DIDACTIC WORKBOOK

The Added Value of Multidimensional IRT Models

Robert D. Gibbons, Jason C. Immekus and R. Darrell Bock

Center for Health Statistics, University of Illinois at Chicago

June 2007

Corresponding Author:
Robert D. Gibbons Ph.D.
Director, Center for Health Statistics
University of Illinois at Chicago
1601 W. Taylor
Chicago IL 60612
(312) 413-7755 (phone)
(312) 996-2113 (fax)
e-mail:rdgib@uic.edu

Acknowledgements: Supported by Contract 2005-05828-00-00 from the National Cancer Institute. The authors are grateful to A. John Rush M.D. for his contribution to this workbook by providing some of the data sets for the applied examples.

1. Introduction

Patient reported outcomes (PRO) measurements are an integrated component of our health care system. This form of assessment refers to the use of patients' evaluation of their own physical and emotional well-being generally in response to medical care that they are receiving for treatment purposes. PRO measurements that yield psychometrically sound scores (reliable, valid) permit health care providers to evaluate directly the impact of a given treatment from the patient's perspective, as well as determine the efficacy of specific pharmaceuticals or medical devices. This requires a comprehensive, flexible, affordable solution for PROs measurement and management completed with the Health Insurance Portability and Accountability Act. Although self-report inventories of health status serve a different purpose than cognitive tests of achievement or aptitude (typically the focus of the various model-based measurements to be described) the psychometric procedures used for the development, maintenance, and scoring of these tests can be readily adapted to issues that may arise in PRO measurement.

A central issue in PRO measurement is whether obtained scores represent the measured trait (e.g., severity of depression). The empirical question is whether the relationships among scale items can be explained by a single underlying trait (e.g., depression), and are thus unidimensional, or form sub-scales to operationalize the trait's multidimensional structure. Factor analysis is a multivariate statistical procedure used to investigate the data structure of a set of observed variables (e.g., test scores, items). As a data analytic technique, factor analysis has a long, rich history in the dimensionality assessment of psychological measures; over the past century, it has served useful in developing and testing theoretical explanations of human abilities and behavior (Harman, 1976). In these roles, factor analytic results have substantial theoretical

and statistical implications. Although the early use of factor analysis to analyze scores from test batteries is now rarely seen, the common factor model incorporated in structural equation modeling (e.g., confirmatory factor analysis) (Jöreskog, 1969) continues to be widely applied. Analysis of item responses to determine the dimensionality of item banks or putative tests has expanded greatly with the introduction of item response theory (IRT) based methods applicable to dichotomously and polytomously scored item-level data (Bock, Gibbons, & Muraki, 1988; Bock, Gibbons, & Schilling, in press; Mislevy, 1986).

The widespread use of PRO measurements across clinical settings makes it imperative to gauge the extent to which these instruments display evidence of construct validity, i.e., measure the construct intended. To promote an understanding of the use of IRT in PRO measurement, this workbook provides a general introduction to full-information item factor analysis (FIFA) procedures. Both exploratory and confirmatory FIFA procedures are presented, including an introduction to item parameter estimation and estimation of factor scores. The contribution of computerized adaptive testing (CAT) to PRO measurement is also discussed. Applied examples are provided to assist practitioners and researchers to implement IRT-based procedures to develop, maintain, and score instruments for PRO measurements.

The outline of this workbook is as follows. (1) A brief review of PRO measurement, including a description of several PRO measurements that serve as didactic examples within this workbook. (2) A conceptual overview of IRT. (3) An introductory technical description of unidimensional IRT models for dichotomous and graded response data. (4) A presentation of multivariate IRT models for conducting unrestricted, exploratory factor analysis and parameter estimation. (5) An introduction to the confirmatory-based, bifactor IRT model (Gibbons & Hedeker, 1992; Gibbons, Bock, Hedeker, et al., 2007a). (6) The theory and use of CAT in the

context of PRO measurement. The workbook concludes with applied examples to show the application of IRT within the context of PRO measurement.

2. Patient Reported Outcomes

    a. Conceptual overview

PRO measurement has become a critical component of our current health care system. The use of patient reports for assessment purposes within this setting is to provide more efficient, cost-effective, and tailored patient care. For instance, self-report depression inventories provide one method to gauge changes in a patient's depressive symptoms since an alteration in medication(s). In this instance, whether or not the patient's obtained score reflects a decline or increase in depressive symptoms depends on the extent to which the scale's data structure reflects the nature of depression. As such, unless scores show evidence of construct validity, they cannot be effectively used for inferential purposes. That is, PRO instruments that provide scores that represent the measured trait permit health care providers a method to directly evaluate the value of a given treatment from a patient's perspective, as well as the efficacy of specific pharmaceuticals or medical devices.

Regrettably, existing PRO measurements rely heavily on antiquated systems of measurement. That is, the methods generally used to develop and score these scales are based on classical test theory. In particular, a patient's scale score is typically reported in the form of a sum or mean score. Therefore, an examinee selecting the highest category (strongly agree) on a five-point scale for every item on a 20-item depression measure would receive a score of 100, indicating severe depression. Characteristically, these scales (a) include a small fixed-length set of items, (b) require all patients to be measured on the same items, and (c) cannot be readily adapted to meet an individual's testing needs (Hambleton, 1989). Furthermore, the classical test

theory statistics (e.g., Cronbach's alpha, point-biserial correlation) used to investigate the psychometric properties of these instruments are generally sample dependent, indicating that they are influenced by respondent characteristics.

IRT can be used to overcome many of these issues as they pertain to test development, maintenance, and scoring. IRT, previously referred to as latent trait theory, represents a broad class of mathematical models that specify the probability of an item response in terms of item and examinee characteristics (Lord, 1980; Lord & Novick, 1968). As is often the interest in PRO measurement, IRT provides clinicians and researchers working within the context of patient care a method to investigate how a particular examinee will respond to a given item. Advantages of IRT throughout the phases of testing (e.g., development, scoring) include: (a) estimating respondents' trait standing independent of the number of items administered, (b) estimating item parameters (e.g., discrimination, difficulty) independent of the sample of respondents from the larger population, (c) comparing test performance on different test forms, (d) predicting examinee performance on items that have not been administered, and (e) obtaining an estimate of the precision of each test score (Hambleton, 1989; Hambleton & Swaminathan, 1985; Yen & Fitzpatrick, 2006), among many.

To facilitate an understanding of the application of IRT models within PRO measurement, particularly those that deal with multidimensional data, a brief presentation of the instruments that are used as examples in this workbook are presented.

3. Example datasets

For didactic purposes, item-level data from several PRO scales are used to demonstrate the application of IRT. The first is the Psychiatric Diagnostic Screening Questionnaire (PDSQ; Zimmerman & Mattia, 2001), a measure of the most common *Diagnostic and Statistical Manual*

of Mental Disorders Fourth Edition (DSM-IV; American Psychiatric Association, 1994) Axis I disorders encountered in outpatient mental health settings. The second is the Post Traumatic Growth Inventory (PTGI; Tedeschi & Calhoun, 1996), a measure of an individual's changes in self-perception related to an experienced traumatic event (e.g., surviving cancer, rape). Data based on the Jenkins Activity Survey (Jenkins, Rosenman, & Zyzanski, 1972) provides the third example. A brief summary of the development and underlying theory of each scale is provided.

Psychiatric Diagnostic Screening Questionnaire

Zimmerman and Mattia (2001) developed the PDSQ to assess current and recent psychiatric symptoms. It was designed to be administered and scored within the clinician's office before a formal diagnostic evaluation. Development of the PDSQ was based on the following factors that occurred over the past two decades: (a) the need to have standardized instruments to reliably assess published criteria for diagnostic decisions, (b) the development of self-report questionnaires to diagnosis specific DSM-IV disorders, (c) importance of diagnosing comorbidity, or the presence of other disorders beyond the primary disorder, (d) the under-recognition of comorbidity in clinical settings due to inadequate measuring instruments, and (e) the need for clinicians to have instruments to administer during the course of routine diagnostic evaluations (Zimmerman & Mattia, 2001).

The final version of the PDSQ was based on the results of several large-scale studies. The aim of these studies was to develop a clinically useful self-report instrument that yielded scores with evidence of reliability and validity. Scale length was based in consideration of the time constraints present in diagnostic evaluations. The final scale includes 139 items sampled from the following 15 domains: Major Depressive Disorder (MDD), Dysthymia (DYS), Post-traumatic Stress Disorder (PTSD), Bulimia Nervosa (BUL), Obsessive Compulsive Disorder

(OCD), panic disorder (PAN), Mania (MANIA), Psychosis (PSYCH), Agoraphobia (AGOR), Social Phobia (SOC), Alcohol Abuse (ALC), Drug Abuse (DRUG), Generalized Anxiety Disorder (GAD), Somatoform (SOM), and Hypochondriasis (HYP). Scale items are dichotomously scored, with respondents indicating "Yes," score of 1, if the item is applicable, or "No," score of 0, otherwise.

Zimmerman and Mattia (2001) report that the diagnostic performance (sensitivity, specificity, and positive and negative predictive values) of the PDSQ sub-scales in outpatient settings varied in a predictable manner according to the cutoff score. Specifically, as the threshold for case identification increased, sub-scale sensitivity decreased and specificity increased. Furthermore, receiver operating curves were determined for each sub-scale and all areas under the curve were significant (Zimmerman & Mattia, 2001). As far as the authors are aware, no study has tested the PDSQ factor structure. PDSQ items are provided in Table 2. Scale data for the PDSQ is based on the item responses of 3,997 respondents.

Post-Traumatic Growth Inventory

Tedeschi and Calhoun (1996) developed the PTGI to measure changes in individuals' self-perceptions related to an experienced traumatic event (e.g., cancer survivor). The scale's theoretical foundation is based on research reporting that individuals may perceive positive outcomes (e.g., self-perception, philosophy of life) due to a traumatic experience, such as cancer (Collins, Taylor, & Skokan, 1990), combat (Sledge, Boydstun, & Rabe, 1980), or rape (Burt & Katz, 1987; Veronen & Kilpatrick, 1983).

The scale consists of 21 items and requires respondents to rate their experience towards positive growth for each item on the following 6-point scale: 0 = "No Change;" 1 = "Very Small Change;" 2 = "Small Change;" 3 = "Moderate Change;" 4 = "Great Change;" 5 = Very Great

Change." Tedeschi and Calhoun's (1996) factor analytic results supported the following five factors: Relating to Others, New Possibilities, Personal Strength, Spiritual Change, and Appreciation of Life, based on a sample of undergraduate college students (n = 604). Table 1 reports PTGI scale items. Data for the PTGI is based on the responses of 801 breast cancer survivors.

Table 1

PTGI (Tedeschi & Calhoun, 1996) Items

Scale	Item
Relating to Others	1. Knowing that I can count on people in times of trouble.
	2. A sense of closeness with others.
	3. A willingness to express my emotion.
	4. Having compassion for others.
	5. Putting effort into my relationships.
	6. I learned a great deal about how wonderful people are.
	7. I accept needing others.
New Possibilities	8. I developed new interests.
	9. I established a new path for my life.
	10. I'm able to do better things with my life.
	11. New opportunities are avail which wouldn't have been otherwise.
	12. I'm more likely to try to change things which need changing.
Personal Strength	13. A feeling of self-reliance.
	14. Knowing I can handle difficulties.
	15. Being able to accept the way things work out.
	16. I discovered that I'm stronger than I thought I was.
Spiritual Change	17. A better understanding of spiritual matters.
	18. I have a stronger religious faith.
Appreciation of Life	19. My priorities about what is important in life.
	20. An appreciation for the value of my own life.
	21. Appreciating each day.

The Jenkins Activity Survey

The Jenkins Activity Survey (JAS; Jenkins et al., 1972) is a 54-item, self-report measure of Type A behavior (e.g., competitiveness, aggressiveness, haste). It was designed to aid in identifying factors that may contribute to diseases (e.g., heart attacks) in individuals ranging in

age from 25 to 65 years. Item responses are multiple-choice and questions ask respondents to indicate the frequency (e.g., Never, Occasionally, or Almost Always) of one's behavior (e.g., speed, competitiveness) related to specific tasks. Scale scores include an overall Type A behavior in addition to the three following subscales: Speed and Impatience, Job Involvement, and Hard Driving and Competitive. (JAS items are not provided in this workbook due to copyright.) Data for the JAS is based on the item responses of 598 respondents.

4. Conceptualization of Item Response Theory

For those already familiar with traditional methods of educational and psychological testing, an understanding that classical and IRT methods of scoring tests are based on entirely different premises is crucial. Consider the following analogy. Imagine a track and field meet in which ten athletes participate in men's 110-meter hurdles race and also in the men's high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined by both the runner's time and the number of hurdles successfully cleared, *i.e*., not tipped over. On the other hand the high jump is conducted in the conventional way: the cross bar is raised by, say, 2 cm increments on the uprights, and the athletes try to jump over the bar without dislodging it.

The first of these two events is like a traditionally scored objective test: runners attempt to clear hurdles of varying heights which is analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles (or answer the questions). On the high jump, ability is measured by a scale in millimeters and centimeters on the upright and the highest scale position of the cross bar the athlete can clear.

IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until he can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. In IRT, ability is measured by a scale point, not a numerical count.

These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: if hurdles are arbitrarily added or removed, number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if the percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures non-comparable. The same is true of traditional number-right scores of objective tests: scores lose their comparability if item composition is changed.

The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2 cm positions, or if one of those positions were omitted, height cleared is unchanged and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected.

This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of educational, psychological, and patient-reported outcome (PRO) measurement. In all applications of objective testing there is an ever-present need to add, delete, or alter test items in active use. When the measurement methods are based on

classical theory, these changes in the item composition of tests require time consuming and often expensive fieldwork to revise the norms for the test or equate alternative forms of the test. When IRT-based methods are used, items can be moved in and out of tests in the course of routine operational testing without affecting scale score interpretation.

5. Classical Test Theory

Classical test theory grew out of the need for methods of analyzing and scoring multiple-item cognitive tests -- for example, tests of intelligence or educational achievement. In that context there are unambiguous answers to the test exercises, and the obvious measure of a respondent's performance is simply the count of correct responses to the test items. When applied to affective assessment, however, -- for example, to descriptions of personality traits, expressions of opinion, or inventories of behaviors and symptoms -- these methods have several limitations. In these applications, the responses are largely a matter of degree and must be assessed on scales of intensity or frequency. Typical scales consist of multiple ordered categories bearing labels such as "disagree strongly," "disagree," "uncertain," "agree," "agree strongly," or "never," "occasionally," "often," "always." For types of response options, classical test theory has nothing more to offer than assigning successive integer values to the categories and summing these values over items to measure the attribute in question. That is, each item equally contributes to the respondents' assigned score, regardless of the strength of its relationship to the measured trait. Furthermore, no evidence of the appropriateness or optimality of this method of scoring the measuring instrument is provided. The greatest problem with assigned values is that they do not provide for different items having different numbers of categories -- either intentionally or in effect because some categories of some items are rarely or never used. In these situations, for

example, the influence of each item on the summary score depends arbitrarily upon its effective number of categories.

6. Item Response Theory

IRT cuts through these problems by fitting mathematical models that give the probability of response in each item category as a function of parameters characteristic of the item and measurements descriptive of the respondent. For analyzing and scoring responses in two or more ordered categories, these models make possible likelihood-based statistical methods with known optimal properties (Bock & Aitkin, 1981). In particular, they use the total amount of information conveyed in individuals' item response patterns to estimate trait standing. These methods are now in wide use for analyzing and scoring essay tests and open-ended exercises that are graded in ordered categories (Thissen & Wainer, 2001).

Technically, IRT embodies a host of probabilistic models to estimate a respondent's probability of selecting a particular item response category (Lord, 1980; Lord & Novick, 1968). This is facilitated by considering factors related to the item and respondent. Item characteristics generally include discrimination and difficulty parameters. Item discrimination refers to how well an item discriminates between examinees with low and high standing on the underlying latent trait (e.g., depression, post-traumatic growth). Within PRO measurement, item difficulty can be regarded as how likely a particular respondent will endorse an item (i.e., respond "yes" on dichotomously scored item). In some instances, a model that also includes a pseudo-guessing parameter is included to model data for multiple-choice items commonly found on achievement tests (Lord, 1980). Patients' standing on the measured trait, or propensity level, is used in the IRT models to account for the aspect of the individual that contributes to how he/she will respond to a given item.

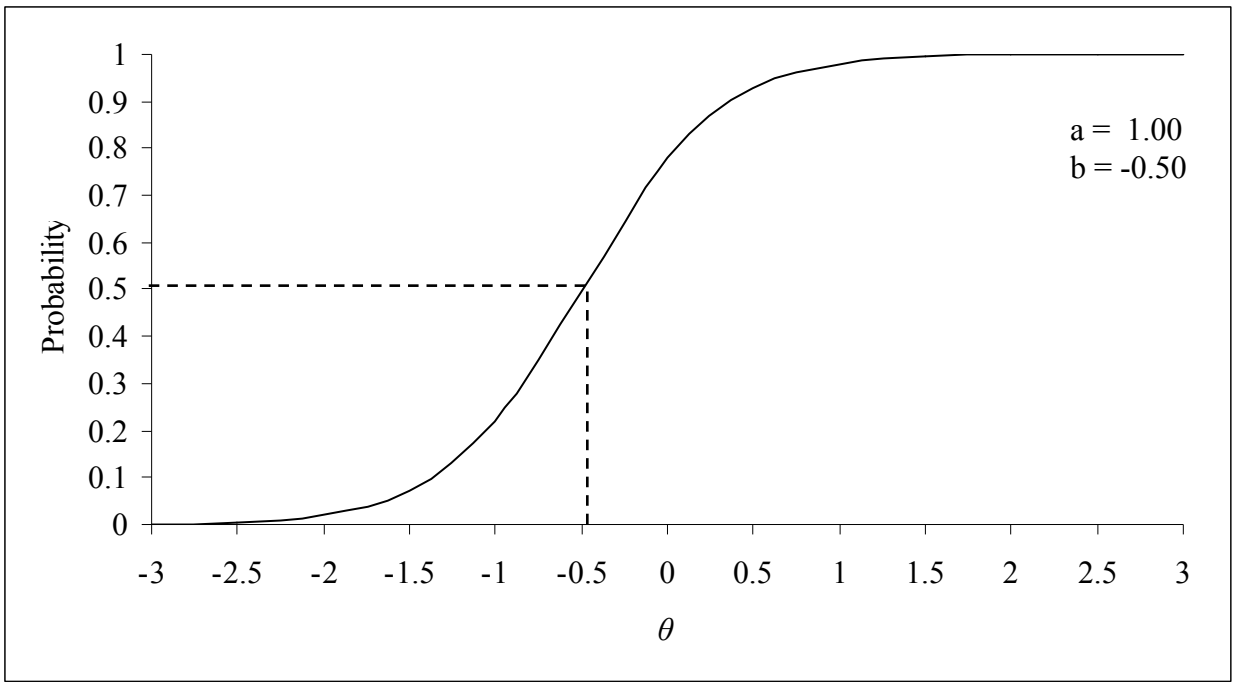
IRT procedures can be applied to a variety of data types. Scale items can be dichotomously (e.g., correct/incorrect, yes/no) or polytomously (e.g., Likert-scored response categories) scored and the categories can be ordered or unordered. Additionally, there is an assortment of IRT models to specify item performance in terms of a single underlying latent trait (e.g., normal ogive, 1- and 2- parameter models). Readers are referred to several informative references to gain an understanding of the available IRT models for dichotomous (Hambleton, 1989; Harris, 1989; Lord, 1980; Lord & Novick, 1968; Yen & Fitzpatrick, 2006), polytomous (Thissen & Steinberg, 1986; Yen & Fitzpatrick, 2006), and multiple-choice (Thissen & Steinberg, 1984) item responses. Several IRT models to handle these data types are presented below.

Item response functions (IRFs), also called *trace lines* (Lazarfeld, 1950), provide a useful graphical description of an item's functioning as modeled in IRT. Figure 1 shows an IRF that models the probability of a positive item endorsement for a dichotomously scored item in terms of an item's discrimination and difficulty parameters, in addition to the underlying latent variable (θ). The latent trait is unobserved and represents a respondent's level of proficiency or propensity. As shown, the IRF models the non-linear relationship between a probability of a positive item endorsement and θ. Inspection of Figure 1 indicates that θ, which typically ranges between -3 and +3 on a z-score metric (mean = 0, standard deviation = 1), is represented on the *x*-axis. Probability estimates of a positive endorsement for a given ability level are reported on the *y*-axis. The threshold parameter (*b*) characterizes the item's level of difficulty and is expressed on the same scale as ability. Its value corresponds to the ability value with a 50% probability of a positive response ("yes" response). Items with a low probability of a positive endorsement have threshold values near -3, whereas items having a high endorsement probability

have values closer to +3. The discrimination parameter (*a*) is proportional to the slope where there is a 50% probability of a correct item response. Flat IRFs indicate poorly discriminating items and steep curves correspond to highly discriminating items. As shown in Figure 1, an assumption of IRT is that an individual's probability of positively endorsing an item is a monotonically increasing function of θ (Lord, 1980; Lord & Novick, 1968).

Figure 1

Hypothetical IRF



a. Unidimensional Models

The unidimensional IRT models for dichotomously scored items are perhaps the most commonly used models. The fundamental model is the normal ogive model; in which the cumulative normal curve serves as the response function (see Lord and Novick, 1968, Chpt. 16). Model assumptions include a single latent trait (e.g., depression) underlies the item responses and the metric of θ for the item response function for each item can be represented as the normal ogive (Lord & Novick, 1968, p. 366). The normal ogive is

$$P_j(\theta) = \Phi(y_j) \tag{1}$$

where Φ is the cumulative normal distribution function and $y_j = a_j(\theta - b_j)$, called the *normal deviate*. Equation (1) models the probability of an individual with a given level of θ positively endorsing item j (or obtaining a correct response). The probability of not endorsing item j is $P_j = 1 - \Phi(y_j)$.

The similar but mathematically more convenient family of probabilistic models is the logistic models (see Birnbaum, 1968). The logistic (cumulative) distribution function is

$$P_j(\theta) = \Psi(1.7z_j) \tag{2}$$

where ψ is the logistic cumulative distribution function, 1.7 is a scaling constant to make the model comparable to the normal ogive model (Camilli, 1994; Birnbaum, 1968), and $z_j = a_j(\theta - b_j)$, referred to as the *logistic deviate*.

The one-parameter model, or Rasch model (Rasch, 1966), is the most restrictive and only includes item difficulty and θ in estimating item performance. The two-parameter (2-PL) model also includes an item's discrimination parameter. The three-parameter (3-PL) model is the least restrictive and also includes a pseudo-guessing parameter in addition to discrimination and difficulty parameters. The 3-PL model may not be readily applicable to mental health measures, as it is typically used for data in which guessing could occur, such as multiple-choice items on achievement tests.

The 1-parameter model was developed by Rasch (1966) to model an individual's probability of a positive item endorsement in terms of item difficulty (level of endorsement) and θ. The logistic model is

$$P_i(\theta) = \frac{1}{1 + \exp^{-(\theta - b_i)}} \tag{3}$$

where $P_i(\theta)$ is an individual's probability of a positive item endorsement with a particular trait level, or theta, and b is item difficulty. The model is the most restrictive of the unidimensional IRT models as it posits equal discrimination across items. Although this is generally an untenable assumption to be met in applied testing contexts (Hambleton & Jones, 1989; Hambleton & Swaminathan, 1985; Traub, 1983), the model is easier to work with because only a single item parameter needs to be estimated.

The 2-PL model relaxes the restrictive assumption of equal discrimination specified in the 1-PL model by also including a discriminatory power parameter in the model. The model is

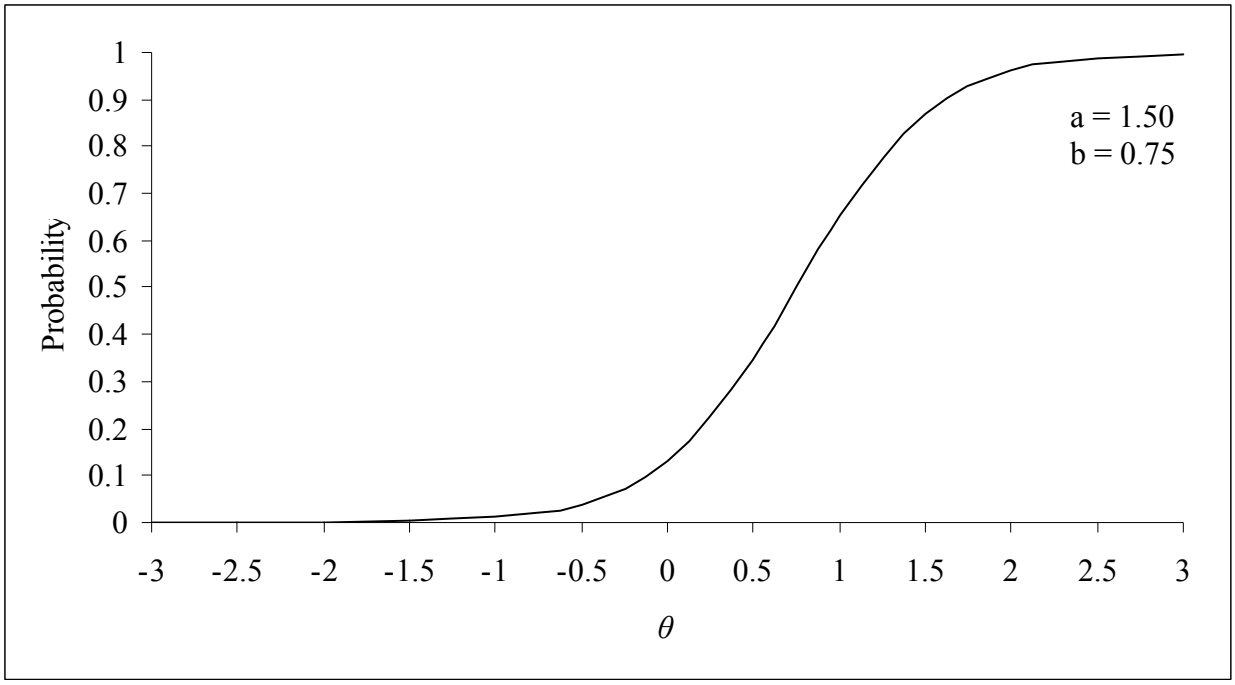
$$P_i(\theta) = \frac{1}{1 + \exp^{-1.7a_i(\theta - b_i)}} \quad (4)$$

where a is item discriminatory power, and the other model parameters can be interpreted as those presented for the 1-PL model. Discrimination parameters typically range from 0 to 2 (Hambleton & Swaminathan, 1985), with high values being more effective with discriminating between respondents with low and high trait levels.

Figure 2 illustrates an IRF for an item based on the 2-PL model. Compared to that shown in Figure 1, the curve is steeper and corresponds to an item that is more strongly related to the measured trait. The threshold (b parameter) for this item is 0.75. The lower asymptote approximates zero, indicating that an examinee with low standing on the measured trait has roughly a zero probability of endorsing a positive response. For example, a non-depressed respondent would likely have a low probability of answering "yes" on an item asking whether he/she has felt helpless over the past several days.

Figure 2

Hypothetical IRF based on 2-PL IRT Model



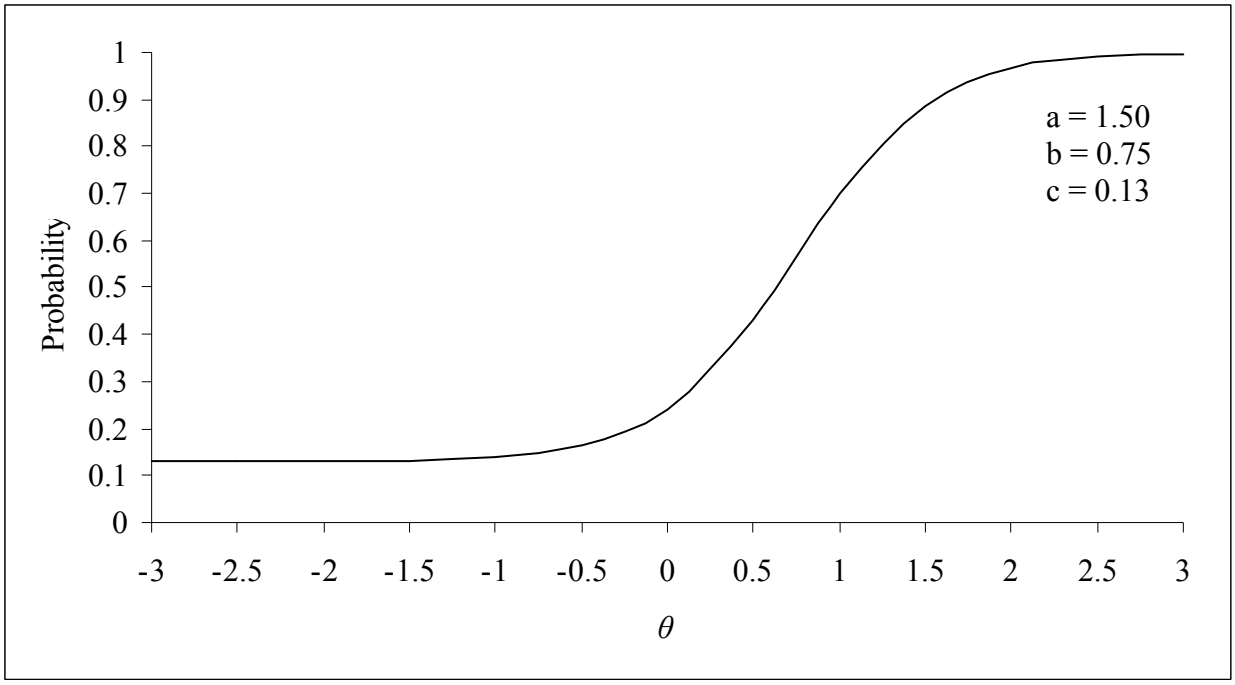
The 3-PL model builds on the 2-PL model by also including a pseudo-guessing parameter, c_i. The form of the model is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp^{-1.7a_i(\theta - b_i)}} \quad (5)$$

where c_i is the lower asymptote of the item characteristic curve, which indicates the lowest probability of a correct response that may occur due to guessing (Lord, 1980). Figure 3 shows an IRF based on the 3-PL IRT model. The lower asymptote is greater than 0 (c = .13), indicating that respondents with varying trait levels have some probability of a positive item endorsement.

Figure 3

Hypothetical IRF based on 3-PL IRT Model



There is also a class of IRT models for polytomously scored items (e.g., Likert scales). These include, for example: Samejima's (1969) graded response model, Bock's (1972) nominal (non-ordered) response model, Master's (1982) partial credit model, and Andrich's (1978) rating scale model, which Muraki (1990) generalized by introducing a discriminating power parameter, and Thissen and Steinberg's (1984) model for multiple–choice items. These models estimate an examinee's probability of selecting a particular response category (e.g., strongly disagree, disagree, neutral, agree, strongly agree) for a given item. For example, a patient with severe depression would most likely have a high probability of answering "strongly agree" on an item asking whether he/she has felt helpless over the past few days.

Samejima's (1969) graded response model is perhaps the most widely used unidimensional IRT model for ordered, polytomous responses (e.g., 1, 2, 3, …, m – 1, where m – 1 is the highest trait level). The categorical response probability is

$$P_{jk}(\theta) = \Phi(y_{jk}) - \Phi(y_{j,k-1}) \tag{6}$$

where $P_{jk}(\theta)$ is the probability of an individual with a given θ selecting category k of item j, and is the difference between the probabilities of selecting successive categories.

The logistic model is

$$P_{jk}(\theta)* = \frac{1}{1 + e^{-a_j(\Theta - b_{jk})}} \tag{7}$$

where, $P_{jk}(\theta)*$ is the probability that person with θ will reach category k or higher on item j, b_{jk} refers to the point on the trait continuum where an examinee has a 50% probability of selecting category k, and a refers to the item's discriminatory power (equal across categories).

Therefore, the probability that individual n will endorse category k is

$$P_{jk}(\theta)* = \frac{1}{1 + e^{-a_j(\Theta - b_{jk})}} - \frac{1}{1 + e^{-a_j(\Theta - b_{j,k-1})}}.$$

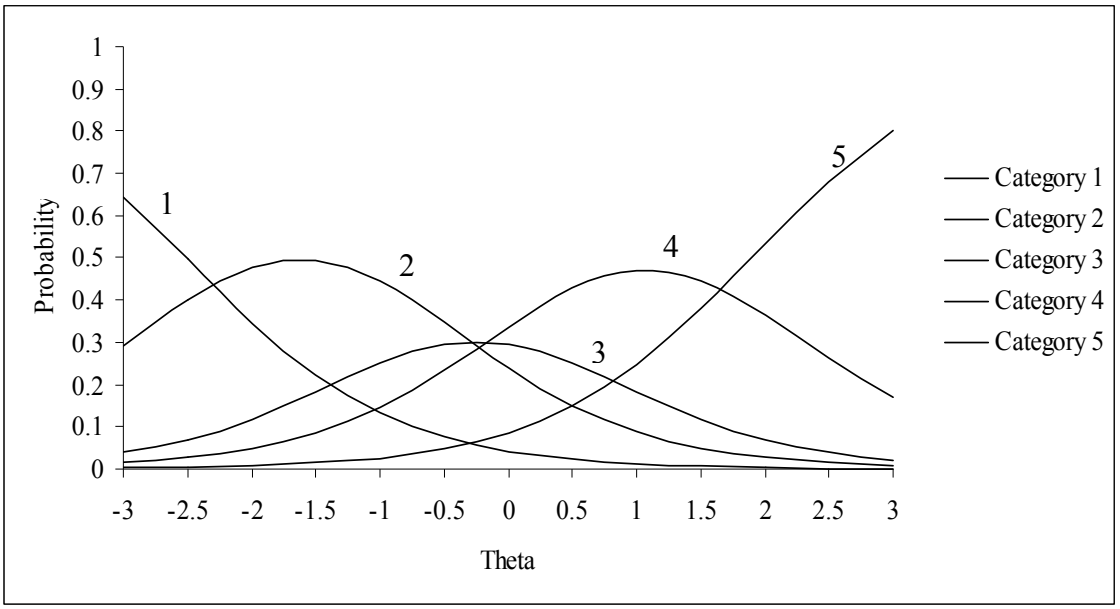
The model specifies that each previous category must be obtained prior to selecting the next highest category (Samejima, 1969).

Figure 4 illustrates the probability of a selecting one of five possible response categories on a Likert scale item based on Samejima's (1969) graded response model. Inspection of the IRFs for each response category indicates that lower trait estimates correspond to higher probabilities of selecting lower response categories (e.g., 1, 2), whereas higher trait estimates correspond to choosing higher response categories (e.g., 3, 4). As specified in the model, the categorical trace lines have equal slopes and unique threshold (difficulty) parameters. The hypothetical trace lines in the figure could correspond to any type of measure in which

respondents select a particular response (e.g., strongly disagree, neutral, strongly agree), including the PTGI (Tedeschi & Calhoun, 1996).

Figure 4

Hypothetical IRFs for Item with Five Category Response



Muraki (1983, 1990) introduced a rating scale version of the graded response model that included category parameters to represent the psychological distance among points on the rating scale. It differs from Samejima's original model in that (a) it requires estimation of $(n-1)m$ fewer parameters, (b) the category parameters associated with the points on the rating scale may be separately estimated from the item parameters, and (c) the items may be unidimensionally ordered by the item intercept. Characteristics of the rating scale model are that (a) items with different numbers of response categories cannot be used, and (b) the model assumes common distances between response categories for all items.

IRT model selection hinges on several considerations. Among the factors include: sample sizes, properties of items, purpose of study, and shape of the score distribution, among many. For example, stable parameters estimates based on the 1-PL model (Rasch model) have been

reported for a test length of 20 items and sample size of 200 (Wright & Stone, 1979). For the 2-PL model, parameters characterizing a 30 item measure can be estimated based on 500 respondents (Hulin, Lissak, & Drasgow, 1982). As for the 3-PL model, Hulin et al. (1982) and Swaminathan and Gifford (1983) found that a sample size of 1,000 would yield acceptable parameter estimates for 60 and 20 item measures. Hambleton (1989) suggests the following sample size recommendations to obtain stable parameter estimates: 200 (1-PL), 500 (2-PL), and 1,000 (3-PL). Larger sample sizes (> 1,000) are required for polytomous items (De Ayala & Sava-Bolesta, 1999). Yen and Fitzpatrick (2006) provide a review of studies addressing the effect of test length, sample size, and parameter estimation on the performance of IRT.

Application of unidimensional IRT models includes meeting the strong assumptions of unidimensionality and local independence (Lord, 1980; Lord & Novick, 1968). Unidimensionality requires that the items measure a single underlying latent trait; local independence is an extension of this principle and suggests that after accounting for ability, item responses are uncorrelated (Lord, 1980).

Advancements in IRT over the past several decades have enabled it to grow as a robust and powerful data analytic strategy for a wide range of testing applications. Areas in which IRT is routinely applied include: (a) test and survey development (Beck & Gable, 2001; Hambleton & Swaminathan, 1985), (b) differential item functioning (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993), (c) test score equating (Cook & Eignor, 1991), (d) test scoring (Thissen & Wainer, 2001), and (e) Computerized Adaptive Testing (CAT - Wainer, Dorans, Eignor et al., 2000), among many.

More recently, IRT has been advanced to model the dimensions underlying scale data in the form of exploratory and confirmatory factor analysis (Bock et al., 1988; Gibbons et al.,

2007a; Gibbons & Hedeker, 1992). This is an important development in the area of IRT, as it provides practical ways to model the inherently multidimensional structure of PRO measures.

7. Multidimensional IRT Models

a. Underlying theory

Many psychological constructs are multidimensional, in that they can be measured as subscales of a more general construct. This is particularly true in the measurement of personality, for example, but many ability and achievement variables can also be measured on multiple indicators. However, unidimensional IRT models have been predominant across social science research (e.g., psychologically, sociology, education) and health related quality of life (HRQOL) measurement primarily because, historically, multidimensional IRT parameter estimation procedures were not fully developed or studied.

A critical problem in the construction of affective and cognitive tests is establishing the number of dimensions of individual differences among respondents that are required to account for response data from a given set of items administered to a sample from some population of respondents. The statistical procedure for this purpose is so-called "item" factor analysis, or FIFA -- that is, application of multiple factor analysis directly to the item responses rather than to test scores.

Several studies have examined the effects on item parameter estimation of applying unidimensional IRT models to item response data that are not strictly unidimensional (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979; Way, Ansley, & Forsyth 1988). Two general finding emerge from these studies: (a) if there is a predominant general factor in the data, and dimensions beyond that major dimension are relatively small, the presence of multidimensionality has little effect on item parameter estimates and the associated ability or

impairment estimates (also called theta estimates); (b) on the other hand, if the data are multidimensional with strong factors beyond the first (as would be the case with a multiple-indicator personality, HRQOL, or achievement instruments) unidimensional parameterization results in parameter and theta estimates that are drawn toward the strongest factor in the set of item responses; this tendency is ameliorated to some extent if the factors are highly correlated.. The first situation has led to the development of procedures for determining "essential unidimensionality" (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996), which can be defined as a set of test items that are not strictly unidimensional, but are "unidimensional enough" that the application of unidimensional IRT estimation procedures will result in parameter and theta estimates that are not seriously distorted by the existing degree of multidimensionality in the data.

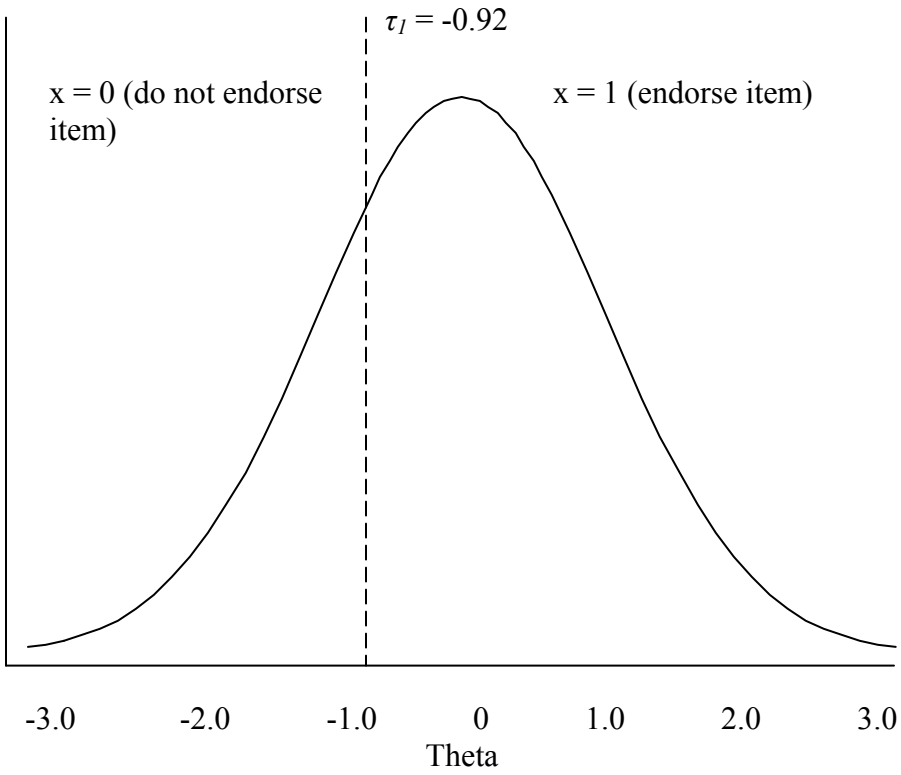
The second situation is more serious since unidimensional parameter estimation procedures applied to such data will result in serious distortion of the measurement characteristics of the instrument. Folk and Green (1989) examined the effects of using unidimensional item parameter estimates with two-dimensional data in the context of both adaptive and conventional tests. Their results indicated that theta estimates were drawn to one or the other of the two traits underlying the data, with the tendency more pronounced for adaptive tests when there is a likely chance that the non-dominant factor will not contribute to the scale score. In addition, the effect was greater when the two dimensions were relatively uncorrelated. Their results suggested that the greater effect on adaptive tests was due to the fact that in the adaptive tests, item discrimination parameter estimates were used both to select items (through item information) and to estimate theta.

b. Factor analysis of tetrachoric correlations

Traditionally, item factor analysis has been carried out on tetrachoric correlations between all pairs of dichotomous item responses. Tetrachoric correlations represent the relationship between two dichotomous variables with underlying continuous distributions. Figure 5 shows a dichotomous variable (item) with a continuous underlying distribution. Within this illustration, the threshold parameter (τ_1) equals -0.92. This value indicates the point on the trait scale in which a respondent would be expected to indicate a positive item endorsement ($x = 1$). Figure 6, for example, provides a graphical depiction of hypothetical bivariate distributions for responses on two depression items (Items 1 and 2, respectively). Each axis represents the underlying latent trait continuums, and the ellipses illustrate several possible shapes of the distributions of the responses to these items. The threshold parameters (τ) represent the point on the trait scale where a positive endorsement would be expected for each item. In this figure, the threshold for Item 1 (τ_1) equals -0.85, where as the threshold for Item 2 (τ_1) is -0.90. This indicates that respondents with trait estimates above -0.85 and -0.90 on Items 1 and 2, respectively, would be more likely to indicate a positive response.

Figure 5

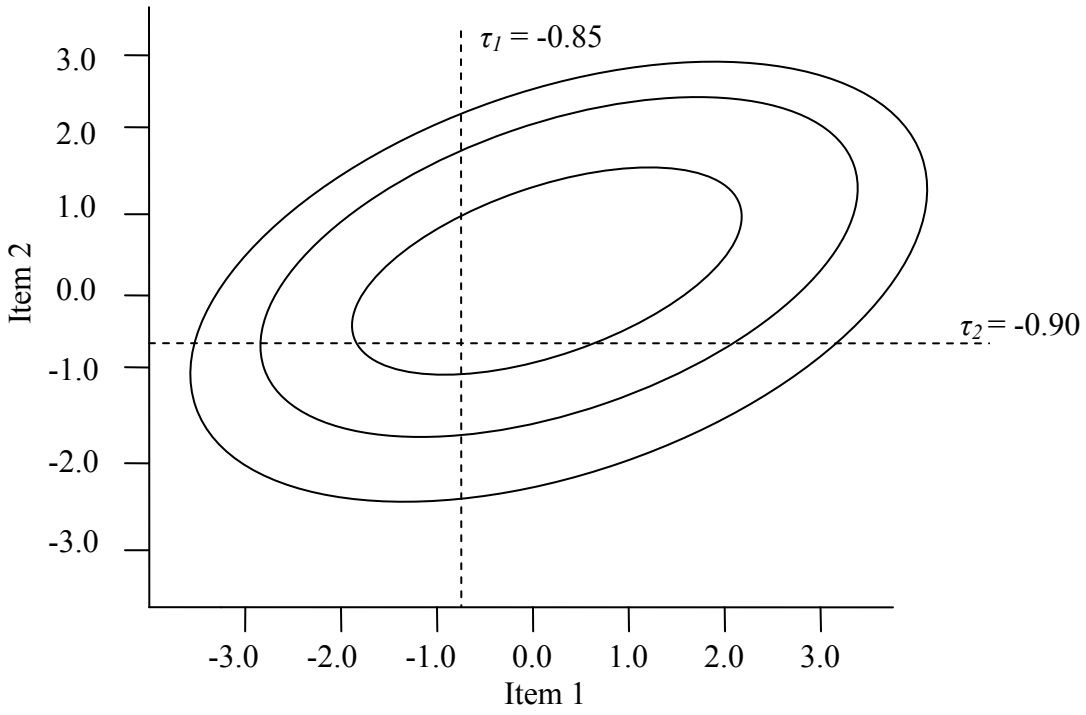
Dichotomous Item Response with Continuous Underlying Distribution



Note. τ_1 represents threshold parameter that dichotomizes the continuous distribution.

Figure 6

Bivariate Distribution of Item Responses for Two Items



Note. τ_1 and τ_2 represent threshold parameters that represent cut-points on underlying continuous scale for Items 1 and 2 where responses become dichotomized.

Factor analysis of the tetrachoric correlation matrix with iteration of communality estimates performs reasonably well when the sample size is very large and the probability of the accepted response is near 50% for all items. Tetrachoric correlations assume that the four-fold frequencies of correct (e.g., scored 1) and incorrect (e.g., scored 0) responses of paired items (e.g., 0, 0; 0, 1; 1, 1) arise from a similar partition of the bivariate normal distribution at points corresponding to the item thresholds. However, problems with the use tetrachoric correlations for factor analysis have been noted by Carroll (1945).

c. Factor analysis of polychoric correlations

This procedure generalizes to ordered-multiple categories by factor analysis of the pairwise polychoric correlations, along with estimates of the thresholds between the categories of each item. This approach to item factor analysis encounters difficulties, however, when the proportions in some cells of the pairwise frequency tables become extreme, which causes the calculation of the correlations to become numerically unstable. In addition, this method of item factor analysis has the disadvantage of not providing a rigorous test of the statistical significance of successive factors as they are extracted from the correlations. Nor does it make use of the information contained in associations of the responses at higher orders than pairwise: all the information used in factor analyzing polychoric correlations is contained in the pairwise correlations, as opposed to each respondent's complete item response vector

d. Full-information item-factor analysis

Item response theoretic factor analysis of responses in two or more ordered categories overcomes these difficulties by avoiding the calculation of pairwise correlations. Instead, it uses the full information contained in all orders of association for a multiple factor model directly to the item response patterns in the data. Highly developed methods for this purpose are now available (see Bock & Gibbons, in press; Schilling & Bock, 2005). If multiple rating categories are graded and are dichotomized near the median, the latter condition (pairwise correlations) will be met and the results may be useful for preliminary examination of the item factor structure. This type of analysis can be extended to multiple rating categories if response probabilities of joint category occurrences for pairs of items are calculated on the assumption of an underlying bivariate normal response process. In more exacting work where a test of additional factors in the model is required, full-information maximum likelihood item factor analysis is preferable (Bock

et al., 1988) and can also handle multiple categories. This approach does not require computation of joint occurrence frequencies and is robust in the presence of items for which the extreme category response probabilities are very low or very high. It evaluates the likelihood of the pattern of item responses from each case and is equivalent to examining item response joint occurrences of all orders. Using adaptive Gaussian quadrature, the likelihoods can be computed with good accuracy with only two or three points per dimension for perhaps as many as ten factors. Simulation studies by Schilling and Bock (2005) reproduced generating parameter values used in eight dimensions with two points per dimension.

If a given set of items is found to be multidimensional, development of the test instrument can proceed in different ways according to the objective of the investigator. If the purpose is to obtain a profile of measurements describing the respondent's position on each dimension, the factor analysis will identify the subsets of items that best represent the corresponding dimensions. The subsets can then be presented to future respondents as separate measures, possibly with separate instructions and separately timed. Alternatively, all of the items can be presented as a single test and the respondents positions on the dimensions estimated directly in the form of factor scores.

Historically, IRT assumed unidimensional item sets, that is, items for which the responses could be accounted for by a single attribute or random effect parameter for each subject. However, empirical Bayes and marginal maximum likelihood methods easily extend the theory to more than one dimension, the approach sketched by Bock and Aitkin (1981) and presented more fully Bock et al. (1988). The basic ideas follow.

Following Thurstone (1947) assume that an individual's response to a test item j is controlled by a latent variable

$$y_{ij} = \sum_{k}^{m} \alpha_{jk}\theta_{ki} + \varepsilon_{ij}\ , \tag{8}$$

where α_{jk} is the loading of item j on factor k, θ_{ki} is the proficiency or propensity of individual i on factor k (e.g., depression), and ε_{ij} is an independent residual. According to the conventions of Thurstonian factor analysis, the variable y and θ_k are assumed standard normal, $N(0,1)$, and the θ_k are uncorrelated. The residuals (ε) are independent and normally distributed with mean 0 and variance $\sigma_j^2 = 1 - \sum_k^m \alpha_{jk}^2$, i.e., ε is NID (0, σ). The quantity $\sum_k^m \alpha_{jk}^2$ is called the common factor variance or *communality* of the item, and σ_j^2 is called the unique variance, or *uniqueness*.

Individual i is assumed to respond positively to item j when y_{ij} is greater than the item threshold γ_j. Thus, the probability that an individual with factor score vector θ_i will respond positively to item j, as indicated by the item score $x_{ij} = 1$ is given by the normal ogive item-response function,

$$\begin{aligned}
\Phi_j(\boldsymbol{\theta}_{ij}) &= P(x_{ij} = 1 \mid \boldsymbol{\theta}_{ij}) \\
&= P(y_{ij} > \gamma_j \mid \boldsymbol{\theta}_i) \\
&= \frac{1}{2\pi\sigma_j}\int_{\gamma_j}^{\infty} \exp\left[-\frac{1}{2}\left(y_{ij} - \sum_k^m \alpha_{jk}\theta_{ki}\right)^2 / \sigma_j^2\right] dy_j \\
&= \Phi\left(\frac{\gamma_j - \sum_k^m \alpha_{jk}\theta_{ki}}{\sigma_j}\right)
\end{aligned} \tag{9}$$

and the probability that the individual will respond negatively, indicated by $x_{ij} = 0$, is the complement,

$$P(x_{ij} = 0 \mid \boldsymbol{\theta}_i) = 1 - \Phi(\boldsymbol{\theta}_i) . \tag{10}$$

Since the multiple factor model implies conditional independence (i.e., the items are uncorrelated conditional on the underlying factors $\boldsymbol{\theta}$), the conditional probability of the item score vector $\boldsymbol{x}_i$ is

$$P(x = \boldsymbol{x}_i \mid \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \prod_{j}^{n_i} [\Phi_i(\boldsymbol{\theta}_i)]^{x_{ij}} [1 - \Phi_j(\boldsymbol{\theta}_i)]^{1-x_{ij}} . \tag{11}$$

For computational purposes it is convenient to express the argument of the response function in terms of an intercept,

$$c_j = -\gamma_j / \sigma_j , \tag{12}$$

and factor slopes

$$a_{jk} = \alpha_{jk} / \sigma_j , \tag{13}$$

rather than threshold and factor loadings.

In the context of Bayes estimation, (11) is the likelihood of $\boldsymbol{\theta}_i$, and the prior, which is multivariate normal, is completely specified. However, because of the nature of this likelihood function, this is an example of a model outside the exponential family for which no closed form of the posterior mean or covariance matrix is available. Note, however, that the unconditional probability of score pattern $\boldsymbol{x}_i$ can be expressed as

$$h(\boldsymbol{x}_i) = \int_{-\infty}^{\infty} P(\boldsymbol{x} = \boldsymbol{x}_i \mid \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} . \tag{14}$$

The integral in (14) can be numerically approximated by m -fold Gauss-Hermite product quadrature. Further details of parameter estimation are provided by Bock and Aitkin (1981), Bock et al. (1988), and Bock and Gibbons (in press).

8. The Bifactor Model

a. Underlying theory

Most typical applications of IRT have assumed unidimensional item sets, that is, items for which the responses could be accounted for by a single attribute for each subject. However, as previously discussed, Bock and Aitkin (1981) and Bock et al. (1988) extended the IRT model to the multidimensional case, where each item is related to one or more underlying latent dimensions, traits, or constructs of interest. In part, however, this multidimensionality is produced by the sampling of items from multiple domains of an overall social or psychological construct. For example, in the measurement of life quality, items are selected from satisfaction domains such as satisfaction with family, income, neighborhood, etc. It is quite natural for such data to appear to be multidimensional, when in fact, they measure a unidimensional construct, *i.e*., quality of life; however, the items within domains are more highly correlated than items between domains.

If the factor pattern shows that the factors are substantially correlated, investigators may wish to estimate a general level of performance over all dimensions, while at the same time taking into account the redundant information within the item subsets that reduces the precision of estimation of the general factor. In that case, the item bifactor model (Gibbons & Hedeker, 1992), consisting of a general factor and independent item group factors, can be fitted to the data. It allows for the effect of so-called "failure of conditional independence" within the item groups on the standard error of measurement for the general factor.

The bifactor model assumes the presence of a general factor involving all items and two or more group factors corresponding to specified mutually exclusive subsets of items. This restrictive model is relevant whenever the item domain contains sub-domains in which items

share a common feature of format or content. Examples are reading comprehension when several questions are asked about the same reading passage, or a health inventory where different aspects of health—physical, emotional, etc.—are probed. The presence of the subgroups of items will typically introduce association of response to items in the subset that is greater than can be attributed to the general factor. If these dependencies are not taken into account when computing scores for the general factor, the standard error of estimation is underestimated. The bifactor model controls for these effects and yields accurate estimates and accurate standard errors of the general factor. Because the group factor loadings have nonzero loadings only within the subgroups, the quadrature required to compute response-pattern likelihoods is two-dimensional regardless of the number of groups. This enables the computation to employ a sufficient number of quadrature points in each dimension to make adaptive quadrature unnecessary.

b. Full-information item bifactor analysis

A plausible s-factor solution (where *s* equals number of factors) for many types of psychological and educational tests is one that exhibits a general factor and $s-1$ group or method related factors. The bifactor solution constrains each item j to have a non-zero loading α_{j1} on the primary dimension and a second loading ($\alpha_{jk}, k = 2,...,s$) on not more than one of the $s-1$ group factors. For four items, the bifactor pattern matrix might be

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix}$$

where the first column of the matrix represents the primary factor, and the second and third columns represents the group factors. This structure, which Holzinger and Swineford (1937) termed the "bifactor" solution, also appears in the inter-battery factor analysis of Tucker (1958)

and is one confirmatory factor analysis model considered by Jöreskog (1969). In these applications, the model is restricted to test scores considered to be continuously distributed. But it is easy to conceive of situations where the bifactor pattern might also arise at the item level (Muthén, 1989). It is plausible for paragraph comprehension tests, for example, where the primary dimension describes the targeted process skill and additional factors describe content area knowledge within paragraphs. Similarly, in the context of mental health measurement, symptom items are often selected from measurement domains and can be related to the primary dimension of interest (*e.g.*, mental instability) and one sub-domain (*e.g.*, anxiety). In these contexts, items would be conditionally independent between paragraphs or domains, but conditionally dependent within paragraphs or domains.

Gibbons and Hedeker (1992) derived an item-response model for binary response data exhibiting the bifactor structure and developed a practical method of item parameter estimation. As they demonstrated, the bifactor restriction leads to a major simplification of likelihood equations that (a) permits analysis of models with large numbers of group factors (*e.g.*, domains), (b) permits one-dimensional conditional dependence among identified subsets of items, and (c) in many cases provides a more interpretable factor solution than an unrestricted full-information item factor analysis (*e.g.*, Bock et al., 1988). Demars (2006) reported that the full-information item bifactor analysis of dichotomous data can be expected to provide accurate parameter and trait estimates under very general conditions. Gibbons et al., (2007a) extended the bifactor model to the case of polytomous items, such as the multi-category rating scales.

In the bifactor case, the rating scale model is

$$z_{jt}(\boldsymbol{\theta}) = c_j + d_t + \sum_{k=1}^{s} a_{jk}(\theta_k), \tag{15}$$

where only one of the $k = 2,...,s$ values of a_{jk} is non-zero in addition to a_{j1}.

Gibbons et al. (2007a) derived the likelihood equations and a method for their solution for bifactor extensions of both the rating scale model and the Samejima model for ordinal response data. The Bock and Aitkin (1981) full-information item-factor analysis may be similarly generalized to the graded response model. Bock, Gibbons and Schilling (in press), have also developed a method for obtaining factor scores for each of the *s-1* group or domain factors in addition to the primary factor.

9. Simulation Study

a. Overview of simulation study

A simulation study was conducted to investigate the effects of applying Samejima's (1969) graded response model in unidimensional and bifactor form to multidimensional data. Conditions varied in the simulation are: (a) test length, 50 items or 100 items, (b) number of dimensions, 5 or 10, (c) primary loadings, .50 or .75, and (d) domain loadings, 0.25 or 0.50. Outcome results include: standard deviation of theta estimates, posterior standard deviations (PSDs, or standard errors) of Bayes EAP scores, log-likelihood (model fit), differences between estimated and actual theta, and percentage change between unidimensional and bifactor models of these variables. The generated data were based on a four-point categorical scale, and the examinee distribution was assumed to be normal, $N(0,1)$, based on 1,000 replications. In the following, we summarize the key findings of this study.

Figure 7 reports the standard deviations of the theta estimates for the unidimensional and bifactor models across the 12 simulated conditions. Inspection of the figure indicates that the theta estimates based on the unidimensional model were more varied across conditions. The magnitude of the difference decreased when the primary and secondary loadings decreased, leading to a more unidimensional solution. As shown, as the number of items increased from 50

to 100, the theta estimates for both models became more varied, but not as severe for the bifactor model.

Figure 7

Mean Standard Deviations of Theta of the Unidimensional and Bifactor Models based on 1,000 Replications per Condition (Number Items [NI] = 50 or 100, Number Dimensions [ND] = 5 or10, Primary loadings [PL] = .50 or .75, Domain Loadings [DL] = .25 or .50)

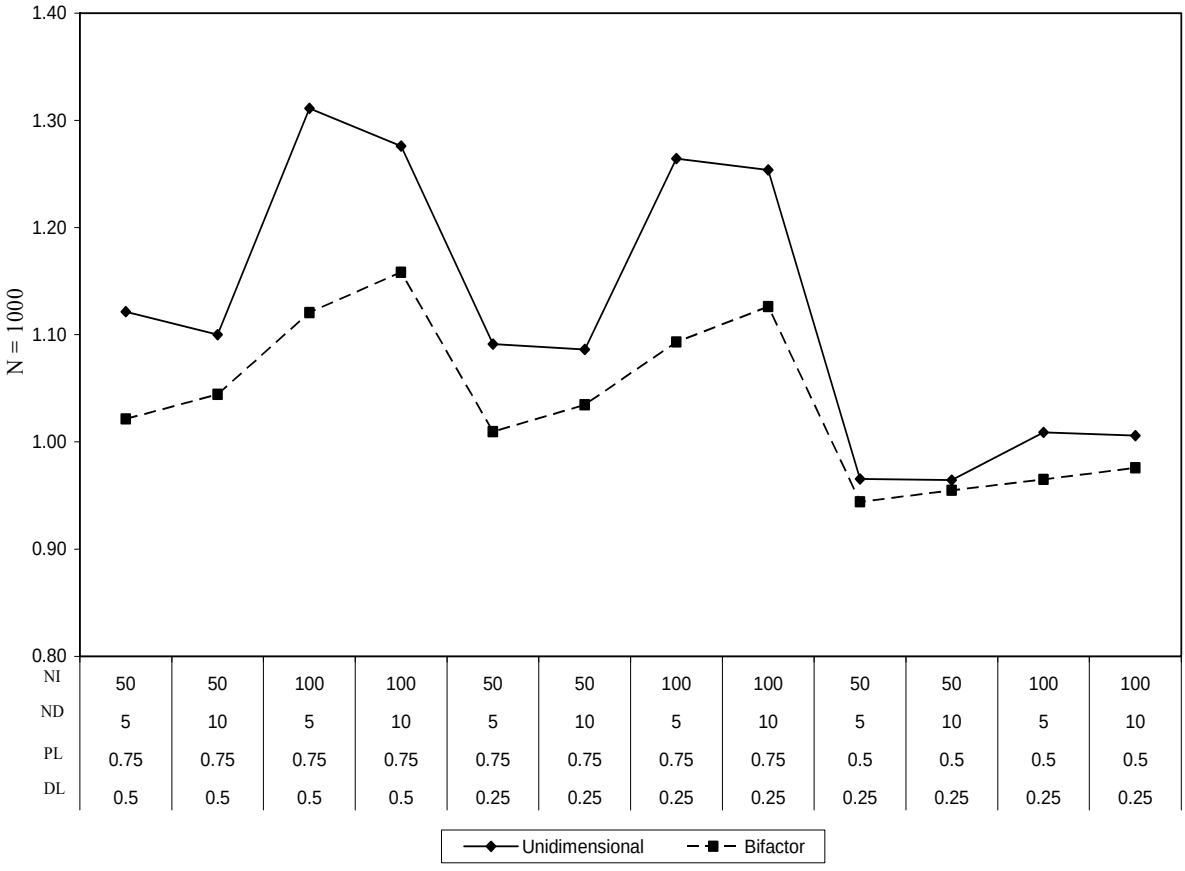


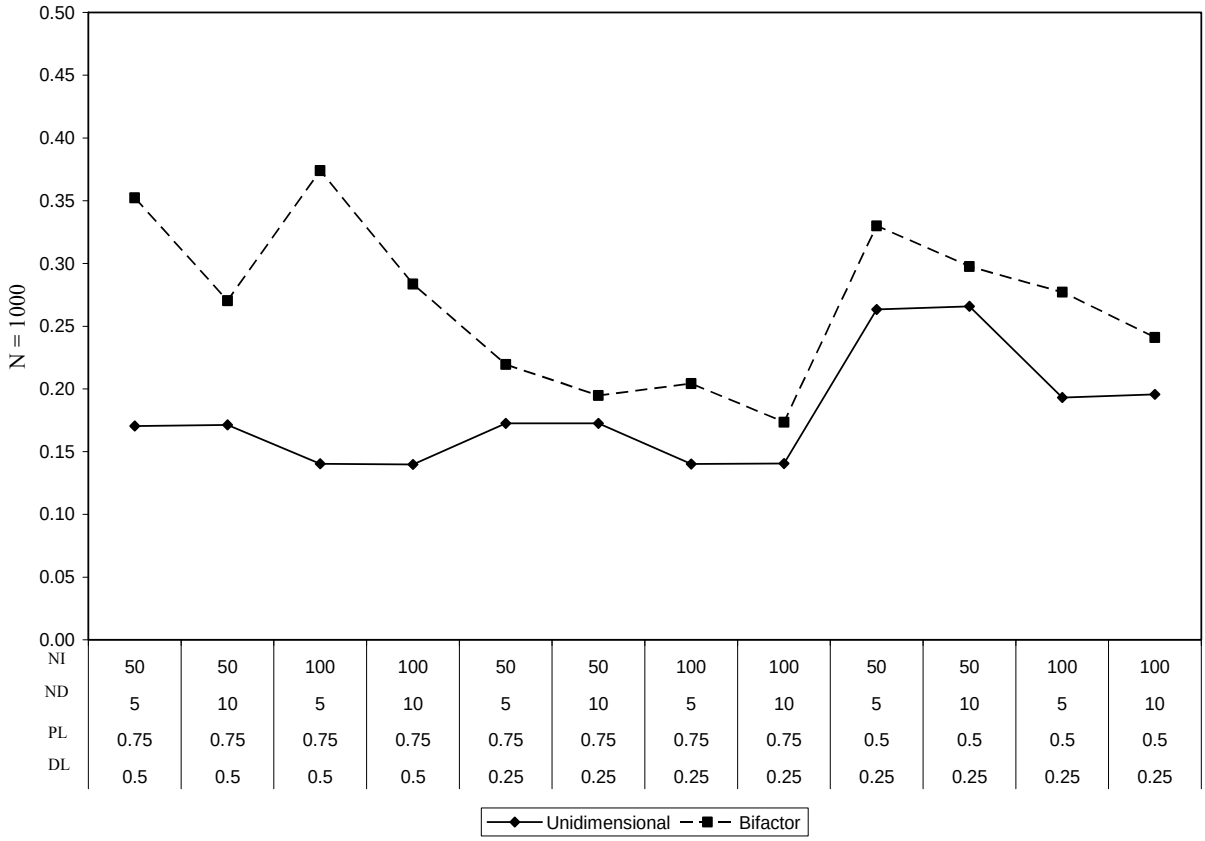
Figure 8 reports the mean posterior standard deviation (PSD) of the Bayes *expected a posteriori* scores (EAP; Bock & Mislevy, 1982). As shown, the differences in the PSD between the models can be dissected in terms of the dimensionality of the underlying data. Specifically, in the conditions in which the primary loadings are 0.75 and the domain loadings are 0.50, the PSD of the unidimensional model substantially underestimates the PSDs from the bifactor model. As

shown, the largest PSD for the unidimensional model occurs with 100 items and 5 dimensions. The PSD estimated by the bifactor model remains fairly consistent across the conditions in which the underlying structure can be regarded as strongly multidimensional (i.e., primary loadings = 0.75, domain loadings = 0.50).

For the conditions in which the primary loadings are 0.75 and the domain loadings are 0.25, the PSD for the unidimensional approaches that for the bifactor model but, nevertheless, continues to underestimate the bifactor result, which is the correct value in this case. The largest discrepancies between the PSD of these models occurs when the number of dimensions is 5 and the number of items is 50 and 100. The smallest difference between the mean PSDs for the unidimensional and bifactor models occurs when the number of dimensions is 10 with 50 items. For the bifactor model, the PSD decreases slightly when the number of items increases from 50 to 100. However, the number of dimensions does not seem to significantly influence the PSD of the bifactor model.

Figure 8

Mean Posterior Standard Deviations of Bayes Expected A Posterior Scores of the Unidimensional and Bifactor Models based on 1,000 Replications per Condition (Number Items [NI] = 50 or 100, Number Dimensions [ND] = 5 or10, Primary loadings [PL] = .50 or .75, Domain Loadings [DL] = .25 or .50)



The results of this study illustrate the consequences attached to applying a unidimensional IRT model to data with varying degrees of multidimensionality compared to the bifactor model. Several results are worth considering in light of using IRT procedures to model health outcomes measurements. The first set of results addressed the variability in estimated theta values, or examinees' standing on the latent trait. Compared to the unidimensional model, the bifactor model yielded theta estimates that were more homogeneous across simulated data

structures. However, EAP estimates become more varied with an increase in test length, which would be expected in practical testing applications due to a given test's ability to discriminate between examinees with different ability levels. That is, longer tests are generally more reliable. The only condition in which the unidimensional and bifactor models provided comparable results was when the data deviated slightly from unidimensionality (primary loadings = 0.50, domain loadings = 0.25).

PSD estimates were found to be underestimated across all conditions for the unidimensional model. For the bifactor model, PSD values were consistently below 0.20 across conditions, except when the total test length was 50 and the primary loadings were 0.50 and the domain loadings were 0.25. One setting in which the underestimation of PSDs could affect test scores is in computer adaptive testing, in which each item is intentionally selected to provide the most information for estimating of examinee ability in the sense of greatest reduction of PSD. Using PSD estimates based on the unidimensional model may therefore lead to suboptimal estimates of examinee ability. Used as measurement error variance, the inverse squared unidimensional PSDs are not valid for weighting observation in statistical analyses using the scores as data.

While not presented in detail here, we also found that (a) the bi-factor model exhibited significantly improved fit over the unidimensional alternative, and (b) root mean square errors (RMSE) between the estimated and actual theta values used to generate the data for the bi-factor model were lower than those reported by the unidimensional model, indicating better fit of the model to the observed data,

10. Application to real data

a. Description of datasets

To promote an understanding of the application of IRT in health outcomes research, several applied examples are provided. Data for these examples are based on the previously discussed PDSQ, PTGI and JAS scales. Results are first presented for the PDSQ, followed by the PTGI and JAS, respectively.

Psychiatric Diagnostic Screening Questionnaire

This first example demonstrates testing the factor structure of the PDSQ in terms of a bifactor model compared to a unidimensional model. Other competing models tested included a symptom domain model, with all but the domain of interest (i.e., fourteen rather than fifteen domains in which each domain is left out in of each analysis). Testing the fit of these competing models provides the basis to determine (a) the factor structure of the PDSQ, (b) which, if any, or all, of the symptom domains add unique test information above and beyond that provided by a unidimensional model, and (c) how PDSQ scores should be used for diagnostic purposes (one overall score or multiple sub-domain scores).

The first model fit the unidimensional model to the data (χ^2 = 488,924.45, df = 3,512, p < .001). Next, the bifactor model including a primary dimension and all 15 symptom domains was fit to the data, and revealed significantly improved model fit compared to the unidimensional model ($\chi^2_{Difference}$ = 79,624.73, $df_{Difference}$ = 139, p < .001). This suggests that the unidimensional hypothesis must be rejected; that is, the scale is multidimensional. Figure 9 presents observed and predicted response proportions for the 15 domain bifactor model and illustrates excellent fit of the model to the observed 139 symptom response proportions (r = 0.9992).

Figure 9

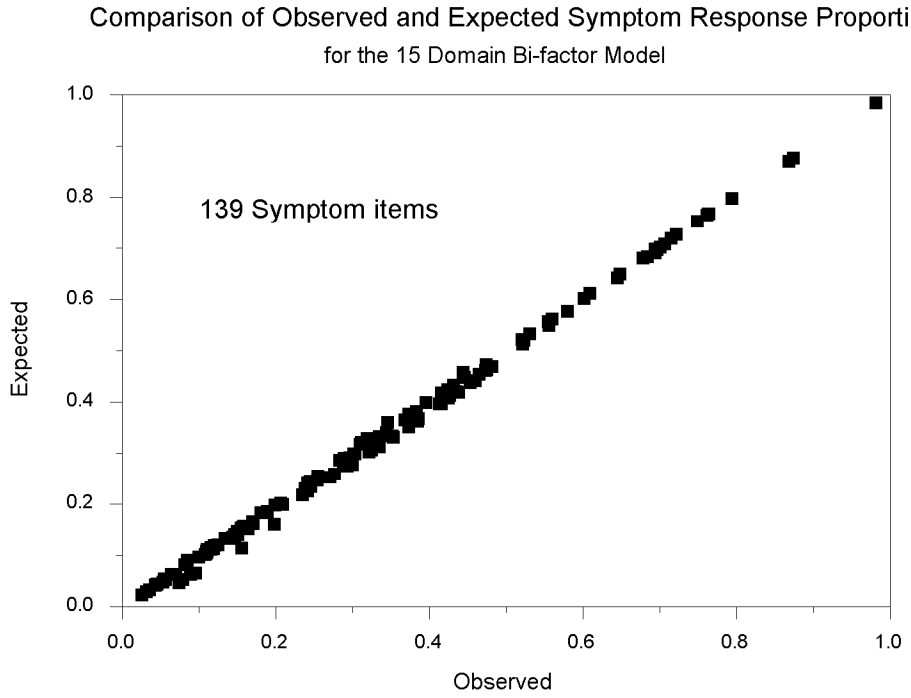


Table 2 presents the parameter estimates for the bifactor model. The item thresholds describe the level on the underlying mental illness dimension that each item shifts from a negative to a positive endorsement. Items with small or negative thresholds are endorsed at lower levels of mental illness (psychiatric symptoms), whereas items with larger positive thresholds are endorsed at higher levels of psychiatric symptoms. For example, the symptom item that is endorsed at the lowest level of mental illness (i.e., smallest threshold = -2.146) is item 1 (Feel sad or depressed past 2 weeks). By contrast, the symptom item that is endorsed only at the highest levels of mental illness (i.e., largest threshold = 1.987) is item 79 (Think had special powers). In terms of domains, Dysthymia and Generalized Anxiety have low thresholds, whereas Mania, Psychosis, Alcohol Abuse, and Drug Abuse have uniformly high thresholds.

In terms of loadings on the primary dimension (interpreted as factor loadings on the overall mental illness dimension), Alcohol Abuse, and Drug Abuse domains have low loadings

on the primary mental illness dimension (on average about 0.20), whereas Post-Traumatic Stress, Obsessive Compulsive, Panic, Psychosis, Agoraphobia, Social Phobia, Generalized Anxiety, and Hypochondiasis have the highest (on average about 0.50). Interestingly, the Major Depressive Disorder items had lower loadings on the primary dimension, indicating that the primary dimension only accounts for a small proportion of item variance. By contrast, loadings on the individual symptom domains, were generally uniform and quite high (0.5 to 0.9), with Alcohol Abuse, and Drug Abuse domains having the highest domain loadings, most likely due to the fact that they had the lowest loadings on the primary dimension and are therefore measuring two domains that are distinct from the rest of the test symptom-items. Thus, symptom domains with the lowest loadings on the primary domain are not reflected in the total PDSQ score.

Table 2

PDSQ Items, Threshold and Factor Loadings for Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
MDD				
1	Feel sad or depressed past 2 weeks	-2.15	0.17	0.52
2	Sad/depressed every day past 2 weeks	-0.72	0.26	0.43
3	Less pleasure from things 2 weeks	-1.13	0.17	0.39
4	Less interest in most activities 2 weeks	-1.13	0.20	0.35
5	Appetite significantly lower 2 weeks	0.18	0.19	0.13
6	Appetite significantly greater 2 weeks	0.66	0.10	0.03
7	Sleep at least 1-2 hours less 2 weeks	-0.13	0.25	0.08
8	Sleep at least 1-2 hours more 2 weeks	0.70	-0.04	0.11
9	Feel jumpy and restless 2 weeks	0.15	0.39	0.09
10	Tired nearly every day past 2 weeks	-1.15	0.18	0.24
11	Feel guilty about things 2 weeks	-0.60	0.36	0.33
12	Negative thoughts about self 2 weeks	-0.68	0.30	0.49
13	Feel like failure past 2 weeks	-0.38	0.34	0.54
14	Problems concentrating every day past 2 weeks	-0.83	0.34	0.22
15	Decision making more difficult 2 weeks	-0.51	0.33	0.28
16	Think of dying in passive ways 2 weeks	0.22	0.26	0.73
17	Wish you were dead 2 weeks	0.56	0.19	0.90
18	Think you're better off dead 2 weeks	0.32	0.22	0.85
19	Thoughts of suicide past 2 weeks	0.26	0.19	0.75
20	Seriously consider taking life 2 weeks	1.23	0.23	0.77
21	Think specific way to take life 2 weeks	0.91	0.17	0.74
DYS				
22	Feel sad/down most days past 2 years	-0.34	0.35	0.79
23	Poor appetite/overeat most days 2 years	-0.03	0.36	0.58
24	Not sleep enough/too much sleep 2 years	-0.49	0.35	0.68
25	Tired most days past 2 years	-0.48	0.32	0.80
26	Problem concentrating/making decisions 2 years	-0.18	0.41	0.71
27	Low self-esteem most days 2 years	-0.53	0.41	0.66
28	Feel hopeless about future 2 years	-0.23	0.42	0.64
PTSD				
29	Ever experienced traumatic event	0.09	0.31	0.48
30	Ever witnessed traumatic event	0.36	0.29	0.40
31	Thoughts of trauma pop into mind	0.19	0.41	0.75
32	Upset because thinking of trauma	0.39	0.43	0.75
33	Bothered by memory/dreams of trauma	0.24	0.45	0.78
34	Reminders of trauma cause distress	0.29	0.48	0.76

Note. MDD = Major Depression, DYS = Dysthymia, & PTSD = Post-Traumatic Stress.

Table 2 (cont.)

PDSQ Items, Threshold and Factor Loadings for Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
35	Block out thought/feeling of trauma	0.12	0.45	0.74
36	Avoid activities remind of trauma	0.46	0.48	0.67
37	Flashbacks of traumatic event	0.63	0.50	0.64
38	Reminders make you shake	0.70	0.57	0.59
39	Feel distant because of trauma	0.52	0.46	0.73
40	Feel numb because of trauma	0.54	0.43	0.70
41	Give up goals because of trauma	0.78	0.46	0.63
42	Keep guard up because of trauma	0.24	0.45	0.70
43	Jumpy because of a trauma	0.67	0.52	0.62
BUL				
44	Often go on eating binges	0.47	0.33	0.86
45	Can't control how much you eat	0.69	0.31	0.85
46	Eat so much uncomfortably full	0.36	0.28	0.87
47	Eat a lot when not hungry	0.44	0.24	0.88
48	Eat alone because embarrassed	0.90	0.28	0.82
49	Feel disgusted after overeating	0.47	0.27	0.90
50	Upset with self because of binges	0.56	0.28	0.89
51	Strict diets, exercise excessively	1.20	0.30	0.55
52	Force self to vomit	1.61	0.30	0.55
53	Weight most important thing	0.10	0.21	0.51
OCD				
54	Worry about dirt, germs	1.25	0.46	0.44
55	Worry something you forgot	0.53	0.55	0.41
56	Worry you'd act/speak violently	0.55	0.60	0.30
57	Compelled to do things over and over	1.21	0.50	0.66
58	Do things over that interfered	0.98	0.47	0.66
59	Wash and clean excessively	1.22	0.47	0.60
60	Excessively check and do things over	0.91	0.51	0.68
61	Count things obsessively/excessively	1.31	0.44	0.55
PAN				
62	Scared because heat beating fast	0.53	0.45	0.72
63	Scared because short of breath	0.67	0.48	0.71
64	Scared because shaky or faint	0.57	0.52	0.66
65	Anxiety attacks for no reason	0.21	0.59	0.49
66	Anxiety attacks, think will go crazy	0.41	0.64	0.45
67	Anxious attacks with 3 or more symptoms	0.30	0.60	0.62
68	Worry about having anxiety attacks	0.69	0.63	0.44
69	Anxiety attacks caused avoid situations	0.45	0.64	0.30
70	Feel excessively cheerful/happy	1.11	0.14	0.84

Note. BUL = Bulimia Nervosa, OCD = Obsessive Compulsive Disorder, & PAN = Panic.

Table 2 (cont.)

PDSQ Items, Threshold and Factor Loadings for Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
MANIA				
71	Feel extremely self-confident	1.18	0.13	0.87
72	So much energy, need less sleep	1.40	0.16	0.84
73	Talk more than usual	1.05	0.35	0.60
74	Thought could do everything	1.01	0.24	0.62
75	Do impulsive things	1.03	0.32	0.49
PSYCH				
76	People tell imagination	1.20	0.49	0.50
77	Convinced others spying	0.85	0.55	0.55
78	Think danger because someone plotting	1.55	0.56	0.48
79	Think had special powers	2.00	0.41	0.51
80	Think some force controlled	1.91	0.49	0.53
81	See/hear things other people didn't	1.64	0.47	0.49
AGOR				
82	Avoid situation because afraid of anxiety attack	0.69	0.64	0.35
83	Anxious going far away from home	1.06	0.59	0.53
84	Anxious being in crowded places	0.54	0.64	0.61
85	Anxious standing in long likes	0.80	0.64	0.54
86	Anxious being on bridge or in tunnel	1.11	0.49	0.50
87	Anxious traveling in bus, train, plane	1.09	0.49	0.57
88	Anxious driving/riding in a car	1.12	0.51	0.49
89	Anxious being home along	1.01	0.52	0.28
90	Anxious being in open spaces	1.68	0.59	0.50
91	Get anxious as soon as in situation	0.59	0.64	0.61
92	Avoid situation because made you anxious	0.44	0.63	0.62
SOC				
93	Worry about embarrassing self	0.08	0.52	0.64
94	Worry you'd say something stupid	-0.03	0.50	0.66
95	Nervous when people pay attention	-0.12	0.46	0.69
96	Nervous in social situations	0.10	0.52	0.65
97	Avoid situations because afraid embarrass self	0.34	0.57	0.64
98	Worry public speaking	0.12	0.39	0.64
99	Worry eating in front of others	0.85	0.47	0.49
100	Worry using public restrooms	1.27	0.45	0.34
101	Worry writing in front of others	1.08	0.43	0.39
102	Worry saying something stupid	0.15	0.46	0.76
103	Worry asking questions around others	0.27	0.44	0.72
104	Worry business meetings	0.73	0.37	0.58

Note. MANIA = Mania, PSYCH = Psychosis, AGOR = Agoraphobia, & SOC = Social Phobia.

Table 2 (cont.)

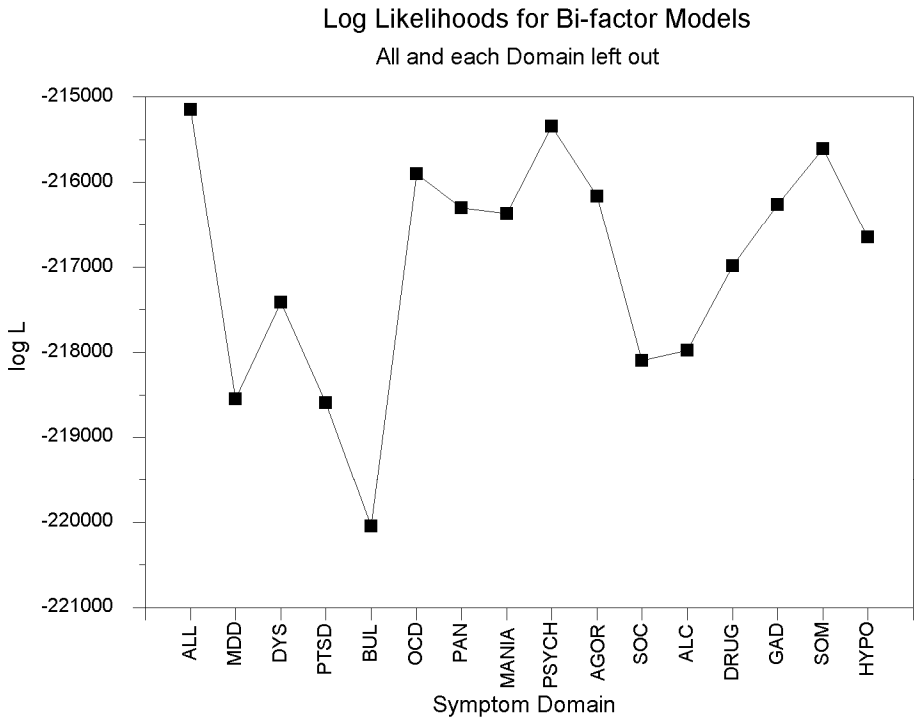
PDSQ Items, Threshold and Factor Loadings for Bifactor Model

Domain/ Item	Question	Threshold	Primary Loading	Domain Loading
105	Worry parties/social gatherings	0.38	0.46	0.74
106	Get anxious as soon as in situation	0.24	0.53	0.66
107	Avoid situations because made you anxious	0.20	0.54	0.62
ALC				
108	Think drink too much	1.21	0.08	0.92
109	Family say drink too much	1.51	0.18	0.86
110	Doctor/friends say drink too much	1.68	0.13	0.88
111	Think about cutting down on drinking	0.99	0.05	0.91
112	Think had alcohol problem	1.55	0.11	0.83
113	Problem with marriage because of drinking	1.64	0.19	0.84
DRUG				
114	Think using drugs too much	1.54	0.19	0.93
115	Family say use drugs too much	1.73	0.28	0.85
116	Doctor/friends say use drugs too much	1.85	0.24	0.89
117	Think about cutting down on drug use	1.34	0.26	0.89
118	Think had a drug problem	1.74	0.17	0.89
119	Problem with marriage because of drug use	1.71	0.29	0.86
GAD				
120	Nervous person most days	-0.05	0.56	0.43
121	Worry bad things happened	-0.08	0.59	0.36
122	Worry about things shouldn't	-0.28	0.52	0.42
123	Worry daily	-0.48	0.52	0.67
124	Feel restless because worrying	-0.53	0.57	0.66
125	Problem falling asleep because anxiety	-0.45	0.49	0.47
126	Tension in muscles because anxiety	-0.57	0.53	0.42
127	Trouble concentrating because worrying	-0.73	0.58	0.57
128	Snappy/irritable because worrying	-0.73	0.45	0.42
129	Hard to control worrying	-0.51	0.54	0.69
SOM				
130	Had a lot of stomach problems	0.19	0.35	0.46
131	Bothered by aches/pains	-0.14	0.41	0.53
132	Get sick more than most people	0.85	0.35	0.75
133	Health been poor most of life	1.26	0.35	0.64
134	Doc not able to find cause for sick	1.02	0.37	0.47
HYPO				
135	Worry might have serious illness	0.43	0.44	0.83
136	Hard to stop worrying about illness	0.74	0.49	0.83
137	Doctor said didn't have illness	1.13	0.48	0.64
138	Worry illness, interfere with activities	1.11	0.55	0.68
139	Visit doctor much because worried about illness	1.26	0.46	0.65

Note.ALC = Alcohol Abuse, DRUG = Drug Abuse, GAD = Generalized Anxiety, SOM = Somatoform, HYPO = Hypochondiasis.

Tests of the statistical difference between models with and without the individual diagnostic domain (e.g., Drug Abuse, Mania) indicated that each of the 15 symptom domains contributed to model-data fit. This finding is supported by the consistently high domain loadings for items within each of the 15 domains. Figure 10 displays the log likelihoods for the model with all domains (*log L* = -215,149) and the 15 other models in which each of the domains was left out. The largest improvement in fit of the ALL model versus a model without a specific domain, was for the Bulimia Nervosa domain (*log L* = -220,044), indicating the strength of this group factor in accounting for the inter-item correlations. The smallest improvement in fit was for the Psychosis domain (*log L* = -215,342). Nevertheless, even for Psychosis the corresponding likelihood ratio chi-square statistic was statistically significant: –2(-215,342-(-215,149)) = 382, *df* = 6, *p*<.0001.

Figure 10



Post-Traumatic Growth Inventory

Samejima's (1969) unidimensional graded response model and unrestricted and restricted multidimensional IRT models were fit to the PTGI scale data (*N*=801). Samejima's (1969) model was fit to the data to compare its results to the bifactor model (Gibbons et al., 2007a) for polytomous data. For the multidimensional IRT models, the PTIG factor structure was first investigated using an IRT-based unrestricted factor analysis to address previous research questioning the stability of the scale's factor structure across diverse samples (e.g., Ho et al., 2004). Subsequently, a bifactor analysis of the original PTGI factor structure and results based on the exploratory analysis were conducted. For the data used in this study, overall scale score internal consistency (Cronbach's alpha) was .96.

Samejima's (1969) unidimensional graded response model was fit to the PTGI data using MULTILOG (Thissen, Chen, & Bock, 2003). (Syntax to fit Samejima's [1969] graded response model to scale data in MULTILOG is provided in Appendix A). Table 3 reports item slope and threshold estimates. Inspection of Table 3 shows that each item has a single slope value and each category (e.g., Strongly Disagree, Agree) has a unique threshold parameter value. As shown in the table, items 5, 10, and 21 were the most discriminating items. Figure 11 shows the IRFs of Item 1. It illustrates that the probability of selecting the next highest category increases monotonically with one's standing on post-traumatic growth.

Table 3

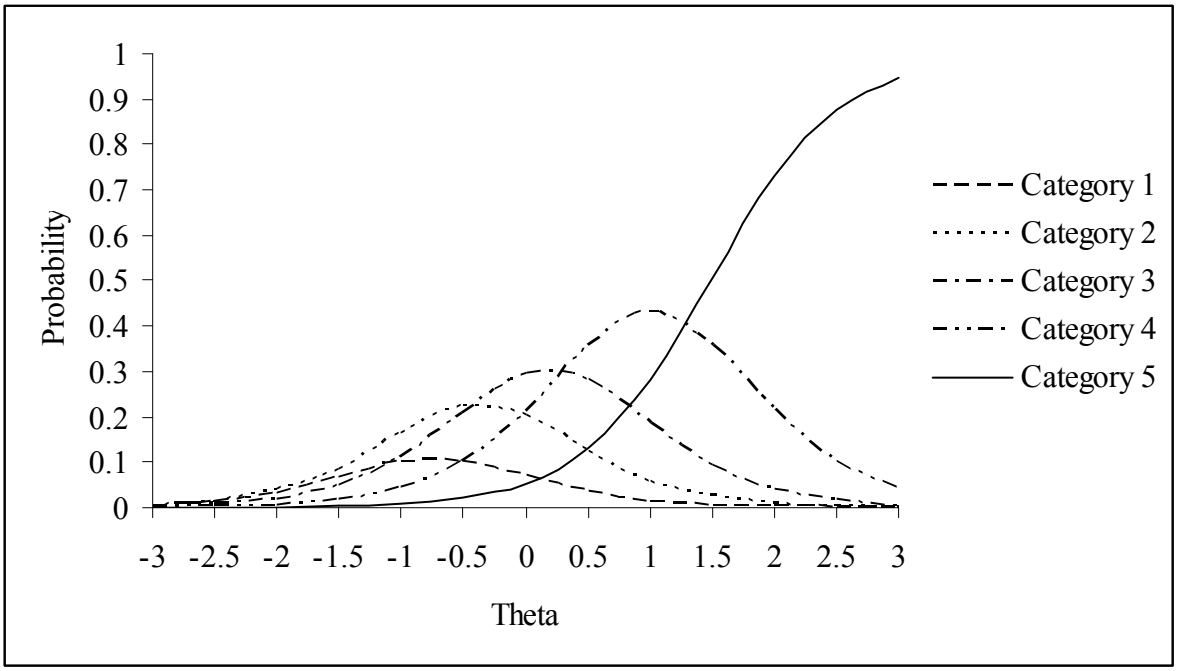
PTGI Parameter Estimates based on Samejima's (1969) Graded Response Model

Item	Slope	Threshold 1	Threshold 2	Threshold 3	Threshold 4	Threshold 5
1	1.92	-0.83	-0.61	-0.13	0.53	1.48
2	2.49	-0.67	-0.40	0.05	0.66	1.53
3	2.44	-0.46	-0.17	0.24	0.96	1.78
4	2.59	-0.58	-0.44	-0.08	0.54	1.38
5	2.72	-0.36	-0.13	0.25	1.03	1.88
6	2.18	-0.74	-0.46	-0.08	0.55	1.38
7	2.11	-0.40	-0.10	0.45	1.27	2.11
8	2.27	0.14	0.15	0.61	1.37	2.06
9	2.61	-0.06	0.18	0.56	1.15	1.72
10	3.32	-0.33	-0.14	0.24	0.94	1.60
11	2.10	0.22	0.47	0.94	1.52	2.13
12	1.91	-0.89	-0.54	-0.07	0.99	2.08
13	2.04	-0.56	-0.35	0.04	0.81	1.82
14	2.69	-0.60	-0.38	0.00	0.57	1.31
15	2.60	-0.46	-0.25	0.18	0.84	1.62
16	2.38	-0.62	-0.35	-0.06	0.50	1.24
17	2.50	-0.52	-0.31	0.06	0.60	1.33
18	2.24	-0.28	-0.06	0.23	0.74	1.38
19	2.06	-0.98	-0.77	-0.41	0.47	1.53
20	2.64	-1.16	-0.89	-0.59	0.13	0.90
21	3.10	-0.78	-0.60	-0.31	0.28	0.94

Note. Thresholds refer to categorical difficulty values, or the point on theta scale in which respondent has 50% probability of category endorsement.

Figure 11

IRFs of PTGI Item 1



e

Based on previous research questioning the robustness of the PTGI factor structure across groups (Ho et al., 2004; Sheikh & Marotta, 2005), a unrestricted full-information item factor analysis using the recently developed POLYFACT program was conducted. (Appendix B provides the syntax to run full-information item factor analysis for polytomous data.) Results based on a promax rotation supported a five factor solution. Specifically, the data seem to be explained in terms of a dominant factor and several minor factors, approximating the scale's original theoretical factor structure (Tedeschi & Calhoun, 1996). Table 4 indicates that each item typically reported a dominant loading, with the exception of items 3, 6, 12, and 13. For comparison purposes, a limited-information exploratory factor analysis was also conducted using Mplus 4.0 (Múthen & Múthen, 1998-2006). As shown, the two approaches yielded similar results, with discrepancies occurring for items with high cross-loadings. Table 5 shows that the empirical factors were moderately correlated.

Table 4

Full Information and Limited Information Unrestricted Item Factor Analysis of PTGI items

Item	Factor 1		Factor 2		Factor 3		Factor 4		Factor 5	
	Full	ULS	Full	ULS	Full	ULS	Full	ULS	Full	ULS
1	-0.019	0.209	**0.864**	**0.820**	0.110	0.296	-0.006	0.136	0.018	0.256
2	-0.097	0.318	**0.714**	**0.677**	0.129	0.285	0.042	0.239	0.237	0.337
3	0.334	0.453	0.126	0.341	0.023	0.304	-0.121	0.127	**0.551**	**0.508**
4	0.140	0.384	0.170	0.368	-0.057	0.256	0.147	0.313	**0.562**	**0.524**
5	**0.484**	**0.535**	0.054	0.293	0.105	0.340	-0.012	0.231	0.346	0.394
6	0.287	**0.431**	**0.366**	**0.500**	-0.251	0.164	0.123	0.289	0.400	0.367
7	**0.784**	**0.651**	0.264	0.377	-0.109	0.202	-0.108	0.135	0.075	0.275
8	**1.018**	**0.731**	-0.018	0.208	0.034	0.274	0.083	0.253	-0.191	0.211
9	**0.947**	**0.732**	-0.113	0.173	0.138	0.358	0.077	0.245	-0.082	0.244
10	**0.554**	**0.594**	-0.118	0.235	0.073	0.310	-0.001	0.260	0.481	0.480
11	**0.782**	**0.628**	-0.100	0.171	0.105	0.332	0.005	0.218	0.113	0.266
12	0.17	0.326	0.053	0.215	**0.910**	**0.744**	-0.079	0.148	-0.097	0.180
13	-0.007	0.264	0.080	0.295	0.464	0.**521**	-0.051	0.165	0.420	0.416
14	-0.131	0.296	0.269	0.454	0.013	0.330	-0.120	0.193	0.881	**0.576**
15	0.072	0.390	-0.031	0.319	-0.006	0.311	-0.110	0.161	0.948	**0.613**
16	0.015	0.412	-0.072	0.306	-0.134	0.226	0.197	0.342	0.898	**0.512**
17	0.071	0.334	0.134	0.276	0.326	0.496	**0.666**	**0.566**	-0.111	0.235
18	0.098	0.366	-0.027	0.207	-0.096	0.274	**0.946**	**0.796**	0.083	0.261
19	0.080	0.312	0.073	0.217	0.939	**0.722**	-0.035	0.201	-0.070	0.232
20	-0.149	0.267	-0.049	0.282	0.597	**0.623**	0.134	0.339	0.476	0.388
21	-0.110	0.311	-0.080	0.304	0.216	0.426	0.088	0.320	**0.859**	**0.585**

Note. Full = full-information item factor analysis. ULS = Unweighted least squares.

Table 5

Factor Correlations

	Factors				
	1	2	3	4	5
1	1.000				
2	0.650	1.000			
3	0.605	0.533	1.000		
4	0.622	0.512	0.619	1.000	
5	0.766	0.769	0.687	0.661	1.000

Next, the bifactor IRT model using Samejima's (1969) graded response IRT model was used to analyze the item responses of the PTGI. The following two models were analyzed: (a) the original five factor model of the PTGI, as per Tedeschi and Calhoun (1996), and (b) a model based on the unrestricted FI item factor analysis. (Appendix C provides the syntax for fitting the bifactor model for graded response data within POLYFACT).

The bifactor model was fit to the data based on the original five-factor structure of the PTGI (Tedeschi & Calhoun, 1996), $\chi^2_{653} = 32{,}104.11$. Table 6 reports primary factor loadings and factor loadings on the five sub-domains. As shown, items reported moderately high loadings on the primary factor (Factor 1), suggesting that the items were related to posttraumatic growth. Within this model, the most discriminating items on the primary factor were Item 21, $\lambda_{21,1} = 0.821$, Item 10, $\lambda_{10,1} = 0.819$, and Item 5 $\lambda_{5,1} = 0.781$; whereas two of the least discriminating items were, for example, Item 1, $\lambda_{1,1} = 0.640$, and Item 12, $\lambda_{12,1} = 0.663$. Notably, similar findings were obtained based on Samejima's (1969) graded response model. Secondary factor loadings were weak to moderate, with an average loading of 0.345. Table 7 reports item thresholds, while Table 8 shows the observed and expected proportions of respondents across categories. The root mean square error value of 0.022 indicates the difference between the observed and expected proportions (across all items and categories) was small, indicating substantial model data fit. Compared to the fit of the unidimensional model ($\chi^2 = 33{,}249.29$, df = 674, $p < .001$), the bifactor model resulted in a statistical improvement in model fit ($\chi^2_{Difference} = 1{,}145.18$, $df_{Difference} = 21$, $p < .001$).

Table 6

Full-Information Item Bifactor Analysis of PTGI Based on Original Five-Factor Model

Item	Primary	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	0.640	0.592				
2	0.725	0.571				
3	0.753	0.124				
4	0.761	0.204				
5	0.781	0.102				
6	0.690	0.325				
7	0.697	0.172				
8	0.703		0.454			
9	0.756		0.494			
10	0.819		0.271			
11	0.683		0.409			
12	0.663		0.095			
13	0.699			0.134		
14	0.761			0.545		
15	0.775			0.191		
16	0.732			0.282		
17	0.751				0.473	
18	0.715				0.619	
19	0.679					0.341
20	0.761					0.595
21	0.821					0.251

Table 7

Item Thresholds Based on Full-Information Item Bifactor Analysis of Original PTGI Five-Factor Model

Item	0-1	1-2	2-3	3-4	4-5
1	-0.601	-0.431	-0.068	0.440	1.172
2	-0.534	-0.299	0.086	0.619	1.337
3	-0.355	-0.101	0.257	0.893	1.605
4	-0.469	-0.343	-0.023	0.541	1.280
5	-0.277	-0.069	0.283	1.000	1.758
6	-0.573	-0.355	-0.040	0.487	1.167
7	-0.293	-0.045	0.410	1.084	1.737
8	-0.076	0.172	0.569	1.210	1.756
9	-0.012	0.217	0.556	1.067	1.547
10	-0.272	-0.086	0.282	0.940	1.575
11	0.231	0.443	0.830	1.291	1.769
12	-0.668	-0.396	-0.034	0.815	1.646
13	-0.413	-0.252	0.068	0.701	1.505
14	-0.500	-0.313	0.022	0.551	1.217
15	-0.383	-0.195	0.201	0.809	1.491
16	-0.502	-0.272	-0.026	0.466	1.109
17	-0.413	-0.237	0.090	0.580	1.198
18	-0.184	-0.012	0.233	0.677	1.189
19	-0.733	-0.574	-0.298	0.423	1.290
20	-0.987	-0.773	-0.504	0.135	0.839
21	-0.697	-0.534	-0.259	0.293	0.926

Table 8

Observed and Expected (in Italics) Proportions From the Original Five-Dimensional Graded Bifactor Analysis of PTGI Scale Data (N = 801)

	0	1	2	3	4	5
1	0.253	0.049	0.122	0.196	0.235	0.145
	0.274	*0.059*	*0.140*	*0.197*	*0.210*	*0.121*
2	0.272	0.067	0.132	0.201	0.215	0.112
	0.297	*0.086*	*0.152*	*0.198*	*0.177*	*0.091*
3	0.332	0.085	0.127	0.223	0.154	0.079
	0.361	*0.099*	*0.142*	*0.213*	*0.132*	*0.054*
4	0.297	0.040	0.107	0.205	0.215	0.136
	0.320	*0.046*	*0.125*	*0.215*	*0.194*	*0.100*
5	0.351	0.072	0.131	0.242	0.141	0.062
	0.391	*0.082*	*0.139*	*0.230*	*0.119*	*0.039*
6	0.262	0.066	0.110	0.200	0.208	0.154
	0.283	*0.078*	*0.123*	*0.203*	*0.192*	*0.122*
7	0.355	0.085	0.165	0.218	0.119	0.059
	0.385	*0.097*	*0.177*	*0.202*	*0.098*	*0.041*
8	0.427	0.091	0.146	0.192	0.087	0.056
	0.470	*0.098*	*0.147*	*0.172*	*0.074*	*0.040*
9	0.443	0.082	0.127	0.167	0.096	0.084
	0.495	*0.091*	*0.125*	*0.146*	*0.082*	*0.061*
10	0.352	0.062	0.132	0.232	0.136	0.085
	0.393	*0.073*	*0.145*	*0.216*	*0.116*	*0.058*
11	0.544	0.079	0.132	0.120	0.069	0.056
	0.591	*0.080*	*0.126*	*0.105*	*0.060*	*0.038*
12	0.241	0.079	0.122	0.310	0.180	0.069
	0.252	*0.094*	*0.140*	*0.306*	*0.158*	*0.050*
13	0.315	0.052	0.114	0.228	0.200	0.091
	0.340	*0.061*	*0.126*	*0.231*	*0.175*	*0.066*
14	0.288	0.056	0.112	0.194	0.202	0.147
	0.308	*0.069*	*0.132*	*0.200*	*0.179*	*0.112*
15	0.325	0.060	0.137	0.211	0.170	0.097
	0.351	*0.072*	*0.157*	*0.211*	*0.141*	*0.068*
16	0.286	0.072	0.085	0.185	0.201	0.171
	0.308	*0.085*	*0.097*	*0.190*	*0.187*	*0.134*
17	0.308	0.055	0.117	0.184	0.187	0.149
	0.340	*0.067*	*0.129*	*0.183*	*0.165*	*0.115*
18	0.382	0.064	0.095	0.161	0.151	0.147
	0.427	*0.068*	*0.097*	*0.159*	*0.132*	*0.117*
19	0.221	0.045	0.085	0.258	0.261	0.130
	0.232	*0.051*	*0.100*	*0.281*	*0.238*	*0.098*

Table 8 (continued)

20	0.165	0.047	0.065	0.213	0.257	0.252
	0.162	*0.058*	*0.087*	*0.247*	*0.246*	*0.201*
21	0.232	0.041	0.076	0.195	0.228	0.227
	0.243	*0.054*	*0.101*	*0.217*	*0.207*	*0.177*

Second, the bifactor model based on the results of the unrestricted FIFA was fit to the data, $\chi^2_{653} = 32{,}054.82$. Table 9 reports primary factor loadings and factor loadings on the five sub-domains. Similar to previous results, items reported moderately high loadings on the primary factor (Factor 1). Approximately half of the items reported slightly higher loadings on the primary factor compared to the results reported in Table 6. Within this model, the three most discriminating items were Item 21, $\lambda_{21,1} = 0.839$, Item 10, $\lambda_{10,1} = .801$, and Item 15 $\lambda_{15,1} = 0.776$; whereas the least discriminating items were Item 11, $\lambda_{11,1} = 0.663$, and Item 1, $\lambda_{1,1} = 0.681$, respectively. On average, secondary loadings were higher for this model than for the original PTGI model, indicating substantial residual association. Average loadings on the secondary factors were 0.375. Individual tests of the added value of each group factor resulted in a significant improvement in model fit with the inclusion of the group factor (*ps* < .001), indicating that each domain contributed to accounting for the relationships among items.

Table 10 reports item thresholds, while Table 11 shows the observed and expected proportions of respondents across categories. The root mean square error value of .018 indicates the difference between the observed and expected proportions (across all items and categories) was small, indicating substantial model data fit. Compared to the fit of the unidimensional model ($\chi^2_{674} = 33{,}249.29$, $p < .001$), the model resulted in a statistical improvement in model fit ($\chi^2_{Difference} = 1{,}194.47$, $df_{Difference} = 21$, $p < .001$).

Table 9

Full-Information Item Bifactor Analysis of PTGI based on Unrestricted Factor Analysis Results

Item	General	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1	0.681		0.550			
2	0.765		0.522			
3	0.747	0.159				
4	0.767					0.216
5	0.770	0.245				
6	0.682					0.340
7	0.682	0.372				
8	0.672	0.527				
9	0.726	0.518				
10	0.801	0.334				
11	0.663	0.449				
12	0.669	0.112				
13	0.714			0.195		
14	0.775					0.329
15	0.776					0.213
16	0.714					0.483
17	0.760				0.465	
18	0.714				0.623	
19	0.691			0.421		
20	0.786			0.434		
21	0.839					0.149

Table 10

Item Thresholds Based on Full-Information Item Bifactor Analysis of Re-specified PTGI Five-Factor Model

Item	0-1	1-2	2-3	3-4	4-5
1	-0.607	-0.436	-0.068	0.446	1.178
2	-0.538	-0.304	0.087	0.620	1.346
3	-0.354	-0.103	0.252	0.881	1.580
4	-0.473	-0.351	-0.033	0.531	1.276
5	-0.276	-0.072	0.274	0.981	1.729
6	-0.580	-0.365	-0.050	0.484	1.162
7	-0.288	-0.039	0.405	1.062	1.703
8	-0.076	0.168	0.559	1.186	1.715
9	-0.016	0.210	0.546	1.047	1.513
10	-0.267	-0.083	0.279	0.825	1.539
11	0.228	0.437	0.819	1.271	1.734
12	-0.663	-0.393	-0.032	0.810	1.631
13	-0.401	-0.246	0.068	0.697	1.495
14	-0.493	-0.308	0.030	0.552	1.206
15	-0.377	-0.192	0.199	0.796	1.475
16	-0.492	-0.269	-0.030	0.457	1.093
17	-0.407	-0.232	0.092	0.576	1.189
18	-0.181	-0.012	0.230	0.672	1.177
19	-0.736	-0.575	-0.298	0.417	1.272
20	-1.015	-0.772	-0.496	0.150	0.822
21	-0.711	-0.540	-0.257	0.291	0.906

Table 11

Observed and Expected (in Italics) Proportions from the Re-specified Five-Dimensional Graded Bifactor Analysis of PTGI Scale Data (N = 801)

	0	1	2	3	4	5
1	0.253	0.049	0.122	0.196	0.235	0.145
	0.272	*0.060*	*0.141*	*0.199*	*0.208*	*0.119*
2	0.272	0.067	0.132	0.201	0.215	0.112
	0.295	*0.085*	*0.154*	*0.198*	*0.178*	*0.089*
3	0.332	0.085	0.127	0.223	0.154	0.079
	0.362	*0.098*	*0.140*	*0.212*	*0.132*	*0.057*
4	0.297	0.040	0.107	0.205	0.215	0.136
	0.318	*0.045*	*0.124*	*0.215*	*0.197*	*0.101*
5	0.351	0.072	0.131	0.242	0.141	0.062
	0.391	*0.080*	*0.137*	*0.229*	*0.121*	*0.042*
6	0.262	0.066	0.110	0.200	0.208	0.154
	0.281	*0.077*	*0.122*	*0.206*	*0.192*	*0.123*
7	0.355	0.085	0.165	0.218	0.119	0.059
	0.387	*0.098*	*0.173*	*0.199*	*0.100*	*0.044*
8	0.427	0.091	0.146	0.192	0.087	0.056
	0.470	*0.097*	*0.145*	*0.170*	*0.075*	*0.043*
9	0.443	0.082	0.127	0.167	0.096	0.084
	0.395	*0.072*	*0.143*	*0.213*	*0.116*	*0.062*
10	0.352	0.062	0.132	0.232	0.136	0.085
	0.395	*0.072*	*0.143*	*0.213*	*0.116*	*0.062*
11	0.544	0.079	0.132	0.120	0.069	0.056
	0.590	*0.079*	*0.125*	*0.105*	*0.061*	*0.041*
12	0.241	0.079	0.122	0.310	0.180	0.069
	0.254	*0.094*	*0.140*	*0.304*	*0.157*	*0.051*
13	0.315	0.052	0.114	0.228	0.200	0.091
	0.344	*0.059*	*0.124*	*0.230*	*0.176*	*0.067*
14	0.288	0.056	0.112	0.194	0.202	0.147
	0.311	*0.068*	*0.133*	*0.197*	*0.177*	*0.114*
15	0.325	0.060	0.137	0.211	0.170	0.097
	0.353	*0.071*	*0.155*	*0.208*	*0.143*	*0.070*
16	0.286	0.072	0.085	0.185	0.201	0.171
	0.311	*0.083*	*0.094*	*0.188*	*0.187*	*0.137*
17	0.308	0.055	0.117	0.184	0.187	0.149
	0.342	*0.066*	*0.128*	*0.1871*	*0.165*	*0.117*
18	0.382	0.064	0.095	0.161	0.151	0.147
	0.428	*0.067*	*0.096*	*0.158*	*0.131*	*0.120*
19	0.221	0.045	0.085	0.258	0.261	0.130
	0.231	*0.052*	*0.100*	*0.279*	*0.237*	*0.102*

Table 11 (continued)

	0	1	2	3	4	5
20	0.165	0.047	0.065	0.213	0.257	0.252
	0.155	*0.065*	*0.090*	*0.250*	*0.235*	*0.206*
21	0.232	0.041	0.076	0.195	0.228	0.227
	0.239	*0.056*	*0.104*	*0.216*	*0.203*	*0.182*

Jenkin's Acitivity Survey

An unrestricted FIFA was conducted on the Jenkin's Activity Survey using the item responses of 600 men from central Finland drawn from a larger survey sample. Items are predominantly rated on three-point scales representing little or no, occasional, or frequent occurrence of the activity or behavior in question (e.g., pace of eating). Wording in the positive and negative direction varies across items. For this example, the item responses were recoded to provide an example of FIFA using dichotomous data. Appendix D provides the POLYFACT syntax for this analysis (including the recode procedure). Results indicated the presence of four distinguishable factors. Table 12 reports the factor loadings, based on promax rotation. Item and factor correspondences are judged based on the magnitude of the loading, irrespective of the loadings sign. Strong, positive loadings indicate that one's endorsement of the item corresponds to an increased standing on the underlying factor. On the other hand, a strong, negative loading suggests that an individual with a low standing on the underlying factor corresponds to an endorsement of the lowest response category. This would occur, for example, on a depression item ("Have you felt depressed over the past few days") administered to a non-depressed respondent. As shown, the numbers of items corresponding to each factor are 12, 9, 6, and 5. Strong consideration of the substantial cross-loadings of several items (e.g., Q269, Q313, Q276) should be ignored due to the collapsing of response categories. The percent of explained variance for each of the four factors was: 33.616, 15.803, 10.020, and 5.306, respectively.

Table 12

Factor Loadings of Jenkin's Activity Survey based on Promax Rotation

		Factor			
	Item	1	2	3	4
1	Q156	**0.325**	0.148	-0.063	-0.055
2	Q157	-0.211	0.131	**0.239**	0.203
3	Q158	-0.120	**-0.560**	0.055	-0.186
4	Q166	-0.124	**-0.767**	-0.082	0.265
5	Q247	0.364	**0.550**	-0.125	0.064
6	Q251	0.295	0.088	-0.166	**-0.349**
7	Q252	0.168	0.133	-0.149	**-0.633**
8	Q253	0.112	**0.549**	-0.217	-0.059
9	Q254	0.323	0.165	**-0.630**	0.080
10	Q257	0.030	-0.152	**-0.784**	-0.250
11	Q258	0.092	-0.121	**0.670**	0.001
12	Q259	0.381	-0.146	**1.050**	0.069
13	Q260	**0.801**	-0.069	0.275	-0.123
14	Q261	**0.798**	0.207	0.225	-0.111
15	Q262	**0.824**	0.226	0.292	-0.127
16	Q263	-0.463	0.054	-0.282	**-0.641**
17	Q265	0.120	0.006	**-0.681**	-0.278
18	Q266	**0.913**	0.024	0.104	0.228
19	Q267	0.197	-0.377	-0.454	**-0.515**
20	Q268	-0.123	-0.372	-0.064	**-0.736**
21	Q269	**0.444**	-0.232	-0.376	-0.367
22	Q270	**0.809**	-0.077	0.058	-0.065
23	Q272	**0.446**	0.212	0.040	-0.139
24	Q273	-0.012	**-0.713**	0.308	-0.206
25	Q275	-0.118	**-0.795**	-0.143	0.254
26	Q276	0.285	**-0.650**	0.273	-0.353
27	Q279	**1.095**	-0.027	-0.078	0.522
28	Q280	**1.115**	-0.161	-0.111	0.586
29	Q307	**0.685**	0.074	-0.364	0.049
30	Q308	**0.737**	-0.113	-0.108	-0.079
31	Q313	-0.046	**-0.797**	-0.085	-0.532
32	Q314	0.246	**-0.730**	-0.157	0.019

Note. Bolded items indicate highest factor loadings.

Table 13 reports factor correlations, indicating the relationship between the underlying traits. As shown, low to moderate correlations were reported between Factors 1 and 2 and

Factors 3 and 4. On the other hand, moderate negative correlations were reported between Factor 1 and Factors 3 and 4.

Table 13

Correlation between Factors underlying Jenkin's Activity Survey Data

	1	2	3	4
1	1.000			
2	0.243	1.000		
3	-0.427	-0.160	1.000	
4	-0.518	-0.053	0.322	1.000

The aforementioned examples demonstrate the use of multidimensional IRT models in health outcomes research. The first example tested the PDSQ factor structure. The theoretically-based 15 sub-domains of the scale provided the basis to conduct a confirmatory-based FIFA using the bifactor model (Gibbons & Hedeker, 1992). The bifactor solution was of interest as the scale is designed to (a) serve as a general psychiatric screening instrument, and (b) provide clinical information on a range of symptom domains (e.g., Mania, Major Depression Disorder, Alcohol Abuse). The second example was based on breast cancer survivor data from the PTGI (Tedeschi & Calhoun, 1996). Empirical evidence questioning the stability of the PTGI factor structure across samples justified conducting an unrestricted factor analysis prior to fitting the bifactor model to the data. Finally, analysis of the Jenkin's activity scale demonstrated the recode of variables to conduct unrestricted analyses on dichotomous data. Implications of these analyses are briefly reviewed.

Results of the first study suggested that the bifactor model provided a plausible description of the PDSQ scale data. The correlation between observed and estimated symptom response items over the 139 items was $r = 0.9992$. Relative to a unidimensional model, the bifactor model with all 15 symptom domains provided a significant improvement in fit relative to

both the unidimensional model, and each of the 15 bifactor models with 14 domains (i.e., leaving one domain out in each). Secondary factor loadings were generally moderate to large, indicating a substantial relationship between the observed and latent variables. These findings indicate that all of the 15 symptom domains are required to adequately model these data. That is, they provide evidence for the validity of these 15 diagnostic symptom domains. While all items are correlated through their relationship to the primary mental illness dimension, loadings on each of the 15 diagnostic symptom domains reveal that they are non-zero and reflect their relative independence. Furthermore, results clearly indicate that inclusion of these domains dramatically improves the fit of the model over a unidimensional model that does not consider the group factors; this does not necessarily mean that they are qualitatively distinct "diseases."

The structure of the PTGI was modeled in terms of unidimensional and multidimensional IRT models. Samejima's (1969) graded response model was fit to the data to provide a basis to illustrate the effect of fitting a unidimensional model to multidimensional data. A subsequent dimensionality assessment of the PTGI based on an unrestricted factor analysis indicated that the scale deviated from its original factor structure postulated by Tedeschi and Calhoun (1996). Specifically, results suggested a primary domain in addition to minor group factors. Suggestion of a primary domain underlying the scale data is not surprising given the sampling of items from sub-domains related by a broad post-traumatic growth factor.

Overall model fit and parameter estimated supported the fit of the bifactor model to the PTGI scale data. Specifically, moderate to high factor loadings were reported on the primary dimension, combined with low to moderate loadings on the group factors. As such, the primary dimension was found to be an important component for accounting for the relationship among scale items. Furthermore, the EAP trait estimates on the primary dimension are adjusted in

consideration of the group factors. Results of this research complement other research that has reported that the bifactor model can contribute to modeling scale data in health outcomes research (e.g., Chen, West, & Sousa, 2006; Gibbons & Hedeker, 1992; Gibbons et al., 2007; Reise, Morizot, & Hays, in press).

An unrestricted FIFA of the Jenkin's Activity Scale (1972) supported a four factor solution. Although currently available POLYFACT program now facilitates analyzing this scale without recoding the data, nevertheless this example serves to demonstrate the program's capabilities. For this analysis, results supported the presence of four factors underlying the scale data. Factor correlations were low to moderate, indicating the distinctiveness of emergent factors.

11. Computer adaptive testing (CAT)

a. Underlying theory

The past thirty years have witnessed an exponential increase in the use of test administration within the framework of CAT. Within health outcomes research, for example, adaptive tests have been found to be more efficient than conventional tests (Fliege, Becker, Walter, Bjorner, Klapp, & Rose, 2005; Ware, Bjorner, & Kosinski, 2000). That is, in an adaptive test a given level of measurement precision can be reached much more quickly than in a test in which all examinees are administered the same items. This situation results from selecting items that are most informative for an individual at each stage of test administration in the adaptive test. Typical adaptive tests result in a 50% average reduction in number of items administered, and some reductions in the range of 80% to 90% have been reported, with no decrease in measurement quality (Brown & Weiss, 1977). In addition, as has been indicated, adaptive tests allow control over measurement precision. Thus, adaptive tests result in measurements that are

both efficient and effective. Recent research (Gibbons et al., 2007b) using the 615 items of the Mood-Anxiety Spectrum Disorders Scale (MASS) has shown test length reductions due to adaptive testing procedures (based on the bifactor model) averaging 95% (615 to and average of 24 items for the general dimension), in both post-hoc simulation and live testing, on the General dimension of the MASS. Actual testing time required to score the General scale and the four content scales was reduced from an average of 115 minutes to 22 minutes.

Research since the 1970s has shown that adaptive testing procedures are most effective when combined with IRT procedures (e.g., Kingsbury & Weiss, 1980, 1983; McBride & Martin, 1983). Thus, an item bank for use in adaptive testing can be calibrated according to an IRT model. The point at which a test is to be started (frequently referred to as the "entry point") can be determined by taking into account individual status variables or other data about an individual (e.g., previous scores, age, gender, clinical evaluations). Explicit procedures for estimating an entry point for an adaptive test are available in conjunction with IRT using Bayesian statistical methods (e.g., Baker, 1992, Chap. 7; Weiss & McBride, 1984). IRT procedures for estimating an individual's trait level are applicable to the adaptive testing process. Procedures of maximum likelihood or Bayesian estimation permit estimation of trait, or in this case impairment levels, based on one or more responses made by a single individual in an adaptive test. Thus, a continuous updating of the impairment level can be accomplished after each item is administered in an adaptive test, and the next item to be administered can be based on the impairment estimate derived from all previous items administered. Item selection rules derived from IRT and adaptive testing can explicitly use concepts of item information (Hambleton & Swaminathan, 1985, Chap. 6; Weiss, 1985). Thus, at a given current impairment estimate the most informative item not yet administered can be chosen for administration. When items are selected using this maximum

information item selection rule, the net effect is an extremely efficient procedure for reducing the error of measurement at each successive stage in the administration of an adaptive test (Weiss, 1985).

Finally, adaptive testing procedures developed in accordance with IRT can take advantage of a number of different procedures for terminating an adaptive test. One procedure frequently applied, however, is to reduce the individualized standard error of measurement (SEM) to an *a priori* specified level before a test is terminated (Weiss & Kingsbury, 1984). This technique allows the number of items administered to an individual to vary, but it also results in control of the resultant level of SEM for the individual tested and through the standard errors the reliability of scores. Thus, for the individual who responds essentially in accordance with the IRT model, a given level of SEM will be achieved more quickly than for an individual who does not respond in accordance with the IRT model, resulting in a slower reduction of the individualized SEM.

b. Case study - PDSQ example

A demonstration of CAT in the context of PRO testing is provided by conducting a post-hoc simulation administration of the PDSQ (Zimmerman & Mattia, 2001). Post-hoc simulation of the PDSQ full scale and subscale items within a CAT environment was conducted using POSTSIM 2.0 (Weiss, 2005). The premise of the program is to simulate the administration of a particular scale within a CAT setting using previously obtained data. Therefore, with examinee scale data and previously estimated item parameters (e.g., discrimination, difficulty) the test is re-administered within the framework of CAT to: (a) estimate the number of items that each individual would need to be administered to obtain an accurate trait estimate, (b) identify the items used to assess each respondent, and (c) determine the point on the trait continuum (e.g.,

depression, drug use) where the instrument provides the most informative measurement (e.g., low, middle, high).

For the PDSQ data, POSTSIM 2.0 (Weiss, 2005) modeled each respondent's probability of a "Yes" response based on a previously estimated item bifactor model. The program provides options for score estimation, item selection, and termination of testing. A range of prior distributions (e.g., -3, -2, 0, 1, etc.) were placed on examinees' initial theta estimates (e.g., level of depression) to investigate the effect this had on item administration. Bayes EAP scores were used to indicate trait levels. Item administration was based on selecting the next item that had the maximum amount of information for trait estimation, which results in the fastest decrease in the standard error in measurement (SEM) (Weiss, 2005). Last, two fixed SEM termination criteria to stop testing were investigated: (a) after the information of the next item to be administered fell below 0.30, and (b) after the information of the next item to be administered fell below 0.40.

Results of Post-hoc simulation of CAT

Post-hoc simulation of PDSQ full scale items was conducted first to determine the number of items required to obtain trait estimates on the general psychiatric dimension. Subsequently, the analysis was conducted for each of the 15 PDSQ subscales, separately. Statistics used to judge the efficiency of the CAT session included: (a) the correlation between the full theta and CAT theta, with acceptable values above 0.90; (b) mean/average signed difference between full theta and CAT theta, with values close to zero desired; (c) mean/average absolute difference between full theta and CAT theta, with values close to zero desired; and (d) mean number of items administered.

Post-hoc simulation of the PDSQ full scale items (n = 139) was conducted. Table 14 reports the CAT results for this analysis based on a fixed SEM termination of .40. As shown,

regardless of the implied prior distribution, the correlation between the CAT theta and full theta were virtually the same, as was the mean number of items administered. This indicated that the initially specified trait distribution did not influence trait estimation. However, the correlation between the CAT theta and full theta estimates was lower than the desired (< 0.90), indicating less than acceptable trait estimates. On average, less than 11 of the total 139 PDSQ items were administered, with only 5 items administered in some cases and all 139 items in other instances.

Table 14

CAT Results of Fixed Standard Error of Measurement Termination of 0.40

General Factor (139 items)	SEM						
Prior θ	0,1	1,1	2,1	3,1	-1,1	-2,1	-3,1
r (θ)	0.838	0.836	0.840	0.840	0.842	0.838	0.842
Mean θ diff	-0.182	-0.186	-0.149	-0.149	-0.160	-0.161	-0.165
Mean abs diff	0.382	0.382	0.371	0.371	0.374	0.378	0.377
Mean no. items	10.117	10.175	10.987	10.987	10.426	10.537	10.887
Range items	5-139	5-139	5-139	5-139	5-139	6-139	6-139

Note: *r* (*θ*) = correlation of full theta and CAT theta; Prior (*θ*) = Prior distribution of examinee trait estimate; Mean θ diff = Mean/average signed difference between full theta and CAT theta; Mean abs diff. = Mean/average absolute difference between full theta and CAT theta; Mean no. items = Mean number of items administered to estimate CAT theta.

Next, post-hoc simulation of the PDSQ full scale items was conducted using a more restrictive fixed SEM termination of 0.30. As reported in Table 15, the prior distribution did not appreciably affect results. Inspection of the table indicates that the correlation between the CAT theta and full theta exceeded the cutoff value of 0.90, with all values nearly equal within rounding. The mean number of items administered was 22, or double the number required when the fixed SEM termination was set at 0.4. Furthermore, acceptable trait estimates were obtained by administering about 117 less items than the total scale. As shown, the number of items administered ranged between 10 and 139.

Table 15

CAT Results of Fixed Standard Error of Measurement Termination of 0.30

General Factor (139 items)	SEM						
Prior θ	0,1	1,1	2,1	3,1	-1,1	-2,1	-3,1
r (θ)	0.925	0.927	0.924	0.924	0.925	0.924	0.925
Mean θ diff	-0.153	-0.155	-0.142	-0.142	-0.147	-0.155	-0.159
Mean abs diff	0.262	0.259	0.260	0.260	0.260	0.264	0.266
Mean no. items	21.239	21.457	22.276	22.276	21.700	21.680	22.044
Item Range	10-139	10-139	11-139	11-139	11-139	11-139	11-139

Note: r (θ) = correlation of full theta and CAT theta; Prior (θ) = Prior distribution of examinee trait estimate; Mean θ diff = Mean/average signed difference between full theta and CAT theta; Mean abs diff. = Mean/average absolute difference between full theta and CAT theta; Mean no. items = Mean number of items administered to estimate CAT theta.

The next analysis conducted a post-hoc simulation of PDSQ subscales. Based on the previous finding that trait estimation was not affected by the imposed prior distribution, a normal prior was used in subsequent analyses. Table 16 reports CAT results for each of the PDSQ subscales based on a fixed SEM termination of 0.4. Inspection of the results indicates that the correlation between the CAT theta and full theta was generally contingent on the number of items within the respective subscale. For instance, correlations were higher for subscales with greater than 10 items, with the exception of the Hypochondriasis subscale. However, these correlations were less than 0.90 for most of the subscales comprised of less than 10 items, including MANIA and PSYCH. For MANIA, an average of 1 item was administered, indicating that no other items contributed to trait estimation based on this fixed SEM termination criteria.

Mean differences between the CAT theta and full theta indicated differences less than a half of a point, suggesting accurate CAT trait estimates. As shown, the mean number of items administered within each subscale was less than half of the total number of subscale items, a 50% reduction in the total number of administered items. For several of the subscales comprised

of less than 10 items, an average of 1 item was administered. Specifically, it appears that one item provided the most information for trait estimation in the MANIA, PSYCH, ALC, and DRUG subscales, each consisting of 6 items. Overall, an average of 56 items is required, which represents a 60% reduction in item administration relative to the full scale.

The next analysis conducted a post-hoc simulation of the subscale items with a fixed SEM termination of 0.30. As reported in Table 17, with the exception of MANIA, all subscale CAT thetas reported correlations at or above 0.90 with full thetas. Compared to the fixed SEM termination of 0.40, the correlations reported in this condition were not substantially larger, except for the subscales with less than 10 items (e.g., PSYCH, ALC). As expected, the mean difference between CAT theta and full theta decreased across subscales with this less stringent termination criterion. Inspection of the table indicates that on average the mean number of items increased by one, or two, items, depending on the number of subscale items. Furthermore, the range of the administered items decreased, as the lower range value generally increased by one or two items, whereas the upper range value typically remained unchanged. Overall, an average of 79 items is required, which represents a 43% reduction in item administration relative to the full scale.

Table 16

POSTSIMS Results of Simulated Administration of Group Factor Items – Terminate When Item Information is below .40

	Group Factor														
	MDD	DYS	PTSD	BUL	OCD	PAN	MANIA	PSYCH	AGOR	SOC	ALC	DRUG	GAD	SOM	HYPO
No. Items	21	7	15	10	8	8	6	6	11	15	6	6	10	5	5
Prior θ	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
r (θ)	.951	.962	.981	.952	.906	.966	.742	.816	.976	.969	.899	.899	.941	.925	.980
Mean θ diff	.150	.015	-.095	-.086	-.191	-.095	-.159	-.280	-.191	-.091	-.038	-.048	.092	-.130	-.041
Mean abs. Diff.	.239	.165	.178	.215	.340	.241	.442	.399	.263	.202	.219	.147	.299	.248	.091
Mean no. items	5.28	4.79	8.83	3.70	2.62	4.15	1.00	1.00	4.64	8.51	1.00	1.00	3.82	2.20	3.00
Item Range	4-7	3-7	4-14	2-8	2-6	3-6	1-1	1-1	3-8	4-13	1-1	1-1	3-5	2-3	2-5

Note: r (θ) = correlation of full theta and CAT theta; Mean θ diff = Mean/average signed difference between full theta and CAT theta; Mean abs diff. = Mean/average absolute Difference between full theta and CAT theta; Mean no. items = Mean number of items administered to estimate CAT theta. 1 = Major Depression (MDD); 2 = Dysthymia (DYS), 3 = Post-Traumatic Stress (PTSD); 4 = Bulimia Nervosa (BUL); 5 = Obsessive Compulsive (OCD); 6 = Panic (PAN); 7 = Mania (MANIA); 8 = Psychosis (PSYCH); 9 = Agoraphobia (AGOR); 10 = Social Phobia (SOC); 11 = Alcohol Abuse (ALC); 12 = Drug Abuse (DRUG); 13 = Generalized Anxiety (GAD); 14 = Somatoform (SOM); 15 = Hypochondriasis (HYPO).

Table 17

POSTSIMS Results of Simulated Administration of Group Factor Items – Terminate When Item Information is Below .30

	Group Factor														
	MDD	DYS	PTSD	BUL	OCD	PAN	MANIA	PSYCH	AGOR	SOC	ALC	DRUG	GAD	SOM	HYPO
No. Items	21	7	15	10	8	8	6	6	11	15	6	6	10	5	5
Prior θ	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
r (θ)	.964	.968	.986	.976	.908	.981	.879	.942	.948	.985	.942	.899	.956	.956	.980
Mean θ diff	.123	-.011	- .082	-.081	-.191	-.053	-.094	-.182	-.038	-.069	-.042	-.048	.076	-.054	-.041
Mean abs. Diff.	.203	.011	.149	.154	.337	.149	.282	.247	.038	.138	.042	.147	.261	.186	.091
Mean no. items	6.07	6.29	9.84	5.35	2.70	5.63	2.53	2.33	5.40	10.15	2.79	1.00	4.46	3.00	3.00
Item Range	5-8	5-7	5-14	3-10	2-6	4-8	2-6	2-6	3-10	6-14	2-6	1-1	4-5	3-3	2-5

Note: r (θ) = correlation of full theta and CAT theta; Mean θ diff = Mean/average signed difference between full theta and CAT theta; Mean abs diff. = Mean/average absolute Difference between full theta and CAT theta; Mean no. items = Mean number of items administered to estimate CAT theta. 1 = Major Depression (MDD); 2 = Dysthymia (DYS), 3 = Post-Traumatic Stress (PTSD); 4 = Bulimia Nervosa (BUL); 5 = Obsessive Compulsive (OCD); 6 = Panic (PAN); 7 = Mania (MANIA); 8 = Psychosis (PSYCH); 9 = Agoraphobia (AGOR); 10 = Social Phobia (SOC); 11 = Alcohol Abuse (ALC); 12 = Drug Abuse (DRUG); 13 = Generalized Anxiety (GAD); 14 = Somatoform (SOM); 15 = Hypochondriasis (HYPO).

12. Review of available software

BILOG-MG

BILOG-MG is a windows-based IRT program designed for the analysis of binary items including multiple-choice or short answer items scored right, wrong, omitted, or not-presented. It is effective for the application of the 1-, 2-, and 3- PL unidimensional models. It is a large-scale production application and capable of handling an unlimited number of items and respondents, thus making it a valuable tool for all stages of scale development and maintenance. It can simultaneously conduct item analysis and scoring for any size subtests or subscales. Parameter estimation is based on the method of Bock and Aitkin (1981). Output for resultant parameter and scoring estimates are in files suitable for purposes of subsequent analyses or score reporting.

Additionally, the program extends the IRT models to multiple group analysis. Applications in educational and clinical settings include: Differential item functioning (DIF), vertical scaling, nonequivalent groups equating to maintain scale comparability as new test forms are designed, detecting item parameter drift over time, calibrating and scoring tests in two-stage testing procedures to reduce total testing-time, and estimating latent ability or proficiency distributions for diverse populations (e.g., students, patients).

MULTILOG

MULTILOG extends the capabilities of BILOG-MG by also modeling MULTIple categories through the use of LOGistic IRT models. In addition to the traditional IRT models, multi-categorical models include: Samejima's (1969) graded response model, Bock's (1972) nominal (non-ordered) response model, and Thissen and Steinberg's (1984) model for multiple – choice items. As such, it provides a powerful IRT program for the analysis of items comprised of multiple response categories, such as Likert-type scales (e.g., strongly disagree – strongly agree).

MML is used for item parameter estimation in the presence of unknown trait values, and ML estimates with known trait values. The likelihood-ratio chi-square statistic provides a measure of model-data fit. The difference between likelihood-ratio chi-square statistics based on models in which specific item parameters are either constrained or freely estimated across two groups (e.g., males vs. females) provides the basis for investigations of differential item functioning, or item bias. The program also is effective for equating analyses.

TESTFACT

The TESTFACT program is capable of implementing all the procedures of classical item analysis, test scoring, and factor analysis of inter-item tetrachoric correlations. Additionally, it can conduct exploratory IRT-based full-information factor analysis (FIFA) for an unlimited number of items and a maximum of fifteen factors. Additionally, it can model data with an underlying bifactor structure. Its simulation feature enables it to simulate various data types, based on user specified parameters (e.g., slopes, thresholds). As with the other IRT programs resultant output files can be used for subsequent statistical analyses and test score reporting.

For the exploratory FIFA, three distinct methods of multidimensional numerical integration for the E-step of the EM algorithm are provided, including: adaptive quadrature, non-adaptive quadrature, and Monte Carlo integration. Both adaptive and non-adaptive methods can be used to estimate scores on each factor. Estimation of the classical reliability of the factor scores is also available.

For the confirmatory full-information bifactor analysis, Bayes estimation of scores on the general factor is included, as well as standard errors that account for variation among responses due to the group factors.

POLYFACT

POLYFACT is newly developed IRT software that expands the capacity of TESTFACT by providing for polytomously scored items. It also includes the dichotomous and polytomous bifactor models of Gibbons and co-workers (2007a).

13. Summary and Conclusions

The emergent use of self-report instruments in health outcomes research settings provides the basis for applying state-of-the-art analyses to determine the extent to which obtained scores can be used for subsequent decision-making purposes. As shown, practitioners and researchers alike are faced with notable decisions when modeling such data. As a psychometric technique, IRT offers a powerful, flexible method to handle PRO measurement data throughout all stages of scale development, maintenance, and scoring. Nevertheless, the use of advanced modeling procedures (i.e., IRT) has so far received comparative little use on psychological research (Borsboom, 2007). Until recently, the unidimensionality and local independence requirements of IRT have largely limited its use with modeling psychological scale data, which is typically multidimensional. This should change as the advancements in IRT discussed here permit its use for dimensionality assessment and scoring of scales in studies that are both exploratory and confirmatory in nature.

For the field of health outcomes research, the variety of IRT models provides a host of data analytic tools to handle both dichotomously and polytomously scored items. Full-information item factor analysis extends the traditional unidimensional models to IRT-based methods to handle complex data structures (e.g., more than one underlying latent variable). These methods are now extended to polytomously scored items. These advancements enable

PRO researchers to apply IRT-based procedure through all phases of test development and scoring.

As was presented, there are a host of IRT models to analyze various types of PRO data. Traditional unidimensional IRT models have received the most extensive treatment across testing contexts. Most promising to modeling PRO scale data are the recently developed multidimensional IRT models. The item factor analytic IRT models overcome the restrictive requirement of a unidimensional test structure, an untenable assumption in most PRO testing situations. Until recently, the IRT-based factor analytic procedures were exploratory in nature. Specifically, they did not (a) rely on a priori information to determine the number of underlying latent traits, and (b) provide researchers the ability to specify the relationships between items and factors. These methods relied on testing the statistical difference between the likelihood values of models with and without a factor to determine the number of underlying latent traitS.

Gibbons and Hedeker (1992) and Gibbons et al. (2007a) derived the full-information item bifactor model for dichotomously and polytomously scored items respectively. The model represents the first confirmatory-based IRT model to test the dimensionality of scale data. It is unique in that it relies on a priori theoretical considerations to test the relationship between the observed and latent variables. Advantages of the bifactor restriction leads to a major simplification of likelihood equations that (a) permits analysis of models with large numbers of group factors (*e.g.*, domains), (b) permits conditional dependence among identified subsets of items, and (c) in many cases provides a more parsimonious factor solution than an unrestricted full-information item factor analysis (e.g., Bock & Aitkin, 1981).

The simulation and applied data studies presented demonstrate some of the advantages (e.g., accurate trait & parameter estimates) that multidimensional IRT has to offer PRO research.

The simulation study showed several significant benefits of applying the bifactor model over Samejima's (1969) unidimensional graded response model to data with varying degrees of multidimensionality. First, compared to the unidimensional model, the bifactor model yielded theta estimates that were more homogeneous across simulated data structures. Second, PSD estimates were found to be underestimated across all conditions for the unidimensional model. This will lead to premature conclusion of CAT testing sessions and a false sense of precision with which the underlying trait is estimated. Third, the larger empirical standard deviations for the unidimensional model lead to decreased statistical power for between group comparisons, and will require larger sample sizes than would be required if the theta values were estimated by the correct multidimensional model. Fourth, the mean log-likelihood values always indicated statistically significant improvement in fit for the bifactor model as compared to the unidimensional model, even when the data had only mild departure from unidimensionality. Overall, multidimensional models can be expected to provide more reliable estimates of the underlying impairment dimension and more accurate estimates of uncertainty, relative to their unidimensional counterparts.

The real data examples illustrated how exploratory and confirmatory-based factor analytic IRT procedures can be used to model health outcomes scale data. Within the applied data examples, the bifactor model was found to provide acceptable model-data fit. For the PDSQ data, items reported varied loadings on the primary dimension and, in general, strong loadings on the group factor they were designed to measure. Analysis of the PTGI data indicated that the scale's original factor structure did not provide the best fit to the data. Although, for the most part, emergent factors were similar to the original factors, the original factor structure reported by Tedeschi and Calhoun's (1996) was less salient that those reported here using IRT-based

factor analysis. However, the finding of dissimilar factor structure in the current study compared to the original PTGI factor structure is not surprising given the heterogeneity of the samples (i.e., undergraduate college students vs. breast cancer survivors).

The fit of the bifactor model to the PTGI data indicated the presence of a general posttraumatic growth factor. Primary factor loadings exceeded 0.65, indicating a strong relationship between the observed variables and the general factor. Inspection of secondary factor loadings indicated high residual association among scale items. Testing the fit between competing bifactor models indicated that the PTGI cannot be considered a unidimensional model.

The results of the post-hoc simulation indicated that the PDSQ can be administered within the context of CAT and be expected to provide accurate trait estimates without administering all full scale or subscale items. A fixed SEM termination of .30 was found to be acceptable for the overall PDSQ scale. CAT trait estimates exhibited correlations with full scale trait estimates above the established cutoff criteria. For the PDSQ primary dimension, testing required the administration of roughly 22 items, or about 16% of the total test items. This clearly demonstrates the efficiency that CAT administration of the PDSQ has to offer to diagnostic evaluation setting. For example, a CAT-based primary dimension theta estimate could be used as a general screening test for depression in a primary care setting, and if positive (i.e., above a clinically relevant threshold), further adaptive testing of sub-domains (such as those associated with depression, including: mania, major depressive disorder) could be pursued. Furthermore, this finding supports previous research regarding the efficiency in which CAT has to offer across testing environments (e.g., Brown & Weiss, 1977; Fliege et al., 2005; McBride & Martin, 1983; Ware et al., 2000). Even testing for all sub-domains resulted in a 50%-60% reduction in test

items administered via CAT relative to traditional full-scale administration. Furthermore, it is very rare for clinicians in primary care settings to measure change in depression over time with their patients. As such, the empirical evidence of this research indicates that CAT could be a very efficient way to accomplish this goal.

There are a plethora of factors to consider when applying IRT to model mental health data. Despite their obvious desirability, clear-cut guidelines to identify the "best" IRT model to use for a given data set are elusive. This is largely attributed to the myriad of unique factors encountered when seeking to model any given dataset. In any given instance, these factors include: availability of the theoretical structure of the scale, sample size, and number of factors, among many. Nonetheless, initial consideration should be leveled at the theory used to guide scale development. This was the general approach to model the PTGI data above. That is, the availability of *a priori* information regarding the nature of the relationships between the observed and latent variables suggests that a confirmatory-based modeling approach is appropriate. Contrary, the absence of theory or the presence of uncertainty regarding the number of factors underlying the data hints at justification for conducting an exploratory-based analysis.

Aside from these considerations, additional research is needed to indicate the critical factors in selecting an appropriate IRT model. For instance, Riese et al. (in press) discuss several added benefits of including a primary dimension in addition to the theoretically *a priori* specified group factors in modeling health outcomes. However, areas in which empirical evidence is needed include (a) sample size, (b) the magnitude of the correlation between factors to be considered distinct dimensions, and (c) the accuracy of parameter and trait estimates for the different models under various conditions (e.g., non-normal data, sparse data, etc.). As such,

considerable research is needed to explore the applicability of multidimensional IRT models to various types of scale data.

The aim of this workbook was to provide researchers information on the added value of multidimensional IRT models over simpler unidimensional alternatives. As demonstrated, there are serious consequences associated with fitting unidimensional models to multidimensional data. Since most PRO measures are inherently multidimensional, investigators should use an appropriate multidimensional IRT model in the analysis and scoring of their data. The FI bifactor model represents one type of multidimensional IRT procedure capable of modeling data with a multidimensional structure. Notably, the use of the bifactor model as a method to describe health outcomes measurements has recently begun to emerge (e.g., Chen et al., 2006; Gibbons et al., 2007; Riese et al. in press). As such, the bifactor model seems like a plausible psychometric modeling technique to test the theoretical structure of various types of PRO instruments. Further research into the application of multidimensional IRT models to PRO data is strongly encouraged.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders, 4th edition.* Washington, DC: Author.

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

Ansley, T. M. & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 39-48.

Baker, F. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Beck, D. T., & Gable, R. K. (2001). Item response theory in affective instrument development: An illustration. *Journal of Nursing Measurement, 9*, 5-22.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). 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. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill. (1985 reprint, Scientific Software International, Chicago.)

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bock, R. D., & Gibbons, R. D. (in press). Item Factor Analysis of Polytomous Response Data. In R. Ostini & M. Nering (Eds.), *Handbook of Polytomous Item Response Theory Models: Development and Applications.*

Bock, R. D., Gibbons, R. D., & Schilling, S. (in press). *POLYFACT User's Manual*.

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

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425-440.

Burt, M. R., & Katz, B. L. (1987). Dimensions of recovery from rape: Focus on growth outcomes. *Journal of Interpersonal Violence, 2*, 57-81.

Camilli, G. (1994). Origin of the Scaling Constant "d" = 1.7 in Item Response Theory. *Journal of Educational & Behavioral Statistics, 19*, 293-95.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items and between tests. *Psychometrika, 26*, 347 – 372.

Collins, R. L., Taylor, S. E., & Skokan, L. A. (1990). A better world or a shattered vision? Changes in life perspective following victimization. *Social Cognition, 8*, 263-285.

Cook, L., & Eignor, D. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37-45.

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

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item Parameter Recovery for the Nominal Response Model. *Applied Psychological Measurement*, *23*, 3-19.

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

Du Toit, M. (Ed.) (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International, Inc.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research, 14*, 2277-2291.

Folk, V. G. & Green, B.F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, 13, 373-389.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychomatrika, 57*, 423-436.

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

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A. Grochocinski, V. J., Bhaumik, D. K., & Stover, A. (2007b). Mental health computerized adaptive testing. *Submitted for publication*.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement, 3rd edition*. (pp. 147-200). Phoenix, AZ: American Council on Education/Macmillan Publishing.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kuwer Academic Publishers.

Harmon, H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: The University of Chicago Press.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational. Measurement: Issues and Practice, 8*, 35-41.

Ho, S. M., Chan, C. L. W., & Ho, R. T. H. (2004). Posttraumatic growth in Chinese cancer survivors. *Psycho-Oncology, 13*, 377-389.

Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika*, 2, 41-54.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249 -260.

Jenkins, C. D., Rosenman, R. H., & Zyzanski, S. J. (1972). *The Jenkins Activity Survey of Health Prediction*. New York: The Psychological Corporation.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.

Lazarfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362-412). New York: Wiley.

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

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

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

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Muraki, E. (1983). *Marginal maximum likelihood estimation for three-parameter polychotomous item response models: Application of and EM algorithm*. Doctoral Dissertation, University of Chicago.

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika,* 54, 557-585.

Múthen, L. K., & Múthen, B. O. (1998-2006). *Mplus user's guide. Fourth Edition*. Los Angeles, CA: Múthen & Múthen.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Reise, S. P., Morizot, J., & Hays, R. D. (in press). The Role of the bi-factor model in resolving dimensionality issues in health outcomes measures, *Medical Care*.

Reckase, M. D. (1979). Unidimensional latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Samejima F (1969), Estimation of latent ability using a response pattern of graded scores, *Psychometrika Monograph Supplement*, 17, 1-68.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533-555.

Sheikh, & Marotta (2005). A cross-validation study of the Posttraumatic Growth Inventory. *Measurement and Evaluation in Counseling and Development, 38*, 66-77.

Sledge, W. J., Boydstun, J. A., & Rabe, A. J. (1980). Self-concept changes related to war captivity. *Archives of General Psychiatry, 37*, 430-443.

Stout, W. F, Habing, B., Douglas, J., Kim, H, Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331-354.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), *New horizons in testing* (pp. 9 – 30). New York: Academic Press.

Tedeschi, R. G., & Calhoun, L. G. (1996). The posttraumatic growth inventory: Measuring the positive legacy of trauma. *Journal of Traumatic Stress, 9*, 455-472.

Thissen, D., Chen, W., & Bock, D. (2003). MULTILOG for Windows (Version 7.0) [Computer Program]. Chicago, IL: Scientific Software International.

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*, 567-577.

Thissen, D. Steinberg, L, & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 149-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response theory. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-114). Hillsdale, NJ: Lawrence Erlbaum.

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

Thurston, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, 23, 111-136.

Veronen, L. J., & Kilpatrick, D. G. (1983). Rape: A precursor of change. In E. J. Callahan & K. A. McCluske (Eds.), *Life-span developmental psychology: Nonnormative life events* (pp. 67-191). New York: Academic Press.

Wainer, H., Dorans, N., Eignor, R., Flaugher, B., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (eds.) (2000). *Computerized adaptive testing: A primer* (Second Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ware, J. E., Bjorner, J. B., Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing. *Medical Care, 38*, 73-82.

Way, W. D., Ansley, T.N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12, 239-252.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 774-789.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37 (2)*, 70-84.

Weiss, D. J. (2005). *Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0*. St. Paul MN: Assessment Systems Corporation.

Weiss, D. J., & Kingsbury, G. G. (1984), Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing, *Applied Psychological Measurement*, 8, 272-285.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed.). Westport, CT: American Council on Education and Praeger Publishers.

Zimmerman, M., & Mattia, J. I. (2001). A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry*, 58, 787-794.

Appendix A

MULTILOG Syntax for fitting Samejima's (1969) Graded Response Model to PTGI (Tedeschi & Calhoun, 1996)

```
MULTILOG Example
Syntax for fitting SAMEJIMA'S (1969) GRADED MODEL to PTGI DATA (N=801)
>PROBLEM RANDOM, INDIVIDUAL, NITEMS=21, NGROUP=1,
      NEXAMINEES=801, DATA= PTGI.DAT';
>TEST    ALL, GRADED,
      NC=(6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6);
>SAVE;
>LABELS  ITEMS=(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21),
NAMES=('I1','I2','I3','I4','I5','I6','I7','I8','I9','I10','I11','I12','I13','Ii14','I15','I16',
            'I17','I18','I19','I20','I21');
>ESTIMATE NCYCLES=100;
>END;
6
123456
11111111111111111111
22222222222222222222
33333333333333333333
44444444444444444444
55555555555555555555
66666666666666666666
(21A1)
```

Appendix B

POLYFACT Syntax for Conducting Unrestricted Full-Information Item Factor Analysis for Polytomous Data on PTGI (Tedeschi & Calhoun, 1996)

```
>TITLE
Unrestricted item factor analysis of PTGI data:
Example Program
>PROBLEM NITEM = 21, RESPONSE = 7, MAXCAT = 6;
>RESPONSE X, 1, 2, 3, 4, 5, 6;
>CATEGORY NCAT = (6(0)21);
>FACTOR NFACT = 5, NROOT = 6, ROTATE = PROMAX;
>FULL CYCLES = 40, QUAD = 3;
>PRIOR;
>INPUT NIDCH = 4, NFMT = 1, FILE= 'PTGI.DAT';
(4A1, T1, 21A1)
>STOP
```

Appendix C

POLYFACT Syntax for fitting Bifactor Model (Gibbons et al., 2007a) for Graded Response Data to PTGI (Tedeschi & Calhoun, 1996)

```
>TITLE
   Polytomous item bifactor analysis of the PTIG data:
     Example Program
>PROBLEM NITEM=21, RESPONSE=8, MAXCATEGORY=6, NOTPRESENTED;
>RESPONSE X, 1, 2, 3, 4, 5, 6, 0;
>CATEGORY NCAT = (6(0)21);
>BIFACTOR NIGROUP = 5, LIST = 3, CYCLES = 6, QUAD = 7,
  GROUPLIST = ((3 5 7 8 9 10 11), (1 2), (12 13 19 20),
          (17 18), (4 6 14 15 16 21));
>SAVE PARM;
>INPUT NIDCH = 4, NFMT = 1, FILE = 'PTGI.DAT';
 (4A1, T1, 35A1)
>STOP
>POLYCHORIC NDEC = 3;
>SCORE LIST = 25, METHOD = 2;
```

Appendix D

POLYFACT Syntax for conducting unrestricted FIFA to Jenkin's Activity Scale (Jenkins et al., 1972

```
>TITLE
DICHOTOMIZE JENKINS ACTIVITY SURVEY RESPONSES AND CONDUCT
UNRESTRICTED FIFA
>PROBLEM NITEMS=48, SELECT=32, RESPONSES=5, MAXCAT=3,NOTPRESENTED;
>COMMENTS
        This example analyzes 32 items selected from the 48-item version
        of the Jenkins Activity Survey for Health Prediction, Form B
        (Jenkins, Rosenman, & Zyzanski, 1972). The data are responses
        of 600 men from central Finland drawn from a larger survey
        sample. Most of the items are rated on three-point scales
        representing little or no, occasional, or frequent occurrence of
        the activity or behavior in question. The Category statement is used to
        recode 0 responses to "1" and 1 and 2 responses to "2." Wording in the
        positive or negative direction varies from item to time as
        follows (item numbers are those of the original pool of items
        from which those of the present form was selected):

   -Q156, -Q157, +Q158, -Q165, -Q166, -Q167, +Q247, +Q248, -Q249, -Q250,
   +Q251, +Q252, +Q253, +Q254, +Q255, +Q256, +Q257, -Q258, -Q259, +Q260,
   +Q261, +Q262, +Q263, +Q264, +Q265, -Q266, +Q267, +Q268, +Q269, +Q270,
   +Q271, +Q272, -Q273, -Q274, -Q275, +Q276, +Q277, +Q278, -Q279, -Q280,
   +Q307, +Q308, +Q309, +Q310, +Q311, -Q312, -Q313, -Q314.

        The item parameters and factor scores will be saved in the
        files EXAMPL4A.PAR and EXAMPL04.FSC, respectively.
        Cases will be scored by EAP (Expected A Posteriori, or Bayes)
        estimation with adaptive quadrature (Method 2).

>NAMES Q156,Q157,Q158,Q165,Q166,Q167,Q247,Q248,Q249,Q250,Q251,Q252,
        Q253,Q254,Q255,Q256,Q257,Q258,Q259,Q260,Q261,Q262,Q263,Q264,
        Q265,Q266,Q267,Q268,Q269,Q270,Q271,Q272,Q273,Q274,Q275,Q276,
        Q277,Q278,Q279,Q280,Q307,Q308,Q309,Q310,Q311,Q312,Q313,Q314;
>RESPONSE  '8','0','1','2','.';
>CATEGORY NCAT=(3(0)48),RECODE;
        CODE = '0', VALUE = (1(0)48);
        CODE = '1', VALUE = (2(0)48);
        CODE = '2', VALUE = (2(0)48);
>SELECT  1, 2, 3, 5, 7, 11(1)14, 17(1)23, 25(1)30, 32, 33, 35, 36, 39(1)42, 47, 48;
>TETRACHORIC LIST, NDEC = 3;
>FACTOR NFAC=4, NROOT = 8, ROTATE = PROMAX;
>FULL QUAD=3, CYCLES=40;
>PRIOR;
```

```
>SCORE LIST=3, METHOD=2;
>TECHNICAL PRECISION=0.005, ITLIMIT=10;
>SAVE PARM, FSCORES ;
>INPUT NIDCHAR=10, SCORES, FILE='EXAMPL04.DAT';
(10A1,T1,48A1)
>STOP
>KEY  00200022002222222002222220222222000222002222200 0;
>SCORE  LIST=2,METHOD=2;
```