Modern psychometric models can be used to test distinct hypotheses about the relations between psychological constructs and item responses, which may or may not be validated with inferential statistics. CTT makes very restrictive hypothesis about these relations, ultimately limiting the practical utility of clinical tests when the model is incorrect.

Perhaps this traditional lack of concern with measurement models is due primarily to psychologists' principal interest in patients, not tests. Diagnoses of psychopathology are meant to aid in treatment. Thus, clinicians' interests in tests are more practical than abstract. Nevertheless, there is a fundamental error in the CTT model that threatens to invalidate the utility of clinical tests' parameters (i.e., the structural relations between variables in a model): test parameters are confounded with person parameters—the two cannot be separated. In such situations, psychologists can only make judgments about examinees and tests relative to each other. Theorists have long acknowledged that something akin to a "psychometric grail" exists in the development and understanding of measurement instruments independent of the object(s) of measurement

(Thurstone, 1928; Wright, 1980). A proper model ought to incorporate separable parameters for both person and test parameters. Clear definitions and quantifications of psychological constructs can emerge only through such models.

The great importance of IRT can be attributed to the development of a class of models where person parameters and item parameters can be separated (Lord & Novick, 1968; Rasch, 1960). This separation has two immediate benefits: (1) item and test parameters can be estimated independent of person parameters; and (2) person parameters can be estimated independent of item and test parameters. IRT dissects the various components of a testing system, allowing psychometricians to study tests without the nuisance of individual differences, and clinicians to study people without the nuisance of test idiosyncrasies. Any system is better understood when all relevant variables are accounted for. Indeed, Georg Rasch (1960), a pioneer of IRT, developed his model by aspiring to mimic the explicit measurement of mass and force in a physical system (Rasch, 1960, pp. 115). Unlike CTT, a collection of models makes up the IRT family; each is characterized by increasingly comprehensive "systems" of measurement.

The humble, and yet remarkably complex goal of IRT, is to provide models that directly quantify psychological phenomena. As the physicist measures mass and force, so to might the psychologist assign concrete value to the otherwise intangible qualities of the mind—the very structure of thought. This is no simple task, and we should expect no simple solution. If nothing else, IRT is complex. Fortunately, there is an intuitive elegance to the models that allows for conceptual understanding without extensive knowledge of the underlying mathematics.

IRT Explained

Contrary to popular belief, IRT models and their foundations have existed for some time. The models gained prominence with the work of Lawley (1943; as cited by Lord & Novick, 1968), Lord (1952; 1953), Birnbaum (1968), and Rasch (1960), but their roots can be traced as far back as the phi-gamma law—a model for the ability to discriminate between physical stimuli based on the integral of the normal distribution—and the logistic function (Verhulst, 1845) developed in the 19th century. Item analyses of Binet and Simon's (1905) test of children's mental development reveal obvious precursors to IRT (see Bock, 1997). And, inspired by what he saw in Binet and Simon's data, Thurstone (1928) used the normal ogive curve—a standard model in IRT—to develop his law of comparative judgments for the measurement of attitude. Guttman (1944) and Coombs (1964) proposed methods for scaling data that can be characterized as deterministic (as opposed to probabilistic) IRT models. The study of biological assays in the first half of the 20th century led to the development of probit analysis and the logistic function as an alternative to the normal ogive model (Finney, 1952; Fisher & Yates, 1938; Hanley, 1952 as cited in Lord & Novick, 1968). But it was not until the 1970s and 80s that IRT began to flourish with the advent of accessible computers and the development of equations with practical applications for scoring data (Bock, 1997).

IRT models tend to be *stochastic* (or at least partially stochastic); that is, examinees' responses (e.g., "True" vs. "False") are assumed to be probabilistic. Also, IRT often models the existence of *latent variables*, constructs that cannot be directly measured, yet are inferred to exist based on peripheral associations among measurable qualities. For example, while eye color is observable, intelligence is not. There is no direct path to measuring intelligence. Instead, the "amount" of intelligence a person possesses is inferred from observed intercorrelated variables (e.g., academic success, processing speed, word knowledge, etc.). The concept of a latent variable is well established in the minds of many psychologists, as the common factor model (but not the principal components model) also assumes the existence of latent variables. In IRT, latent variables are symbolized with the Greek letter θ ("theta"), which can represent any psychological construct under investigation (e.g., depression, intelligence, conscientiousness, etc.). It is normally assumed in IRT that a test measures only one θ (i.e., unidimensionality); however, extensions of IRT to multidimensional data exist. Fundamentally, it must be assumed that local independence can be achieved. That is, after removing all common sources of item covariance (i.e., all latent variables), pairs or patterns of items must demonstrate no remaining correlation (*weak* and *strong* local independence respectively). IRT models have historically been applied to dichotomous response items (e.g., pass vs. fail), but applications to polytomous data have become common.

IRT models rest on the assumption that the probability of an examinee passing an item (where "passing" may refer to responding correctly or affirmatively) is a function of two sets of parameters: (1) their standing on θ, the person parameter; and (2) the characteristics of the item, the item parameters. The conditional probability of a particular response (e.g., $X = 1$; where 1 = "True") is a function of a person parameter (θ) and a set of item parameters (ξ; the Greek letter "ksi"):

$$P(X = 1 | \theta, \xi) = f(\theta, \xi). \tag{2}$$

This conditional probability is fundamental to all IRT models. In essence, the function is the regression of the probability of an item response [$P(X = 1)$] on the person and item parameters (θ, ξ). It is a logical proposition for how variables in a system affect one another. Unfortunately, the relation is not linear, and thus it is not possible to employ the typical linear regression form. Fortunately, the relation does tend to take a lesser-known form, the *normal ogive* (the integral or summation of the normal curve). The relation is also well represented by the *logistic ogive*, an "S"-shaped function known for modeling the exponential rate of natural growth followed by saturation. It is somewhat intuitive that both functions would closely model the probability of success as each is inherently related to the mathematical constant "*e*" (i.e., Euler's number), sometimes referred to as "the magic number of growth".

An example of a logistic ogive is presented in Figure 1. The x-axis represents a normally distributed latent variable (e.g., depression) with $\mu = 0$, and

σ = 1; however, it should be noted that non-normal as well as non-parametric data can be accommodated within the IRT framework. The y-axis represents the probability of a particular response (e.g., “True”), which can range from zero to one. Figure 1 is called an *item characteristic curve* (ICC), a graph of the probability of passing an item conditional on the continuum of latent variable values or θ. In Figure 1, an individual with a θ score of 0 would have a .50 probability of answering the item affirmatively, while a person with a θ score of 3 would have a .99 probability of answering the item affirmatively. Almost all IRT models “link” the probability of an item response with person and item parameters using a normal or logistic ogive. The choice between the two concerns a disparity in scaling; models can be scaled in either logistic or normal metric. Including a multiplicative constant to the logistic model produces results nearly identically to the normal ogive model (Haberman, 1974). The normal ogive model is more intuitive, but the logistic model is easier to work with mathematically as the latter does not require integration. Therefore, most presentations of IRT are confined to logistic IRT models. The exact shape of the logistic functions (and the normal ogives) will vary depending upon the item parameters.

Explanations of IRT are most easily understood for unidimensional scales with dichotomous items. The three most common are the Rasch or one-parameter model, the two-parameter model, and the three-parameter model. Each successive model estimates more parameters and can be considered a more accurate representation of the data. However, identification of the models and interpretation of their results becomes increasingly complex as parameters are added. The three primary item parameters estimated in IRT models are item difficulty, item discrimination, and item lower asymptote (pseudo-guessing). More generally, the item parameters from IRT are directly related to item parameters from factor analysis with categorical variables. The models were developed in relative isolation, and hence take on superficially distinct forms (see Heinen [1996] for an overview of the relations between latent trait models); however, the equivalence of some IRT models and factor analysis with categorical variables has been demonstrated (Kamata & Bauer, 2008; Takane & de Leeuw, 1987). The decision to analyze data with the common factor model or the IRT model is a bit arbitrary. The decision will ultimately be made by the purpose of the analysis. IRT provides very clear parameterizations of item characteristics. Factor analysis provides factor loadings, parameters familiar to psychologists because of their direct relation to correlation coefficients, but less useful for item analysis.

An item's *difficulty* parameter (b) is the θ value along the latent variable continuum (i.e., the range of all possible θ values) at which an individual has a .50 probability of passing that item (the parameter takes on a slightly different meaning for more complex models). Graphically, the difficulty parameter is the ICC's point of inflection. For example, if an item has a difficulty parameter equal to 0 (as in Figure 1), examinees whose latent variable values fall above 0 have greater than a .50 probability of passing the item, while examinees whose latent variable values fall below 0 have less than a .50 probability of passing the item. In the study of psychophysics this parameter is known as the threshold or limen and in biological assay research it is known as the median lethal dose. In psychological research, the latent trait distribution is often assumed to take the form of a normal distribution, and thus *b* is unbounded (i.e., $-\infty < b < \infty$). In relation to factor analysis, the item difficulty parameter (*b*) is a function of an item's threshold (*v*; the Greek letter "nu") and an item's standardized loading (λ; the Greek letter "lambda"):

$$b = \frac{\nu}{\lambda}. \tag{3}$$

An item's *discrimination* parameter (*a*) is related to the slope of its ICC at its difficulty value (the ICC's point of inflection). Items with higher discrimination values are more discriminating between distinct levels of *θ*. The discrimination parameter is also bound by infinity, but it normally takes on an absolute value of less than 4. (We typically limit our discussion to the odds of passing an item, and thus *a* will generally be positive). In relation to factor analysis, the item discrimination parameter is a function of an item's standardized loading (*λ*):

$$a = \frac{\lambda}{\sqrt{1-\lambda^2}}. \tag{4}$$

Thus, the discrimination parameter is both conceptually and mathematically related to a factor loading.

Finally, the *lower asymptote* parameter (*c*), also known as the *pseudo-guessing parameter*, is so named to account for the fact that with some types of response formats (e.g., multiple choice tests), examinees can pass items simply by guessing. The pseudo-guessing parameter might also be thought of as the probability of an examinee passing an item when they are void of the latent trait (i.e., when $\theta = - \infty$). It is referred to as "pseudo-guessing" because examinees rarely guess based on odds alone. For example, examinees that are distracted by an attractive incorrect response option could do worse than chance. On clinical or personality tests, the pseudo-guessing parameter has occasional been thought of as indicative of a response style (e.g., social desirability, true response bias, etc.). In bioassay toxicology research, the *c* parameter functions as an estimate of natural mortality. The lower asymptote parameter does not have an equivalent parameter in factor analysis.

Before IRT models and application of those models are presented, it is useful to discuss the language of IRT, which has long been the domain of those who study ability and achievement. Such literature has often served as the driving force for advancements in psychometric theory. This is true even with CTT, where researchers and test developers studying personality and mental health domains have mimicked test development methods from the measurement of ability and achievement (e.g., Jackson, 1976). A similar adaptation of IRT is driving the current revolution in psychometric theory. Unfortunately, this means that much of the IRT language is uniquely suited to the assessment of ability and achievement. For example, item responses are commonly noted as passes or failures, latent variables are often referred to as abilities, and the lower asymptote parameter is considered to be a guessing parameter. It should be realized, however, that such terms are simply labels used to describe mathematical variables in the model. Latent “abilities” can be thought of as latent “traits”, “factors”, or any other general variable, “passing” an item can be thought of as “affirming” or “endorsing” an item, and item “difficulty” can be thought of as “severity” (e.g., the severity of depression required to endorse the item, “I’ve often thought of ending my own life.”).

IRT Models

The Rasch model. Georg Rasch, a Danish mathematician who took a special interest in statistics and measurement in cognitive testing, developed his model—known as the *Rasch model*—in an effort to achieve objective measurement in the social sciences (Rasch, 1960). Specifically, Rasch was concerned with the relativistic nature of psychological measurement. He felt that measures of psychological variables ought to behave like measures of physical variables (e.g., using a ruler to measure height has the same meaning whether we are measuring an elephant or a pencil). Rasch named this property of measurement *specific objectivity*: interpretation of measurement units independent of the object of measurement. Such objectivity was only achievable with a model that estimated person parameters and item parameters independently, an IRT model. Rasch (1960, p. 19) used Poisson process models to demonstrate how if certain conditions are met, and person parameters are known (or are sufficiently estimated), test parameters can be estimated with specific objectivity.

In the Rasch model, item difficulty is estimated separately for each item. Thus, a test may contain any combination of low, medium, and high difficulty items. However, the Rasch model demands that all items in a test have the same discrimination value. Each must be equally related to the latent variable (this is required in order to produce a sufficient statistic for estimating person parameters). This implies that all items must have equivalent factor loadings and biserial correlations. The Rasch model is also sometimes referred to as the *one-parameter model*, as only item difficulty is allowed to vary. This is apparent in the formula, where the probability of a person passing an item is conditional only on their ability (θ) and the item's difficulty (b):

$$P(X = 1 | \theta, b) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}} \, . \qquad (5)$$

The form $e^x/1+ e^x$ is the logistic link function mentioned earlier (used in non-linear regression), and can largely be ignored. Instead, the reader should instead focus on the θ - b term in parentheses. This part of the equation conveys that whether or not a person endorses an item is a function of the difference between the item's difficulty and the person's ability. The ICCs for three hypothetical test items that fit a Rasch model are displayed in Figure 2. The ICCs for each item take on identical S-shaped logistic functions, which reflect the fact that each has the same item discrimination value (equal slopes); however, the items are separated along the x-axis. Specifically, each of the three items has a unique difficulty value (i.e., the point at which the probability of passing each item is .50). Item 1 has a low difficulty value, Item 2 has a medium difficulty value, and Item 3 has a high difficulty value.

Creating a test composed of items with identical discrimination values is not a simple task. Despite this, the Rasch model has enjoyed great popularity. To understand why, one very important feature of a Rasch model must be noted: the total number of items an examinee answers correctly is a sufficient statistic for knowledge of the θ distribution. This means that latent trait estimates are neither person nor item specific. Thus, a test fitting the Rasch model exhibits a strict form of specific objectivity that cannot be accomplished with other IRT models. Some argue that only items that the fit a Rasch model should be selected for tests because of the benefits the model provides (e.g., Wright, 1992). A counterargument is offered that forcing test items to conform to the Rasch model limits a test’s validity (e.g., Hambleton, 1992). The argument has taken on somewhat of a philosophical tone, and cannot likely be resolved through empirical means.

Rasch models enjoy great popularity in Europe, and have seen moderate use in the United States. They have been fit to measures and diagnostic criteria for anxiety (e.g., Ludlow & Guida, 1991), compulsive smoking (Breteler, Hilberink, Zeeman, & Lammers, 2004), depression (e.g., Bouman & Kok, 1987; Chambon, Cialdella, Kiss, & Poncet, 1990; Chang, 1996; Cole, Rabin, Smith, & Kaufman, 2004; Maier & Philipp, 1986), general psychopathology (Olsen, Mortensen, &

Bech, 2004), health (Andrich & Van Schoubroeck, 1989), pain (Kalinowski, 1985), paranoia (e.g., Kreiner, Simonsen, & Mogensen, 1990), schizophrenia (e.g., Bell, Low, Jackson, & Dudgeon, 1994; Lewine, Fogg, & Meltzer, 1983), and self esteem (e.g., McRae, 1991). The model has even been applied to a projective test (Tuerlinckx, De Boeck, & Lens, 2002). However, because the Rasch model demands identical item discrimination parameters, it often fails to fit scales developed with CTT technology (e.g., Tenenbaum, Furst, & Weingarten, 1985).

The two-parameter model. The two-parameter model is a more general case of the Rasch model. It is often estimated using the logistic function, and thus is usually referred to as the *two-parameter logistic* (2PL) model. As with the Rasch model, item difficulty is estimated separately for each item. However, the 2PL model estimates unique item discrimination parameters. Thus, two or more items may have different or identical difficulty values and different or identical discrimination values. This is apparent in the formula, where the probability of a person passing an item is conditional on their ability (*θ*), the item's difficulty (*b*), and the item's discrimination (*a*):

$$P(X = 1|\theta, b, a) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}. \qquad (6)$$

The 2PL model is also based on the difference between the item's difficulty and the person's ability, but the impact of the difference is weighted by item discrimination parameters (i.e., not all items are equally relevant to the construct). The ICCs for three hypothetical test items that fit a 2PL model are displayed in

Figure 3. As can be seen, the ICCs for each item take on non-identical S-shaped logistic functions. The differences in the shapes of the ICCs reflect varying item discrimination values. The steeper curves of items 1 and 3 reflect higher discrimination values. The flatter curve of item 2 reflects a lower discrimination value. As with the Rasch model, the ICC's may vary along the x-axis as each item's difficulty parameter is estimated separately. In general, higher discrimination parameters equate with more accurate assessment at an individual item's difficulty value.

Some curious situations can arise with the 2PL model. In particular, the item ICCs can cross along the latent variable continuum. Crossed ICCs indicate that the ordering of local item difficulty is dependent on the θ distribution. In Figure 2, item 2 has the highest probability of endorsement for a person with a θ value of -4, but has the lowest probability of endorsement for a person with a θ value of +4. This occurrence seems rather illogical for those who have not examined the 2PL model's formula in great detail. Yet these results are reasonable from a measurement perspective. The relation between an item with a low discrimination value and θ is more ambiguous than the relation between an item with a high discrimination value and θ. Items with low discrimination parameters may be passed or failed for a variety of reasons unrelated to θ (e.g., another, unaccounted for, latent trait). Thus, passing an item with a high discrimination parameter tells us more about a person's standing on θ than does passing an item with a low discrimination parameter.

Interpretations of the 2PL model are more ambiguous than interpretations of the Rasch model. Namely, total scores are not sufficient statistics in the 2PL model due to varying discrimination parameters. In order to estimate a person's θ value, we must have knowledge of the specific items answered correctly or incorrectly so that they can be weighted by a. Therefore, item parameters cannot be estimated in accordance with Rasch's strict concept of specific objectivity. The 2PL model does have the distinct advantage of being much more flexible than the Rasch model. Fitting a 2PL model to data allows researchers to accept items that vary in difficulty and discrimination. Thus, not only is it easier to create tests based on the 2PL model, it is also easier to fit a 2PL model to existing tests.

Because of such flexibility, the 2PL model is more congruent with existing clinical measures than is the Rasch model. It has been fit to measures and diagnostic criteria for adult attachment (e.g., Fraley, Waller, & Brennan, 2000), anxiety (Ietsugu, Sukigara, & Furukawa, 2007; Rodebaugh et al., 2004), children's moods and feelings (Sharp, Goodyer, & Croudace, 2006), depression (e.g., Childs, Dahlstrom, Kemp, & Panter, 2000; Clark, Cavanaugh, & Gibbons, 1983; Dorus, Kennedy, Gibbons, & Ravi, 1987; Gibbons, Clarke, VonAmmon Cavanaugh, & Davis, 1985), personality (e.g., Ferrando, 1994; Grayson, 1986; Haans, Kaiser, & de Kort, 2007; Kamakura & Balasubramanian, 1989; Waller, Thompson, & Wenk, 2000), job satisfaction (e.g., Hulin, Drasgow, & Komocar, 1982; Parsons & Hulin, 1982), job selection (e.g., Raju, Steinhaus, Edwards, & DeLessio, 1991), modernity (e.g., Hui, Drasgow, & Chang, 1983), psychopathy (e.g., Cooke & Michie, 1997; Cooke & Michie, 1999), sexual harassment (e.g., Donovan & Drasgow, 1999), stress (e.g., Smith & Reise, 1998), and substance use (Martin, Chung, Kirisci, & Langenbucher, 2006; Saha, Chou, & Grant, 2006).

The three-parameter model. The three-parameter model is a more general case of the 2PL model. As with the 2PL model, the three-parameter model is often estimated using the logistic function, and thus is typically referred to as the *three-parameter logistic* (3PL) model. The model adds the lower asymptote or pseudo-guessing parameter, which can be set to a constant or freely estimated for each item. As mentioned earlier in this paper, the pseudo-guessing parameter accounts for potential guessing or response bias. In the formula for the 3PL model, the probability of passing an item is conditional on θ, b, a, and c:

$$P(X = 1 | \theta, b, a, c) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}} . \qquad (7)$$

The ICCs for three hypothetical test items that fit a 3PL are displayed in Figure 4. As with the 2PL, the ICCs for each item take on non-identical S-shaped logistic functions. Also, as with the Rasch model and the 2PL model, the ICCs are separated along the x-axis as each item’s difficulty parameter is estimated separately. Unlike the Rasch and 2PL models, the additional pseudo-guessing parameter (c) creates a non-zero lower asymptote for some ICCs. Specifically, item 3 has a non-zero lower asymptote.

While the 3PL model is more general than both the Rasch model and the 2PL model, it also adds mathematical complexity that makes item parameters more difficult to estimate. In addition, item parameters in the model have a more ambiguous meaning. The difficulty parameter in the 3PL model no longer has the same interpretation as it did with the Rasch and 2PL models (specifically, the inflection point of the ICC will be greater than a .50 probability of passing an item for $c > 0$).

The 3PL model has only rarely been applied to clinical assessment (e.g., Harvey & Murry, 1994; Rouse, Finger, & Butcher 1999). Conceptualizing the impact of “pseudo-guessing” on items related to personality and psychopathology is difficult. Most applications of the lower asymptote to non-cognitive measures have occurred under the pretext of a *social desirability* parameter. For example, if examinees are unwilling to respond openly to an item about sexual practices, drug use, mental health, etc., that item could bias all examinees’ responses towards a more conservative response option. Note, however, that this strategy assumes uniform response bias among examinees. The logic would suggest that items, not examinees, are biased towards particular response options. Therefore, the strategy cannot be used to differentiate between examinees with different response styles, and serves only to uniformly alter the probability of all examinees’ responses. Rouse et al. fit a 3PL model to 5 scales from the second edition of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, & Kaemmer, 2001) designed specifically to identity personality disorders. The authors found a substantial correlation between estimates of c and indices of social desirability.

Models for polytomous data. The Rasch (1PL), 2PL, and 3PL models are all designed to model the odds of dichotomous item responses (e.g., correct vs. incorrect, true vs. false, yes vs. no). In contrast, clinical tests often present examinees with multiple response options (e.g., a Likert-type scale). Fortunately, extensions of IRT to polytomous item response formats emerged soon after dichotomous IRT models. As with the models presented thus far, polytomous IRT models tend be hierarchically nested; each makes more assumptions than the previous. The *nominal response model* (Bock, 1972), and an extension that accounts for guessing (Thissen & Steinberg, 1984), are the most general, simply requiring nominal response options. Polytomous Rasch models include the *rating scale model* (RSM; Andrich, 1978a, 1978b) and the *partial credit model* (Masters, 1982), special cases of the nominal response model that assume ordered categorical response options and invariant item discrimination parameters. The models estimate *thresholds* or *step difficulties* between response categories: the points at which examinees endorse, or “step into”, higher response categories. The rating scale model assumes equidistant intervals between item response categories; the partial credit model does not. More closely related to the 2PL model, the *generalized partial credit model* (Muraki, 1992) allows for varying discrimination parameters.

The *graded response model* (Samejima, 1969, 1996) is the oldest and best-known polytomous IRT model. In the formula, an examinee’s probability of scoring in each of the specific ordered categorical response options is a function of the varying item threshold parameters (similar to item difficulty parameters but specific to each response option) and the varying item discrimination parameters. The model is unique in that it comes from the Thurstone tradition of comparative judgments (sometimes referred to as *difference models*) rather than the Likert tradition of dominance ratings (sometimes referred to as *divide-by-total models*; see Thissen & Steinberg, 1986). For example, the graded response model can be used to determine the boundaries at which examinees' become more likely to endorse successive response categories on an item offering multiple response options (e.g., never, rarely, sometimes, or always). Figure 5 presents the *category response curves* for an item with four response categories. The lines represent the probability of an examinee with a given θ value endorsing a particular option. In the example, an examinee with a θ value of -2 is most likely to endorse the "Never" category, while an examinee with a θ value of +2 is most likely to endorse the "Sometimes" category. Muraki (1990) developed a modification of the graded response model in which the distance between response categories is assumed to be constant within items, while the difficulty of each item (set of category responses) is estimated separately. Samejima (1969) demonstrates that the graded response model produces better estimates of θ values than do dichotomous IRT models. Indeed, existing tests can improve reliability by changing to polytomous response formats (e.g., Lanyon & Thomas, 2009).

The graded response model has been applied to measures and diagnostic criteria for affect and emotion (e.g., Baker, Zevon, & Rounds, 1994), anxiety

(Rodebaugh et al., 2004; Rodebaugh, Woods, Heimberg, Liebowitz, & Schneier, 2006), childhood behavior (Lambert et al., 2003), criminal behavior (e.g., Osgood, McMorris, & Potenza, 2002), depression (e.g., Orlando, Sherbourne, & Thissen, 2000), inattention and impulsivity (Gomez, 2008), job satisfaction (e.g., Hanisch, 1992), obsessive-compulsive disorder (Williams, Turkheimer, Schmidt, & Oltmanns, 2005), personality (e.g., Bejar, 1977), post-traumatic stress (e.g., Orlando & Marshall, 2002), quality of life (e.g., Uttaro & Lehman, 1999), and self-esteem (e.g., Gray-Little, Williams, & Hancock, 1997).

Models for multidimensional data. The IRT models discussed so far have all assumed that a single latent variable accounts for the observed interercorretlations among items (i.e., unidimensional scales), a limiting requirement. The need for multidimensional models in psychological assessment has long been recognized (e.g., Thurstone, 1947). Researchers and test developers who make incorrect assumptions of unidimensionality are either forced to remove misfitting items from tests, or carry through with analyses despite the violations. While doing so can be acceptable (i.e., when it does not drastically alter results), there are tests and items for which unidimensional models of latent variables simply do not accurately account for empirical data. For example, in clinical assessment it can be difficult to create a depression item that is not also related to anxiety. While multidimensionality has often been considered something of a nuisance, researchers have begun to model the unaccounted for variance in the development of more fine-tuned concepts of latent traits. Figure 6 presents the *item characteristic surface* for an item that is dependent on two latent variables (θ_1 and θ_2). A multidimensional system requires more parameters. Thus, the estimation of such models is much more complex than with those discussed previously.

Multidimensional Rasch models have been developed (e.g., McKinley & Reckase, 1982; Fischer & Seliger, 1997), but may be too restrictive for clinical tests. Extensions of the 2PL and 3PL models to a multidimensional framework are available for dichotomous items (see Reckase, 1997) and polytomous items (Kelderman, 1997). Several authors (Christoffersson, 1975; McDonald, 1967; Muthén, 1978) have developed multidimensional normal ogive models for dichotomous items by extending common factor theory. The common factor model can serve as an approximation of the normal ogive model through the analysis of tetrachoric or polychoric correlation matrices. These models are directly related to factor analysis—through the equivalencies discussed earlier—greatly facilitating psychologists' understanding of results. However, there are some computational difficulties associated with analyzing tetrachoric correlation matrices. As an alternative, researchers can use a normal ogive model for multidimensionality data known as *full-information item factor analysis* (see Bock, Gibbons, & Muraki, 1988), and its extension to polytomous data (Muraki & Carlson, 1995), as more sound procedures. Full-information refers to the direct analysis of item responses instead of item correlations. For researchers interested in understanding the underlying structure of measured constructs—the "building blocks” of psychiatric disorders—componential IRT models can be used to analyze the related components of item difficult along with the traits required for endorsement (e.g., the *multicomponent latent trait model*; Whitely, 1980*;* Embretson, 1984).

Hierarchical models are routinely employed in psychological research through the use of structural equation modeling (SEM; see Bollen, 1989), but are less pervasive in IRT. Gibbons and Hedeker’s (1992) *bifactor IRT model*, and an extension to polytomous data (Gibbons et al., 2007), are appropriate when examinees’ observed responses are a function of their standing on a pervasive general trait as well as a series of domain specific traits. *Higher-order IRT models* are appropriate when examinees’ observed responses are a function of several lower-order traits, which are themselves a function of higher-order traits (Sheng & Wikle, 2008). Figure 7 presents an example of a bifactor model and Figure 8 presents an example of a higher-order model. The models are mathematically equivalent under certain conditions; the higher-order model can be thought of as a constrained bifactor model (Chen, West, & Sousa, 2006; McDonald, 1999). However, the bifactor model is generally preferred, particularly when researchers want to examine predictive relations between domain specific factors and external criteria (Chen et al., 2006). Both add complexity to parameter estimation and interpretation, but likely provide more accurate representations of data. Kamata (2001) demonstrates how the hierarchical generalized linear model provides an IRT framework for 2 or more higher-order levels, but work in this area has been sparse. In educational domains, the researchers sometimes employ the *testlets* (Wainer & Kiely, 1987), in which locally dependent items are combined. The testlet model can also be thought of constrained versions of the bifactor model (DeMars, 2006).

Researchers have generally found that multidimensional and hierarchical models improve measurement precision in comparison with unidimensional models for tests of personality and psychopathology (e.g., Cabrero-García & López-Pina, 2008; Cole, Rabin, Smith, & Kaufman, 2004; Gardner, Kelleher, & Pajer, 2002; Gibbons, Rush, & Immekus, 2009; Gibbons et al., 2008; Michie & Cooke, 2006; Wang, Chen, & Cheng, 2004; Yao & Boughton, 2007). Using simulated data, DeMars (2006) demonstrated how using an independent items model (unidimensional) to estimate data with a true bifactor structure can lead to inaccurate parameter estimates. In addition, such models offer clinicians a glimpse into the underlying structure of clinical disorders. For example, Smits and De Boeck (2003) used a componential IRT model to identify three components contributing to the psychological experience of guilt: norm violation, worrying about what one did, and a tendency to restitute. Such explicit mathematical modeling can be used to enrich clinical descriptions of patients' symptoms. The bifactor model can also be used to determine if underlying multidimensionality meaningfully alters interpretations of scales (Parsons & Hulin, 1982; Reise, Morizot, & Hays, 2007). For example, some researchers have found that even when bifactor models fit data, the variance contributed by domain specific factors is too small to dismiss a more parsimonious unidimensional framework (Brouwer, Meijer, Weekers, & Baneke 2008; Reise et al., 2007).

The ever-increasing list of multidimensional IRT models provides for sophisticated modeling of cognitive processes through mathematical equations (van der Linden & Hambleton, 1997). Such modeling is conceptually similar to, and at times mathematically equivalent with, SEM. Unfortunately, many of the models are simply too complex to be of practical use to clinical researchers. Nonetheless, the advancement and unification of various latent trait analyses promises to increase access to such models, perhaps one day making their use routine.

Other extensions of IRT models. IRT models for nonparametric data (Mokken, 1971), nonmonotone responses (Andrich, 1988), multiple groups (manifest, Mislevy, 1984; and latent, Rost, 1990), and locally dependent data (Jannarone, 1997) have all emerged. Many of these extensions pose formidable challenges to those wishing to use them (e.g., complex formulas, lack of accessible software for performing the analyses, and unknown proprieties of the models). Technological and theoretical advances must be made before the wide array of IRT models becomes accessible. Some researchers argue that such models provide better fit for clinical data, as non-parametric IRT models have been successfully fit to measures of psychopathology (Meijer & Baneke, 2004; Roberson-Nay, Strong, Nay, Beidel, & Turner, 2007) and personality (Stark, Chernyshenko, Drasgow, & Williams, 2006).

Comparing and Contrasting Applications of CTT and IRT

While it may be apparent that IRT comprises a class of models substantially more complex than CTT, it may not be clear what practical advantages these models provide. Lord and Novick (1968) detail how item parameters in the CTT model can be used to predict and/or alter the behavior of an entire measurement instrument. Wiggins' (1973) treatment of personality assessment emphasizes the practical implications of CTT on test development and use. An example of the more sophisticated use of CTT comes from Jackson's (1976) utilization of internal consistency procedures in the development of a personality inventory. And, an extension of the CTT model mentioned earlier, generalizability theory (see McDonald, 1999), has added even more sophistication to test development and evaluation.

CTT should not quickly be discounted in test development. Indeed, it has served psychologists well for over 100 years of test development (though it should be noted that the techniques have often been used in tandem with factor analysis). However, CTT and its extensions share the common flaw mentioned earlier: an inability to separate item and person parameters. The estimation of item difficulty, item discrimination, and test validity and reliability coefficients using CTT are all dependent on specific populations and specific tests. For example, an item on an intelligence test may be very difficult for a population of developmentally delayed children while simultaneously very easy for a population of developmentally advanced children. CTT gives us no way to determine which difficulty estimate is “correct”. Typically, test developers have circumvented this problem by collecting normative data. And, in many circumstances (e.g., when test examinees are well matched with the normative data), the strategy has sufficed. But it can also be limiting. For example, the practice of comparing the scores of ethnic minorities to the scores of a normative sample primarily made up of the ethnic majority can be inaccurate and unethical.

In addition, several odd mathematical consequences are the results of, and ill-explained by, the CTT model. For example, when items are not passed at a rate of 50% in the calibration sample, item-test correlations are attenuated. This results in a practical advantage to choosing items of “average” difficulty. In addition, researchers often find that tests with high item intercorrelations have poor predictive validity. Yet adding items to a test in order to increase validity will in turn lower reliability: the *attenuation paradox* (Lord & Novick, 1968, p. 344). While such occurrences are not paradoxical from a mathematical point of view, they certainly do not make intuitive sense. Most would expect that adding items to a test ought to add some useful information, or at worst, add no information at all (i.e., neither helpful nor harmful). These limitations arise because the CTT model proposes no explicit relation between observed scores and the underlying cause of those scores (i.e., the latent variable). It should again be noted that the CTT model can be viewed as a less explicit latent variable model. Guttman-Cronbach alpha, for example, is a lower bound estimate of the more explicit common factor model’s reliability coefficient “omega” (ω ;McDonald, 1999). IRT allows for more fine-grained evaluations of existing measures, which are related to, but far superior than CTT-based evaluations of existing measures. Notable examples are described below.

Reliability, information, and standard error. The CTT-based concept of reliability and the IRT-based concept of *information* are both inversely related to standard error of measurement. Higher information equates with higher reliability, lower standard error, and more precise latent trait estimates. However, whereas the traditional CTT-based notion of reliability is assumed to be constant for all examinees (although a conditional standard error of measurement equation does exist), the IRT-based notion of information is assumed to differ between examinees. Specifically, information is a function of θ called the *item information function*. Item information functions peak at item difficulty parameters (e.g., an item with a b parameter equal to 1 will produce the most information for examinees with θ values equal to 1). Intuitively, most would suspect that asking a kindergartner to solve a calculus equation would provide very little information about the child's achievement in mathematics; however, asking a college student to solve the same question could be more informative. Questions that are too hard or too easy for an examinee provide little information about their ability.

The information of an entire measure is called the *test information function*. Unlike reliability, information is additive when local independence holds. An item's absolute contribution to a test is not dependent on the group of items already contained in the test (Lord, 1980, p. 72). The test information function is simply the sum of all item information functions. Figure 9 is an example of a test information function from a measure that is most accurate at the higher end of the θ distribution, and least accurate at the lower end. The measure is well suited for use in populations with high severity of a disorder, but ill-suited for use in populations with low severity. The *standard error of measurement function*, also displayed in Figure 9, is inversely related to the test information function.

Information functions are used to evaluate the precision of existing items (e.g., Marshall, Orlando, Jaycox, Foy, & Belzberg, 2002) and scales (e.g., Flannery, Reise, & Widaman, 1995; Frazier, Naugle, & Haggerty, 2006). For example, researchers have evaluated test information functions to determine where existing clinical measures are most accurate. Young, Halper, Clark, and Scheftner (1992) evaluated the Beck Hopelessness Scale (Beck, Weissman, Lester, & Trexler, 1974) and concluded that the test makes the most accurate θ estimates for mid to high θ values. Consequently, the authors concluded that the scale is of little diagnostic use for individuals low on the construct of "Hopelessness". Researchers have found that the Diagnostic and Statistical Manual of Mental Disorders' (DSM; see American Psychiatric Association, 2000) diagnostic criteria for substance dependence (Langenbucher et al., 2004), depression (Aggen, Neale, & Kendler; 2005), and borderline personality (Feske, Kirisci, Tarter, & Pilkonis, 2007) each have highly peaked information functions. This suggests that the DSM's criteria are best used to make dichotomous classifications of individuals (e.g., dependent vs. non-dependent, depressed vs. non-depressed, and borderline vs. non-borderline) rather than continuous classifications. Similarly, Harvey and Murry (1994) found that the Myers-Briggs Type Indicator (Myers, 1962) functions best as an indicator of dichotomized θ values (i.e., *types*) rather than placing people on continuums.

Tests can be reorganized into shorter and/or more informative versions by trimming away items that provide little information (e.g., Cooper & Gomez, 2008; Kim & Pilkonis, 1999). For example, Grayson (1986) chose the ten most informative items from the Eysenck Personality Questionnaire subscales (Eysenck & Eysenck, 1975) in order to create shorter, more informative scales. Duncan-Jones, Grayson, and Moran, (1986) did the same for the General Health Questionnaire (Goldberg, 1972). Lee and Smith (1988) fit a Rasch model to the California Psychological Inventory (Gough, 1956) and created a tailored version with 71% fewer items than the original measure. More deliberate attempts at test construction involve the use of *target information functions* (Luecht & Hirsch, 1992), where test developers choose items that will maximize information for a predetermined range of θ values based on need, law, or precedent. The methodology is similar to criterion-based test construction, where the criterion is the target information function (Hambleton & de Gruijter, 1983). For example, *mastery tests* (Lord, 1980) or *screening tests* can be designed to provide peak information at a chosen threshold used to classify cases from non-cases (Biranbaum, 1968). Kessler et al. (2002) created a screening measure for psychological distress by choosing items that contributed information in the 90-99th percentile of the θ distribution.

Researchers have found that many existing screening or categorically diagnostic measures have information functions that dramatically peak at a diagnostic cutoff typically on the "impaired" end of the θ distribution. For example, a screening version of the original Psychopathy Checklist-Revised (Hare, 1991) contributes less overall information than its full-version counterpart, but nearly as much information at the diagnostic cutoff (Cooke, Michie, Hart, & Hare, 1999). The mental health scales of Psychological Screening Inventory provide peak information near the test developer's recommended cutoffs for concern versus no concern, but are relatively inaccurate for individuals with low symptoms of distress (Lanyon & Thomas, 2009). Curiously, this implies that test developers have achieved their hypothetical target information functions without the benefits of IRT. Why does this happen? Reise and Waller (2009) suggest that information functions peak on the higher ends θ distributions because clinical constructs are meaningful in one direction only. For example, they point out that the low end of depression is lack of depression, not happiness. Clinicians rarely ask questions on this end of the spectrum because they lack clinical relevance (e.g., "Have you ever felt sad?"). In addition, items have historically been chosen based on their point-biserial correlations with a criterion group variable (e.g., individuals classified as "normal" vs. "abnormal"). Items with the highest point-biserial correlations, those endorsed frequently by the "normal" group and infrequently by the “abnormal group”, have difficulty parameters directly at the criterion point. Thus, test developers who chose items that best differentiated between criterion groups, had, by default, chosen items with overlapping difficulties values near the underlying criterion θ value: a target information function.

Perhaps the most opportunistic use of information functions comes through *computer adaptive tests* (CATs; e.g., Walter, Becker, Bjorner, Fliege, Klapp, & Rose, 2007). CATs tailor item administration to produce peak information for individual examinees’ θ estimates. Computers must be used to generate real-time iterative estimates of examinees’ θ values, as they are unknown prior to test administration. Once an initial estimate has been produced, subsequent items are chosen from a pre-calibrated item pool to maximize information. This normally involves administering a slightly harder item when an examinee answers correctly and administering a slightly easier item when an examinee answers incorrectly. To appreciate the strategy, consider a hypothetical situation where a clinical graduate student comes across their first depressed patient. After initial introductions, the student asks the patient, “Are you feeling hopeless?” to which the patient replies, “Yes, I am.” And then, given the student’s naïve understanding of depression, they follow up by asking the next question on their list, “And have you been feeling a little blue lately?” It does not take a veteran clinician to realize that a person who is hopeless is almost certainly “feeling a little blue lately”. The question contributes almost no useful information to the assessment. A wiser clinician might have asked a more appropriate follow-up question, “Have you thought of ending your own life?”

CATs function like wise clinical interviewers. They make ongoing estimates of an examinees’ θ value, and choose to administer items that will provide the greatest amount of information. Doing so can drastically reduce testing time and burden. For example, Forbey and Ben-Porath, (2007) reduced MMPI-2 testing by 20% with an experimental simulation CAT. Waller and Reise (1989) simulated a CAT version of the Absorption Scale of the Multidimensional Personality Questionnaire (Tellegen, 1982) and were able to reach accurate estimates of latent trait values using on average only 25% of the original items. Kamakura and Balasubramanian (1989) did the same for socialization subscale of the California Psychological Inventory using on average only 33% of the original items. Reise and Henson’s (2000) simulated CAT version of revised Neuroticism-Extroversion-Openness Inventory (Costa & McCrae, 1992) reduced item administration by half.

Functional CATs require a large collection of items with precalibrated parameters (e.g., Lai, Cella, Chang, Bode, & Heinemann, 2003). Unfortunately, such complex “item banks” are difficult to develop without the aid of heavily funded research projects or professional testing services. Private interest in developing CATs for industrial/organizational applicants has spurred some development (e.g., Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001; Schneider, Goff, Anderson, & Borman, 2003). Several authors have recommended national efforts to create large-scale item banks to facilitate the effort (e.g., Hahn, Cella, Bode, Gershon, & Lai, 2006; Revicki & Cella, 1997; Revicki & Sloan, 2007), and the National Institutes of Health is currently funding the most ambitious attempt to do so in the realm of clinical assessment. The Patient-Reported Outcomes Measurement Information System (PROMIS) is a multi-site collaborative research effort to standardize a large bank of physical and psychoemotional health items (Cella et al., 2007). While the PROMIS system currently invites researchers to utilize developmental CATs in collaborate research efforts, the scales are not yet available for clinical practice.

Scaling and equating. IRT's explicit measurement models facilitate meaningful scaling of item and person parameters. CTT-based total scores do not directly quantify psychiatric disorders, nor are they directly related to the behavioral, cognitive, and/or emotional symptoms of distress. Nonetheless, a common assumption is that endorsement of more symptoms equates with a higher likelihood or severity of a disorder. That is, total scores are assumed to maintain ordinal properties with respect to latent variables. Percentile rankings are the only "permissible statistic" that can be used to meaningfully summarize ordinal data (Stevens, 1946). Unfortunately, this leaves clinicians in the unenviable position of explaining to patients' how their responses compare with normative patients' responses: the epitome of relative measurement. In contrast, it can be argued that IRT simultaneously scales the relations between item parameters and person parameters on what approaches an interval scale of measurement. That is, increases in θ values equate with additive or linear increases in the log-odds of item endorsement. Patients' scores can be described in direct relation to symptoms of distress.

To understand why this is important, consider Juan and Erica's hypothetical scores on measure of depression. Within a CTT framework, the relation between Juan and Erica's scores can only be described in reference to a normative population (or to each other). For example, because Juan's 99th percentile total score is higher than Erica's 7th percentile total score, we can conclude that Juan is more depressed. However, we can say nothing beyond this comparison. In IRT, the difference between Juan and Erica's θ values can additionally be interpreted with respect to the behavioral, cognitive, and/or emotional symptoms related to the latent trait under investigation. Odds and/or probabilities of symptom endorsement can be provided because logistic and normal ogive models link θ values with the probability of affirming various items. For example, we might conclude that Juan's standardized θ score of +2.50 equates with a 99% chance he is feeling sad, a 90% chance he is feeling hopeless, and a 30% chance he is contemplating suicide. On the other hand, Erica's standardized θ score of -1.50 equates with a 20% chance she is feeling sad, a 5% chance she is feeling hopeless, and a 1% chance she is contemplating suicide. We could even determine how much of a decrease in θ is required to reduce Juan's suicidal ideation to 5%. (See "Wright Maps" for another example of simultaneous scaling of parameters). Such meaningful descriptions of examinees' scores are not only useful in the diagnosis of psychiatric disorders, but also in the development of symptom oriented treatment plans.

In research situations, simultaneous scaling of person and item parameters contributes to the theoretical understanding of latent variables. For example, analyses of the Psychopathy Checklist-Revised have revealed that individuals with a “callous/lack of empathy” are more psychopathic than individuals with a “need for stimulation” (Cooke & Michie, 1997). The interpersonal and affective features of psychopathy have higher thresholds than the impulsive and antisocial features (Cooke et al., 1999). Analyses of the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) reveal that items related to suicide have higher difficulty values than items related to crying (Gibbons, Clarke, VonAmmon Cavanaugh, & Davis, 1985). Activities of daily living related to mobility, bathing, dressing, and eating are more commonly impaired than activities related to communication, bowel continence, and orientation (Teresi, Cross, & Golden, 1989). Most of these results are somewhat obvious to both lay and professional observers; however, it should be appreciated that CTT cannot make such distinctions.

IRT-based θ values are meant to estimate constructs, not total scores. Therefore, two or more tests of the same construct can be calibrated on the same scale of measurement irrespective of each scale’s content. For example, a clinician may wonder how the Beck Depression Inventory is related to the MMPI-2 Depression scale. Within a CTT framework, these comparisons are complicated by the nonidependence of person and item parameters. The CTT-based notion of item difficulty (proportion of examinees who pass the item within a sample) must diverge for the same item if administered to distinct groups of examinees with non-equivalent average abilities. Item difficulty is population specific. In addition, item reliability is typically evaluated based on each item's correlation with all other items on a test. Item reliability is also test specific. Item parameters in IRT have the same meaning regardless of the population being assessed or the test being evaluated.

Test and item equating is simplified by only requiring that item parameters and person parameters for two or more tests of the same construct are calibrated on the same scale of measurement. Item calibration of this sort can be methodologically and computationally complex (see Stocking & Lord, 1983), but quite valuable when successful. For example, by equating distinct tests of depression (e.g., Carmody et al., 2006; Orlando, Sherbourne, & Thissen 2000) and general health (e.g., Martin et al., 2007) researchers have been able to simplify the diagnostic process while improving accuracy. Such gains occur because the information provided by two or more distinct measures of the same construct are additive. Thus, instead of having two related but psychometrically distinct measures (as with CTT), IRT allows clinicians to combine information into a single estimate.

Differential item functioning. Item parameters derived from unique populations cannot always be calibrated on the same scale of measurement. This occurs when item parameters are dependent on a specific population, not because of failure to model parameters separately, but because of group biases. Non-biased items are those for which the probabilities of examinees from different populations passing them are equal when the examinees' θ values are equal. Mathematically, the following equality must hold:

$$P(X=1|\theta,C)=P(X=1|\theta)\,, \tag{8}$$

where C denotes membership within a given population or class of individuals. Thus, the probability of an item response is dependent on the person parameter but not on the population. Items that do not maintain the above property are said to display *differential item functioning* (DIF); items that do are said to display *measurement invariance*. Figure 10 presents two hypothetical item characteristic curves for the same item estimated within a male sample and again within a female sample. Clearly, the item's parameters are dependent on the population of examinees; the item is less difficult for males than for females. Women and men of the same ability find the item unequally difficult. CTT makes discovering this bias more complicated when legitimate group differences in ability are confounded with item parameters.

Examinations of DIF have become common in multicultural settings. For example, Hulin, Drasgow, and Komocar (1982) found DIF for items in the English and Spanish versions of a job satisfaction measure due to discrepancies in translation. DIF has been evaluated across age (Balsis, Gleason, Woods, & Oltmanns, 2007; Kim, Pilkonis, Frank, Thase, & Reynolds, 2002), gender (e.g., Bejar, 1977; Carle, Millsap, & Cole, 2008; Donovan & Drasgow, 1999; Jane, Oltmanns, South, & Turkheimer, 2007; McRae, 1991; Reise, Smith, & Furr, 2001; Santor, Ramsay, & Zuroff, 1994; Smith & Reise, 1998), language (e.g., Hulin, Drasgow, & Komocar, 1982; Orlando & Marshall, 2002), patient population (Bedi, Maraun, & Chrisjohn, 2001), race, ethnicity, and culture (e.g., Bolt, Hare, & Neumann, 2007; Bontempo, 1993; Cooke & Michie, 1999; Ellis, Becker, & Kimmel, 1993; Hui, Drasgow, & Chang, 1983; Liu & Zhang, 2006; Williams, Turkheimer, Schmidt, & Oltmanns, 2005), and testing medium (Chuah, Drasgow, & Roberts, 2006). And, IRT has the additional advantage of examining in great detail potential bias in the predication of outcomes (e.g., Leung & Drasgow, 1986).

Longitudinal research. Differential item functioning can also be found in the context of longitudinal within group comparisons. Researchers who study change (e.g., due to time or treatment) must also be cautious of differential item functioning (Horn & McArdle, 1992). Changes in θ estimates ought to be caused by changes in examinees' true θ values rather than changes in item parameters—lack of metric invariance—or changes in the structure of constructs relating to the items—lack of configural invariance. IRT and confirmatory factor analysis are both valuable in the investigation of measurement invariance. However, confirmatory factor analysis is better suited for the assessment of configural invariance and IRT is better suited for the assessment of metric invariance (Meade, Lautenschlager, & Hecht, 2005). As mentioned earlier, item parameters in IRT are directly related to item parameters in factor analysis. But, IRT presents a more explicit item parameterization, and thus better facilitates examinations of metric invariance.

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

Model and person fit. One does not typically discuss model fit in the context of CTT. Indeed, as mentioned multiple times thus far, CTT does not specify the relation between true scores and θ values. Yet the study of reliability in CTT does involve rarely tested assumptions about true scores (i.e., strict parallelism, parallelism, or tau-equivalence). McDonald (1999) has demonstrated how such assumptions can be thought of as special cases of the Spearman single-factor model, an assumption that all items measure a common symptom. Thus, model fit should be evaluated for CTT, even if it is not standard practice. Researchers employing IRT models, on the other hand, commonly test model assumption using global and local fit indices. For example, researchers have questioned whether unidimensional, monotone IRT models are appropriate for clinical constructs. It seems likely, they argue, that item characteristic curves for achievement questions (e.g., “If Billy has five apples, how many will he have left if he gives three to Tommy?”) will look different than item characteristic curves for personality questions (e.g., “I am as sociable as most people?”). If the manner of response is different, the normal ogive and logistic functions may not accurately model observed data. As mentioned earlier, a variety of IRT models have been developed to account for such situations.

Reise and Waller’s (1990) successful fit of the 2PL model to personality data is commonly cited as evidence for the applicability of the logistic function. Later analyses suggested that the 3PL model provides only negligible improvement in fit over the 2PL model (Reise & Waller, 2003). But, some have found only limited success with logistic models. Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) could not successfully fit a 2PL model, a 3PL model, or a GRM to two separate measures of personality. The authors suggest that the misfit was likely due to two possible sources: (1) uncounted for multidimensionality; or (2) the inability of the logistic function to correctly model examinees’ responses to non-cognitive items. Stark, Chernyshenko, Drasgow, and Williams (2006) point out that the monotonic 2PL model explicitly assumes *dominance* (i.e., Likert’s model) as opposed to *ideal response* (i.e., Thurstone’s model). Dominance models assume that as θ increases the probability of a response must also increase. Ideal response point models assume that as θ increases, an ideal response option (i.e., for which the person has the highest probability of responding) is approached. Once that ideal point has passed, the probability of responding decreases (i.e., a peaked ICC). Stark et al. argue that the unfolding model (an ideal response point model) is superior to the 2PL model because it can accommodate all types of data.

An unavoidable limitation of all assessment instruments is the potential for erroneous diagnostic outcomes. However, while most psychologists are willing to accept that true scores or θ estimates are accurate only within the limits of standard error, a more concerning situation arises when the relation between item parameters and person parameters are intentional or unintentionally distorted for particular examinees. That is, there are some individuals for whom the IRT model may not fit. In CTT, the only method available for detecting such aberrant results is to identify total scores falling in the extremes of the normative distribution. For example, if examinees endorse too many infrequently endorsed items on the MMPI-2, it is recommended that their test scores should not be interpreted. Using IRT, a researcher aiming to assess such distortion can employ two possible methods: (1) look for extreme θ values (similar to CTT methodology but with IRT sophistication; e.g., Zickar, Gibby, & Robie, 2004); or (2) look for improbable response patterns.

Analyses of *person fit* serve to identify examinees for whom the response model does not accurately predict their performance. Specifically, the strategy is like an analysis of DIF at the person level. Take for example a measure that fits a Rasch model. For such a test, high difficulty items should be endorsed less often than low difficulty items. The likelihood of an examinee endorsing high difficulty items but not low difficulty items is highly improbably given the item parameters. Deviant item response patterns suggest that the test is not accurately estimating examinees' θ values. Possible explanations for deviant response patterns include poor person-model fit, poor effort, or cheating/misrepresentation. Drasgow, Levine, and Williams (1985) developed a *z*-score index for maximum likelihood estimates (*lz*) that can be used to determine how deviant a response pattern is in comparison to an assumed normal distribution of response patterns. The maximum likelihood of θ estimates for examinees may differ even when each receives the same θ estimate. Reise and Waller (1993) applied the *lz* person fit statistic to items from a personality measure and concluded that it does have potential for identifying aberrant response styles, but was too unreliable to provide meaningful moderation of other variables. Reise and Due (1991) warn that the characteristics of item parameters (e.g., item difficulty) and characteristics of the entire measure (e.g., length) can affect the validity of the *lz* statistic. In addition, Reise (1995) expressed concerns about the distribution assumption and power of *lz*. The use of *person characteristic curves*, an alternative approach to examining person fit, can be used in conjunction with *lz* to help explore the causes of aberrant responses (Nering & Meijer, 1998).

The accuracy of such person fit statistics with actual data is mixed. Zickar and Drasgow (1996) used person fit algorithms to assess misrepresentation on personality tests and found only limited success. Ferrando and Chico (2001) compared IRT person fit analyses of misrepresentation to traditional measures of misrepresentation (i.e., response bias scales) and found the IRT person fit approach to be less accurate. Perhaps the limitation with the approach is that it can only be used to identify sequentially improbably response options. If an examinee responds to a measure in an unusually exaggerated manner, but endorses items in the correct sequence, the person fit statistics will not identify the response style as aberrant. In the extreme example, the response pattern of an examinee who endorses 59 of 60 items from the *F* scale on the MMPI-2 would not be considered aberrant even though the CTT based T-score for such a pattern would literally be off the page. Nonetheless, whether indicative of misrepresentation or not, aberrant response patterns do lead to poor classifications of examinees (Frost & Orban, 1990; Hendrawan, Glas, & Meijer, 2005), and thus the identification of such patterns is warranted.

Modern Psychometric Theory and Test Validity

Decades of research have established that modern psychometric theory—inclusive of both common factor theory and IRT—is the preeminent statistical framework for reducing measurement error in psychological and educational testing. Modern psychometric theory's impact on measurement validity, however, is less clear. The natural assumption that IRT improves test validity by improving test reliability may be premature. Reliability is a necessary but not sufficient condition for what is classically considered to compose construct validity: The extent to which a test measures the subject material it purports to measure, predicts outcomes, converges with measures of similar constructs, and diverges from measures of dissimilar constructs (Rosenthal & Rosnow, 1991; Strauss & Smith, 2009). Strictly speaking, IRT models do not fully address such forms of test validity. Latent variables are mathematical entities, perhaps nothing more than vectors within a coordinate system. The extent to which a mathematical variable serves as a marker of disease or represents a meaningful psychological construct must be established through empirical investigation. Some accounts refer to the validity of a test with respect to internal structure (e.g., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999); however, this alone cannot validate a test instrument.

Traditional treatment of measurement validity. Historically, validity has been at the forefront of clinical measurement; the majority of self-report inventories were developed using rational-theoretical and empirical methodologies (Lanyon & Goodstein, 1997). That is, after creating items in accordance with logic and theory (i.e., content validity), scales were refined by selecting only those items that accurately distinguished between groups with and without clinical conditions (i.e., criterion validity). The Beck Depression Inventory, California Psychological Inventory, Millon Clinical Multiaxial

Inventory (Millon, 1983), and MMPI were all influenced by such approaches. Measurement error was examined using typical CTT-based analyses of reliability (e.g., Guttman-Cronbach alpha coefficient); however, reduction of measurement error clearly took a secondary role in test development. Clinical psychologists deemphasized the reasons for item endorsement, instead focusing on outcomes. Despite good intent—a blatant emphasis on criterion validity sometimes referred to as "blind" or "dust bowl" empiricism—it was soon realized that neglecting measurement structure had detrimental consequences for practical use of measures. On the MMPI, for example, item and scale overlap contributed to poor discrimination between clinical criterion groups (e.g., depressed patients vs. anxious patients), rendering individual scale interpretations inadequate for differential diagnosis (e.g., Rubin, 1948). Scales on the Millon Clinical Multiaxial Inventory overlapped by as much as 60%, sharing more than 90% of variance (Wetzler, 1990).

Further complicating matters, empirical data suggested that self-report inventories did not coincide with the discrete categories of psychiatric nosology (i.e., the "neo-Kraepelinian" nomenclature). That is, both items and scales were not disorder specific. Depressed patients did not solely endorse depression items and anxious patients did not solely endorse anxiety items. Researchers soon realized, however, that configural patterns of clinical scale elevations were more often associated with unique disorders. For example, a number of MMPI "cookbooks" were published reporting clinical disorders most commonly associated with code-types (e.g., Lachar, 1974). Interpretations based on multiple scales can provide greater discrimination than interpretations based on individual scales (see Meehl, 1950), but the practice highlights an apparent contradiction between the simplistic unidimensional models commonly employed in the evaluation of measurement error and the more complex multidimensional interpretations given to patients' responses. Not surprising, configural patterns are inherently unreliable when evaluated by the CTT model (e.g., Vincent, 1990).

Clinicians do not often question the validity of tests and scoring procedures developed using methodologies based on content and criterion validity. These simple and direct approaches to test development are, if nothing else, conveniently intertwined with observed psychiatric disorders. Nonetheless, unacceptably large measurement error, poor discrimination between criterion groups, and the atheoretical nature of configurable pattern interpretation have long jeopardized the scientific footing of clinical assessment. Strauss and Smith (2009) reviewed construct validity in psychological measurement, concluding that the practice of combining distinct facets of psychological distress into unitary scales unnecessarily confounds domains of psychopathology. Emphasizing test validity at the expense of test reliability has proven to be an untenable option.

Modern treatment of measurement validity. As psychometric theory progressed, researchers turned to quantitative tools based on common factor theory to create new tests, refine existing scales, or simply to confirm theorized internal structure (see Comrey, 1988; Floyd & Widaman, 1995). While applications of IRT to clinical test development and evaluation have been less prevalent, equivalence of the two models implies mutual utility in clinical domains. By combining rational-theoretical approaches to test development with modern statistical approaches, test developers began to embrace modern mathematics while accommodating some clinical theory. Specifically, test developers focused on the reduction of measurement error through the creation of unidimensional scales. Unitary total scores originating from unidimensional scales are more reliable than those produced from multidimensional scales (McDonald, 1999). Ostensibly, this movement seems aligned with prevailing conceptualizations of psychopathology in the DSM: distinct disorders. Thus, it might be expected that unidimensional test construction would increase both the reliability and the validity of clinical tests.

Yet some clinical researchers and practitioners remain indifferent and/or skeptical of modern technologies. Some test developers, for instance, have been hesitant to alter existing test structure due to fears that major revisions will drastically change test properties and render preexisting research obsolete (e.g., Hare, 2003; see Knowles & Condon, 2000; Silverstein & Nelson, 2000). For this reason, only a select few have been willing to make dramatic changes to popular tests. Moreover, some authors have criticized unidimensional scale construction as being limited in clinical domains (e.g., Fava, Ruini, & Rafanelli, 2004). Ben-Porath and Tellegen’s (2008) repackaging of items on the MMPI-2 Restructured Form (MMPI-2-RF) with the aid of both CTT and factor analytic techniques, for example, has been the target of significant criticism. Nichols (2006) warned that such factor analytic refinements caused MMPI scales to drift far from clinically relevant constructs. Caldwell (2006) argued that the scales may represent nothing more than narrow or content specific areas of illness. Gordon (2006) criticized the "assumption that psychopathology can be reduced to the pure and distinct atoms of personality traits" (p. 870). And, early evidence suggests that such concerns may be warranted (Binford & Liljequist, 2008).

Despite obvious improvements in test reliability, unidimensional factor analyses may not lead to overwhelming improvements in test validity. To appreciate this phenomenon, recall the apparent paradox within CTT whereby improved reliability can decrease validity. Incorporating diverse symptoms into scales will improve diagnostic accuracy by sampling a wider domain of predictors. This is particularly true when the target of assessment is a multidimensional variable. Yet, the reliability or information provided by a unidimensional scale is, in contrast, improved by amplifying shared variance among items (i.e., reducing diversity). This simple fact should not be taken for granted: unidimensional scales—through the very process of improving their reliability—cannot provide optimal prediction of multidimensional constructs. If psychiatric disorders are unidimensional in nature, we should expect unidimensional scales to provide strong construct validity. If psychological disorders are multidimensional in nature, we should not.

The latent structure of self-reported psychopathology. Beneath the conceptual veneer of diagnostic categories lies a persistent finding that complicates clinical assessment: many psychiatric disorders stem from multiple causes—common and unique (Krueger, 1999, 2002). It is not uncommon for patients to receive multiple diagnoses. Most prominently, the comorbidity of mood and anxiety disorders (Maser & Cloninger, 1990; Merikangas et al., 1996; Mineka, Watson, & Clark, 1998), as well as conduct, personality, and substance disorders (Armstrong & Costello, 2002; Waldman & Slutske, 2000) has been well documented. A brief glance through the DSM reveals that most disorders share common symptoms. For example, impulsivity is a symptom of Bipolar Disorder and Attention Deficient/Hyperactivity Disorder, memory impairment is a symptom of Major Depressive Disorder and Dementia, and disinterest in social interaction is a symptom of Autistic Disorder and Schizoid Personality Disorder.

Such blurring of diagnostic categories and symptoms has led some to suggest the existence of underlying continuums uniting many psychological disorders (e.g., Angst, & Dobler-Mikola, 1985; Cassano, Michelini, Shear, & Coli, 1997). Krueger and Finger (2001), for instance, found that the DSM's diagnostic categories of Major Depressive Episode, Dysthymia, Simple Phobia, Agoraphobia, Social Phobia, Panic Disorder, and Generalized Anxiety Disorder could all be located on a continuum of shared variance—an underlying variable the authors labeled internalizing. This construct was later found to be positively correlated with patients' social functioning, missed days of work, and number of lifetime psychiatric hospitalizations (McGlinchey & Zimmerman, 2007). Indeed, researchers have long been aware of telltale signs of underlying multidimensional structure in clinical data (e.g., Derogatis, Klerman, & Lipman, 1972). Specifically, both general and domain specific variables appear to play prominent roles in self-reported symptoms of psychological distress. Empirical confirmations of such structure have grown common (e.g., Brown, Chorpita, & Barlow, 1998; Cassano et al., 2009; Michie & Cooke, 2006; Tackett, Quilty, Sellbom, Rector, & Bagby, 2008; Waller, 1999).

These findings will come as no surprise to clinical researchers and practitioners, for whom ignoring multidimensional structure might be considered professionally incompetent. For example, in the assessment of cognitive ability, the Wechsler Adult Intelligence Scale (WAIS-IV; Psychological Corporation, 2008) is structured to assess four domain specific abilities (verbal comprehension, perceptual reasoning, working memory, and processing speed) in addition to a general ability (full-scale intelligence quotient). Multiple dimensions of interpersonal behavior have long been acknowledged in the assessment of personality (Leary, 1957). And, veteran clinicians will recall that early editions of the DSM hierarchically categorized mental illnesses into the psychodynamic-based concepts of “Psychoses” and “Neuroses” (American Psychiatric Association, 1968; for a history of such terms see Beer, 1996). It is difficult to find any well-known clinical test that does not contain embedded higher-order constructs. The Child Behavior Checklist (Achenbach & Edelbrock, 1983), MMPI, Personality Assessment Inventory (Morey, 1991), and Personality Inventory for Children (Lachar & Gruber, 2001) all make use of general and domain specific traits.

In conceptual models of psychopathology, three general categories of distress are often proposed: (1) internalizing or neurotic distress – pervasive worry, melancholy, and general emotional difficulties; (2) thought or psychotic distress – a departure from reality characterized by delusions, hallucinations, or poor cognitive control; and (3) externalizing or character distress – aggressiveness, substance abuse, and immoral behavior. Watson (2005) has proposed a hierarchical model of emotional distress, integrating common and specific components of mania, depression, and anxiety under the umbrella of internalizing disorders. Krueger, Markon, Patrick, Benning, and Kramer (2007) describe a model in which substance use and aggressive behavior are thought to represented lower-order traits, and general externalizing is thought to represent a higher-order trait. There are, of course, seemingly infinite additional lower- and higher-order conceptual tiers that could be recognized (e.g., psychological vs. physiological distress). From a modeling perspective, Box and Draper (1987) notably argued that, "…all models are wrong, but some are useful" (p. 424); model selection is a process of weighing the relative benefits and limitations of coherence with reality, statistical and conceptual parsimony, and confirmation of existing theory.

The true latent structure of the human psyche notwithstanding, there has long been a consensus among researchers and practitioners that multidimensional structure is intrinsic to examinees' self-reported psychopathology. This explains the diagnostic value of configural patterns: analyzing multiple variables in a multidimensional system produces greater validity than interpreting individual scales. Unidimensional models, whether conceptualized under traditional or modern measurement theory, do not appear to be consistent with the structure of self-reported psychological distress. Unidimensional scales should not be expected to accurately assess psychopathology beyond homogenous symptoms (i.e., content specific distress). Nor should the value of multidimensional measurement be dismissed without empirical investigation. Emphasizing test reliability at the expense of test validity has also proven to be an untenable option.

Towards a more valid measurement model. Progress in the development and refinement of clinical scales has been limited by poor integration of measurement reliability and validity. Namely, overt focus on validity has sacrificed test reliability, while overt focus on reliability has sacrificed test validity. This paper examines an integrated application of reliability and validity in the development, evaluation, and use of clinical measures. That is, purposeful as opposed to haphazard use of traditional and modern psychometric techniques.

The common first step in test development is a thorough consideration of relevant content. Indeed, it has been noted that content validity is a necessary condition for measurement validity (Lanyon & Goodstein, 1997). The researcher’s task, in this instance, is to define and sample the relevant universe of content; that is, to select a representative subset of items from the domain(s) the researcher is interested in measuring. A review of empirical, clinical, and theoretical literature is in order. It is at this point that a researcher can take advantage of previous psychometric analyses by selecting items with desirable parameters. If, for example, the intent is to measure depression, previous analyses of related inventories can be used to select items with appropriate characteristics.

The second step in test development typically involves evaluation of item properties, test score reliability, and internal structure. It is at this point where the critical error of reducing complex psychological syndromes to narrow symptom clusters can be made. Researchers may sacrifice items that could have otherwise improved diagnostic accuracy in order to achieve scale homogeneity. To understand why, it is important to appreciate that both unique and common variance components contribute to item variability. Unique variance is due to both idiosyncratic characteristics of items and unpredictable measurement error. On a unidimensional test, common variance is due to a single latent factor. When tests are truly multidimensional, however, additional sources of common variance contribute to scale and item variability (see Rindskopf & Rose, 1988). This unaccounted for common variance, a violation of local independence, can bias parameter estimates (see DeMars, 2006). To avoid such bias, multidimensional items are often removed from scales during test development. Unfortunately, this strategy will reduce test validity in instances where the construct being assessed is truly multidimensional.

Mislevy, Levy, Kroopnick, and Rutstein (2008) note that the true value of modern psychometric theory lies in the ability to communicate increasingly complex psychological narratives. The ability to precisely partition item variance into multiple latent factors represents a clear advantage over previous methodology. The tools of multidimensional common factor theory and IRT allow for the creation of reliable multidimensional tests. While it cannot be denied that multidimensional models will complicate clinical assessment, it is clear that such complexity is not without purpose. Test developers' reluctance to embrace alterative structures in clinical data may explain why some clinical psychologists have recoiled from modern psychometric theory. In order to move beyond the measurement of symptoms but not syndromes, we must learn to integrate clusters of items and scales into clinically relevant disorders. That is, we must be willing to explore these complex psychological narratives.

Researchers from diverse fields have demonstrated great potential in multidimensional modeling. de la Torre and Patz (2005), for example, derived more precise θ estimates by allowing correlated educational variables (i.e., mathematics, spelling, and social studies skills) to work together in multidimensional frameworks. Wang, Chen, and Cheng (2004) increased the reliability of θ estimates by allowing education and personality factors to correlate through multidimensional IRT. And, in a simulation study, Yao and Boughton

(2007) demonstrated improved classification accuracy (i.e., false negatives and false positives for discrete levels of simulated mathematics proficiency) with multidimensional IRT models.

The tools for multidimensional modeling of clinical constructs are at the field's disposal. Evidence reviewed here suggests that the validity of clinical measures can be improved though the use of multidimensional models. A logical method for choosing an appropriate framework would be to include criterion-related validity variables in the evaluation of tests; in essence, a type of empirical keying for model selection. This can be accomplished most readily through the evaluation of structural relations between latent constructs and measured validity variables. That is, validity variables can be treated much like scale items. A model's ability to account for variance in validity variables, be it unidimensional or multidimensional, should be given significant weight in the selection of a measures' internal structure. By allowing validity variables to influence this selection process, the apparent disconnect between modern psychometric theory and clinical practice can be diminished.

Once an appropriate and valid model is selected, IRT's well-established tools for analyzing items and measurement instruments, over and above traditional models, need only be applied to the data. Model derived θ estimates allow for comprehensive analysis and reduction of measurement error, the creation of computer adaptive tests, meaningful scaling of latent variables, objective calibration and equating, evaluation of test and item bias, greater accuracy in the assessment of change due to therapeutic intervention, and the evaluation of model and person fit. As reviewed previously, IRT provides clear benefits over and above current practice.

An additional benefit of IRT comes into effect when estimating patients' disease status; total scores can give way to more accurate IRT-based θ estimates. Figure 11 demonstrates the relation between θ values and total scores through the use of a *test characteristic curve*: predicted total scores plotted against a range of θ values. As in the example, predicted total scores and θ values are generally related by a monotonically increasing function (Baker, 2001). Unfortunately, the association is nonlinear. As can be seen in Figure 11, total scores produce very imprecise θ estimates at the extremes of the distribution (see Dumenci & Achenbach, 2008)—troubling information for clinicians, who regularly assess and treat abnormal behavior. Mathematically, it can be demonstrated that total scores are inefficient estimates of θ values (McDonald, 1999). In essence, weighted item sums (i.e., IRT-based θ estimates) are more accurate than unweighted item sums (i.e., total scores); total scores cannot be more reliable or more valid than θ estimates when latent variable models accurately reflect observed data.

Total score efficiency is directly related to item parameters. As discrimination parameters diverge, total scores become less efficient estimators. Fortunately, these parameters do not tend to vary a great deal for existing unidimensional scales, as items were typically chosen for their high factor loadings or biserial correlations. Therefore, item weighting should not be expected to significantly improve diagnostic accuracy with respect to scoring patients' responses. Researchers have long noted that the choice between weighted and unweighted item sums makes little practical difference in psychological assessment (e.g., Aiken, 1966; Guilford, Lovell, & Williams, 1942; Potthoff & Barnett, 1932; Retzlaff, Sheehan, & Lorr, 1990). However, great divergence between weighted and unweighted item sums is not required for them to produce different validity coefficients (McCornack, 1956). And, it would be quite unlikely to find equivalent item discrimination parameters when items are allowed to load onto multiple variables. To the extent that such variability in the relations between items and θ values increases, the efficiency of total scores decreases. By using appropriate model parameters to estimate patients' θ values, researchers and clinicians can avoid unnecessary and limited assumptions of the relation between total scores and patients' θ values.

An Example: The Brief Symptom Inventory

The intent of this analysis is to demonstrate improved diagnostic accuracy through purposeful integration of test reliability and validity using modern psychometric techniques. To accomplish this task, analyses will be conducted on a data set containing item level responses to a general measure of psychopathology and clinically relevant criterion-related validity variables.

Outpatient mental health clinics routinely utilize patient self-report screening measures to assess for symptoms of psychological distress. The Brief Symptom Inventory (BSI; Derogatis, 1993) was specifically designed to assess the “psychological symptom status of psychiatric and medical patients…” (Derogatis & Melisaratos, 1983, p. 596). The test consists of 9 primary symptom dimensions: (1) Somatization (SOM) – psychological distress manifested through bodily dysfunction; (2) Obsessive-Compulsive (O-C) – unwanted, irresistible impulses; (3) Interpersonal Sensitivity (I-S) – feelings of inadequacy and inferiority; (4) Depression (DEP) – depressed affect, hopelessness, and withdrawn behavior; (5) Anxiety (ANX) – persistent nervousness, fear, and panic; (6) Hostility (HOS) – aggressive and irritable thoughts, feelings, and actions; (7) Phobic Anxiety (PHOB) – fear of specific objects or general situations; (8) Paranoid Ideation (PAR) – suspicion, distrust, and hostility directed towards others; (9) Psychoticism (PSY) – alienated life style and disturbed thought. The BSI is a condensed version of the Symptom Checklist–90 (SCL-90; Derogatis, & Cleary, 1977) which itself has an extensive lineage (Derogatis, Lipman, & Covi, 1973). The Symptom Checklist–90 was developed using the rational-empirical tradition along with factor analytic techniques to identify primary symptom dimensions. Scale interpretations are based on percentile rankings of total scores in comparison to normative samples of psychiatric inpatients or outpatients.

Researchers have generally been able to recover most of the 9 primary symptom dimensions of the BSI through unidimensional factor models (Derogatis & Melisaratos, 1983; Hayes, 1997; Heinrich & Tate, 1996; Kellett, Beail, Newman, & Hawes, 2004) and IRT models (Long, Harring, Brekke, Test, & Greenberg, 2007), though secondary evidence suggests variation in the number of factors and patterns of loadings (e.g., Holcomb, Adams, & Ponder, 1983). In addition, sizeable correlations between scales' total scores (Boulet & Boss, 1991), and between factors (Hayes, 1997), have led some to question the proposed dimensionality of the test. Some researchers, for example, have argued that covariance on the BSI can be attributed to just one dominant variable (Cyr, McKenna-Foley, & Peacock, 1985; Loutsiou-Ladd, Panayiotou, & Kokkinos, 2008). Such inconsistencies might better be explained by multidimensional latent structure. Finding such structure on the BSI would be consistent with empirical research on related measures of self-reported psychopathology.

There are numerous models that might reasonably account for non-unidimensional latent structure on the BSI. From a predictive standpoint, it can be argued that complexity in such models is beneficial. All things being equal, accounting for additional relevant sources of variance should increase diagnostic accuracy. However, such complexity comes at a price. Estimation of complex models, as well as meaningfully interpretations of their parameters, becomes increasingly difficult. It is commonly advised that structural models should be informed by prevailing theory. Evidence reviewed previously suggests that clinical measures likely maintain three sources of multidimensionality: (1) disorders may be correlated; (2) symptoms of distress may load onto multiple disorders; and (3) self-reported psychopathology may maintain a bifactor structure. It seems plausible that all such sources of multidimensionality are present in clinical data. Unfortunately, a model becomes increasingly saturated as additional layers of complexity are added. One must search for a balance between complexity and parsimony. That is, some restrictions must be made in order to help organize the data―the very purpose of modeling.

Of the multidimensional models considered in this study, the bifactor model seems most aligned with prevailing theories of psychopathology (e.g., Simms, Grös, Watson, & O’Hara, 2008). Applying a bifactor model to the BSI would suggest that each item is related to a general factor (e.g., internalizing) and to a domain specific factor (e.g., depression). In Figure 7, for example, an examinee's response to item 1 (symptom 1) is a function of their severity of depression and their severity of internalizing. Such a model preserves precision in measurement at the level of symptom clusters while accounting for meaningful dimensionality at the level of syndromes.

Aims of the Present Study

Modern psychometric theory will be used to study the construct validity of distinct modeling frameworks for the BSI—a general measure of psychological distress. It is hypothesized that all multidimensional models will provide better fit for the measure than will a unidimensional model. Specifically, models that allow for crossloadings and/or correlated latent variables will provide the most accurate representations of the BSI’s internal structure. Variance accounted for in criterion-related variables will be compared for the multidimensional versus unidimensional frameworks. It is hypothesized that multidimensional models will explain more variance in criterion-related variables than will the unidimensional model. These analyses will be used to select a single model that best accounts for the BSI’s internal structure while maintaining strong criterion-related validity. This model will then be examined more thoroughly using IRT techniques.

Finally, the diagnostic accuracy of the chosen latent variable model will be compared to the diagnostic accuracy of the unidimensional total score model by comparing sensitivity and specificity estimates for each. It is hypothesized that the chosen latent variable model will provide better diagnostic accuracy than the unidimensional total score model. Such findings would support the combined use of modern and traditional psychometric techniques in the development and refinement of psychological measures.

Chapter 2

Method

Participants

Archival data from 688 outpatients seeking psychological counseling (68%) or assessment (32%) between 1999 and 2009 at an Arizona State University mental health clinic were utilized in this study. The sample included students from the university and persons from the community. Table 1 provides demographic information for the sample.

Item Data

Patients’ item level responses to the Brief Symptom Inventory (BSI; Derogatis, 1993) provided the basis for the IRT analyses. The 53-item BSI instructs respondents to rate how much they have been distressed by each symptom within the past 7 days on a 5-point scale (0 = not at all, 1 = a little bit, 2 = moderately, 3 = quite a bit, 4 = extremely). All patients seeking counseling or treatment at the clinic were asked to complete the BSI as part of a routine intake procedure.

Validity Variables

Of the 688 participants in the sample, 515 (75%) were assigned formal diagnoses based on the criteria from the text revised fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; American Psychiatric Association, 2000). Doctoral students with 1 or more years of training in clinical psychology assigned DSM-IV-TR multi-axial diagnoses under the supervision of a licensed clinical psychologist. The DSM-IV-TR provides specific diagnostic criteria for empirically established psychiatric disorders through a multi-axial system including five levels of disorder and disability. Of primary interest in the current study were disorders classified under Axis I – principal or major disorders.

To facilitate data analyses, diagnoses were hierarchically collapsed into the following general categories of distress: (1) mood disorders – including depressive disorder (*n* = 139), bipolar disorder (*n* = 21), mood disorder not otherwise specified (*n* = 4), and mood disorder due to a general medical condition (*n* = 1); anxiety disorders – including phobias and panic disorders (*n* = 41), generalized anxiety disorder (*n* = 71), obsessive-compulsive disorder (*n* = 19), post-traumatic stress disorder (*n* = 15), and anxiety disorder not otherwise specified (*n* = 4); (3) somatoform disorders – including conversion disorder (*n* = 1), hypochondriasis (*n* = 1), body dismorphic disorder (*n* = 1), pain disorder (*n* = 7), and undifferentiated somatoform disorder (*n* = 2); and (4) other DSM-IV-TR disorders (e.g., attention-deficit/hyperactivity disorder, schizophrenia, etc.; *n* = 264). These three primarily categories of psychological distress were chosen due to their relative high frequency within the sample, along with the relative scarcity of any other category of psychological distress. Multiple DSM-IV-TR diagnoses can be given to a single patient; thus, it became beneficial to treat each general category of distress as a single dichotomous variable: presence or absence of a mood, anxiety, or somatoform disorder. One supplementary diagnostic variable—presence of just an anxiety disorder (*n* = 78) versus just a mood disorder (*n* = 112)—was created to examine the ability of the models to differentiate between clinical groups (somatoform disorders could not evaluated in this manner due to too few cases).

In addition, patients' self-reported primary concerns in seeking counseling or assessment were coded by undergraduate research assistants. The author trained three undergraduate raters to categorize patients' existing intake or assessment summaries working in pairs of two. As with DSM-IV-TR diagnoses, presenting concerns were hierarchically collapsed into three general categories of distress: mood, anxiety, and health concerns. Fifty files were re-coded to compute an estimate of inter-rater agreement; agreement between groups of raters was moderately high (κ = .68). Table 2 provides sample frequencies and percentages for each presenting concern and diagnostic category.

The validity of patients' self-reported primary reasons for seeking counseling and DSM-IV-TR diagnoses cannot be directly compared against a gold-standard for accuracy. Indeed, DSM-IV-TR diagnoses are themselves often considered to hold a gold-standard position in clinical research. To help clarify the nature of these criterion-related validity variables, patients' scale level responses to the second edition of the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, & Kaemmer, 2001) were examined for concurrent validity. The 567-item MMPI-2 is the most widely used measure of psychopathology in psychological assessment (results from MMPI-2 extended score reports were used to code validity and clinical scales' T-scores). Of the 688 participants in the study, 142 (21%) were administered the MMPI-2 based on treatment or diagnostic need.

Table 2 provides means and standard deviations for each scale within the sample. MMPI-2 configural profiles for patients presenting with a mood, anxiety, or health concern are presented in Figure 12. MMPI-2 configural profiles for patients diagnosed with a mood disorder, anxiety disorder, somatoform disorder, or no DSM-IV-TR disorder are presented in Figure 13. Unfortunately, just 15 patients presenting with a mood concern, 13 presenting with an anxiety concern, 3 presenting with a health concern, 38 diagnosed with a mood disorder, 34 diagnosed with a anxiety disorder, 3 diagnosed with a somatoform disorder, and

20 diagnosed with no DSM-IV-TR disorder also competed the MMPI-2. As such, the data presented in Figures 12 and 13 are greatly limited by sample size. Nevertheless, some generalizations can be made.

The profiles in Figure 12 reveal that patients presenting with a mood concern tended to have their highest elevations on the Depression scale (T = 69), the Psychasthenia scale (T = 68), and the Schizophrenia scale (T = 68). As can be seen in Figure 13, patients diagnosed with a mood disorder also had high elevations on the Depression, Psychasthenia, and Schizophrenia scales (T = 72 for each). Such profiles can be referred to as 2-7-8 code types, and are typically associated with neurotic traits and diagnoses of mood and/or anxiety disorders (Friedman, Lewak, Nichols, & Webb, 2001). Patients presenting with an anxiety concern (Figure 12) tended to have their highest elevations on the Psychasthenia scale (T = 68) and the Schizophrenia scale (T = 68). Patients diagnosed with an anxiety disorder (Figure 13) had their highest elevations on the Psychasthenia scale (T = 74), the Depression scale (T = 72), and the Schizophrenia scale (T = 69). Such profiles can be referred to as 2-7 or 2-7-8 code types, and are also typically associated with neurotic traits and diagnoses of mood and/or anxiety disorders (Friedman et al., 2001). Patients presenting with a health concern (Figure 12) tended to have no scale elevations. Patients diagnosed with a somatoform disorder (Figure 13) had their highest elevations on the Depression scale (T = 71), followed a number of lesser elevations on the Hysteria (Hy) scale (T = 67), the Psychopathic Deviate (Pd) scale (T = 67), the Paranoia (Pa) scale (T

= 68), the Hypochondriasis scale (T = 66), and the Psychasthenia scale (T = 66). Such profiles can be referred to as spike 2 code types (Friedman et al., 2001); however, the average profile for this group was poorly differentiated due to co-occurring elevations on multiple clinical scales.

Two primary conclusions can be gleaned from Figures 12 and 13. First, patients in the criterion groups (i.e., those presenting or diagnosed with some form of emotional distress) tended to endorse an elevated number of clinical symptoms on the MMPI-2. Second, the MMPI-2 configural patterns did not provide clear differentiation between clinical groups. This is not a surprising finding, as it has already been mentioned that clinical measures often perform poorly in differentiating between clinical groups. While the descriptive results presented here were likely influenced by small sample size, they also reflect the inherent ambiguity in clinical data. These results suggest that the MMPI-2 profiles of the diagnostic groups are generally consistent with previous findings; unfortunately, they do not confirm the validity of the criterion-related variables considered in the present analyses. As such, the reader is encouraged to consider these variables as fallible indicators of latent constructs; variables that are consistent with, although not diagnostic of, psychopathology. Random error in such variables, while unfortunate, can be modeled in latent variable analyses.

Data Analyses

Model comparisons. No consensus exists on the latent dimensionality of the BSI. As mentioned earlier, some researchers have found evidence supporting a factor structure similar to the heuristic unidimensional scales proposed by the BSI’s authors, while others have found evidence that just one dimension accounts for the majority of variance on all scales. In essence, it has been argued that orthogonal (uncorrelated factors) factorially-simple structure cannot account for observed item covariance on the measure. In the present study, it was hypothesized that several types of multidimensional models could account for these discrepancies: oblique (correlated factors) factorially-simple structure, oblique and orthogonal factorially-complex structures, or oblique and orthogonal bifactor structures. However, there remains significant ambiguity with respect to the patterns of loadings and number of factors within each model. The purpose of the present analysis was not to “prove” any one model of psychopathology, but rather to examine the structural relations between measurable constructs on the BSI and criterion-related validity variables related to the diagnosis of mood, anxiety, and somatoform disorders. Thus, it was beneficial to consider clusters of symptoms on the BSI primarily related to depression, anxiety, and somatization. Previous research on the dimensionality of the BSI suggests that the DEP scale primarily captures symptoms of depression, the ANX and PHOB scales primarily captured symptoms of anxiety (with each representing a subtype), and the SOM scale primarily captures symptoms of somatization.

All proposed orthogonal structures are depicted in Figure 14. The absence of double-headed arrows between the latent variables in Figure 14 conveys that covariance between these nodes was constrained to 0. All proposed oblique structures are depicted in Figure 15. The presence of double-headed arrows between the latent variables in Figure 15 conveys that covariance between these nodes was freely estimated. For the simple structure models, orthogonal (left panel of Figure 14) and oblique (left panel of Figure 15) structures were specified in which items from the DEP scale were allowed to load onto a single latent variable (depression), items from the ANX and PHOB scales were allowed to load onto a single latent variable (anxiety), and items from the SOM scale were allowed to load onto a single latent variable (somatization).

For the complex structure models, orthogonal (center panel of Figure 14) and oblique (center panel of Figure 15) structures were specified in which all but one item from each scale could load onto three possible latent variables (depression, anxiety, and somatization). One item from each scale was constrained for convergence; these items were chosen based on their strong conceptual association with three primary domains of interest (i.e., content validity).

For the bifactor structure models, orthogonal (right panel of Figure 14) and oblique (right panel of Figure 15) structures were specified in which items from all scales loaded onto a general latent variable—referred to hereafter as internalizing—and items from the DEP scale were additionally allowed to load onto a unique domain specific latent variable (depression), items from the ANX and PHOB scales were additionally allowed to load onto a unique domain specific latent variable (anxiety), and items from the SOM scale were additionally allowed to load onto a unique domain specific latent variable (somatization). For the oblique bifactor model, only the domain specific latent variables were allowed to correlate (as is necessary for convergence).

Confirmatory factor analyses were conducted to evaluate the proposed models. The metrics of the latent factors were defined by fixing their variances to unity. Correlations between measurement errors were all fixed at zero. The software package *Mplus* (Muthén & Muthén, 2006) was used to estimate all models. Two types of estimators are commonly used with ordered polytomous data in the program: (1) a robust limited-information weighted least squares estimator with a pairwise present approach for missing data (78 out of 16,512 item responses; 0.005%); and (2) a full-information marginal maximum likelihood estimator with robust standard errors using all available data under missing data theory (missing at random or missing completely at random). Limited-information estimators are not as efficient as full-information estimators, yet full-information estimators cannot be used to evaluate the fit of ordered polytomous data with a large number unobserved response patterns in the program (i.e., due to sparse contingency tables). However, the full-information estimator can be used for model comparisons. For these reasons, the limited-information estimator was used to evaluate model fit, while the full-information estimator was used to estimate model parameters and for comparisons of likelihood values.

Overall goodness-of-fit was evaluated using root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI). Hu and Bentler (1999) suggest that a RMSEA ≤ .06, a CFI ≥ .95, and a TLI ≥ .95 are indicative of acceptable fit. The Akaike information criterion (AIC; Akaike, 1987) and Bayesian information criterion (BIC; Schwarz, 1978) statistics—both decreasing as model fit improves—were used to compare models. In addition, the difference between the log-likelihoods for models—distributed as a chi-square (adjusted for Satorra-Bentler scaling) with $DF_1 – DF_2$ degrees of freedom—was used to determine if the more constrained simple structure models significantly worsened fit in comparison the less constrained bifactor and complex structure models.

These direct model comparisons are possible because the simple structure models are generally nested within both the complex and bifactor structure models (DeMars, 2006; Rindskopf & Rose, 1988). With respect to both orthogonal and oblique models, the simple structure model is a special case of the complex structure model in which all crossloadings have been constrained to zero. The simple structure model is also a special case of the bifactor model in which the item loadings onto a general factor have all been constrained to zero. It is also apparent that orthogonal models are nested within oblique models of similar structures. That is, the orthogonal simple structure model is nested within the oblique simple structure model, the orthogonal complex structure model is nested within the oblique complex structure model, and the orthogonal bifactor structure model is nested within the oblique bifactor structure model. This can be seen most readily by noting that the structures in Figure 14 can all be made more general by allowing for covariance between the factors (i.e., as is depicted in Figure 15). Unfortunately, the nesting of all possible model comparisons is not clear. For example, it is unclear whether orthogonal complex or bifactor structure models are necessarily more general than the oblique simple structure model. Therefore, comparisons of this sort were not pursued in the present study.

Validity estimates. Validity coefficients were estimated in a two-step process. First, models were re-estimated, but with their item parameters fixed to values from previous analyses (i.e., parameter estimates from the model comparisons). Second, criterion-related validity variables were added to each model, and the loadings of each onto the previously identified latent factors were freely estimated. That is, validity variables were treated as “items” (i.e., observed, fallible indicators of the latent constructs). It is often noted that latent variable models can be thought as regression models where the dependent variables are observed and the independent variables are latent. In this sense, the validity variables in the present analyses were regressed onto the latent factors. Doing so allowed for the estimation of communalities (i.e., variance accounted for) in patients’ self-reported presenting concerns and DSM-IV-TR diagnoses. Each loading (i.e., regression weight) was estimated within a separate analysis in order to prevent the variables from affecting one another.

In multiple regression analyses it is typically to either explore the influence of all predictors on an outcome, the influence of just the most impactful predictors on the outcome (i.e., through some iterative strategy), or the influence of a substantively meaningful subset of predictors on the outcome. In the present analysis, the third strategy was chosen to explore criterion-related validity. That is, the ability of the depression factors to explain mood concerns/diagnoses, the ability of the anxiety factors to explain anxiety concerns/diagnoses, and the ability of the somatization factors to explain health/somatoform concerns/diagnoses was examined. When differentiating between mood and anxiety disorder diagnoses, both the depression and anxiety factors were used to explain variance in the variable (i.e., the validity variable is conceptually related to both factors). In the bifactor models, both the domain specific factors (i.e., depression, anxiety, or somatization) as well as the general factor (i.e., internalizing) were used to explain variance in the validity variables.

IRT parameter estimates. Multidimensional IRT parameter estimates were generated under a graded response model in logistic metric for the model shown to demonstrate the greatest construct validity (i.e., acceptable internal structure and strong criterion-related validity). Parameter estimates were generated for the criterion-related validity variables by (again) treating these variables as “items”. IRT parameters were found by converting confirmatory factor analysis parameter estimates from *Mplus* into an IRT metric using the formulas given by Kamata and Bauer (2008). The authors show that discrimination (*a*) and loading (λ) parameters for item *j* can be made equivalent so that

$$a_j = \lambda_j \ , \tag{9}$$

by constraining the residual variance of the latent response variates to unity and by standardizing the common factor variance. Under these same scaling constraints, the authors show that

$$b_j = \frac{\nu_j}{a_j} \ . \tag{10}$$

That is, the IRT difficulty parameter (*b*) for item *j* is equal to the common factor threshold parameter (*v*) divided by the IRT discrimination parameter.

It is traditional in multidimensional IRT to present item parameters in the slope/intercept form. In the expression *a* (*θ* – *b*), multiplying through by *a* results in *aθ* – *ab*. From this, the item intercept is defined as *d* = – *ab* (note that ν in *Equation 10* has the same meaning as -*d*). Using this notation, a two-parameter multidimensional IRT model for dichotomous items is given as

$$P(X = 1 | \theta_{ik}, a_{jk}, d_j) = \frac{e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right) + d_j}}{1 + e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right) + d_j}} \ , \tag{11}$$

where a_{jk} is the discrimination parameter for the j^{th} item on the k^{th} latent variable, d_{j} is the intercept for the j^{th} item, and θ_{ik} is ability parameter for the i^{th} person on the k^{th} latent variable (Reckase, 2009). The formula is similar to that of a two-parameter logistic model with the exception that the probability of item endorsement is now dependent on multiple *θ* values. In the case of a bifactor model, only two of the a_{jk} parameters are allowed to be non-zero for the j^{th} item (Gibbons & Hedeker, 1992). Thus, items are associated with multiple discrimination parameters—one for each relevant latent variable. Generalizations to polytomous items are easily made (see Reckase, 2009). The multidimensional graded response model in logistic metric is given by

$$P(X = u | \theta_{ik}, a_{jk}, d_{jt}) = \frac{e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right)+d_{jt}}}{1+e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right)+d_{jt}}} - \frac{e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right)+d_{jt+1}}}{1+e^{\left(\sum_{k=1}^{m} a_{jk}\theta_{ik}\right)+d_{jt+1}}} , \tag{12}$$

where a_{jk} and θ_{ik} have been previously defined, and d_{jt} is the intercept for the t^{th} category of the j^{th} item. In the case of a bifactor model, again, only two of the a_{jk} parameters will be non-zero for the j^{th} item (Gibbons et al., 2007).

The multidimensional discrimination parameter (MDISC; A) is defined as

$$A_j = \sqrt{\sum_{k=1}^{m} a_{jk}{}^2} \ . \tag{13}$$

A is related to the overall discriminability of an item, and conveys the maximum slope of the item characteristic surface. The multidimensional difficulty parameter (MDIFF; B) is defined as

$$B_j = \frac{-d_j}{A_j} \ . \tag{14}$$

B is related the overall difficulty of an item, and conveys the multidimensional location of an item in the direction of maximum discrimination. Multiple B parameters exist in polytomous multidimensional IRT models—one for each transition between response categories.

The relations between latent variables and items in the model can be made explicit through item characteristic surfaces and equi-probably contour plots. Due to the compensatory nature of the bifactor IRT model (as given in *Equations 11 and 12*), the probability of endorsing an item can increase due to either domain specific or general latent variables. That is, an infinite combination of low and high θ values from each latent variable can result in the same probability of item endorsement. *Equi-probable contour plots* take advantage of the trigonometric relations between multidimensional IRT parameters to show the combinations of θ values that result in equal probabilities of outcome. Equi-probable contour plots were produced in the current analyses to explore the predictive value of the bifactor IRT model.

Diagnostic accuracy. Finally, the classification accuracy of latent variable estimates generated under the selected multidimensional model were compared to total scores (observed test data) with respect to predicting DSM-IV-TR diagnoses. Total scores for depression were computed by summing items from the DEP scale, total scores for anxiety were computed by summing items from the ANX and PHOB scales, and total scores for somatization were computed by summing items from the SOM scale. Estimates of latent factors for each participant were produced with the expected a posteriori method. These person parameters were then used to compute the expected probability of positive diagnoses—using *Equation 11*—for each participant on all DSM-IV-TR diagnoses.

Sensitivity and specificity estimates for the total score model were then compared to sensitivity and specificity estimates for the multidimensional model as summarized through receiver operating characteristic (ROC) curves. ROC curves plot sensitivity (x-axis) by the inverse of specificity (y-axis) in relation to increasing values of predictor variables (i.e., scores from the total score vs. bifactor models). Doing so allows researchers to visualize the relative balance between true positives and false positives for various cutoffs of a predictor. Area under the curve (AUC)—a one number summary equal to the probability that a patient with a positive diagnosis will be rated higher than a patient with a negative diagnosis—was used to summarize the results for each model. DSM-IV-TR diagnoses were used as the criteria for the ROC curve analyses (i.e., the predicted criteria); however, as mentioned previously, considering DSM-IV-TR diagnoses as criteria (i.e., gold-standards) is likely a flawed assumption. Thus, the ROC curves should be interpreted with caution. They are included in the present analysis only to give a sense of the practical benefits of each model.

Chapter 3

Results

Model Comparisons

Model fit results are presented in Table 3. The solution for the orthogonal complex structure model returned an improper parameter estimate when using robust weighted least squares estimation. However, the model successfully converged after constraining this parameter to its closest real value. All remaining models converged without difficulty. Absolute and relative fit statistics suggest that the orthogonal simple and complex structure solutions provided poor fit for the observed data. However, it should be noted that the results for the orthogonal complex structure model may be in error due to convergence problems. The oblique simple structure, oblique complex structure, oblique bifactor structure, and orthogonal bifactor structure models all provided marginal to good fit for the observed data.

The chi-square difference tests indicated significantly worse fit when comparing the orthogonal simple structure model to the oblique simple structure model (diff χ^2 (3) = 418.65, $p < .001$), the orthogonal complex structure model to the oblique complex structure model (diff χ^2 (3) = 18.76, $p < .001$), and the orthogonal bifactor structure model to the oblique bifactor structure model (diff χ^2 (2) = 158.61, $p < .001$). In addition, the orthogonal and oblique simple structure models provided significantly worse fit than the orthogonal and oblique complex structure models (diff χ^2 (42) = 728.17, $p < .001$ and diff χ^2 (42) = 169.09, $p < .001$ respectively) and the orthogonal and oblique bifactor structure models (diff χ^2 (25) = 219.78, $p < .001$ and diff χ^2 (24) = 144.48, $p < .001$ respectively). The AIC and BIC statistics give the same conclusions regarding model comparisons.

Factor loadings for the orthogonal and oblique simple structure models are presented in Table 4. The items in both models loaded highly onto their respective dimensions. It should be noted that the correlations between factors in the oblique model are very high, and in agreement with previous research on the BSI. It appears as though constraining the shared variance between these latent factors to zero in the orthogonal solution greatly contributed to the model's poor fit.

Factor loadings for the orthogonal and oblique complex structure models are presented in Table 5. The items in both models tended to load highly onto their respective dimensions (e.g., items from the DEP scale tended to load most highly onto the depression dimension). However, it is notable that items from the ANX scale loaded poorly onto the anxiety dimension in the oblique solution. The authors of the BSI note that the anxiety-related scales from the BSI appear to assess sub-domains of the more general anxiety construct. Thus, it seems plausible that the anxiety factor should be split into separate dimensions. However, attempts to allow items from the scale to load onto separate factors were unsuccessful. It is also noteworthy that despite allowing items to crossload, the three factors continued to correlate highly in the oblique structure solution (albeit less highly than in the oblique simple structure model). Crossloadings alone could not account for residual correlations between items in the orthogonal model (yet again, convergence problems made it difficult to evaluate fit in this instance).

Factor loadings for the orthogonal and oblique bifactor structure models are presented in Table 6. The items in both models loaded highly onto their respective domain specific factors as well as a common general factor. An exception to this can be found in the orthogonal model, where items from the ANX scale tended to load poorly onto the domain specific anxiety dimension. As with the oblique complex structure model, this can be taken as evidence that the anxiety factor comprises multiple dimensions. Yet, again, attempts to allow items from the scale to load onto separate factors were unsuccessful. In contrast to the orthogonal simple and orthogonal complex structure models, the orthogonal bifactor structure model provided acceptable fit for the data. This result can be better understood by examining the item loadings onto the general dimension. In the oblique bifactor model, the depression, anxiety, and somatization factors correlated highly with one another. However, loadings onto the general factor, in this solution, were modest. In the orthogonal bifactor model, where the depression, anxiety, and somatoform factors were not allowed to correlate, item loadings onto the general factor were much stronger. It appears as though local dependencies between items caused by correlated domain specific factors were effectively accounted for by the introduction of a general factor. That is, it became possible to form orthogonal dimensions by explaining item variance with respect to both domain specific and general factors.

The results of model fit suggest that the BSI does not comprise orthogonal factors with simple structure. Allowing factors to correlate (i.e., oblique solutions), whether through simple, complex, or bifactor structure solutions, improved model fit. Allowing items to crossload, either through complex or bifactor structure solutions, also generally improved model fit. Four multidimensional models were found to provide acceptable fit for the internal structure of the BSI: the oblique simple structure model, the oblique complex structure model, the oblique bifactor structure model, and the orthogonal bifactor structure model. Standard practice is to select the least complex, adequately fitting model of internal structure (i.e., the oblique simple structure solution). However, it has been asserted in the present paper that examining evidence related to criterion-related validity variables should be used to determine the overall value of a latent variable model. Therefore, the four models were next subjected to criterion-like validation in order to determine which model provided the greatest overall construct validity.

Validity Estimates

Factor loadings for the criterion-related validity variables onto the oblique simple, oblique complex, oblique bifactor, and orthogonal bifactor structure models are presented in Table 7. Estimated communalities (variance accounted for) in each validity variable are also presented in Table 7. For ease of interpretation, communality estimates for the four models are shown graphically in Figures 16 and 17. As can be seen in Table 7 and Figure 16, the models were generally comparable in their ability to account for variance in presenting concerns and DSM-IV-TR diagnoses. Notably, however, the orthogonal bifactor structure model outperformed all other models in the assessment of health concerns presentations and somatoform disorder diagnoses. While these results appear promising, it should be noted that the estimates were associated with relatively large standard errors. Thus, while a trend can be noted in the data, some skepticism is warranted.

Table 7 also reveals that the bifactor models (both oblique and orthogonal) outperformed the oblique simple and oblique complex structure models in differentiating between mood disorder diagnoses and anxiety disorder diagnoses. Figure 17 shows the relationship graphically. As can be seen, communalities were estimated in three separate runs for each model: first, using just the depression factor; second, using just the anxiety factor; and third, using both the depression and the anxiety factors. (Note that the internalizing factor was always used to differentiate between the two disorders in the bifactor models as well.) When differentiating between the two disorders using just the depression factor or just the anxiety factor, the oblique and orthogonal bifactor models demonstrated clear superiority in comparison to the oblique simple and oblique complex structure models. These findings appear to be the direct result of modeling common variance with a general (internalizing) factor.

These results can be better understood by noting that the oblique simple and oblique complex structure models performed as well as the bifactor structure models in differentiating between mood and anxiety disorders when the variable was allowed to load onto both the depression and the anxiety factors simultaneously (Figure 17). This likely occurred because the depression and anxiety factors in the oblique simple and complex structure models comprised amalgams of general and domain specific variance. Thus, the individual factors alone could not separate the three apparent sources of measurement variance on the BSI: that due specifically to depression, specifically to anxiety, and specifically to internalizing. However, a multiple regression-like weighted composite of both factors (i.e., similar to the current practice of using configural patterns of scales to differentiate between clinical groups) successfully teased apart the various sources of variance (i.e., similar to partial regression weights). The bifactor models—through their initial separation of depression specific, anxiety specific, and internalizing specific variance—successfully differentiated between disorders by incorporating the internalizing factor.

The oblique and orthogonal bifactor structures appear to have provided the greatest construct validity for the BSI. Both models successfully represented the internal structure of the measure, and both demonstrated superiority with respect to their relationships with criterion-related validity variables. However, in examining the validity variable loadings presented in Table 7, it becomes apparent that the orthogonal bifactor model provided clearer conceptual organization for the BSI (i.e., better content validity). In the oblique bifactor solution, it can be observed that the internalizing factor tended to overlap with the depression factor. Note, for example, that diagnoses of a mood disorder loaded more highly onto the internalizing factor than the depression factor in the oblique bifactor solution. In the orthogonal bifactor solution, conversely, the internalizing, depression, anxiety, and somatization factors appear to represent more distinct constructs.

IRT Parameter Estimates

The orthogonal bifactor model provided good fit for the internal structure of the BSI, demonstrated advantages with respect to predicting criterion-related validity variables, and made for relatively parsimonious interpretations of the latent constructs. Accordingly, the model was chosen for a subsequent multidimensional IRT analysis. Item parameter estimates from a bifactor graded response model in logistic metric are presented in Table 8. In some respects, the item parameters for the bifactor graded response model in Table 8 convey much of the same information as the factor loadings presented above. Indeed, there is a direct relation between the strength of item loading parameters and the strength of item discrimination parameters. However, the *A* and *B* parameters provide unique perspectives on the overall relations between items and sets of latent variables. It can readily be observed that some items were highly discriminating overall in the bifactor model (e.g., item 17), while others were less discriminating (e.g., item 49). Nevertheless, allowing each item to load onto multiple latent variables generally assured some meaningfully discrimination (i.e., the *A* parameters are consistently above 1). Most of the *B* parameters in Table 8 are positive. This implies that items on the BSI were difficult to endorse. In other words, the symptoms were endorsed only by individuals with more severe levels of psychological distress.

The bottom rows in Table 8 present parameter estimates for the dichotomous criterion-related validity variables. Both the presenting concerns and DSM-IV-TR diagnoses variables had smaller *A* parameters than did items from the BSI. This implies that scale items were more closely related to the bifactor model than were validity variables. In addition, the presenting concerns variables tended to have smaller *A* parameters in comparison to the DSM-IV-TR diagnosis variables; the latent structure of the bifactor model was more closely related to patients' diagnoses than patients' self-reported concerns. It can also be observed that *B* parameters for the mood and anxiety presenting concerns variables were larger than the *B* parameters for the mood and anxiety diagnosis variables. That is, patients were more likely to be diagnosed with a mood or anxiety disorder than present with a mood or anxiety concern. This likely reflects the high number of patients within the current sample who primarily sought assessment for a neurocognitive disorder (e.g., Attention-Deficit/Hyperactivity Disorder and Learning Disorders) rather than emotional distress.

The bifactor model was most discriminating for the health concerns presentation and somatoform disorders diagnosis variables; however, both events were quite rare. Indeed, the *B* parameters for both are substantially higher than those for mood and anxiety concerns/diagnoses. The *A* parameters for the mood concerns presentation and mood disorders diagnosis variables were also relatively high. This suggests that the bifactor model adequately assessed mood-related distress. The anxiety concerns presentation and anxiety disorders diagnosis variables, on the other hand, were less accurately assessed by the bifactor model. These results, along with the diminished factor loadings for ANX items onto the anxiety factor (Table 6), suggest that anxiety-related distress was not well captured by the bifactor model. Parameter estimates for discriminating between diagnoses of an anxiety disorder and a mood disorder are also provided in Table

8. As can be seen, the *B* parameters for each variable were near 0. This reflects a relative overall balance between the two (pure) disorders within the current sample. The *A* parameter estimates suggest that the disorders were well differentiated by the combination of the depression and internalizing latent variables, but poorly differentiated by the combination of anxiety and internalizing latent variables.

It is of interest to consider the relative contribution of each latent variable towards the probability of being diagnosed with a DSM-IV-TR disorder. This can be accomplished by examining the *a* values in Table 8—the item/validity variable specific discrimination parameters. With respect to the mood and anxiety disorders diagnosis variables, domain specific (depression or anxiety) and general (internalizing) factors exerted relatively balanced influences on the validity variables (i.e., the ratios of the *a* parameters are all less than 2:1). That is, the probability of patients being diagnosed with a mood disorder increased along with both depression and internalizing, and the probability of patients being diagnosed with an anxiety disorder increased along with both anxiety and internalizing. This was less true for the somatoform disorders diagnosis variable, where internalizing contributed relatively little towards the probability of a positive diagnosis. As such, the results suggest that mood and anxiety disorders are closely related to internalizing, but somatoform disorders are less influenced by the construct.

These relations can be visualized through item characteristic surfaces for the mood disorders diagnosis variable in Figure 18, the anxiety disorders diagnosis variable in Figure 19, and the somatoform disorders diagnosis variable in Figure 20. As can be seen in the figures, the probability of patients being diagnosed with a mood disorder (Figure 18) or an anxiety disorder (Figure 19) increased as both the domain specific (depression or anxiety) and the general (internalizing) factors increased. This result is confirmed in Figure 21, the .50 equi-probable contour for the diagnosis of a mood disorder. Careful study of Figure 21 confirms that both domain specific depression and internalizing played prominent roles in the diagnosis of a mood disorder. This is less true in Figure 20, where the probability of a somatoform disorder diagnosis was almost exclusively related to somatization.

The final two rows of Table 8 concern parameter estimates for the bifactor model’s ability to differentiate between patients diagnosed with just an anxiety disorder and patients diagnosed with just a mood disorder. As can be seen, when differentiating between the disorders by means of the depression and internalizing latent variables or the anxiety and internalizing latent variables, the internalizing latent variable played a small role. Indeed, the item characteristic surfaces in Figures 22 and 23 suggest that patients’ levels of internalizing had almost no impact on differentiating between the disorders. The nearly vertical equi-probable contour in Figure 24 for a .50 probability of discriminating between the disorders by means of the depression and internalizing latent variables also conveys that the latter plays almost no role in the probability of diagnosis. This is precisely what one would expect from the orthogonal bifactor solution. The model appears to have redistributed common variance between factors (i.e., the strong positive correlation in the oblique solution) into an orthogonal internalizing dimension. These results imply that domain specific depression and domain specific anxiety do indeed capture the unique components of their respective disorders. Internalizing contributed to the probability of a patient being diagnosed with a mood or anxiety disorder, but played almost no role in discriminating between disorders—it is a construct shared by both groups.

Diagnostic Accuracy

ROC curves for prediction of a mood disorder diagnosis are presented in Figure 25. As can be seen, the total score model and the bifactor model were very similar in their diagnostic accuracy. The AUCs for the mood disorders diagnosis variable predicted from the total score model and from the bifactor model were both .78. Thus, it appears as though the bifactor model provided no better diagnostic accuracy in the assessment of mood disorders. The ROC curves for prediction of the anxiety disorders diagnosis variable are presented in Figure 26. Again, the total score model and the bifactor model were very similar in their diagnostic accuracy. The AUC for the total score model was .69, slightly better than the .68 AUC for the bifactor model. Thus, it appears as though the bifactor model provided no better diagnostic accuracy in the assessment of anxiety disorders. The ROC curves for prediction of the somatoform disorders diagnosis variable are presented in Figure 27. As can be seen, the bifactor model appears to have provided better diagnostic accuracy than the total score model. This was confirmed by the AUC for total scores of .74 in comparison to the AUC for the bifactor model of .77. It appears as though the bifactor model provided better diagnostic accuracy in the assessment of somatoform disorders. It should be noted, however, that these particular ROC curves are relativity imprecise due to the rarity of somatoform diagnoses (this is reflected in the jagged contour of the curves).

The ROC curves for differentiating between just an anxiety disorder diagnosis and just a mood disorder diagnosis by means of the depression and internalizing variables in the bifactor model versus depression scores in the total score model are presented in Figure 28. It can be observed that the bifactor model tended to provide better diagnostic accuracy than did the total score model. This was confirmed by the AUC for the total score model of .72, which was lower than the AUC for the bifactor model of .76. The ROC curves for differentiating between the disorders by means of the anxiety and internalizing variables in the bifactor model versus anxiety total scores in the total score model are presented in Figure 29. Neither model provided great accuracy in differentiating between the two disorders. However, the total score model ROC curve was particularly poor, suggesting that some patients diagnosed with just an anxiety disorder actually had lower anxiety total scores than patients diagnosed with just a mood disorder. The ROC curve for the bifactor model, in contrast, was consistently positive. In addition, the AUC for total scores of .52 was lower than the .58 AUC for the bifactor model. Thus, it appears as though the bifactor model provided better, but not good, diagnostic accuracy in comparison to the total score model.

Chapter 4

Discussion

Findings

Confirmatory factor analysis and IRT techniques were applied to a clinical data set to explore modern psychometric theory’s impact on test validity in clinical assessment. Specifically, the analyses demonstrated purposeful integration of traditional and modern psychometric techniques in model selection. The hypotheses of this study can generally be accepted. To summarize, in comparison to the unidimensional/total score model, multidimensional models provided better representation of the BSI’s internal structure and demonstrated improved diagnostic accuracy by explaining more variance in some criterion-related variables.

Criterion-related validity variables consisted of DSM-IV-TR diagnoses and patients’ self-reported reasons for seeking psychological treatment or assessment. Configural MMPI-2 profiles revealed that patients diagnosed with DSM-IV-TR psychiatric disorders reported more symptoms of psychological distress than did patients who were not; however, the profiles could not meaningfully differentiate between disorders. Therefore, criterion-related validity variables were treated as fallible indicators of latent constructs, and were analyzed much like scale items in latent variable analyses. The variables were treated as criteria in the ROC curve analyses for the purposes of estimating familiar statistics related to the diagnostic accuracy of clinical measures (i.e., sensitivity and specificity).

Confirmatory factor analyses revealed that all multidimensional latent factor models provided better fit for the BSI than did a unidimensional model. The oblique simple structure, oblique complex structure, oblique bifactor structure, and orthogonal bifactor structure models all provided acceptable overall fit for the measure. However, the orthogonal complex structure model was not successfully estimated, and thus could not be accurately evaluated. In addition, variance in the ANX scale was poorly captured by the oblique complex and orthogonal bifactor structure models. It appears that the ANX scale requires its own domain specific factor (e.g., a bifactor within bifactor model). Although not reported with the present results, attempts to account for such latent sub-structures were made (e.g., correlated errors, correlated factors, and bifactor within bifactor models), but were met with convergence problems and peculiar parameter estimates. As such, the anxiety factor, in this instance, more prominently represents phobic distress rather than general anxiety.

All models were generally comparable with respect to explaining variance in criterion-related validity variables; however, two exceptions were noted in the data. First, the bifactor models (both oblique and orthogonal) explained more variance in health concerns presentations and somatoform disorder diagnoses in comparison to simple or complex structure models. ROC curve analyses confirmed that the orthogonal bifactor structure model provided better diagnostic accuracy in comparison to unidimensional total scores. Second, the bifactor models explained more variance when used to differentiate between patients diagnosed with just an anxiety disorder and patients diagnosed with just a mood disorder. ROC curve analyses again confirmed that the orthogonal bifactor structure model provided better diagnostic accuracy in comparison to unidimensional total scores. However, this was only true when the depression and anxiety factors were used in isolation. When both factors (and total scores) were used simultaneously to differentiate between disorders, there were no observed differences in variance accounted for and diagnostic accuracy among the models.

Of the multidimensional structures considered in the present analyses, the orthogonal bifactor model appeared to provide the greatest construct validity: good representation of internal structure, strong criterion-related validity, and relatively parsimonious partitioning of variance with respect to content validity. IRT analyses revealed that the orthogonal bifactor model formed four distinct constructs: domain specific depression, domain specific anxiety, domain specific somatization, and general internalizing. Diagnoses of mood disorders and anxiety disorders were both strongly related to their respective domain specific latent variables (i.e., depression and anxiety respectively) and to the general internalizing latent variable. Diagnoses of somatoform disorders were strongly related to domain specific somatization, but were weakly related to internalizing. When differentiating between patients diagnosed with just an anxiety disorder and patients diagnosed with just a mood disorder, the depression latent variable and the anxiety latent variable (to a lesser extent) provided meaningful discrimination, but the internalizing factor did not. The results suggest that domain specific latent variables can effectively be used to discriminate between related psychiatric disorders.

Limitations

The results of this study are limited by the validity of the criterion-related variables. Uncertainty in these variables could have resulted from three primary issues. First, DSM-IV-TR diagnoses were made by clinical psychology graduate students with one or more years of training. It is likely that these diagnoses were, on occasion, inaccurate. Indeed, even if the DSM-IV-TR diagnoses had come from expert clinicians, it is likely that some false positives and some false negatives would have been present. Compounding this, using multiple clinical psychology graduate students to assign psychiatric diagnoses is itself a source of error variance (i.e., variance between raters). Second, DSM-IV-TR diagnoses were collapsed into three general categories of psychological distress: mood disorders, anxiety disorders, and somatoform disorders. This was done, in part, to provide clarity in the results. It was beneficial, for example, to have the depression latent variable predict the mood disorders diagnosis variable. Having the depression latent variable predict multiple unique mood disorder variables (e.g., dysthymic disorder, major depressive disorder, mood disorder due to a general medical condition, etc.) could have obscured the results. Third, the DSM-

IV-TR is itself a latent variable model. Like most models, it does not flawlessly represent reality. Indeed, the DSM is regularly revised with the goal of making the model more congruent with evolving clinical theories. Therefore, the DSM-IV-TR, as well as the more general process of psychiatric diagnosis, should be regarded as imperfect.

These limitations in validity variables complicated the process of criterion validation. Fortunately, latent variable models are designed to account for this very issue. Although DSM-IV-TR diagnoses used in the present study were undoubtedly flawed, they likely maintained some degree of validity. In such instances, it is useful to look for converging evidence. Indeed, philosophers of science would assert that these sorts of conclusions are both natural and unavoidable. In the present situation, it can be argued that even with considerable error being introduced into the criterion-related variables, there is no reason to expect systematic bias. That is, all models should have been equally affected by the error. The questions posed in the present study concerned the relative, not the absolute, validity of each model. Therefore, the topic could still be addressed despite attenuations in validity coefficients.

The results of the current study were also limited by small sample size. While 688 participants is relativity large for a study involving clinical populations, it is not ideal for the types of analyses conducted in this study. Indeed, latent variable models, particularly those that employ complex or bifactor structures, place heavy demands on the information provided by samples. Also, researchers who conduct latent variable analyses without explicit hypotheses about the latent structure of their data are at risk of capitalizing on sample specific nuances. That is, chosen models may mimic the characteristics of a particular sample rather than the larger population that the sample is meant to represent. In such situations, it is always best to split the sample in two for the purposes of cross validation. In the current study, specific hypotheses were made about the latent structure of the BSI; however, the work was clearly both exploratory and confirmatory in nature. Thus, it would have been best to replicate the results in a second sample. However, splitting an already limited sample into two subgroups would have seriously threatened most conclusions regarding the models.

Limitations in sample size also restricted the operationalization of criterion-related validity variables. As already mentioned, specific DSM-IV-TR disorders were collapsed into general categories of psychological distress. This was done for clarity, but also out of necessity. A sample of 688 patients in an outpatient setting is far too small to produce large frequencies of specific DSM-IV-TR disorders; sample sizes within most of the specific diagnostic categories were not large enough for statistical inference. Even with collapsed categories, the somatoform disorders variable had very few cases (*n* = 12), leaving all associated parameter estimates subject to considerable error.

It should also be noted that two major subgroups of patients were present in the data: a group primarily seeking psychological counseling and a group primarily seeking psychological assessment. There is potential for measurement variance in samples comprising multiple populations. That is, differential item functioning may have been present with respect to the treatment and assessment subgroups. It is notable that diagnostic rates differed between groups. In the assessment group, 25% of patients were diagnosed with a mood disorders, 24% were diagnosed with an anxiety disorder, 1% were diagnosed with a somatoform disorder, and 82% were diagnosed with some other DSM-IV-TR disorder. In the counseling group, 36% of patients were diagnosed with a mood disorders, 29% were diagnosed with an anxiety disorder, 3% were diagnosed with a somatoform disorder, and 89% were diagnosed with some other DSM-IV-TR disorder. These discrepancies pose the possibility of differential item functioning, but certainly do not assure it. It is likely that the treatment and assessment groups differed on meaningful psychological characteristics as well (possibly explaining the differential rates of diagnoses). Only explicit evaluations of the latent variable models' parameters could have adequately addressed this question. However, the subgroups were not large enough to perform meaningful differential item functioning analyses for the complex latent structures considered in this study. That is, differential item functioning would have been difficult to detect even if present in the data.

Hypotheses Revisited

There is little question that the conclusions rendered in the present study are complicated by methodological limitations; however, some conclusions seem apparent despite the challenges. For example, it does seem clear that the internal structure of the BSI was not accurately represented by the orthogonal simple structure model. Model fit, in this instance, was exceedingly poor, and left little doubt that that sample data were best represented with some type of multidimensional latent structure. Models that allowed for crossloadings and/or correlated latent variables appeared to provide much better representations of the BSI's internal structure. However, conclusions related to criterion-related validity of the models were less robust. Ambiguity in the validity criteria and limitations in sample size left these results subject to doubt.

An examination of participants from a less diverse population could have diminished some of the problems arising from the previously mentioned limitations. Indeed, much of the work on modern psychometric modeling involves the use of simulation data and homogeneous samples. In contrast, an explicit goal of the present study was to examine the value of modern psychometric theory in clinical practice. Thus, while the present results were limited by ambiguity, they were also reinforced by generalizability. These issues reflect the inevitable trade-off between highly controlled experimental research, which can lack real world meaning, and archival research of existing clinical data, which can lack control and precision. It is noteworthy that the orthogonal bifactor model demonstrated improved construct validity despite poorly defined and measured criterion variables. Furthermore, the conclusions are bolstered by their convergence with emerging clinical and measurement theories. While caution is necessary in interpreting the results, there is cause to draw meaningful conclusions.

Implications

The results of the present study suggest that orthogonal simple structure models are not appropriate for the latent structure of depression, anxiety, and somatization as measured by self-report inventories. Whether or not these results apply more generally to all types of clinical measures is an empirical question; however, literature reviewed earlier in this paper suggests that this is likely true. It is noteworthy that the internal structure of the BSI was adequately modeled with the oblique simple structure model. Thus, allowing factors to correlate may dramatically improve the fit of simple structure models for some types of clinical instruments—a technique that has long been used to improve factor analytic work (see Mulaik, 2010).

McDonald (1999) noted that factorially-simple (independent clusters) solutions greatly simplify parameter estimation and model interpretations in comparison to factorially-complex solutions. Allowing items to load onto multiple factors can create confusion. Because of this, McDonald recommends that composite measures be deconstructed into their more basic unidimensional, and potentially correlated subtests. That is, researchers should alter tests to fit desirable psychometric models. Doing so would allow psychologists to maintain useful content domains of psychological distress while improving the psychometric properties of their clinical measures. In some respects, the authors of the BSI—as well as the authors of other well-known psychological measures—have taken this very path. Years of factor analytic refinements have focused on reducing psychological measures to their more basic components. And, although the strategy has improved the dimensionality of clinical measures, it has also left a notable byproduct—oblique factors. By focusing on specific symptoms or clusters of psychological distress, models have not explicitly accounted for the higher-order and/or hierarchical associations in clinical data. These unexplained covariances could represent meaningful elements of clinically theory.

In the current study, the oblique simple structure model provided a parsimonious representation of the BSI's internal structure. However, strong positive correlations between the depression, anxiety, and somatization factors in the model suggest that further complexity underlies the data. Just as observed correlations between test items are interpreted as indications of latent variables, model-implied correlations between latent variables can be interpreted as indications of additional unexplained latent structure. Thus, while we may choose to interpret the depression, anxiety, and somatization factors of the BSI as unidimensional-like entities, correlated factors within the model suggest that there remains a substantial amount of unexplained common variance. This common variance appears to have manifested itself on the BSI through strong positive correlations between anxiety-related items, depression-related items, and, to a lesser extent, somatization-related items.

An alternative strategy to scale refinement is to alter psychometric models to fit desirable tests. That is, a test developer may instead choose to fit more complex latent structures to scales already shown to provide useful clinical measurement. In the present study, complex models of latent psychopathology—the complex and bifactor structure solutions—outperformed the oblique simple structure model. In particular, the orthogonal bifactor structure model emerged as an approach to scale refinement that provides a good balance between accurate representation of internal structure and strong criterion-related validity.

The orthogonal bifactor structure model has the ability to tease apart the latent relations between otherwise correlated factors and/or items. On the BSI, correlations between the depression, anxiety, and somatization factors were effectively accounted for through the inclusion of a general internalizing construct. Unfortunately, this more complex solution also altered the meaning of the depression, anxiety, and somatization factors within the model: each represents domain specific variance that is free of internalizing. The depression factor, for example, represents depression-specific variance that is independent of internalizing variance. The meaning of domain specific latent constructs such as these is likely foreign to most clinical psychologists. What would it mean, for example, for a patient to be low on domain specific depression but high on general internalizing? The answer is not apparent. McDonald's (1999) recommendation for factorially-simple models would appear to avoid these complex interpretation issues.

Users of the MMPI-2-RF have run into similar challenges in interpreting patients' scores. The measure differs from the MMPI-2 in that the clinical scales have been restructured so that each represents a domain specific construct that is independent of a more general "demoralization" construct. Critics of the MMPI-2-RF argue that these content specific scales are "alien" to clinical psychologists (Nichols, 2006). Yet, the MMPI-2-RF developers did not use factorially-complex models to create domain specific factors; rather, the authors deconstructed the MMPI-2 clinical scales into their more basic (oblique) unidimensional subtests—the very recommendation offered by McDonald (1999). Thus, there may be no avoiding that refinements of clinical scales based on modern psychometric theory will alter meaning. This may be true whether done through factorially-simple models or through factorially-complex models.

It is offered that refinements of scales based on these two seemingly opposing methods are, as it is said, two sides of the same coin. Both methods attempt to account for multidimensionality on supposedly unidimensional scales. Oblique factorially-simple models do so by deconstructing measures into more basic unidimensional, correlated components. Orthogonal factorially-complex models do so by deconstructing conglomerations of variance into more basic, uncorrelated domains. In sum, a primary difference between these multidimensional factorially-simple models and multidimensional factorially-complex models is the formation of oblique versus orthogonal factors. The othrogonal bifactor solution represents a factorially-complex model where latent relations between factors are explained by a general source of variance shared by all items (McLeod, Swygert, & Thissen, 2001). Such models seem appropriate when general domains of distress are relevant to psychological constructs. The oblique simple structure model represents a factorially-simple model where latent relations between factors are not explicitly explained, but are acknowledged through correlations among factors. Such models also seem appropriate when general domains of distress are relevant for various forms of psychopathology, but are, for obvious reasons, less comprehensive and less accurate. The results of the present study suggest that there are important differences between the manner in which external validity is maximized in the oblique simple structure model and the orthogonal bifactor structure model.

When used to diagnose psychiatric disorders—to differentiate between impaired versus non-impaired populations—the orthogonal bifactor model and the oblique simple structure model performed equally well. The reason for this was revealed in the multidimensional IRT analyses—both general and domain specific sources of variance contributed to diagnoses. Thus, while the latent variables formed in the oblique simple structure model confounded combinations of common and unique variance, both sources of variances were pertinent to the matter at hand. Althouh imprecise, it was not necessary to separate the two sources of variance with respect to modeling the overall probability of positive diagnoses.

The oblique simple structure model and the orthogonal bifactor structure model also provided equivalent diagnostic accuracy when all latent variables were used simultaneously to differentiate between patients diagnosed with just an anxiety disorder and patients diagnosed with just a mood disorder. However, the oblique simple structure model performed less well when the factors were used in isolation. The reason for this was again revealed in the multidimensional IRT analyses. When differentiating between clinical groups, only domain specific variance contributed to accurate diagnoses. Thus, the latent variables in the oblique simple structure model confounded combinations of useful domain specific variance and useless general variance. By using all factors simultaneously to differentiate between clinical groups, a regression-like multivariate prediction equation emerged that suppressed the general (useless) variance and expressed the domain specific (useful) variance. The bifactor model, on the other hand, separated these distinct sources of variance from the onset. A complex regression-like equation is not needed when meaningful variance has already been partialed out of the observed variables. It is somewhat ironic that factorially-simple models lead to complex prediction schemes, while factorially-complex models lead to simple prediction schemes.

Combining principles of modern and traditional psychometric theory leads to a richer understanding of clinical measures and improves measurement validity. Yet, there remains an obvious question as to which multidimensional latent variable model should be selected for scale development, refinement, and interpretation. It must be realized that this is a question without a definitive answer. The multidimensional models considered in this study are all but certain to be incorrect approximations of true latent structure. A more appropriate question to ask is which model provides the best overall construct validity within specific assessment circumstances. It has been argued that the oblique simple structure model and the orthogonal bifactor structure model hold great promise in general for improving measurement in clinical psychology. The models represent related methods of solving the same problem—complexity in clinical data. Some general comments can be made about the appropriateness of each.

First, it is likely that the orthogonal bifactor model more closely approximates the true latent structure of psychopathology. This statement rests on the assumption that truth is complex. The orthogonal bifactor model, which appears to be a more general structure of latent psychological distress, poses fewer assumptions about reality. When a model is made simpler—constrained—it is likely true that the simplification was done in error. As such, generality is typically closer to truth. If a researcher or clinician whishes to employ the model that most closely approximates truth, the orthogonal bifactor solution is to be preferred.

Second, it must also be acknowledged that complexity comes at a price. Factorially-complex, multidimensional latent variable models have the potential to greatly complicate psychometric modeling and the interpretability of clinical measures. Indeed, such complications were observed in the present study with the orthogonal bifactor solution. The formation of domain specific continuums of variance led to unrecognizable constructs of psychopathology as well as unexpected patterns of item loadings. Simple structure models are made more palatable by their ability to maintain recognizable constructs. In addition, the simple structure model represents a riskier prediction of reality. The purpose of proposing models in science is to give structure to the world. As models become increasingly vague, so too does our knowledge. There is good reason to favor simplicity when simple models make accurate predictions. Such models make scientists more efficient in their work. To the extent that the oblique simple structure model can simplify the world while making accurate predictions, it is to be preferred.

Finally, it should also be apparent that some simplifications can be made to models without drastically diminishing their validity. As a more general structure, the bifactor model should provide validity coefficients that are at least equivalent to those provided by the oblique simple structure model (Mulaik & Quartetti, 1997). Divergence among the item discrimination parameters for the internalizing latent variable would likely diminish the value of the oblique simple structure model. However, the orthogonal bifactor structure model's ability to improve accuracy may not excuse its added complexity. Necessity of the model is dependent on particular assessment circumstances. In the present study, it was found that the oblique simple structure model generally performed as well as the bifactor structure model in identifying psychiatric disorders among outpatients.

However, the bifactor model demonstrated superiority when differentiating between disorders—a more complex task. Thus, it can be argued that the simple structure model is appropriate for the identification of any psychiatric disorder, while the bifactor model is appropriate for the identification of specific psychiatric disorders. Yet there was convergence in the diagnostic accuracy of the models when all available sources of variance were called upon. Much like the practice of using confirgural patterns to evaluate MMPI-2 scales, multivariate interpretations of oblique simple structure models can be used in lieu of the bifactor structure model. The decision as to which structure is preferred is perhaps best made by test users and substantive theorists rather than test developers and psychometricians.

Future Directions and Conclusion

Simulation work is needed to compare the diagnostic benefits between the oblique simple and orthogonal bifactor structures. Specifically, comparisons of diagnostic sensitivity and specificity between models under various conditions would help test developers and clinicians decide if the added complexity of the bifactor model is warranted. Such work must be guided by modern and traditional psychometric techniques. As has been demonstrated, both are important in the evaluation of construct validity. Finding that the orthogonal bifactor model better represents the internal structure of a measure is not necessarily sufficient for it to replace an oblique simple structure model. Rather, it should be demonstrated that the model meaningfully improves diagnostic accuracy. It is of interest to determine if scales can be designed to more effectively account for domain specific and general variance. Computer adaptive testing represents an area of growth where the bifactor model could be utilized to maximize diagnostic accuracy in this regard.

In the interim, there is good reason for test developers and test users to consider complex models of psychological distress in their work. Clinical theorists have long proposed models of psychopathology that are much richer than the supposedly unidimensional constructs measured by existing tests. As measurement theory progresses, so too will clinical psychologists' ability to measure psychiatric populations in accordance with complex theories. Exploration of complex latent variable models—particular those in harmony with clinical theory—is encouraged. Traditional practice of selecting just a single model to represent the internal structure of a measure may no longer be appropriate. Rather, objective analyses of the overall construct validity of multiple and diverse models of internal structure are warranted. For example, users of the BSI should consider the measure as having several potential multidimensional latent structures. Patients' responses can be interpreted through an orthogonal bifactor structure model or through an oblique simple structure model dependent upon the circumstances of assessment. Users of the MMPI-2-RF might consider these same interpretive strategies. For example, it is probable that diagnoses of psychiatric disorders will be more accurate when assigned based upon patients' endorsement of content specific items (i.e., items from a specific Restructured

Clinical scale) as well as general items (i.e., items from the Emotional/Internalizing Dysfunction [EID] scale or the Demoralization [RCd] scale).

Work must also progress in determining the suitability of the bifactor model for various forms of psychological distress. The model is but one of many possible multidimensional structures that could account for latent psychopathology on clinical measures. It is important to note that the bifactor model is not a panacea for all possible psychiatric disorders. There are numerous forms of psychological distress, and the structure of each is likely distinct. Latent psychopathology might consist of combinations of unidimensional and multidimensional structures for both continuous and discrete latent variables. While the task of organizing this latent structure may seem daunting, modern psychometric theory is primed for this goal. A former instructor of the author was fond of citing the philosopher of science Otto Neurath, who likened "the overall process of science-building…to the process of rebuilding a boat, plank by plank, not in dry dock but at sea" (Rosenthal, 1997, p. 121). Scientists must not be rendered stagnant by the difficult tasks confronting them. Complexity, in and of itself, is not a valid reason to avoid modern psychometric theory. Rather, psychologists should move forward confident that they possess the tools to sort through the intricacy of the human mind. We must improve our measures, plank by plank, steadfast in the goal of complete scientific description of psychiatric disorders.

IRT and related latent variable models have the potential to dramatically change and improve psychological assessment. Past methodologies in scale development have often focused primarily on criterion validity or on reliability and internal structure, at the expense of the other. Comprehensive models of latent psychopathology can meet the requirements of both modern and traditional psychometric theories, thereby maximizing overall construct validity. To that end, we must continue the arduous process of evaluating and refining models of psychological phenomena.

References

Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

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

Aiken, L. R., Jr. (1966). Another look at weighting test items. *Journal of Educational Measurement, 3*, 183-185.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, D.C.: Author.

American Psychiatric Association (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: APA.

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

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

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12,* 33-51.

Andrich, D., & Van Schoubroeck, L. (1989). The General Health Questionnaire: A psychometric analysis using latent trait theory. *Psychological Medicine, 19,* 469-485.

Angst, J., & Dobler-Mikola, A. (1985). The Zurich study: VI. A continuum from depression to anxiety disorders? *European Archives of Psychiatry & Neurological Sciences, 235*, 179-186.

Armstrong, T. D., & Costello, E. J. (2002). Community studies on adolescent substance use, abuse, or dependence and psychiatric comorbidity. *Journal of Consulting and Clinical Psychology, 70*, 1224-1239.

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

Baker, J. G., Zevon, M. A., & Rounds, J. B. (1994). Differences in positive and negative affect dimensions: Latent trait analysis. *Personality and Individual Differences, 17*, 161-167.

Balsis, S., Gleason, M. E. J., Woods, C. M., & Oltmanns, T. F. (2007). An item response theory analysis of DSM-IV personality disorder criteria across younger and older age groups. *Psychology and Aging, 22*, 171-185.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561-571.

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology, 42*, 861-865.

Bedi, R. P., Maraun, M. D., & Chrisjohn, R. D. (2001). A multisample item response theory analysis of the Beck Depression Inventory-1A. *Canadian Journal of Behavioural Science, 33*, 176-185.

Beer, M. D. (1996). Psychosis: A history of the concept. *Comprehensive Psychiatry, 37*, 273-291.

Bejar, I. I. (1977). An application of the continuous response level model to personality measurement. *Applied Psychological Measurement, 1*, 509-521.

Bell, R. C., Low, L. H., Jackson, H. J., & Dudgeon, P. L. (1994). Latent trait modeling of symptoms of schizophrenia. *Psychological Medicine, 24,* 335-345.

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

Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191-244.

Binford, A., & Liljequist, L. (2008). Behavioral correlates of selected MMPI-2 Clinical, Content, and Restructured Clinical scales. *Journal of Personality Assessment, 90*, 608-614.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bolt, D. M., Hare, R. D., & Neumann, C. S. (2007). Score metric equivalence of the Psychopathy Checklist-Revised (PCL-R) across criminal offenders in North America and the United Kingdom: A critique of Cooke, Michie, Hart, and Clark (2005) and new analyses. *Assessment*, *14*, 44-56.

Bontempo, R. (1993). Translation fidelity of psychological scales: An item response theory analysis of an individualism-collectivism scale. *Journal of Cross-Cultural Psychology, 24*, 149-166.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S.J., Stark, S. and Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965– 973.

Boulet, J., & Boss, M. W. (1991). Reliability and validity of the Brief Symptom Inventory. *Psychological Assessment, 3*, 433-437.

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

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Oxford, England: John Wiley & Sons.

Breteler, M. H. M., Hilberink, S. R., Zeeman, G., & Lammers, S. M. M. (2004). Compulsive smoking: The development of a Rasch homogeneous scale of nicotine dependence. *Addictive Behaviors, 29*, 199-205.

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the Dispositional Hope Scale. *Psychological Assessment*, *20*, 310-315.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107*, 179-192.

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

Cabrero-García, J., & López-Pina, J. A. (2008). Aggregated measures of functional disability in a nationally representative sample of disabled people: Analysis of dimensionality according to gender and severity of disability. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 17*, 425-436.

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

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

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

Cassano, G. B., Michelini, S., Shear, M. K., & Coli, E. (1997). The panic-agoraphobic spectrum: A descriptive approach to the assessment and treatment of subtle symptoms. *American Journal of Psychiatry, 154*, 27-38.

Cassano, G. B., Mula, M., Rucci, P., Miniati, M., Frank, E., Kupfer, D. J., et al. (2009). The structure of lifetime manic--hypomanic spectrum. *Journal of Affective Disorders, 112,* 59-70.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45,* S3-S11.

Chambon, O., Cialdella, P., Kiss, L., & Poncet, F. (1990). Study of the unidimensionality of the Bech-Rafaelsen Melancholia Scale using Rasch analysis in a French sample of major depressive disorders. *Pharmacopsychiatry, 23,* 243-245.

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

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

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.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M., & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 Depression scale. *Assessment, 7,* 37-54.

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

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40,* 359-376.

Clark, D. C., Cavanaugh, S. v., & Gibbons, R. D. (1983). The core symptoms of depression in medical and psychiatric patients. *Journal of Nervous and Mental Disease, 171,* 705-713.

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

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754-761.

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

Cooke, D. J., & Michie, C. (1999). Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology, 108*, 58-68.

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

Coombs, C. H. (1964). *A theory of data*. Oxford, England: Wiley.

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

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). T*he dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cyr, J. J., McKenna-Foley, J. M., & Peacock, E. (1985). Factor structure of the SCL-90-R: Is there one? *Journal of Personality Assessment, 49*, 571-578.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30,* 295-311.

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

Derogatis, L. R. (1993). *BSI, Brief Symptom Inventory. Administration, scoring, and procedures manual (4th Ed.).* Minneapolis, MN: National Computer Systems.

Derogatis, L. R., & Cleary, P. A. (1977). Confirmation of the dimensional structure of the SCL-90: A study in construct validation. *Journal of Clinical Psychology, 33*, 981-989.

Derogatis, L. R., Klerman, G. L., & Lipman, R. S. (1972). Anxiety states and depressive neuroses: Issues in a nosological discrimination. *Journal of Nervous and Mental Disease, 155*, 392-403.

Derogatis, L. R., Lipman, R. S. & Covi, L. (1973). The SCL-90: An outpatient psychiatric rating scale. *Psychopharmacology Bulletin, 9*, 13-28.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine, 13*, 595-605.

Donovan, M. A., & Drasgow, F. (1999). Do men's and women's experiences of sexual harassment differ? An examination of the differential test functioning of the sexualexperiences questionnaire. *Military Psychology, 11*, 265-282.

Dorus, W., Kennedy, J., Gibbons, R. D., & Ravi, S. D. (1987). Symptoms and diagnosis of depression in alcoholics. *Alcoholism: Clinical and Experimental Research, 11*, 150-154.

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

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

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

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24*, 133-148.

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

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Eysenck, H.J., & Eysenck, S.B.J. (1975). *Manual of the Eysenck Personality Questionnaire*. Hodder & Stoughton: London

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

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

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

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

Finney, D. J. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve* (2nd ed.). London: Cambridge University Press.

Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural, and medical research*. New York: Hafner Pub. Co.

Flannery, W. P., Reise, S. P., & Widaman, K. F. (1995). An item response theory analysis of the General and Academic scales of the Self-Description Questionnaire II. *Journal of Research in Personality, 29*, 168-188.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286-299.

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment, 19*, 14-24.

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.

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

Friedman, A. F., Lewak, R., Nichols, D. S., & Webb, J. T. (2001). *Psychological assessment with the MMPI-2.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

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

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

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

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

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

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

Goldberg, D.P. (1972) *Detection of psychiatric illness by questionnaire*. London: Oxford University Press.

Gomez, R. (2008). Item response theory analyses of the parent and teacher ratings of the DSM-IV ADHD rating scale. *Journal of Abnormal Child Psychology, 36*, 865-885.

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

Gough, H. G. (1956). *California Psychological Inventory*. 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.

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

Guilford, J. P., Lovell, C., & Williams, R. M. (1942). Completely weighted versus unweighted scoring in an achievement examination. *Educational and Psychological Measurement, 2*, 13-21.

Gulliksen, H. (1950). *Theory of mental tests*. Oxford, England: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Haans, A., Kaiser, F. G., & de Kort, Y. A. W. (2007). Privacy needs in office environments: Development of two behavior-based scales. *European Psychologist, 12*, 93-102.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Hahn, E. A., Cella, D., Bode, R. K., Gershon, R., & Lai, J. (2006). Item banks and their potential applications to health status assessment in diverse populations. *Medical Care, 44*, S189-S197.

Hambleton, R. K. (1992). *Hambleton's 9 Theses.* Opening remarks in his invited debate with Benjamin Wright, Session 11.05, AERA Annual Meeting 1992. http://www.rasch.org/rmt/rmt62d.htm Retrieved June 4, 2008

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

Hanisch, K. A. (1992). The Job Descriptive Index revisited: Questions about the question mark. *Journal of Applied Psychology, 77*, 377-382.

Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised*. Toronto, Ontario, Canada: Multi-Health Systems.

Hare, R. D. (2003). *Manual for the Revised Psychopathy Checklist (2nd ed.)*. Toronto, ON, Canada: Multi-Health Systems.

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

Hayes, J. A. (1997). What does the Brief Symptom Inventory measure in college and university counseling center clients? *Journal of Counseling Psychology, 44*, 360–367.

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

Heinrich, R. K., & Tate, D. G. (1996). Latent variable structure of the Brief Symptom Inventory in a sample of persons with spinal cord injuries. *Rehabilitation Psychology, 41*, 131–147.

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

Hergenhahn, B. R. (2001). *An introduction to the history of psychology (4th ed.)*. Belmont, CA, US: Wadsworth/Thomson Learning.

Holcomb, W. R., Adams, N. A., & Ponder, H. M. (1983). Factor structure of the Symptom Checklist—90 with acute psychiatric inpatients. *Journal of Consulting and Clinical Psychology, 51*, 535–538.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

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

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

Ietsugu, T., Sukigara, M., & Furukawa, T. A. (2007). Evaluation of diagnostic criteria for panic attack using item response theory: Findings from the national comorbidity survey in USA. *Journal of Affective Disorders, 104,* 197-201.

Jackson, D. N. (1976). *Jackson Personality Inventory*. Port Huron, MI: Sigma Assessment Systems.

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for personality disorders: An item response theory analysis. *Journal of Abnormal Psychology, 116,* 166-175.

Jannarone, R. J. (1997). Models for locally dependent data. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 465-479). New York: Springer.

Kalinowski, A. G. (1985). Measuring clinical pain. *Journal of Psychopathology and Behavioral Assessment, 7,* 329-349.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

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

Kelderman, H. (1997). Loglinear multidimensional item response models for polytomously scored items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Kellett, S., Beail, N., Newman, D. W., & Hawes, A. (2004). The factor structure of the Brief Symptom Inventory: Intellectual disability evidence. *Clinical Psychology and Psychotherapy, 11*, 275–281.

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

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

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging, 17*, 379-391.

Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12*, 245-252.

Kreiner, S., Simonsen, E., & Mogensen, J. (1990). Validation of a personality inventory scale: The MCMI P-scale (paranoia). *Journal of Personality Disorders, 4*, 303-311.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56*, 921-926.

Krueger, R. F. (2002). Psychometric perspectives on comorbidity. In J. E. Helzer, & J. J. Hudziak (Eds.), *Defining psychopathology in the 21st century: DSM-V and beyond*. (pp. 41-54). Arlington, VA, US: American Psychiatric Publishing, Inc.

Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment, 13*, 140-151.

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology, 116*, 645-666.

Lachar, D. (1974). *The MMPI: Clinical assessment and automated interpretation.* Los Angeles: Western Psychological Services.

Lachar, D., & Gruber, C. P. (2001). *Personality Inventory for Children, second edition (PIC-2) manual: Standard form and behavioral summary*. Los Angeles: Western Psychological Services.

Lai, J., Cella, D., Chang, C., Bode, R. K., & Heinemann, A. W. (2003). Item banking to improve, shorten and computerize self-reported fatigue: An illustration of steps to create a core item bank from the FACIT-Fatigue scale. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 12,* 485-501.

Lambert, M. C., Schmitt, N., Samms-Vaughan, M. E., An, J. S., Fairclough, M., & Nutter, C. A. (2003). Is it prudent to administer all items for each Child Behavior Checklist cross-informant syndrome? Evaluating the psychometric properties of the Youth Self-Report dimensions with confirmatory factor analysis and item response theory. *Psychological Assessment, 15,* 550-568.

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

Lanyon, R. I. (2007). *Utility of the Psychological Screening Inventory: A review*. Journal of Clinical Psychology, 63, 283-307.

Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment (3rd ed.).* Oxford, England: John Wiley & Sons.

Lanyon, R. I., & Thomas, M. L. (2009). Comparability of the Psychological Screening Inventory and Psychological Screening Inventory-2. Manuscript in Preparation.

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and* *methodology for personality evaluation*. New York: Ronald Press

Lee, C. W., & Smith, G. A. (1988). The efficiency of a tailored procedure in predicting CPI scale scores. *Australian Psychologist, 23*, 25-30.

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

Lewine, R. R., Fogg, L., & Meltzer, H. Y. (1983). Assessment of negative and positive symptoms in schizophrenia. *Schizophrenia Bulletin, 9*, 368-376.

Liu, X., & Zhang, J. (2006). Application of differential item functioning in clinical questionnaire. *Chinese Journal of Clinical Psychology, 14,* 349-351.

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

Lord, F. M. (1952). *A theory of test scores. Psychometric Monographs, No. 7*. Chicago: University of Chicago Press.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-549.

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

Loutsiou-Ladd, A., Panayiotou, G., & Kokkinos, C. M. (2008). A review of the factorial structure of the Brief Symptom Inventory (BSI): Greek evidence. *International Journal of Testing, 8*, 90-110.

Ludlow, L. H., & Guida, F. V. (1991). The Test Anxiety Scale for Children as a generalized measure of academic anxiety. *Educational and Psychological Measurement, 51*, 1013-1021.

Luecht, R. M., & Hirsch, T. M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16*, 41-51.

Maier, W., & Philipp, M. (1986). A polydiagnostic scale for dimensional classification of endogenous depression: Derivation and validation. *Acta Psychiatrica Scandinavica, 74,* 152-160.

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

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

Martin, C. S., Chung, T., Kirisci, L., & Langenbucher, J. W. (2006). Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: Implications for DSM-V. *Journal of Abnormal Psychology, 115*, 807-814.

Maser, J. D., & Cloninger, C. R. (1990). *Comorbidity of mood and anxiety disorders*. Washington, DC: American Psychiatric Association.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology*, *40*, 343-344.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs, 15*, 167.

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

McGlinchey, J. B., & Zimmerman, M. (2007). Examining a dimensional representation of depression and anxiety disorders' comorbidity in psychiatric outpatients with item response modeling. *Journal of Abnormal Psychology, 116*, 464-474.

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

McKinley, R. L., & Reckase, M. D. (1982). *The Use of the General Rasch Model with Multidimensional Item Response Data*. Iowa City, IA: American College Testing.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 189-216). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McRae, J. A. (1991). Rasch measurement and differences between women and men in self-esteem. *Social Science Research, 20,* 421-436.

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

Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology, 14,* 165-171.

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

Merikangas, K. R., Angst, J., Eaton, W., Canino, G., Rubio-Stipec, M., Wacker, H., et al. (1996). Comorbidity and boundaries of affective disorders with anxiety disorders and substance misuse: Results of an international task force. *British Journal of Psychiatry, 168,* 58-67.

Michie, C., & Cooke, D. J. (2006). The structure of violent behavior: A hierarchical model. *Criminal Justice and Behavior, 33*, 706-737.

Millon, T. (1983). Millon Clinical Multiaxial Inventory manual. Minneapolis, MN: National Computer Systems.

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

Mineka, S., Watson, D., & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology, 49*, 377-412.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359-381.

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis. With applications in political research*. New York: Walter de Gruyter, Mouton.

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

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

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling, 4*, 193-211.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59-71.

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

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

Muthén, L., & Muthén, B. (1998-2006). *Mplus User's Guide. Fourth Edition*. Los Angeles, CA: Muthén & Muthén.

Myers, I. B. (1962). *The Myers-Briggs Type Indicator manual*. Princeton, NJ: Educational Testing Service.

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

Nichols, D. S. (2006). The trials of separating bath water from baby: A review and critique of the MMPI-2 Restructured Clinical scales. *Journal of Personality Assessment, 87*, 121-138.

Olsen, L. R., Mortensen, E. L., & Bech, P. (2004). The SCL-90 and SCL-90R versions validated by item response models in a Danish community sample. *Acta Psychiatrica Scandinavica, 110,* 225-229.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment, 14*, 50-59.

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

Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology, 18*, 267-296.

Parsons, C. K., & Hulin, C. L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Psychology, 67*, 826-834.

Potthoff, E. F., & Barnett, N. E. (1932). A comparison of marks based upon weighted and unweighted items in a new-type examination. *Journal of Educational Psychology, 23*, 92-98.

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

Psychological Corporation (2008). *Wechsler Adult Intelligence Scale—Fourth Edition: Technical and interpretive manual*. San Antonio, TX: Pearson.

Raju, N. S., Steinhaus, S. D., Edwards, J. E., & DeLessio, J. (1991). A logistic regression model for personnel selection. *Applied Psychological Measurement*, *15*, 139-152.

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

Reckase, M. D. (2009). *Multidimensional Item Response Theory.* New York: Springer.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217-226.

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

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16*, 19-31.

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. *Applied Psychological Measurement, 14*, 45-58.

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

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

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

Retzlaff, P. D., Sheehan, E. P., & Lorr, M. (1990). MCMI--II scoring: Weighted and unweighted algorithms. *Journal of Personality Assessment, 55*, 219-223.

Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation. 6*, 595-600.

Revicki, D. A., & Sloan, J. (2007). Practical and philosophical issues surrounding a national item bank: If we build it will they come? *Quality of Life*

Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 16, 167-174.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*, 51-67.

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

Rodebaugh, T. L., Woods, C. M., Heimberg, R. G., Liebowitz, M. R., & Schneier, F. R. (2006). The factor structure and screening utility of the Social Interaction Anxiety Scale. *Psychological Assessment, 18*, 231-237.

Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment, 16*, 169-181.

Rosenthal, R. (1997). Some issues in the replication of social science research. *Labour Economics*, *4*, 121-123.

Rosenthal, R., & Rosnow, (1991). *Essential of behavioral research: Methods and data analysis* (2nd ed.). Boston: McGraw-Hill.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychometric Measurement*, *12*, 397-409.

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

Rubin, H. (1948). The Minnesota Multiphasic Personality Inventory as a diagnostic aid in a veterans hospital. *Journal of Consulting Psychology, 12*, 251-254.

Saha, T. D., Chou, S. P., & Grant, B. F. (2006). Toward an alcohol use disorder continuum using item response theory: Results from the national epidemiologic survey on alcohol and related conditions. *Psychological Medicine, 36*, 931-941.

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

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

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

Schneider, R. J., Goff, M., Anderson, S., & Borman, W. C. (2003). Computerized adaptive rating scales for measuring managerial performance. *International Journal of Selection and Assessment, 11*, 237-246.

Sharp, C., Goodyer, I. M., & Croudace, T. J. (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. *Journal of Abnormal Child Psychology, 34,* 379-391.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68,* 413-430.

Silverstein, M. L., & Nelson, L. D. (2000). Clinical and research implications of revising psychological tests. *Psychological Assessment, 12*, 298-303.

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

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75,* 1350-1362.

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

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

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

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

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

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Tackett, J. L., Quilty, L. C., Sellbom, M., Rector, N. A., & Bagby, R. M. (2008). Additional evidence for a quantitative hierarchical model of mood and anxiety disorders for DSM-V: The context of personality structure. *Journal of Abnormal Psychology, 117,* 812-825.

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

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

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

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

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

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

Thurstone, L. L. (1947). *Multiple-Factor Analysis: A Development and Expansion of the Vectors of Mind*. Chicago: The University of Chicago Press.

Thurstone, L.L. (1959). *The Measurement of Values*. Chicago: The University of Chicago Press.

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology, 82*, 448-461.

Uttaro, T., & Lehman, A. (1999). Graded response modeling of the quality of life interview. *Evaluation and Program Planning, 22*, 41-52.

van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1-28). New York: Springer.

Verhulst, P-F. (1845). Recherches mathématiques sur la loi d'accroissement de la population [Mathematical research on the law of population increase]. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles, 18*, 1-41.

Vincent, K. R. (1990). The fragile nature of MMPI code types. *Journal of Clinical Psychology, 46*, 800-802.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Waldman, I. D., & Slutske, W. S. (2000). Antisocial behavior and alcoholism: A behavioral genetic perspective on comorbidity. *Clinical Psychology Review, 20*, 255-287.

Waller, N. G. (1999). Searching for structure in the MMPI. In S. E., Embretson & S. L. Hershberger (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know* (pp. 185-217). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*, 125-146.

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

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

Watson, D. (2005). Rethinking the mood and anxiety disorders: A quantitative hierarchical model for DSM-V. *Journal of Abnormal Psychology, 114*, 522–536.

Wetzler, S. (1990). The Millon Clinical Multiaxial Inventory (MCMI): A review. *Journal of Personality Assessment*, *55,* 445-464.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, Massachusetts: Addison-Wesley.

Williams, M. T., Turkheimer, E., Schmidt, K. M., & Oltmanns, T. F. (2005). Ethnic identification biases responses to the Padua Inventory for obsessive-compulsive disorder. *Assessment, 12*, 174-185.

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

Wright, B. (1992). *IRT in the 1990s: Which models work best?* Opening remarks in his invited debate with Ron Hambleton, Session 11.05, AERA Annual

Meeting 1992. http://www.rasch.org/rmt/rmt61a.htm Retrieved June 4, 2008

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

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

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

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7,* 168-190.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551-563.

APPENDIX A

TABLES

Table 1

Demographic Information

Variable	*n*	Valid %	%
Age			
18-20	91	13	13
21-30	309	45	45
31-40	158	23	23
41-50	74	11	11
51-60	35	5	5
61-70	11	2	2
71 +	2	< 1	< 1
Missing	8		1
Gender			
Male	272	40	40
Female	415	60	60
Transgender	1	< 1	< 1
Missing	0		0
Ethnicity			
White	470	78	68
Latino	68	11	9
Asian	22	4	3
Black	14	2	2
American Indian	5	1	1
Multicultural	10	2	2
Other	12	2	2
Missing	87		13
Total	688	100	100

Table 2

Diagnostic Information

Variable	*n*	Valid %	%	*M*	*SD*
Presenting Concern					
Mood	173	26	25		
Anxiety	152	23	22		
Health	45	7	7		
Other	370	55	54		
Missing	13		2		
Diagnosis					
Mood Disorder	165	32	24		
Anxiety Disorder	134	26	20		
Somatoform Disorder	12	2	2		
Other Diagnosis	264	51	38		
No Diagnosis	72	14	11		
Missing	173		25		
MMPI-2					
Hypochondriasis	142			57.50	11.29
Depression	142			62.30	13.30
Hysteria	142			57.72	12.15
Psychopathic Deviate	142			60.56	12.17
Masculinity-Femininity	142			51.59	10.33
Paranoia	142			58.19	13.07
Psychasthenia	142			66.09	12.31
Schizophrenia	142			65.43	12.21
Mania	142			57.41	12.78
Social Introversion	142			55.36	12.02

Table 3

Model Fit with Robust Limited-Information Weighted Least Squares (WLSMV) and Full-Information Marginal Maximum Likelihood (MLR) Estimation

	WLSMV			MLR			
Model	CFI	TLI	RMSEA	ln*L*	Free Parameters	AIC	BIC
Simple Structure							
Orthogonal	0.47	0.51	0.35	-16109.99	120	32459.9	33004.0
Oblique	0.94	0.98	0.07	-15752.04	123	31750.0	32307.7
Complex Structure							
Orthogonal	0.88	0.96	0.11	-15670.95	162	31665.9	32400.3
Oblique	0.96	0.99	0.06	-15641.38	165	31612.7	32360.8
Bifactor Structure							
Orthogonal	0.96	0.99	0.06	-15969.98	145	32229.9	32887.3
Oblique	0.99	0.99	0.06	-15647.10	147	31588.2	32254.6

Note. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; ln*L* = log-likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 4

Standardized Loadings onto Orthogonal and Oblique Simple Structure Models

		Factor					
		Orthogonal			Oblique		
Scale	Item	Dep	Anx	Som	Dep	Anx	Som
DEP	9	.64			.67		
	16	.88			.88		
	17	.91			.91		
	18	.79			.80		
	35	.81			.82		
	50	.82			.83		
ANX	1		.69			.72	
	12		.80			.82	
	19		.73			.77	
	38		.64			.69	
	45		.80			.81	
	49		.48			.51	
PHOB	8		.74			.75	
	28		.73			.74	
	31		.74			.74	
	43		.70			.71	
	47		.71			.72	
SOM	2			.71			.74
	7			.64			.65
	23			.57			.65
	29			.75			.80
	30			.71			.72
	33			.72			.72
	37			.78			.79
	Factor						
	Dep						
	Anx	.00			.79		
	Som	.00	.00		.61	.83	

Note. Dep = depression; Anx = anxiety; Som = somatization; PHOB = phobia.

Table 5

Standardized Loadings onto Orthogonal and Oblique Complex Structure Models

		Factor					
		Orthogonal			Oblique		
Scale	Item	Dep	Anx	Som	Dep	Anx	Som
DEP	9	.60	.12	.09	.55	-.06	.18
	16	.87	.05	.08	.90	-.25	.20
	17	.92	.04	.12	.91	-.28	.26
	18	.76	.12	.08	.74	-.04	.10
	35	.81	.00	.00	.80	.00	.00
	50	.79	.15	.09	.73	-.04	.15
ANX	1	.39	.36	.39	.14	.07	.55
	12	.48	.51	.26	.28	.33	.28
	19	.63	.37	.12	.52	.27	.09
	38	.49	.20	.39	.29	-.13	.57
	45	.40	.60	.27	.18	.43	.28
	49	.29	.21	.30	.12	.01	.42
PHOB	8	.00	.79	.00	.00	.81	.00
	28	.15	.73	.20	-.02	.73	.08
	31	.28	.71	.11	.13	.75	-.04
	43	.38	.62	.05	.28	.65	-.10
	47	.42	.50	.15	.26	.38	.17
SOM	2	.17	.25	.63	-.15	-.14	.91
	7	.00	.00	.61	.00	.00	.59
	23	.32	.24	.43	.06	-.11	.69
	29	.22	.36	.59	-.12	.06	.79
	30	.15	.21	.65	-.18	-.20	.97
	33	.14	.16	.67	-.19	-.25	.98
	37	.35	.11	.68	.03	-.41	1.07
	Factor						
	Dep						
	Anx	.00			.57		
	Som	.00	.00		.63	.81	

Note. Dep = depression; Anx = anxiety; Som = somatization; PHOB = phobia.

Table 6

Standardized Loadings onto Orthogonal and Oblique Bifactor Structure Models

		Factor							
		Orthogonal				Oblique			
Scale	Item	Int	Dep	Anx	Som	Int	Dep	Anx	Som
DEP	9	.52	.37			.41	.53		
	16	.65	.61			.79	.40		
	17	.70	.60			.86	.40		
	18	.61	.49			.59	.51		
	35	.62	.51			.55	.64		
	50	.66	.57			.52	.71		
ANX	1	.72		-.05		.41		.58	
	12	.79		.15		.47		.64	
	19	.76		.05		.58		.52	
	38	.72		-.21		.53		.43	
	45	.77		.21		.38		.72	
	49	.51		-.07		.30		.39	
PHOB	8	.64		.56		.20		.79	
	28	.62		.53		.21		.72	
	31	.66		.44		.27		.71	
	43	.65		.38		.34		.63	
	47	.68		.23		.38		.60	
SOM	2	.57			.41	.26			.66
	7	.46			.46	.17			.61
	23	.59			.20	.41			.48
	29	.63			.43	.22			.73
	30	.53			.47	.24			.64
	33	.48			.58	.20			.67
	37	.61			.50	.41			.66
	Factor								
	Int								
	Dep	.00				.00			
	Anx	.00	.00			.00	.70		
	Som	.00	.00	.00		.00	.45	.73	

Note. Dep = depression; Anx = anxiety; Som = somatization; Int = internalizing; PHOB = phobia.

Table 7

Standardized Loadings and Communalities for Validity Variables onto the Oblique Simple, Oblique Complex, Oblique Bifactor, and Orthogonal Bifactor Structure Models

	Oblique Simple				Oblique Complex				Oblique Bifactor					Orthogonal Bifactor				
Variable	Dep	Anx	Som	R^2	Dep	Anx	Som	R^2	Int	Dep	Anx	Som	R^2	Int	Dep	Anx	Som	R^2
Concern																		
Mood	.44			.19	.44			.19	.41	.18			.20	.24	.41			.23
Anxiety		.29		.08		.29		.08	.04		.26		.07	.22		.14		.07
Health			.34	.11			.34	.11	.07			.40	.17	.11			.56	.32
Diagnosis																		
Mood	.57			.33	.57			.33	.41	.40			.33	.38	.45			.34
Anxiety		.45		.21		.45		.21	.07		.43		.19	.35		.24		.18
Somatoform			.46	.21			.46	.21	-.01			.57	.32	.20			.68	.49
Mood vs. Anxiety	-.43			.19	-.43			.19	-.36	-.29			.28	-.09	-.61			.39
Mood vs. Anxiety		.17		.03		.17		.03	-.45		.26		.21	-.09		.31		.10
Mood vs. Anxiety	-.75	.56		.40	-.75	.56		.40	-.34	-.78	.64		.44	-.11	-.59	.22		.41

Note. Dep = depression factor; Anx = anxiety factor; Som = somatization factor; Int = internalizing factor.

Table 8

Item Parameter Estimates for the Bifactor Graded Response Model in Logistic Model

		Int	Dep	Anx	Som									
Scale	Item	a_1	a_2	a_3	a_4	d_1	d_2	d_3	d_4	A	B_1	B_2	B_3	B_4
DEP	9	1.22	0.88			1.56	3.19	4.21	5.97	1.50	1.04	2.12	2.80	3.97
	16	2.54	2.40			-2.60	0.15	2.08	4.61	3.49	-0.74	0.04	0.59	1.32
	17	3.27	2.81			-3.19	0.17	2.76	6.21	4.31	-0.74	0.04	0.64	1.44
	18	1.80	1.43			-0.71	1.11	2.36	4.34	2.29	-0.31	0.48	1.03	1.89
	35	1.86	1.52			-1.12	0.88	2.37	4.12	2.40	-0.47	0.37	0.98	1.71
	50	2.05	1.45			-0.47	1.34	2.79	4.25	2.51	-0.19	0.53	1.11	1.69
ANX	1	1.89		-0.12		-1.83	0.20	1.74	3.97	1.90	-0.96	0.10	0.92	2.09
	12	2.38		0.45		1.20	2.81	4.07	6.20	2.42	0.50	1.16	1.68	2.56
	19	2.10		0.13		-0.34	1.36	2.59	4.53	2.11	-0.16	0.65	1.23	2.15
	38	1.96		-0.58		-2.13	-0.21	1.27	3.19	2.04	-1.04	-0.10	0.62	1.56
	45	2.34		0.63		1.29	2.66	3.72	5.16	2.43	0.53	1.09	1.53	2.13
	49	1.07		-0.15		-0.35	0.99	1.96	3.13	1.08	-0.32	0.92	1.82	2.90
PHOB	8	2.16		1.89		2.71	4.58	5.53	7.36	2.87	0.94	1.60	1.93	2.56
	28	1.94		1.64		3.25	4.39	5.36	6.15	2.54	1.28	1.73	2.11	2.42
	31	1.98		1.32		1.52	2.86	4.04	5.06	2.38	0.64	1.20	1.70	2.12
	43	1.80		1.03		0.96	2.12	3.04	4.27	2.08	0.46	1.02	1.46	2.05
	47	1.76		0.59		1.17	2.49	3.55	4.74	1.85	0.63	1.34	1.91	2.55
SOM	2	1.46			1.04	0.82	2.91	4.39	7.13	1.79	0.46	1.63	2.45	3.98
	7	1.11			1.10	0.89	2.50	3.58	5.57	1.56	0.57	1.60	2.29	3.56

23	1.36			0.47	0.06	1.48	2.40	3.66	1.44	0.04	1.03	1.67	2.55
29	1.77			1.22	1.58	3.17	4.46	6.36	2.15	0.74	1.47	2.08	2.96
30	1.38			1.21	0.97	2.55	3.71	5.02	1.83	0.53	1.39	2.03	2.74
33	1.30			1.58	0.89	2.39	3.60	5.40	2.05	0.43	1.17	1.76	2.64
37	1.79			1.46	0.63	2.25	3.75	5.36	2.31	0.27	0.97	1.62	2.32
Concern													
Mood	0.49	0.85			1.26				0.98	1.28			
Anxiety	0.42		0.26		1.30				0.49	2.64			
Health	0.23			1.23	3.26				1.25	2.61			
Diagnosis													
Mood	0.58	1.00			1.00				1.31	0.76			
Anxiety	0.71		0.48		1.17				0.85	1.37			
Somatoform	0.50			1.72	4.89				1.79	2.72			
Depression vs. Anxiety	0.21	1.42			-0.36				1.44	-0.25			
Depression vs. Anxiety	-0.18		0.58		0.39				0.37	0.64			

Note. Dep = depression; Anx = anxiety; Som = somatization; Int = internalizing; *a* = item discrimination; *d* = item intercept; *A* = multidimensional discrimination parameter (MDISC); *B* = multidimensional difficulty parameter (MDIFF).

APPENDIX B

FIGURES

Figure 1. Example of an item characteristic curve (ICC).

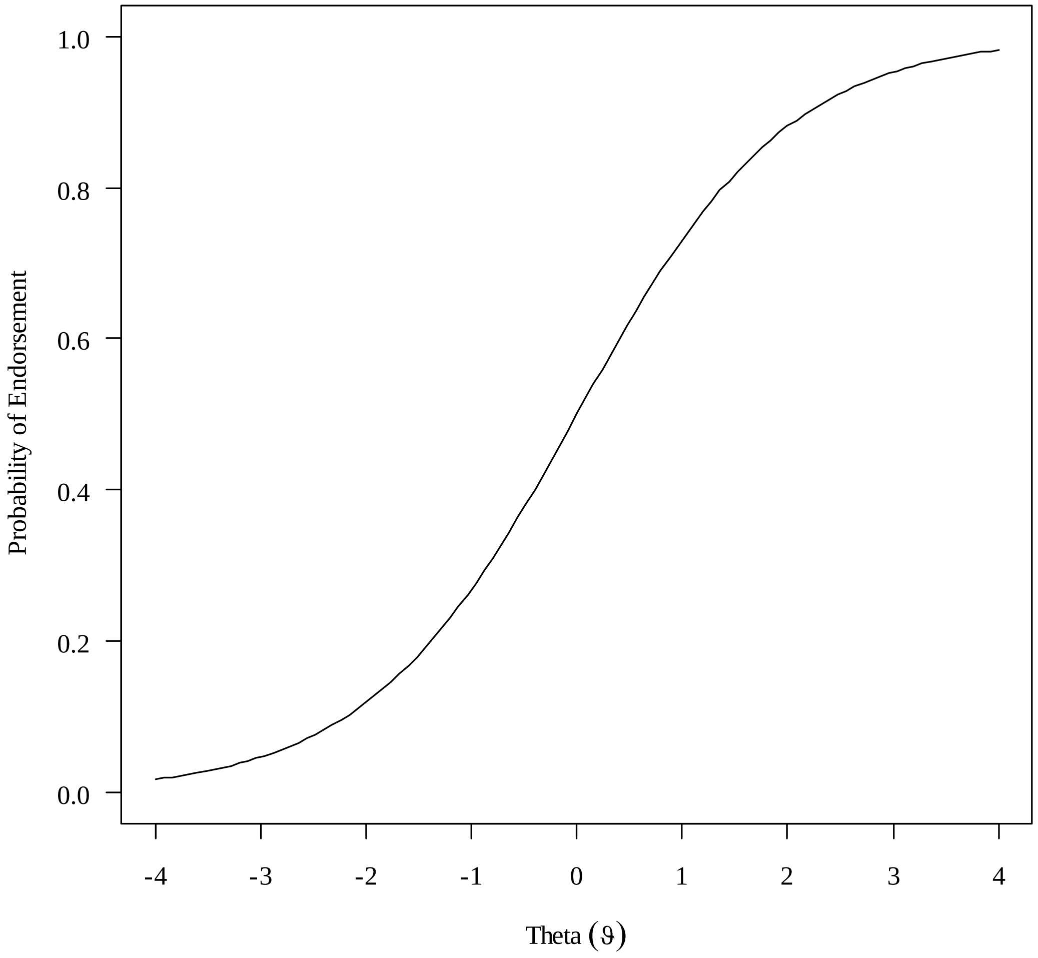


Figure 2. Item characteristic curves (ICCs) for three items fitting the Rasch or one-parameter logistic model.

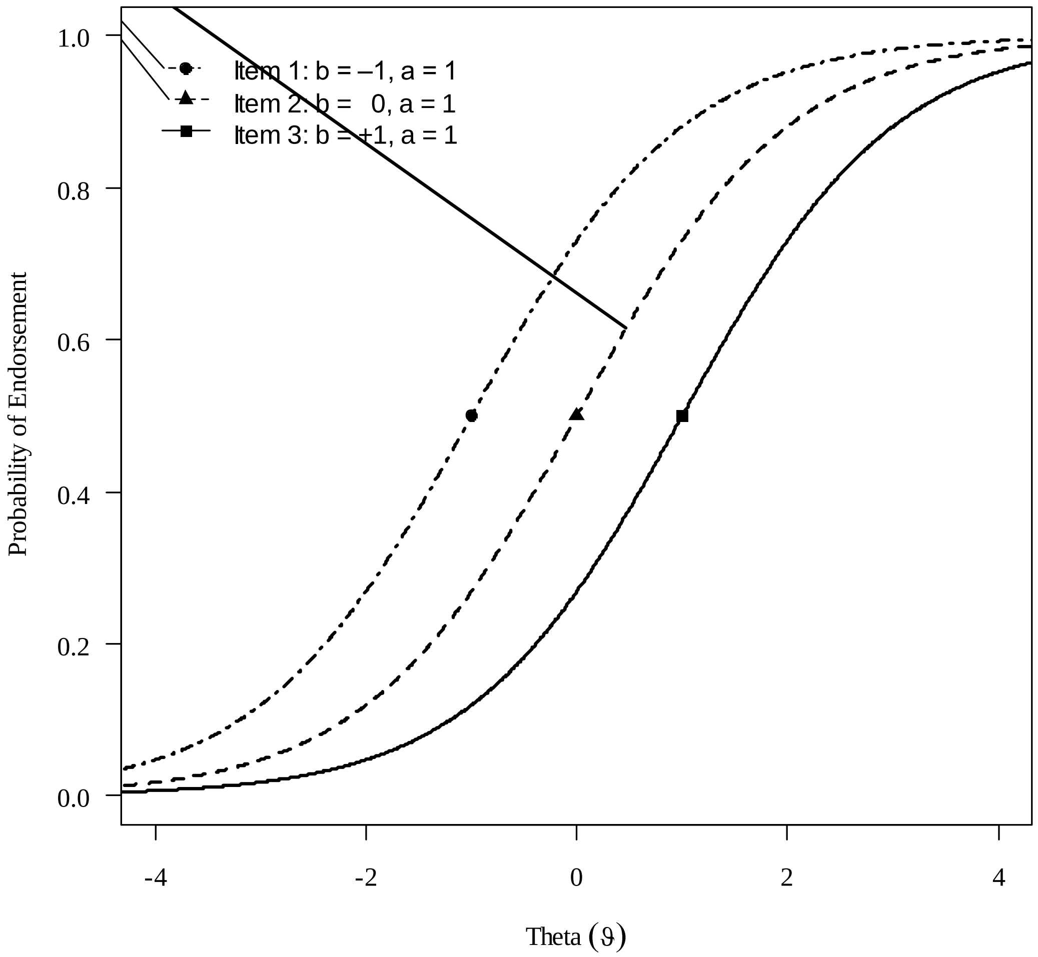


Figure 3. Item characteristic curves (ICCs) for three items fitting the two-parameter logistic (2PL) model.

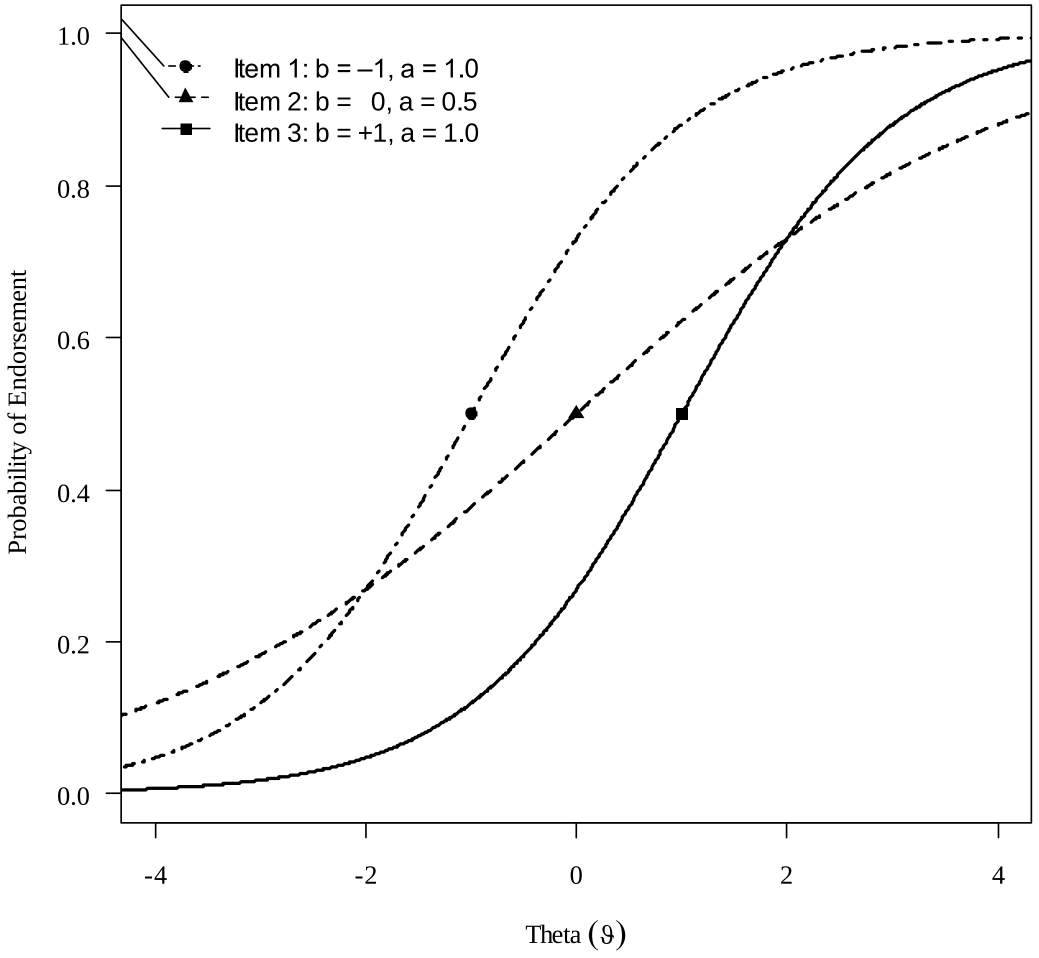


Figure 4. Item characteristic curves (ICCs) for three items fitting the three-parameter logistic (3PL) model.

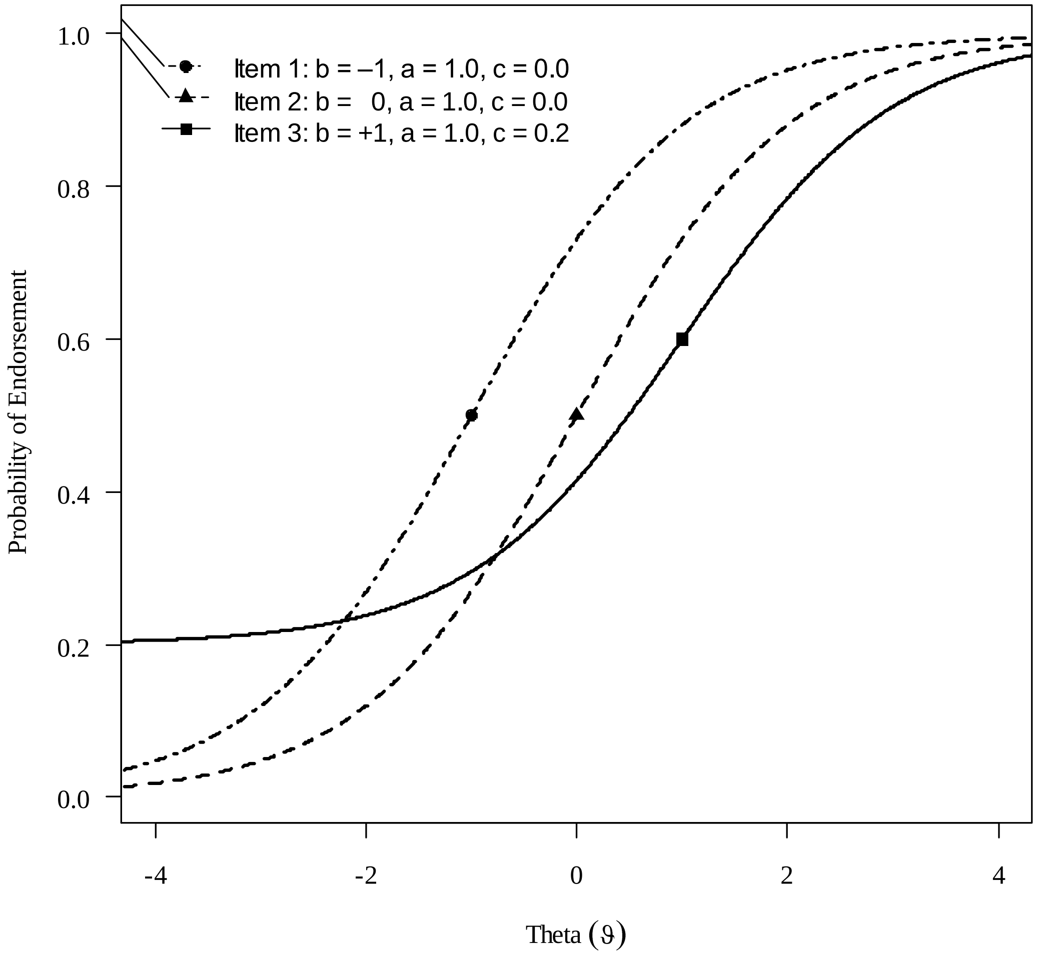


Figure 5. Category response curves for a single item with four response categories fitting the graded response model (GRM).

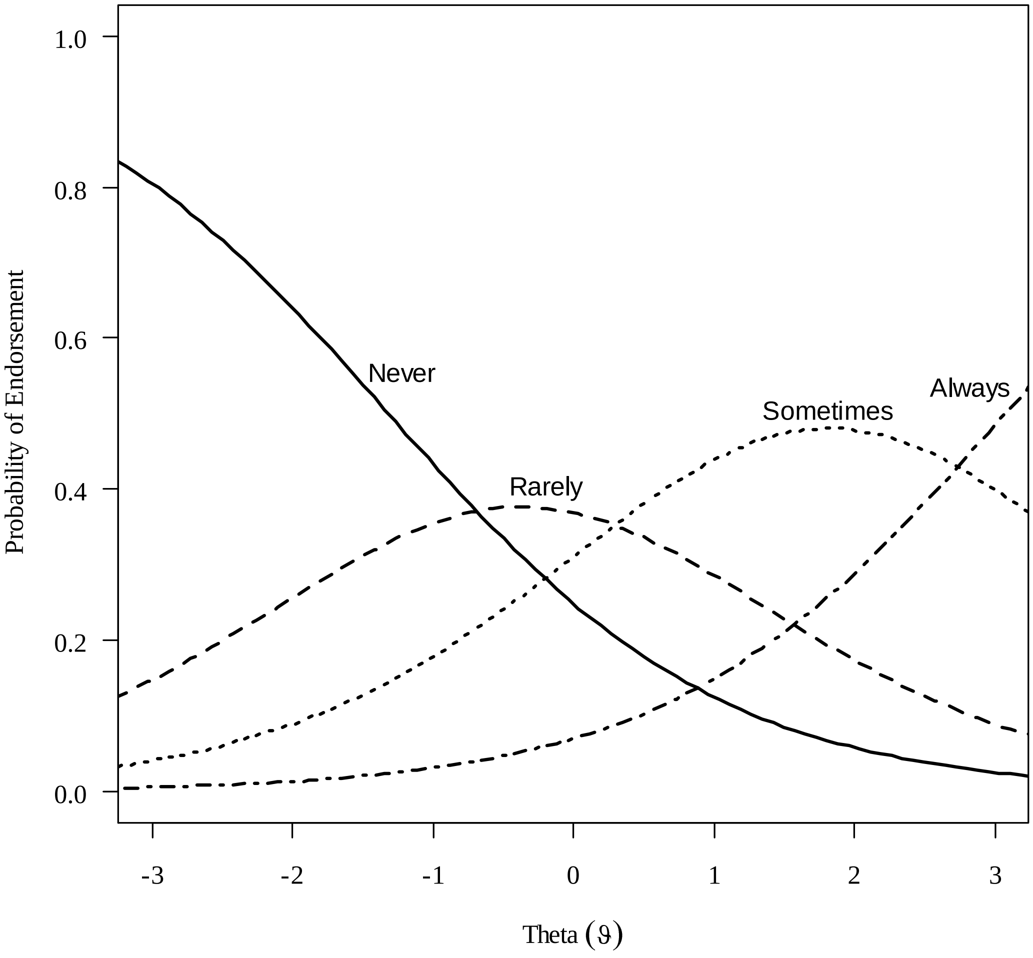


Figure 6. Item characteristic surface for a multidimensional item response model.

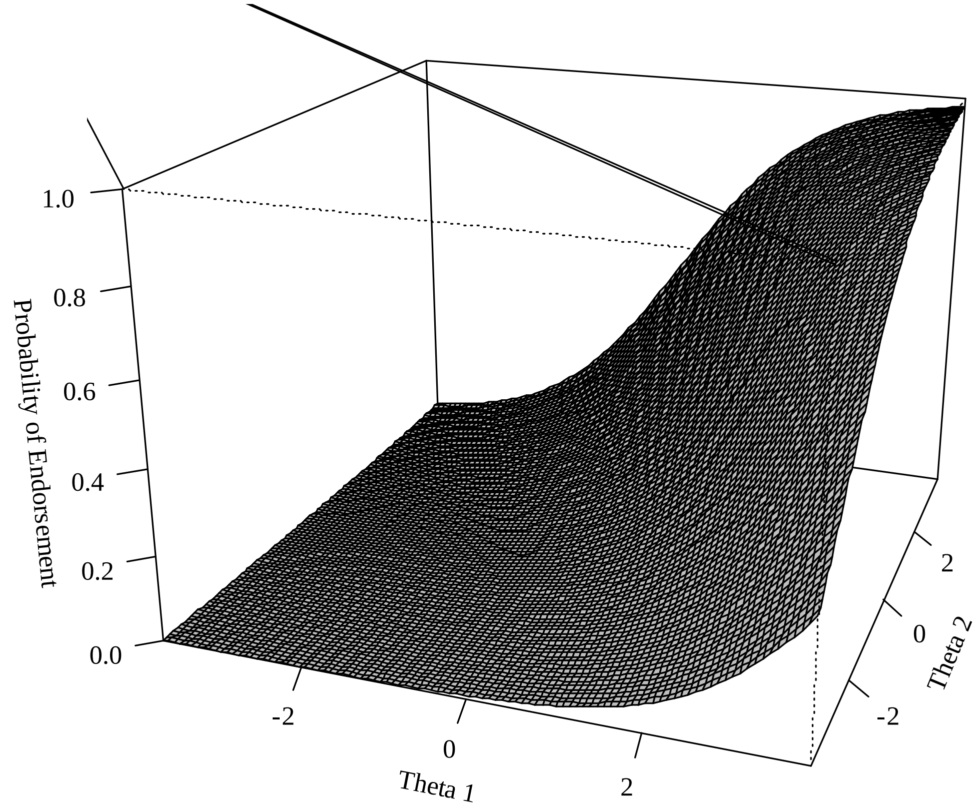


Figure 7. Bifactor model of an internalizing general factor with anxiety and depression domain specific factors.

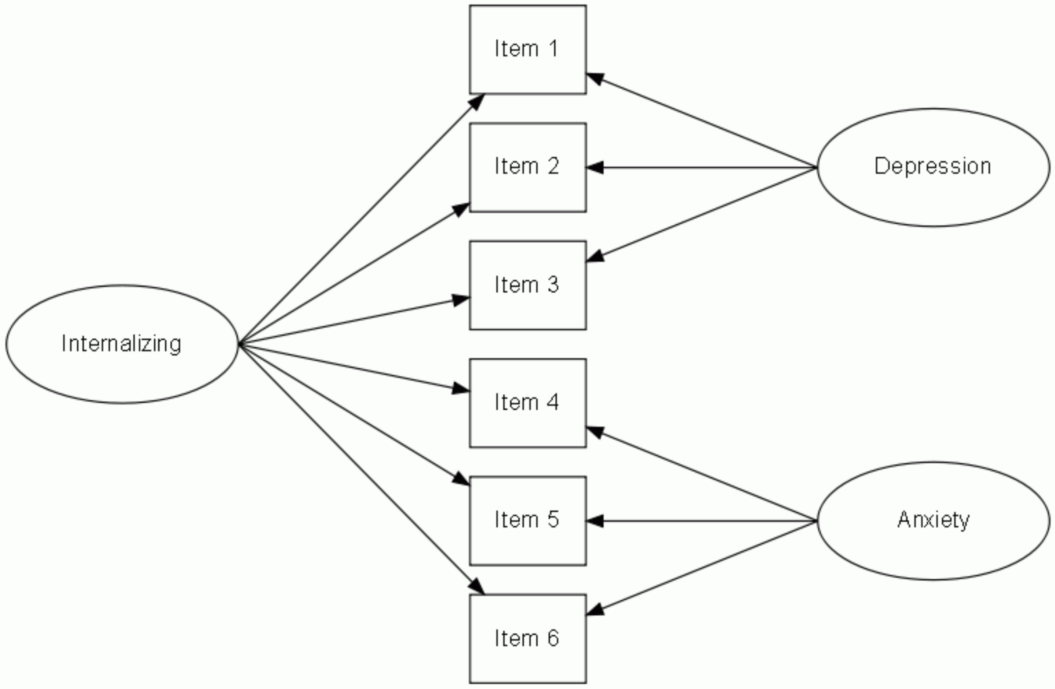


Figure 8. Second-order model of an internalizing higher-order factor with anxiety and depression lower-order factors.

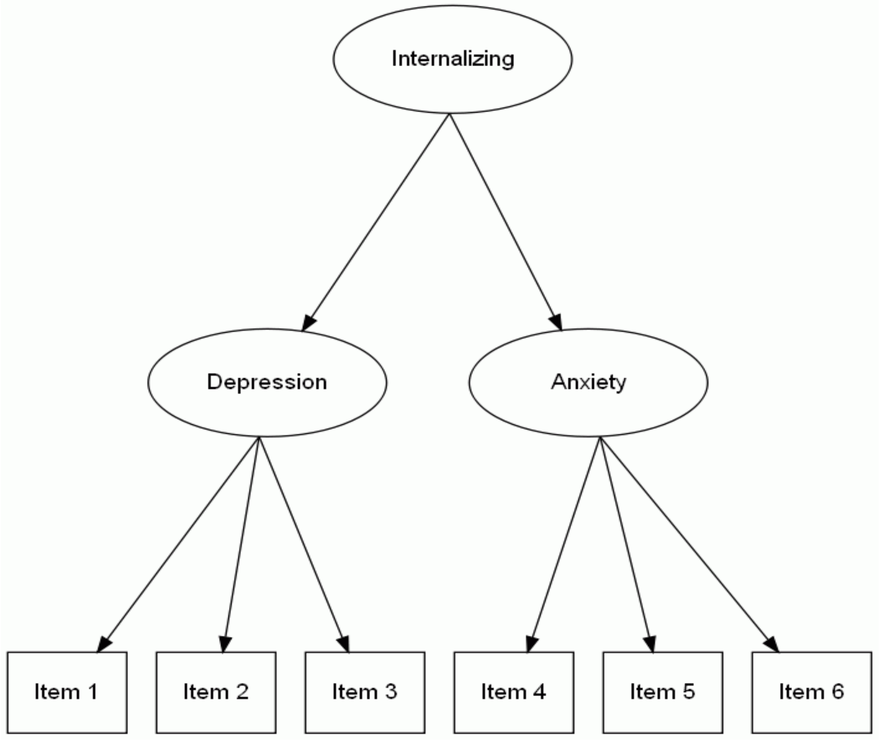


Figure 9. Test information and standard error functions.

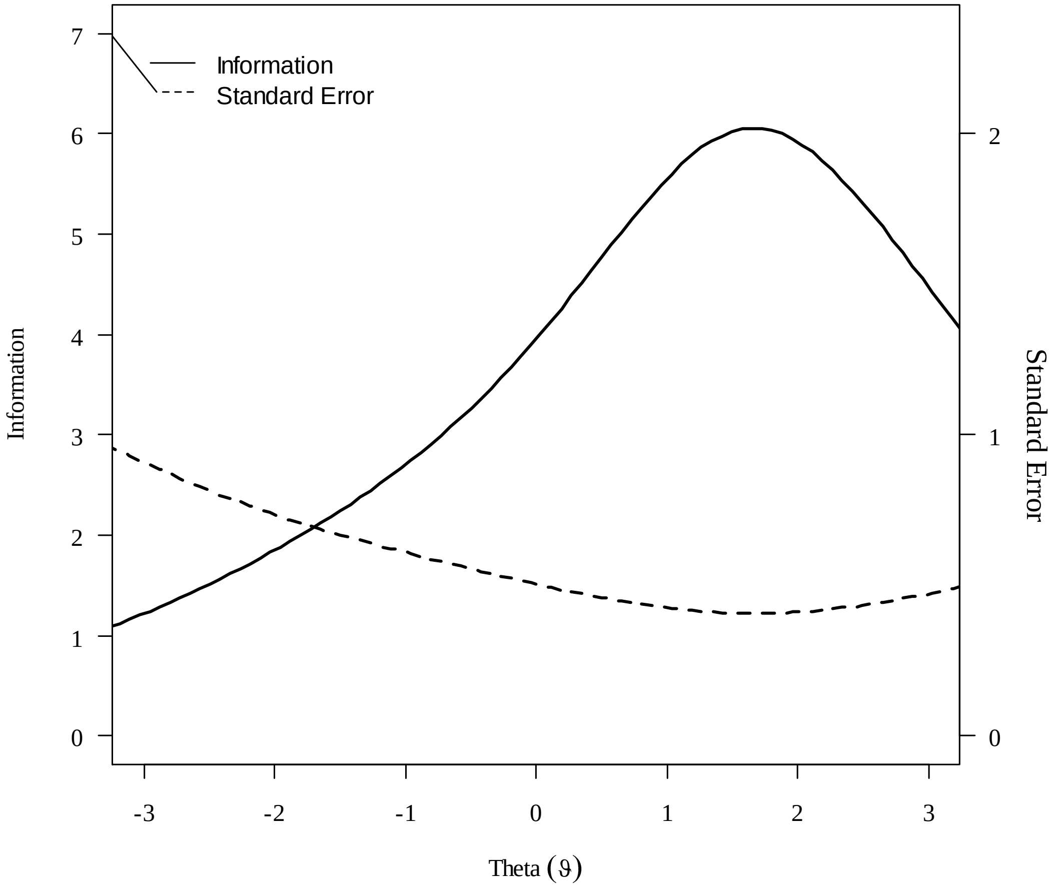


Figure 10. Differential item functioning for the same item with parameters estimated within a male sample and a female sample.

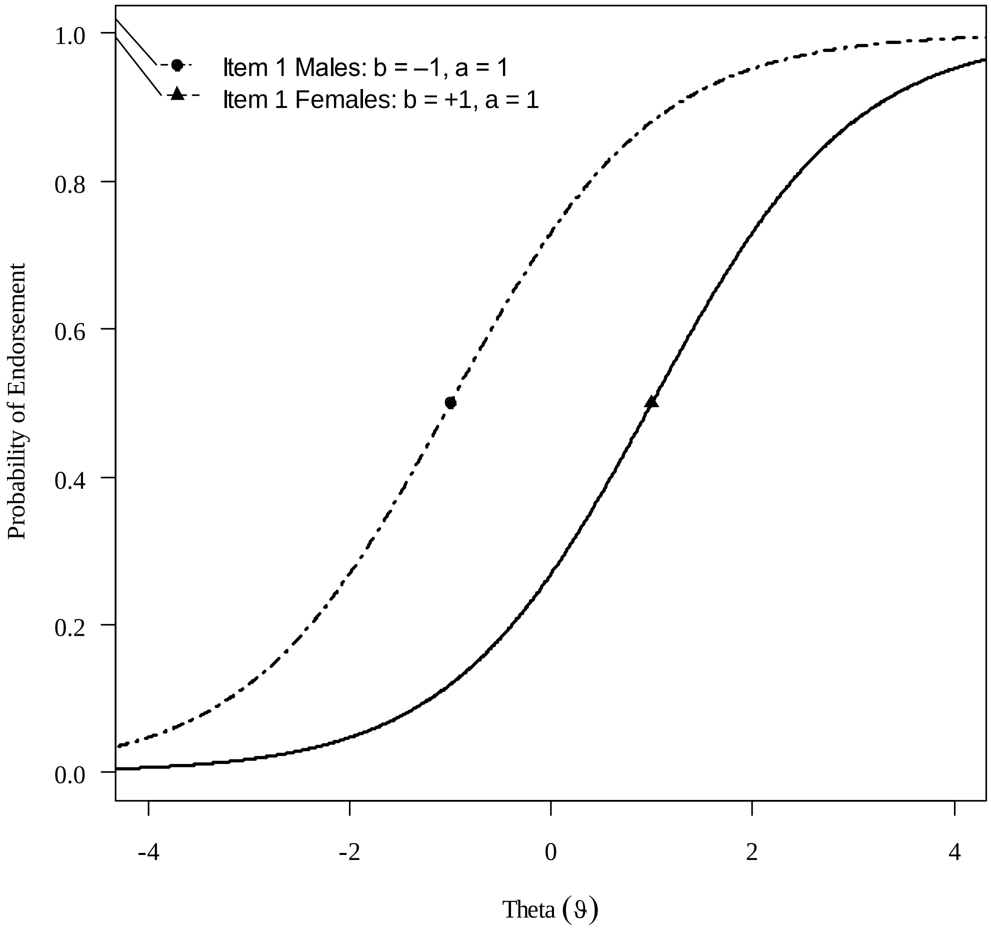


Figure 11. Example of a test characteristic curve (TCC).

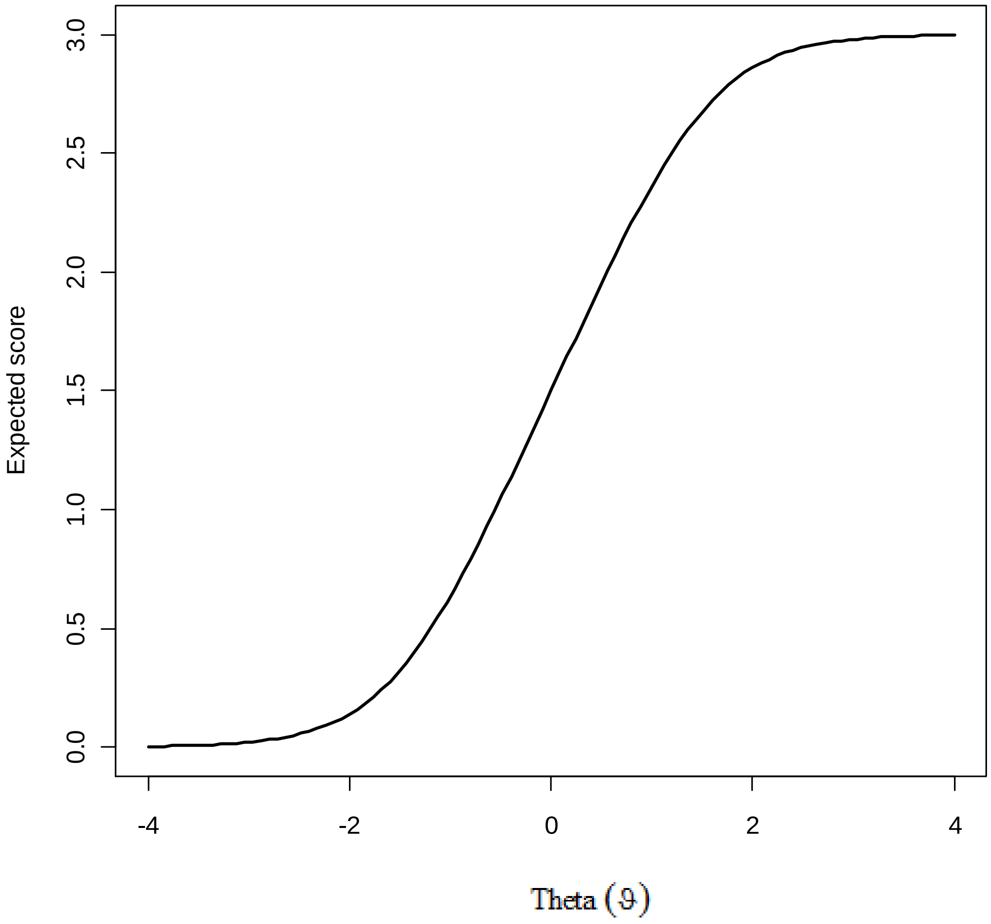


Figure 12. MMPI-2 configural patterns for patients presenting with a mood, anxiety, or health concern. T-scores have a mean or 50 and a standard deviation of 10.

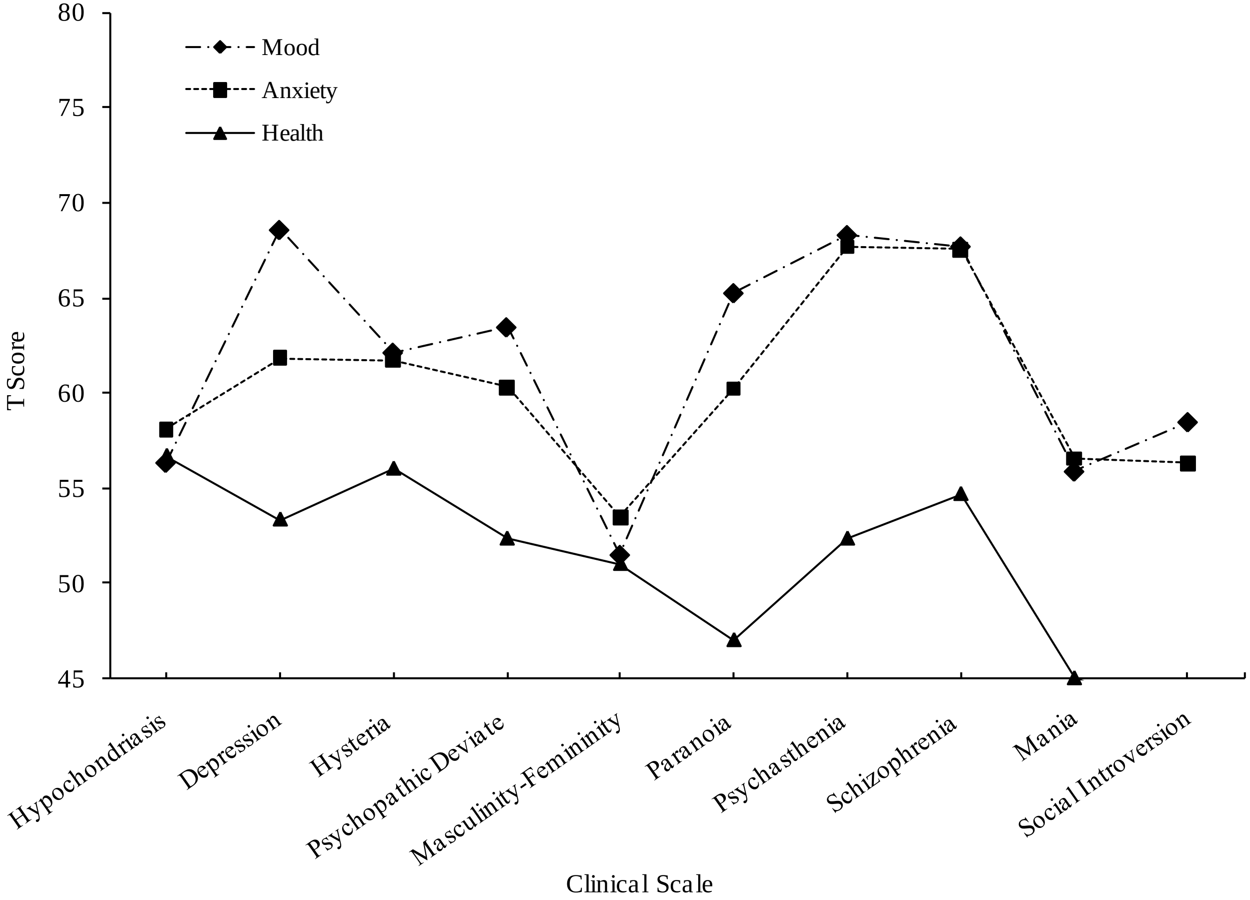


Figure 13. MMPI-2 configural patterns for patients diagnosed with a DSM-IV-TR mood disorder, anxiety disorder, somatoform disorder, or no diagnosis. T-scores have a mean or 50 and a standard deviation of 10.

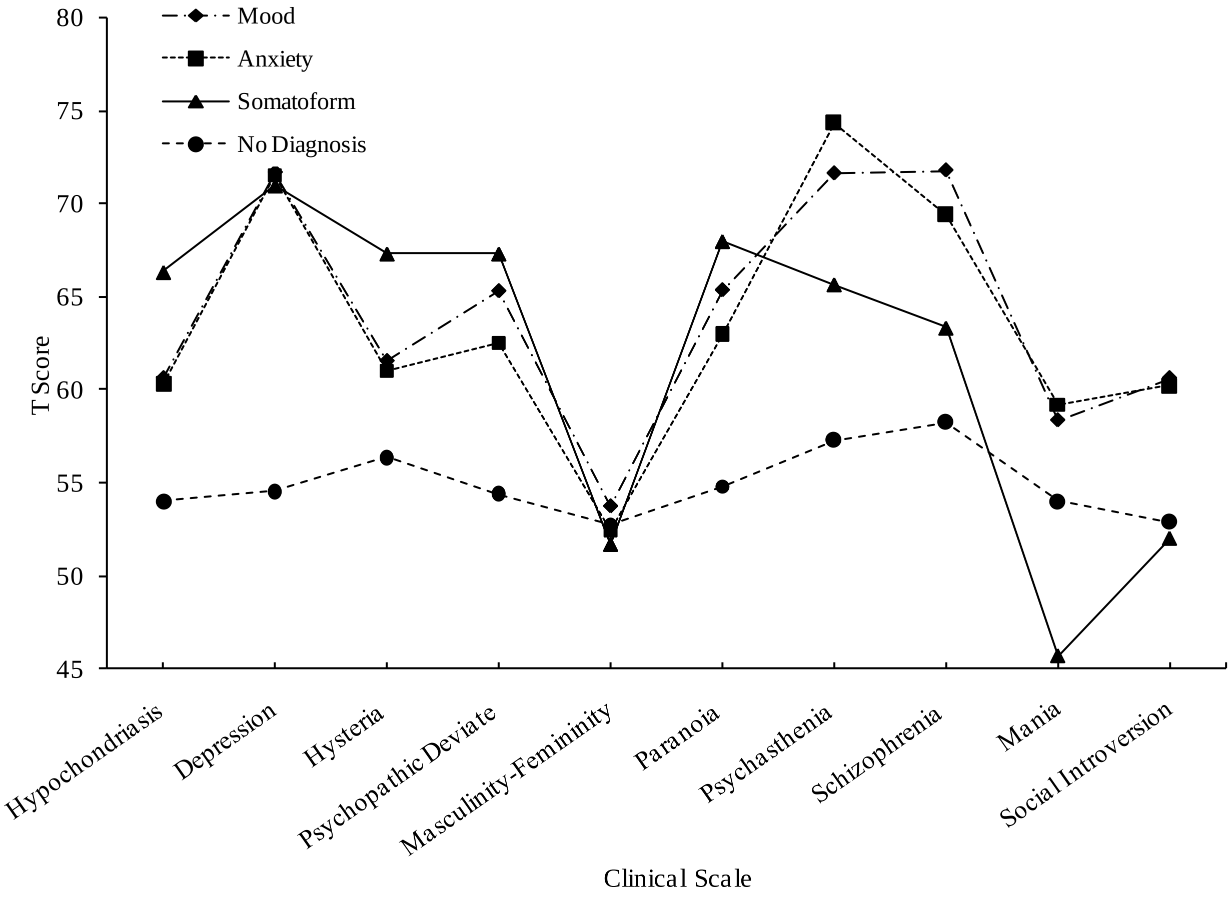


Figure 14. Proposed orthogonal simple, complex, and bifactor structure models. INT = internalizing; SOM = somatization; DEP = depression; ANX = anxiety.

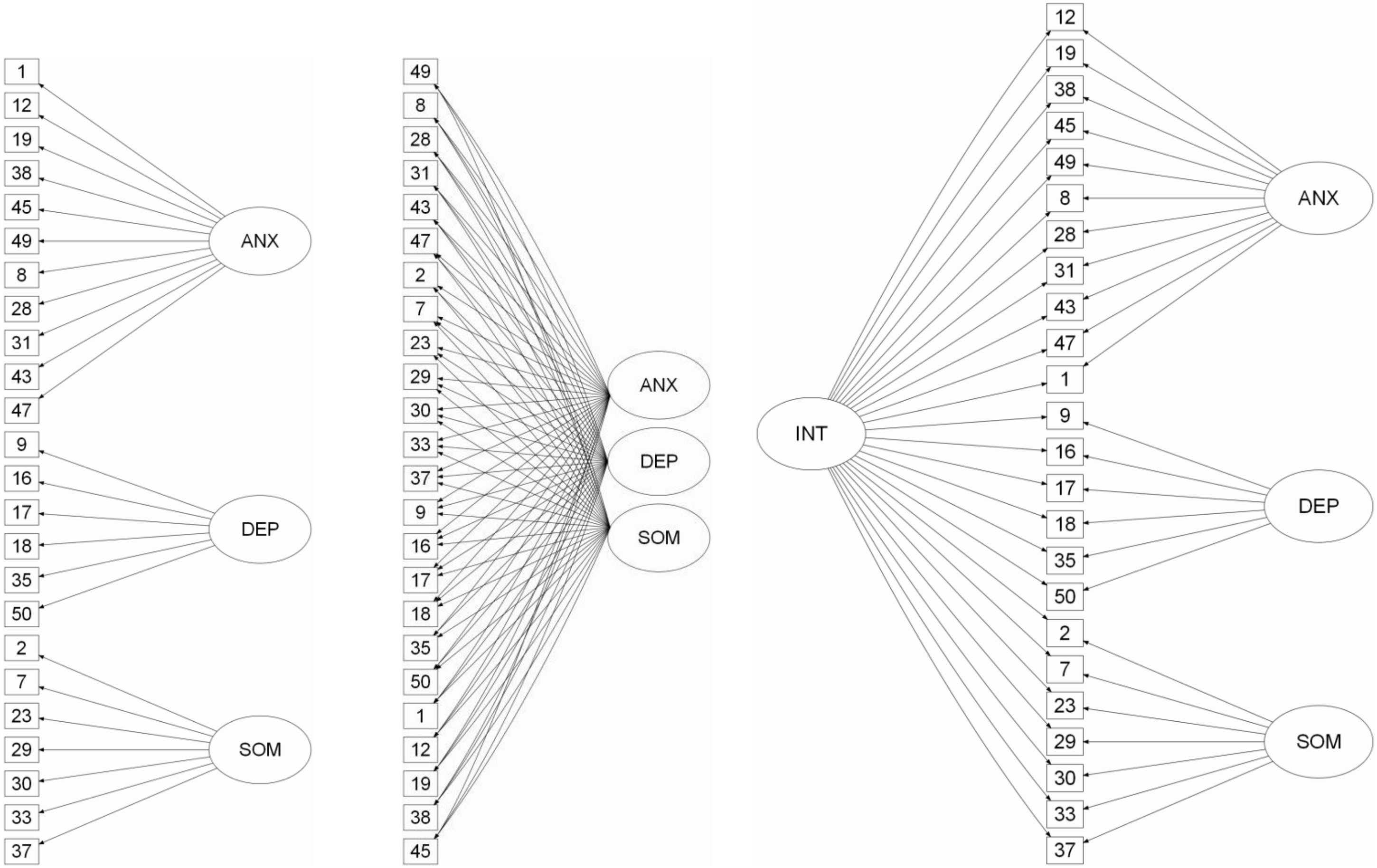


Figure 15. Proposed oblique simple, complex, and bifactor structure models. INT = internalizing; SOM = somatization; DEP = depression; ANX = anxiety.

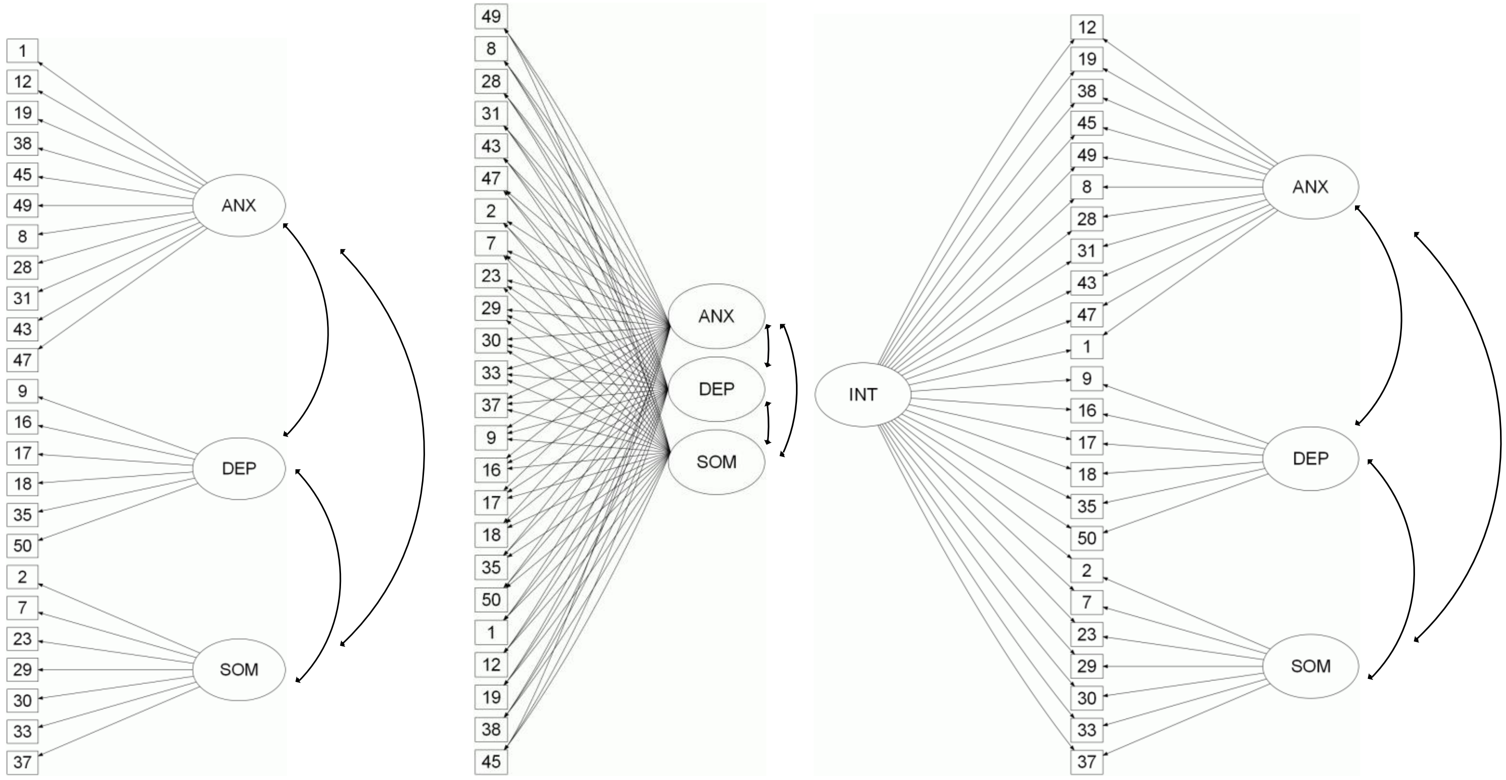


Figure 16. Communalities for validity variables loaded onto the oblique simple, oblique complex, oblique bifactor, and orthogonal bifactor structure models.

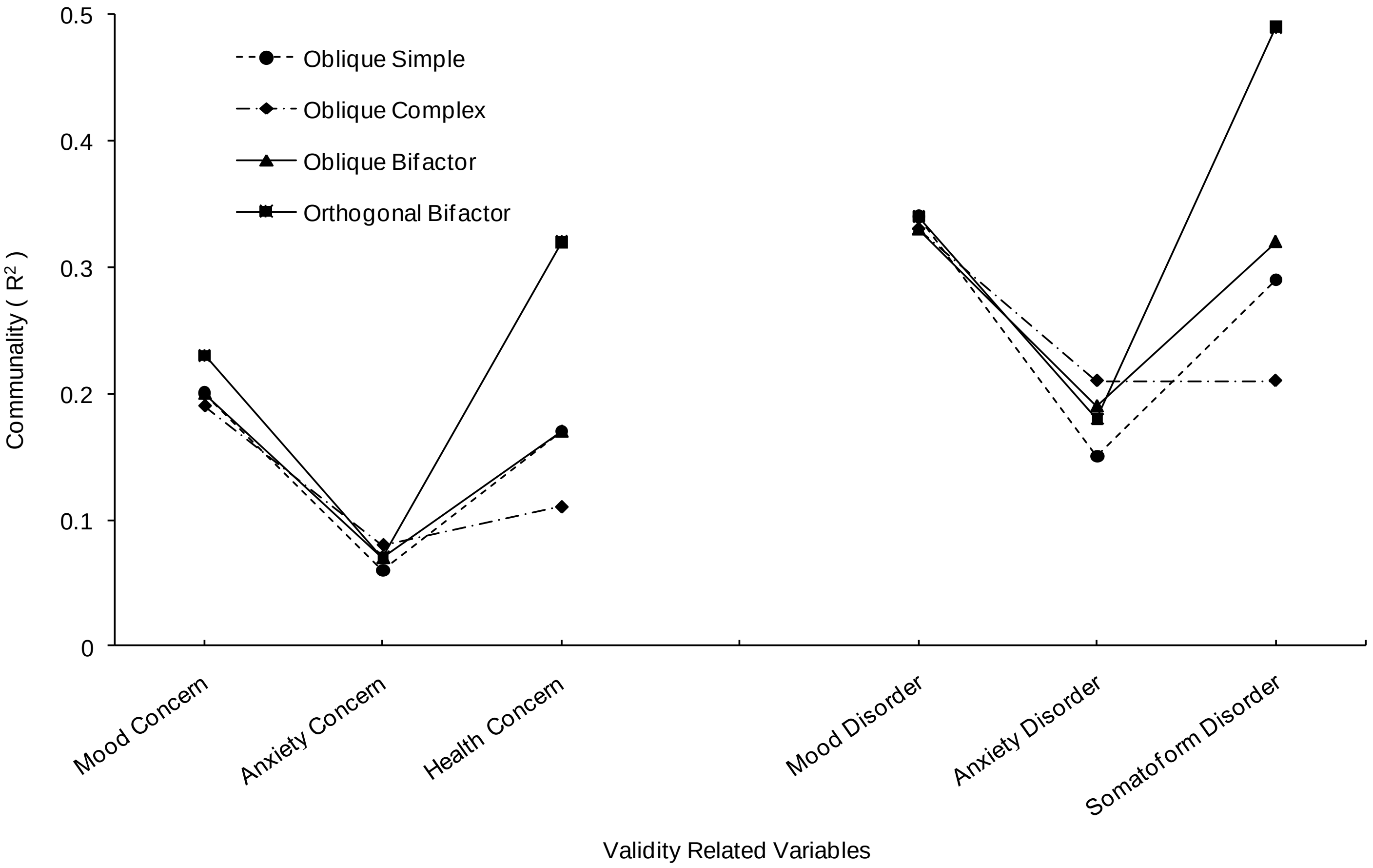


Figure 17. Communalities for the diagnosis of just a mood versus just an anxiety disorder variable loaded onto the oblique simple, oblique complex, oblique bifactor, and orthogonal bifactor structure models. The x-axis displays the factor(s) from each model that the variable was allowed to load onto. (Note that the variable was also allowed to load onto the internalizing factor in the bifactor models).

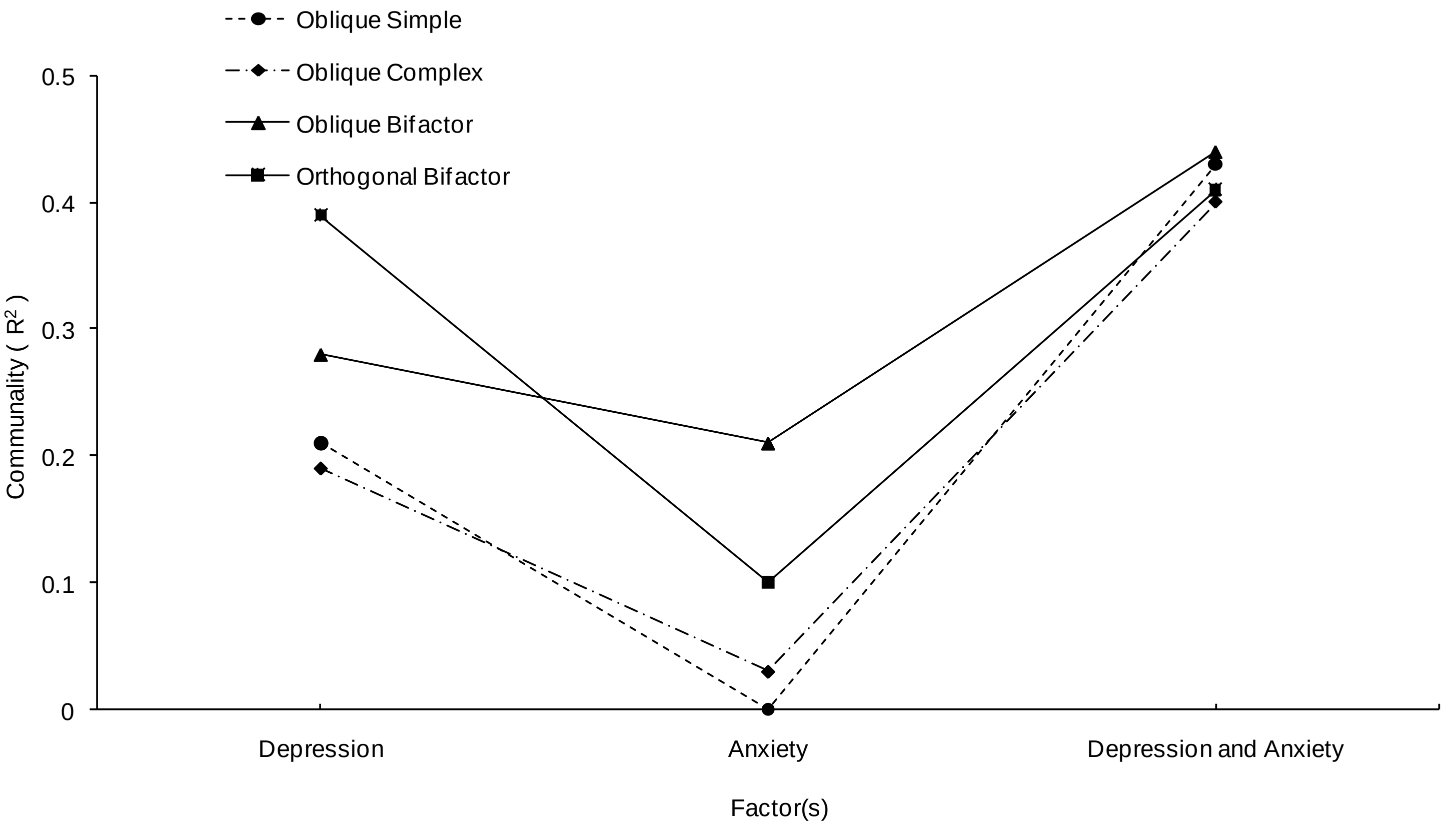


Figure 18. Item characteristic surface plotting the probability of a mood disorder diagnosis by depression and internalizing within an orthogonal bifactor model.

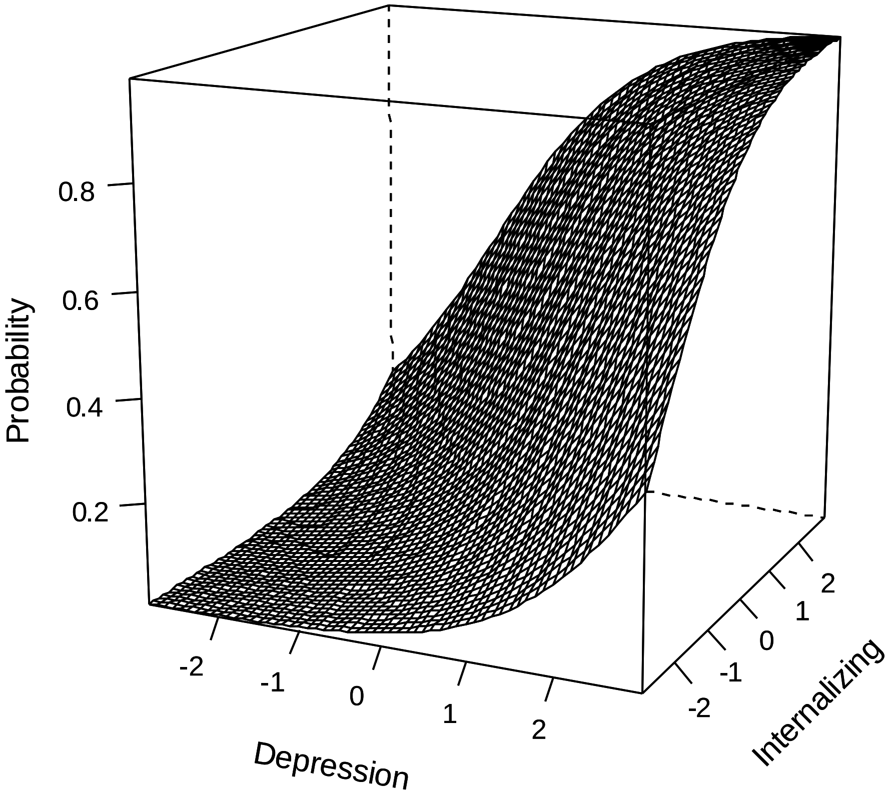


Figure 19. Item characteristic surface plotting the probability of an anxiety disorder diagnosis by anxiety and internalizing within an orthogonal bifactor model.

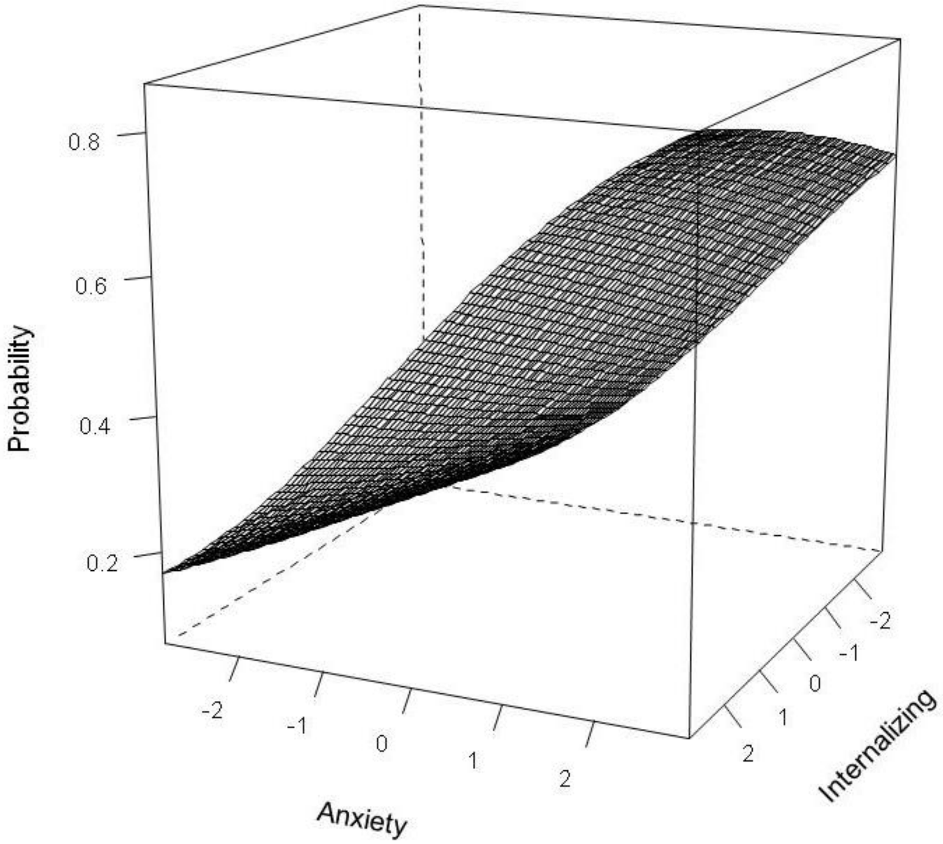


Figure 20. Item characteristic surface plotting the probability of a somatoform disorder diagnosis by somatization and internalizing within an orthogonal bifactor model.

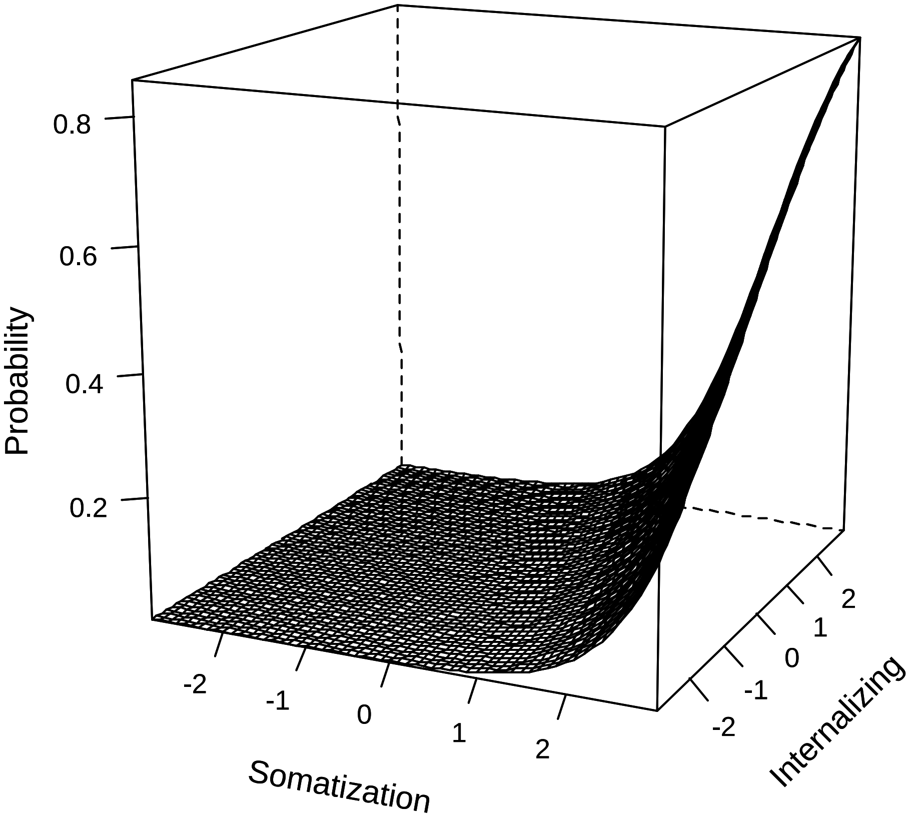


Figure 21. Equi-probable contour for a .50 probability of a mood disorder diagnosis by depression and internalizing within an orthogonal bifactor model.

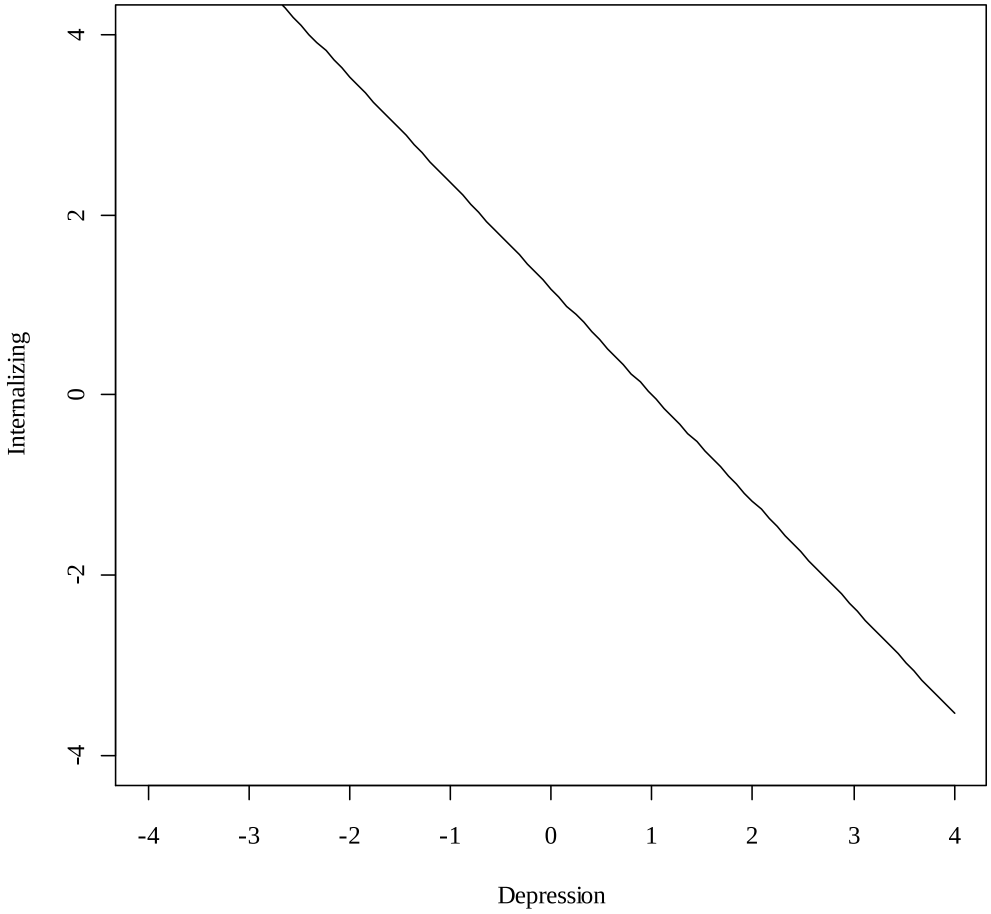


Figure 22. Item characteristic surface plotting the probability of being diagnosed with just an anxiety disorder versus just a mood disorder by depression and internalizing within an orthogonal bifactor model.

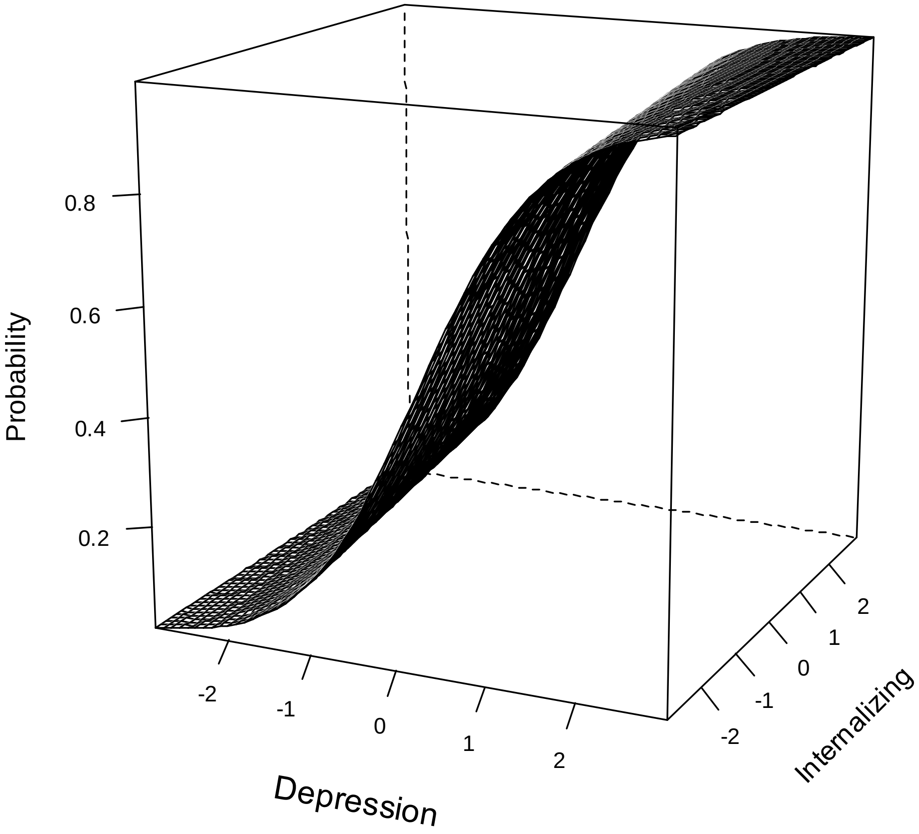


Figure 23. Item characteristic surface plotting the probability of being diagnosed with just an anxiety disorder versus just a mood disorder by anxiety and internalizing within an orthogonal bifactor model.

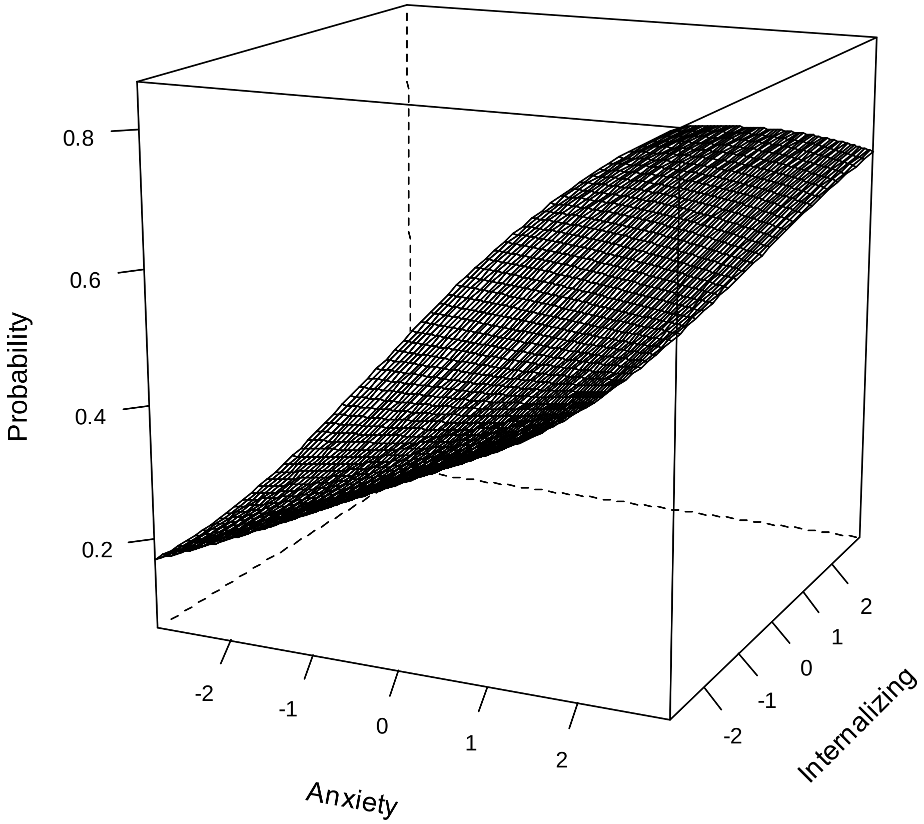


Figure 24. Equi-probable contour for a .50 probability of just an anxiety disorder vs. just a mood disorder diagnosis by depression and internalizing within an orthogonal bifactor model.

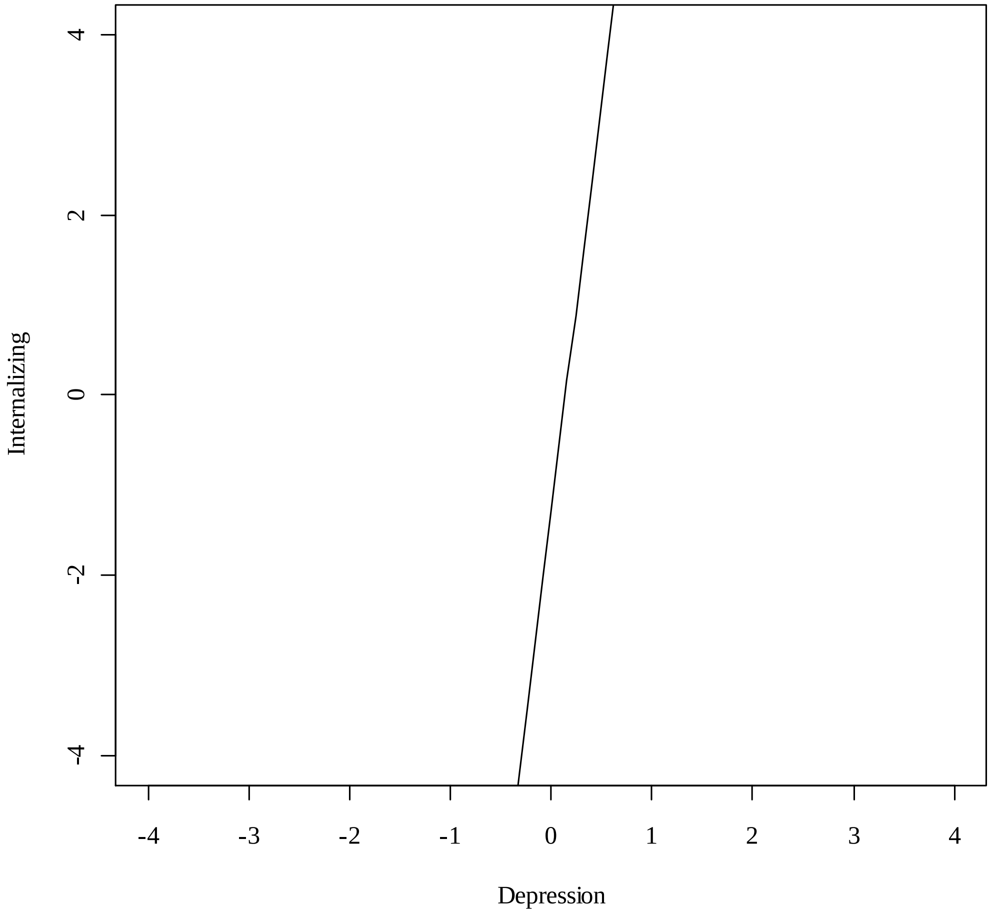


Figure 25. Receiver operating characteristic curve for the diagnosis of a mood disorder by the orthogonal bifactor model and total scores.

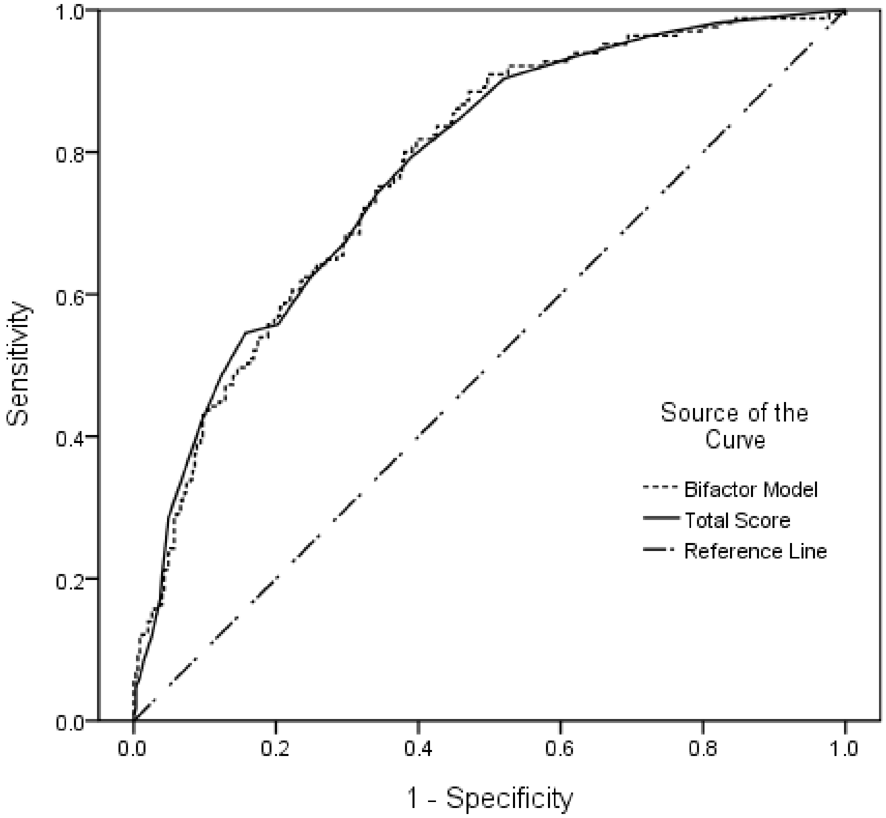


Figure 26. Receiver operating characteristic curve for the diagnosis of an anxiety disorder by the orthogonal bifactor model and total scores.

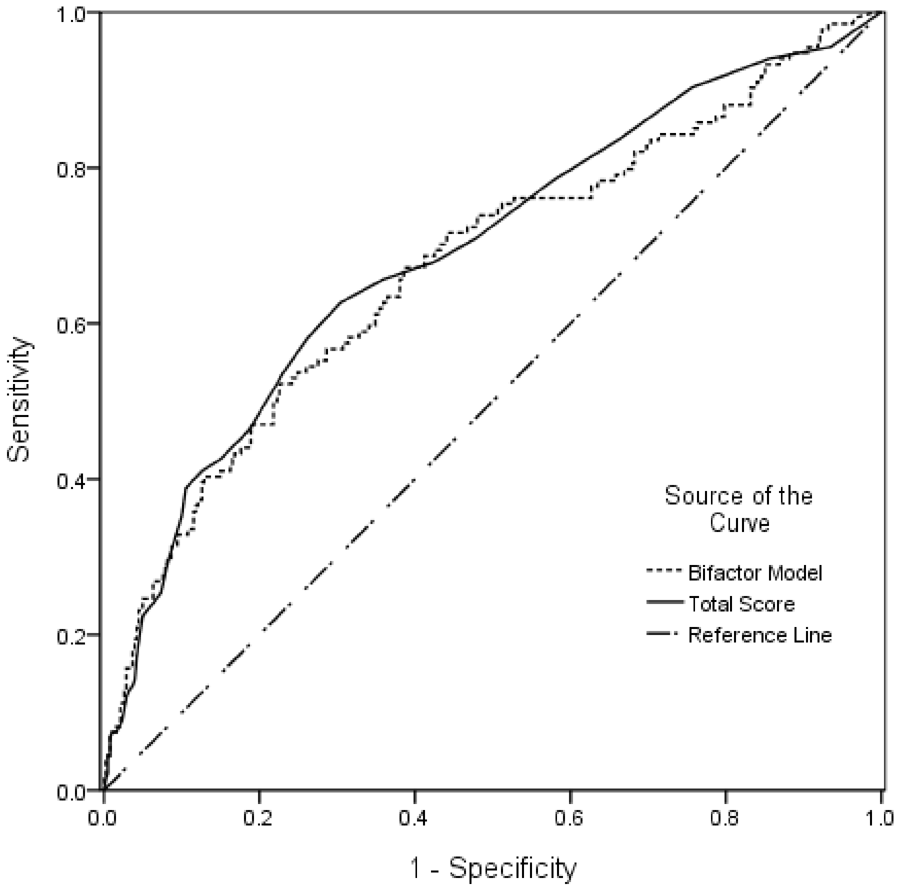


Figure 27. Receiver operating characteristic curve for the diagnosis of a somatoform disorder by the orthogonal bifactor model and total scores.

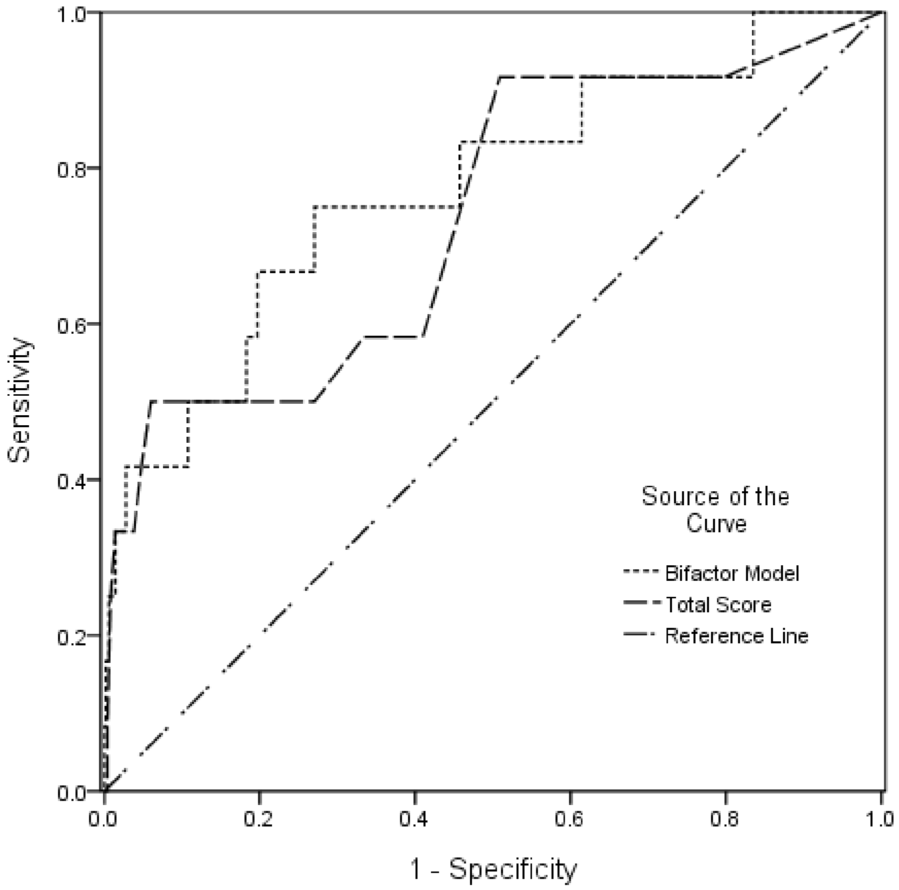


Figure 28. Receiver operating characteristic curve for the diagnosis of just an anxiety disorder vs. just a mood disorder by the orthogonal bifactor model and total scores.

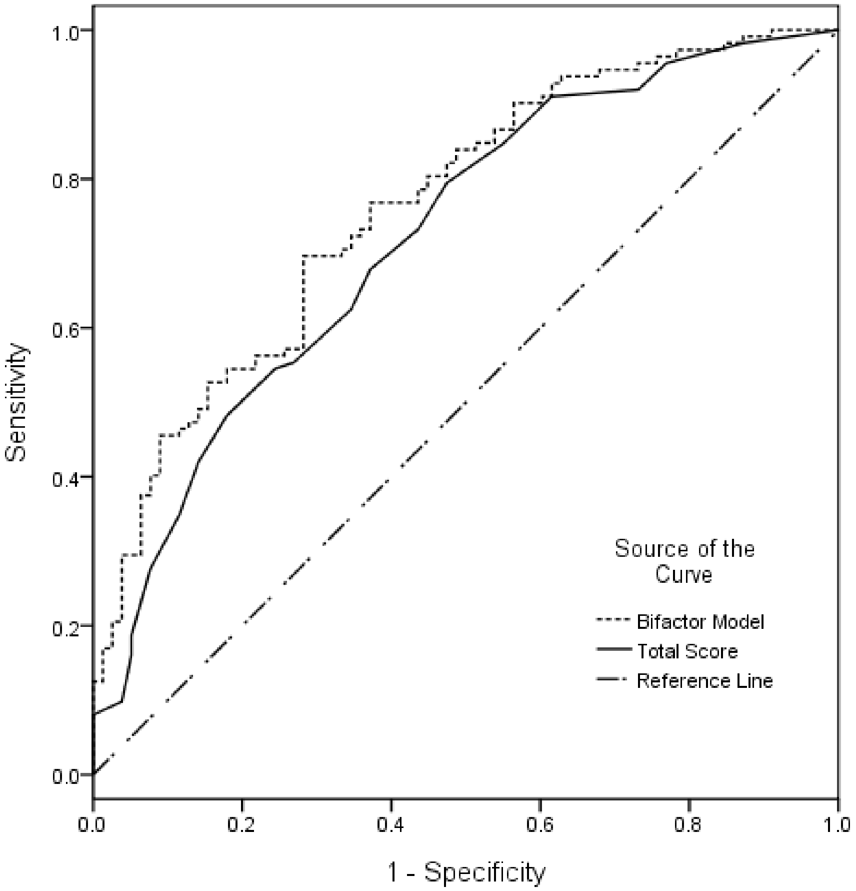
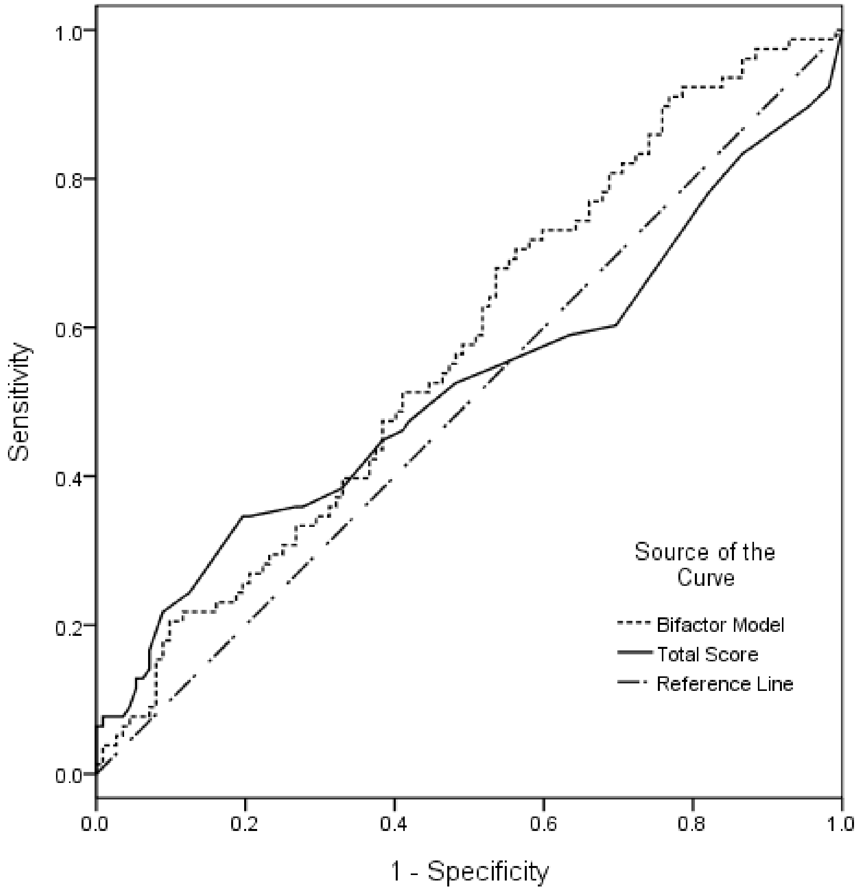


Figure 29. Receiver operating characteristic curve for the diagnosis of just an anxiety disorder vs. just a mood disorder by the orthogonal bifactor model and total scores.



APPENDIX C

DOCUMENTATION OF THE APPROVAL OF

RESEARCH USING HUMAN SUBJECTS

To: Richard Lanyon
PSY

From: Mark Roosa, Chair
Soc Beh IRB

Date: 10/10/2008

Committee Action: **Expedited Approval**

Approval Date: 10/10/2008

Review Type: Expedited F7

IRB Protocol #: 0809003270

Study Title: Item Analysis of Two Psychopathology Screening Inventories

Expiration Date: 10/09/2009

The above-referenced protocol was approved following expedited review by the Institutional Review Board.

It is the Principal Investigator’s responsibility to obtain review and continued approval before the expiration date. You may not continue any research activity beyond the expiration date without approval by the Institutional Review Board.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Soc Beh IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Soc Beh IRB. The new procedure is not to be initiated until the IRB approval has been given.

Please retain a copy of this letter with your approved protocol.