

## TECHNICAL REPORT

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

**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 extend their appreciation to A. John Rush, Ph.D., for his generosity with sharing the applied data used in this study.

## **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 scale items can be accounted for by a single underlying trait (e.g., depression), and are thus unidimensional, or form sub-scales to represent the trait's theoretical, 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).

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.

Technically, IRT embodies a host of probabilistic models to estimate a respondent's probability of selecting a particular item response category. 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 used 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 to account for the aspect of the individual that contributes to how he/she will respond to a given item.

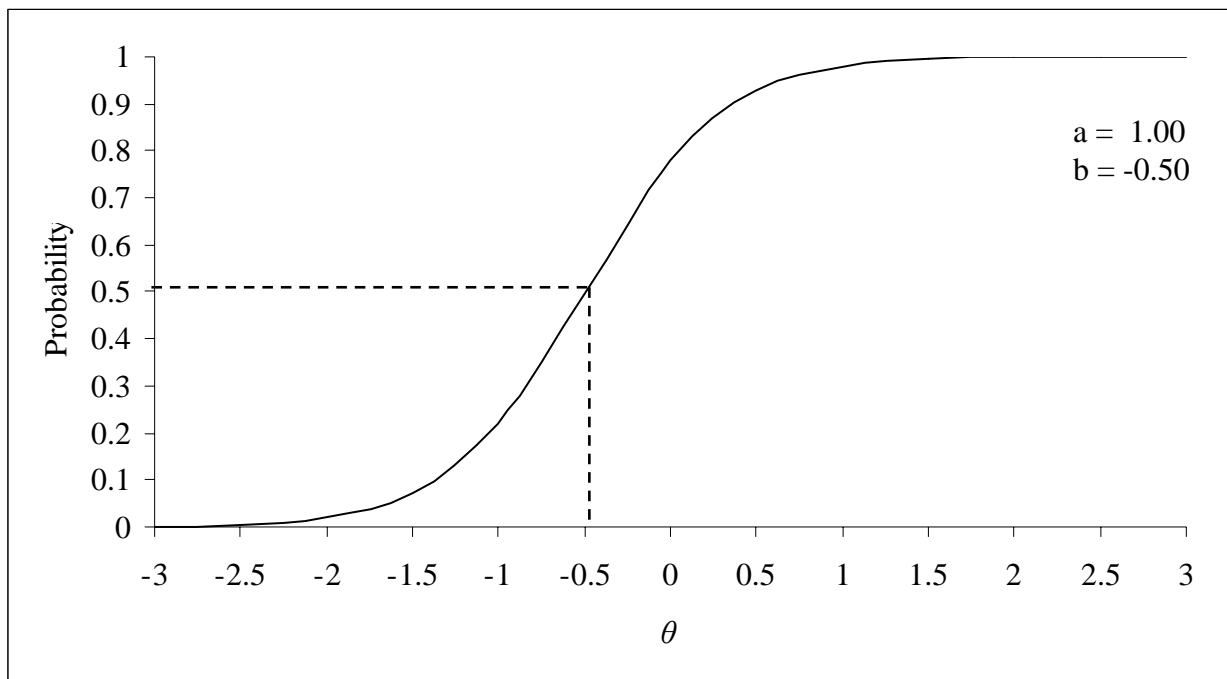
IRT procedures can be applied to a variety of data types. Scale items can be scored dichotomously (e.g., correct/incorrect, yes/no) or polytomously (e.g., Likert-scored response categories) and the categories can be ordered or unordered. Additionally, there is an assortment of IRT models to specify item 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 regarding the available IRT models for dichotomous (Hambleton, 1989; Harris, 1989; Lord, 1980; Lord & Novick, 1968), polytomous (Thissen & Steinberg, 1986), and multiple-choice (Thissen & Steinberg, 1984) item responses.

*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 latent variable ( $\theta$ ). The latent trait is unobserved and represents a respondent’s level of proficiency or propensity (e.g., depression). As shown, the IRF models the non-linear relationship between a probability of a positive item endorsement and the latent trait. Inspection of Figure 1 indicates that  $\theta$ , 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 and 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 the measured trait (Lord, 1980; Lord & Novick, 1968).

**Figure 1**

Hypothetical IRF



### 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 & Novick, 1968, Chpt. 16). Model assumptions include a single latent trait (e.g., depression) underlies the item responses and the metric of  $\theta$  for the item response function for each item can be represented as the normal ogive (Lord & Novick, 1968, p. 366), represented as

$$P_j(\theta) = \Phi(y_j) \quad (1)$$

where  $\Phi$  is the cumulative normal distribution function and  $y_j = a_j(\theta - b_j)$ , referred to as the *normal deviate*. Equation (1) models the probability of an individual with a given level of  $\theta$  positively endorsing item  $j$ . 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

$$P_j(\theta) = \Psi(1.7z_j) \quad (2)$$

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

The one-parameter model, or Rasch model (Rasch, 1966), is the most restrictive and only includes item difficulty and  $\theta$  for 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  $\theta$ . The logistic model is

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

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

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

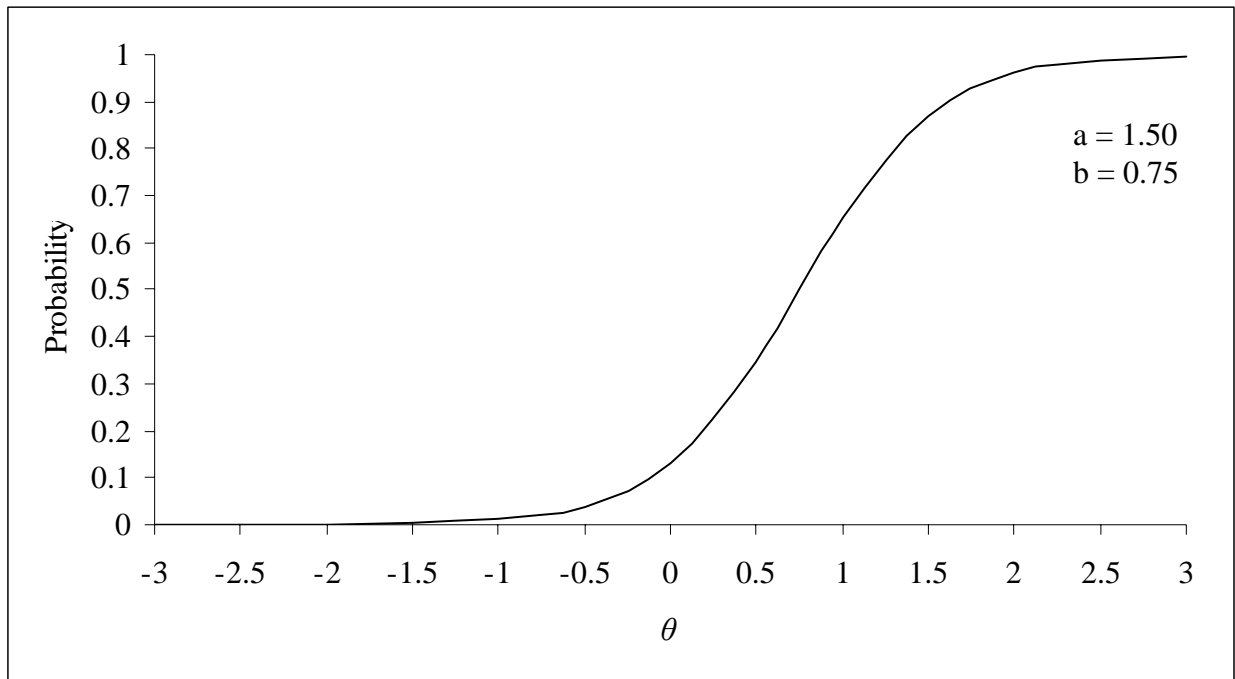
where  $a$  is item discriminatory power, and the other model parameters can be interpreted as those presented for the 1-PL model. Discrimination parameters typically range from 0 to 2 (Hambleton & Swaminathan, 1985), with high values being more effective with discriminating between respondents with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



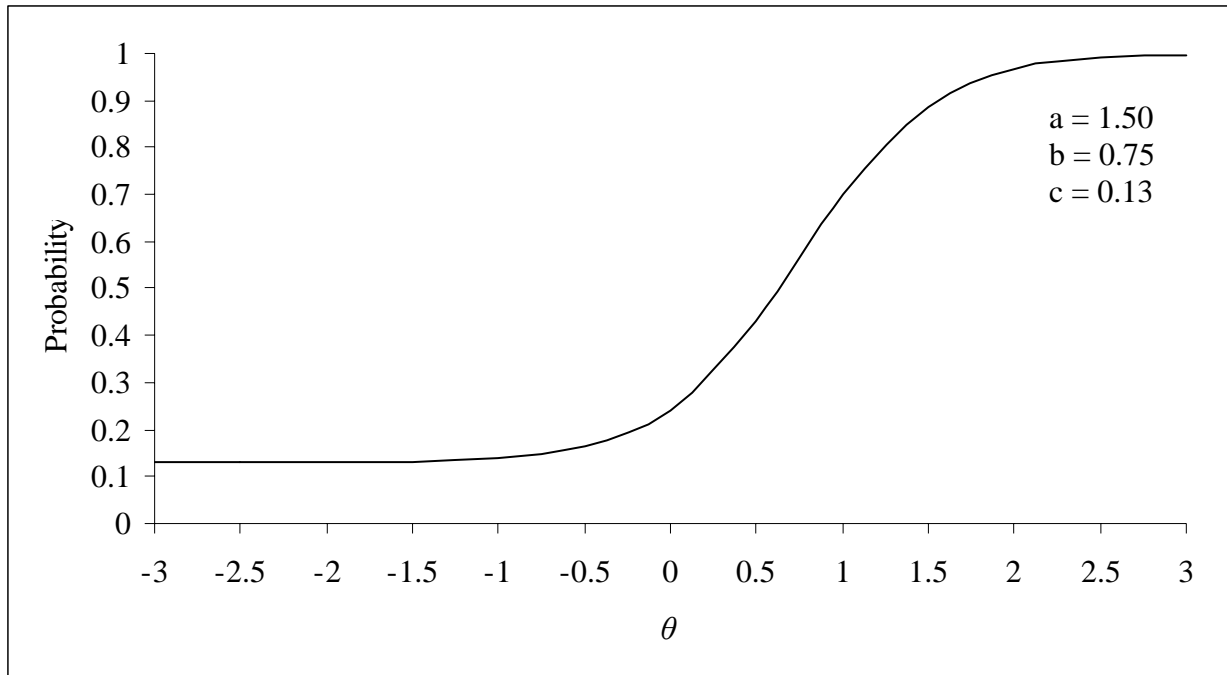
The 3-PL model builds on the 2-PL model by also including a pseudo-guessing parameter,  $c_i$ . The form of the model is

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

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 model. The lower asymptote is nonzero ( $c = .13$ ), 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 discriminatory power parameter, and Thissen and Steinberg's (1984) model for multiple-choice items. Each model is designed to 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

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

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

The logistic model is

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

where,  $P_{jk}(\theta)^*$  is the probability that person with  $\theta$  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).

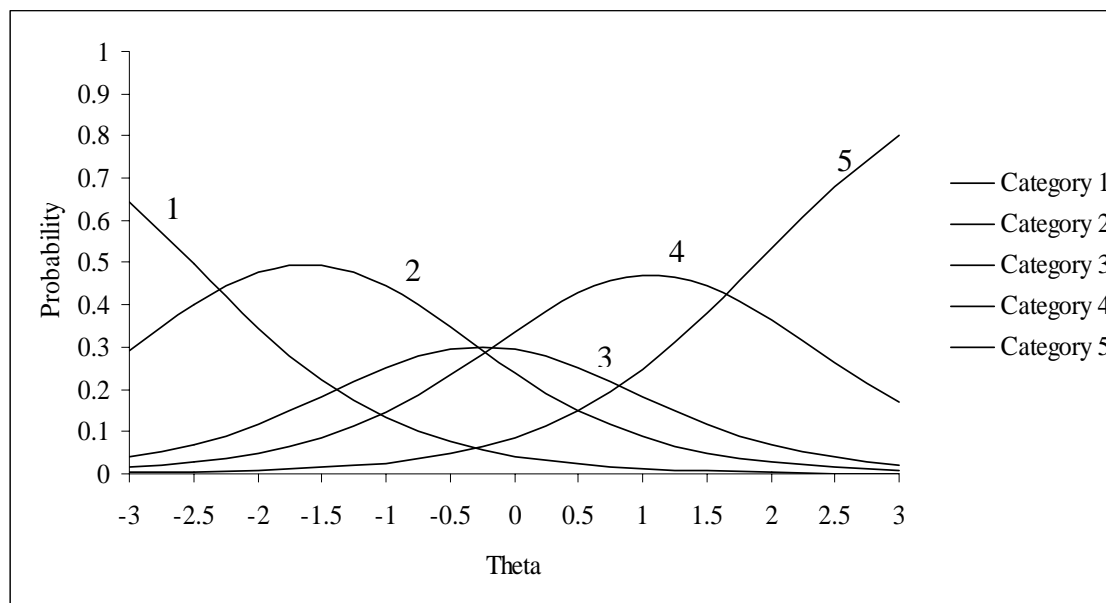
Therefore, the probability that individual  $n$  will endorse category  $k$  is

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

The model specifies that each previous category must be obtained prior to selecting the next highest category (Samejima, 1969, 1997).

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 (i.e., 1, 2), whereas higher trait estimates correspond to choosing higher response categories (i.e., 3, 4, 5). As specified in the model, the categorical trace lines have equal slopes and unique threshold 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), such as the Post-Traumatic Growth Inventory (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. The major advantage of the rating-scale model over Samejima's original model is 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. Model selection can be based on testing model assumptions.

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 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.

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) CAT (Wainer, Dorans, Eignor et al., 2000), among many.

Application of the aforementioned models to scale data 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).

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: (1) 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 theta estimates; (2) on the other hand, if the data are multidimensional with strong factors beyond the first (as may occur with a multiple-indicator personality or achievement instrument, 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, however, 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 the adaptive tests. 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.

Bock and Aitkin (1981) and Bock, Gibbons, and Muraki (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, perceived quality of life can be assessed from items that assess satisfaction with family, income, neighborhood, etc.

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

$$y_{ij} = \sum_k^m \alpha_{jk} \theta_{ki} + \varepsilon_{ij}, \quad (8)$$

where  $\alpha_{jk}$  is the loading of item  $j$  on factor  $k$ ,  $\theta_{ki}$  is the proficiency or propensity of individual  $i$  on factor  $k$  (e.g., depression), and  $\varepsilon_{ij}$  is an independent residual. According to the conventions of Thurstonian factor analysis, the variable  $y$  and  $\theta_k$  are assumed standard normal,  $N(0,1)$ , and that  $\theta_k$  are uncorrelated. The residuals ( $\varepsilon$ ) are independent and normally distributed with mean 0 and variance  $\sigma_j^2 = 1 - \sum_k^m \alpha_{jk}^2$ , i.e.,  $\varepsilon$  is NID  $(0, \sigma)$ . The quantity  $\sum_k^m \alpha_{jk}^2$  is called the common factor variance or *communality* of the item, and  $\sigma_j^2$  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  $\gamma_j$ . Thus, the probability that an individual with factor score vector  $\theta_i$  will respond positively to item  $j$ , as indicated by the item score  $x_{ij} = 1$  is given by the normal ogive item-response function,

$$\Phi_j(\theta_{ij}) = P(x_{ij} = 1 | \theta_{ij})$$

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

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

$$P(x_{ij} = 0 | \theta_i) = 1 - \Phi(\theta_i). \tag{10}$$

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

$$P(x = \mathbf{x}_i | \theta, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \prod_j^{n_i} [\Phi_j(\theta_i)]^{x_{ij}} [1 - \Phi_j(\theta_i)]^{1-x_{ij}}. \tag{11}$$

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  $\theta_i$ , 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  $\mathbf{x}_i$  can be expressed as

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

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).

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.

Specifically, 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  $\alpha_{j1}$  on the primary dimension and a second loading ( $\alpha_{jk}, k = 2, \dots, s$ ) on not more than one of the  $s - 1$  group factors. For four items, the bifactor pattern matrix might be

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

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). For example, 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 domains, but conditionally dependent within domains.

Gibbons and Hedeker (1992) and Gibbons, Bock, Hedeker et al. (2007) derived full-information item bifactor models for dichotomously and polytomously scored items, respectively. The bifactor solution constrains each item to have a non-zero loading on the primary dimension and a second loading on no more than one of the domain factors (Holzinger & Swineford, 1937). Their estimation method permits any number of item domains, and provides a single estimate of the primary dimension.

Gibbons et al. (2007) derived the likelihood equations and a method for their solution for bifactor extensions of both the rating scale model and the Samejima (1969) model for ordinal response data. The ordinal generalization of the Bock and Aitkin (1981) FI item-factor analysis model follows directly from the generalization of the ordinal bifactor model described by Gibbons et al. (2007). 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.

In the bifactor case, Samejima's (1969) graded response model is

$$z_{jk} = c_j + \sum_{k=1}^s a_{jk}(\theta) \quad (15)$$

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

To promote mental health outcome practitioners and researchers' understanding of multidimensional IRT models, a simulation and applied study are provided. The aim of the simulation study is to investigate the implications of applying Samejima's (1969) unidimensional graded response model and in the bifactor form to multidimensional, categorical data.

Subsequently, these models are fit to breast cancer survivor data from the Post-Traumatic Growth Inventory (PTGI; Tedeschi & Calhoun, 1996). The motivation for these studies reflects the inadequacies of existing methods to capture the multi-dimensional characteristics of outcomes data. Currently, both the classical and modern measurement approaches commonly used to analyze and score outcomes data assume that one dominant factor underlies each domain and accounts for most of the variation in scores; the assumption of unidimensionality. If more than one factor exists, the domain must be divided into sub-domains to apply these methods. Under this restrictive set of assumptions, efforts to summarize these data into broader constructs suffer from the lack of clear statistical and analytical decision-rules for combining these data; measures are often simply added together to create a combined score, weighted in terms of their contribution to a criterion variable (e.g., global health measure, see Bozzette et al., 1994), or, alternatively, patient or expert judgment is used to weight these factors within or across domains (Neuman, Goldie, & Weinstein, 2000).

## Method

### *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 *expected a posteriori* scores (EAP; Bock & Mislevy, 1982), 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.

### *Real Data Illustration*

The FI bifactor model for polytomous data, based on Samejima's (1969) model, was fit to data obtained from the PTGI (Tedeschi & Calhoun, 1996) to illustrate its use in real data applications. The PTGI is a widely used measure of an individual's posttraumatic growth. It is based on the premise that traumatic events (e.g., rape, heart attacks) can ultimately lead to positive self-perceptions, such as improved interpersonal relationships. The scale consists of 21 items and requires respondents to rate their experience towards positive growth for each item on a 6-point scale (i.e., 0 = No Change; 1 = Very Small Change; 2 = Small Change; 3 = Moderate Change; 4 = Great Change; 5 = Very Great Change). PTGI scale items are provided in Table 1.

**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.

Note. Response categories: 0 = No Change; 1 = Very Small Change; 2 = Small Change; 3 = Moderate Change; 4 = Great Change; 5 = Very Great Change.

In consideration of empirical evidence suggesting the PTGI factor structure may not be robust across samples (Ho, Chan, & Ho, 2004; Sheikh & Marotta, 2005), an unrestricted FI item factor analysis was conducted prior to fitting the bifactor model. Data for this study was obtained from breast cancer survivors ( $n = 801$ ) (mean age = 57.2,  $SD = 10.1$ ).

Based on the results of the unrestricted FI item factor analysis, the bifactor model, based on Samejima's (1969) graded response model, was fit to the data. Specifically, the analysis was designed to test whether the 5 PTGI diagnostic domains (e.g., New Possibilities, Spiritual Change) (a) provide an improvement in fit over a unidimensional model, and (b) which of the 5

domains significantly improve the fit of the model to the data. These analyses were applied to the 21 items from the PTGI. First, a unidimensional model was fit to the data. Second, a bifactor model with all postulated subdomains was fit to the data. Third, separate bifactor models with one less domain than the total number of domains identified in the FI unrestricted factor analysis was fit to the data. In this way, it can be determined whether the additional domain provided a statistical improvement in model data fit, thus supporting the presence of each domain. To make this determination, the log likelihood ( $\log L$ ) is computed for each model. Minus two times the difference in  $\log L$ 's for the two models is distributed as chi-square on degrees of freedom equal to the difference in number of parameters in the two models. For example, a comparison of the unidimensional model to the 5 diagnostic domain bifactor model is distributed on 21 degrees of freedom ( $df$ ), which is the additional number of parameters for the bifactor model (i.e., one for each item representing the items loading on its domain). For a test of the significance of a single diagnostic domain (e.g., Relating to Others), the likelihood ratio chi-square statistic is distributed on degrees of freedom equal to the number of items in that particular domain (e.g., for Relating to Others,  $df = 7$  since there are 7 items that purportedly relate to Relating to Others).

For each item, the bifactor model computes a threshold, primary factor loading, and a domain-specific factor loading. The threshold describes the point on the underlying primary symptom/impairment dimension (characterized by all items) at which 50% of the sample can be expected to endorse the response category. For example, in the area of depression, an item with a high threshold (e.g., suicidal ideation) is endorsed by the most severely depressed patients, whereas an item with a low threshold (e.g., depressed mood) is endorsed by patients with both high and low underlying levels of depression. The primary loadings can be interpreted as a factor loading (correlation with the underlying primary dimension) that is appropriate for an ordinal

response measure. The domain-specific factor loading represents the correlation of the item with the underlying domain that the item was sampled from.

### *Simulation Results*

Figure 5 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 all 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 both models became more varied, but not as severe for the bifactor model.

**Figure 5**

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 or 10, Primary loadings [PL] = .50 or .75, Domain Loadings [DL] = .25 or .50)

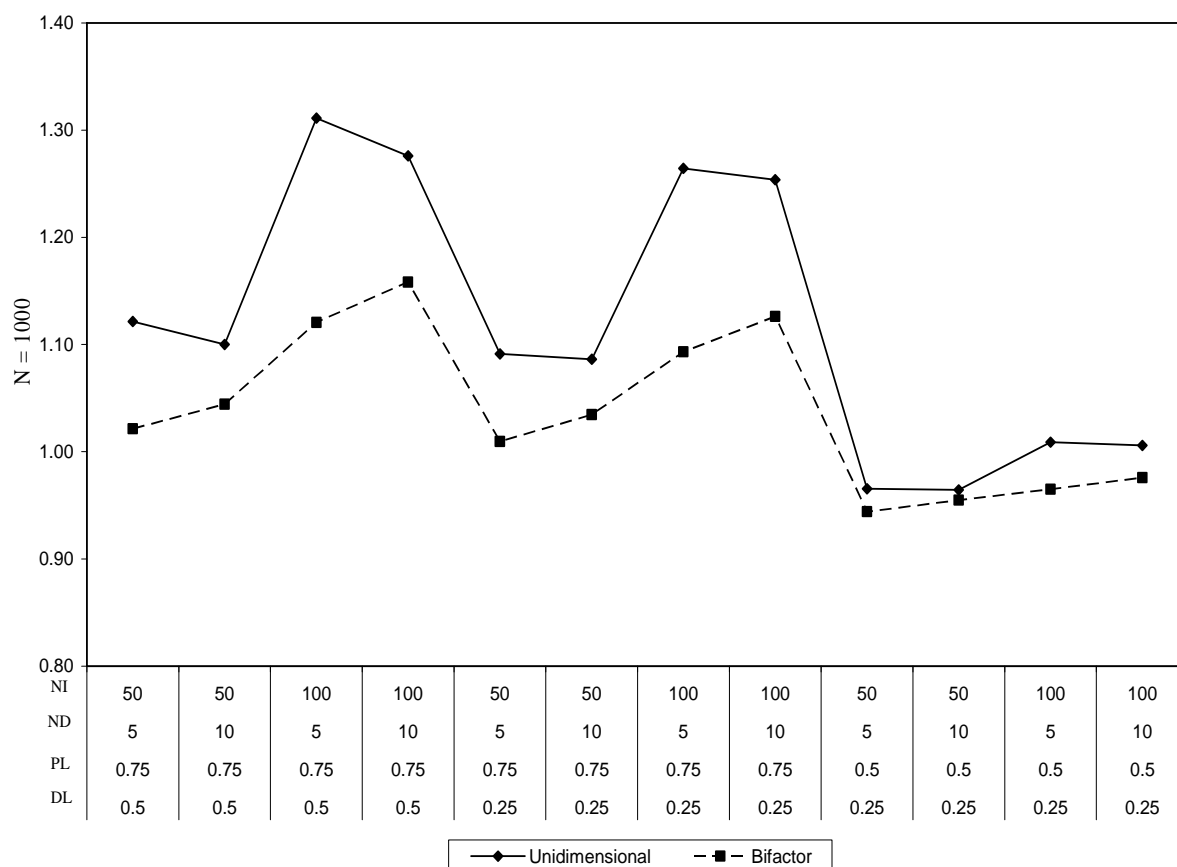


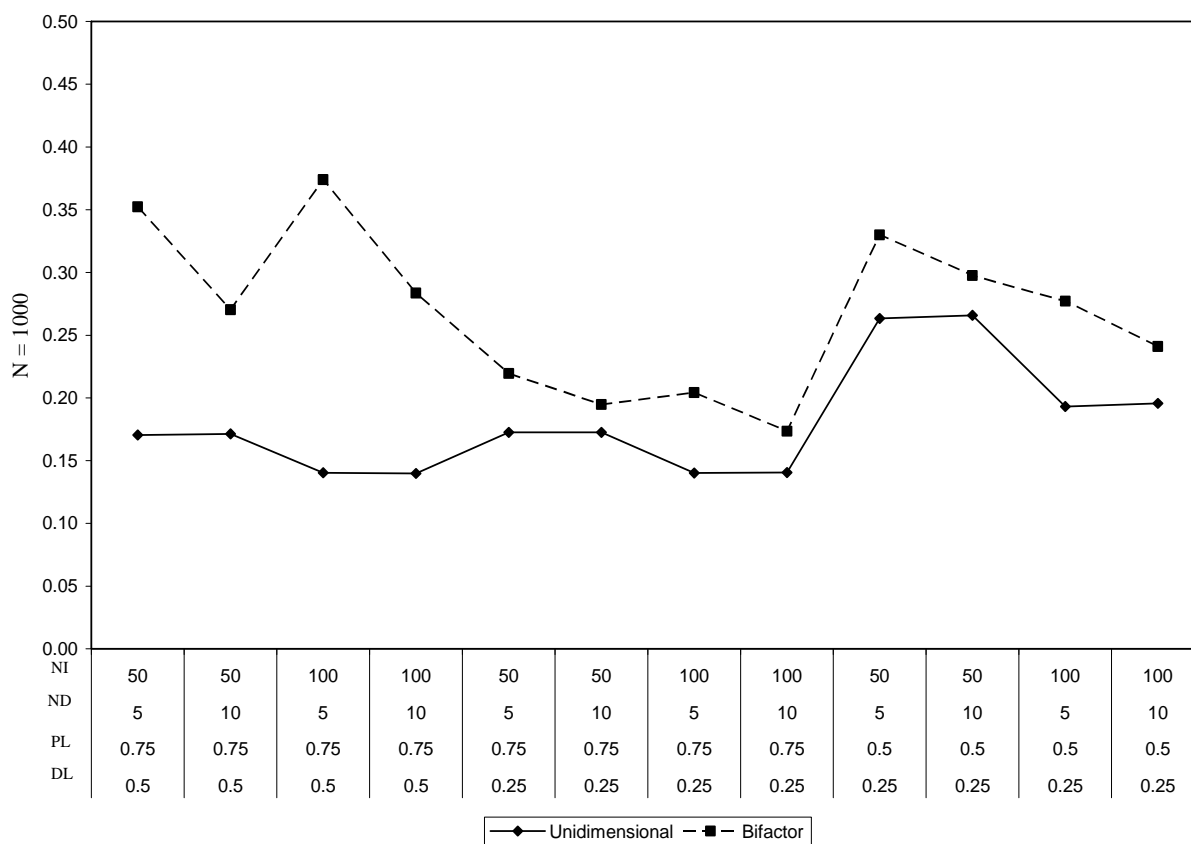
Figure 6 reports the mean posterior standard deviation (PSD) of the Bayes EAP. As shown, the differences in the PSD between the models can be dissected in terms of the dimensionality of the underlying data. Specifically, in the conditions in which the primary loadings are 0.75 and the domain loadings 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 6**

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 or 10, Primary loadings [PL] = .50 or .75, Domain Loadings [DL] = .25 or .50)



### *PTGI Results*

For didactic purposes, 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.

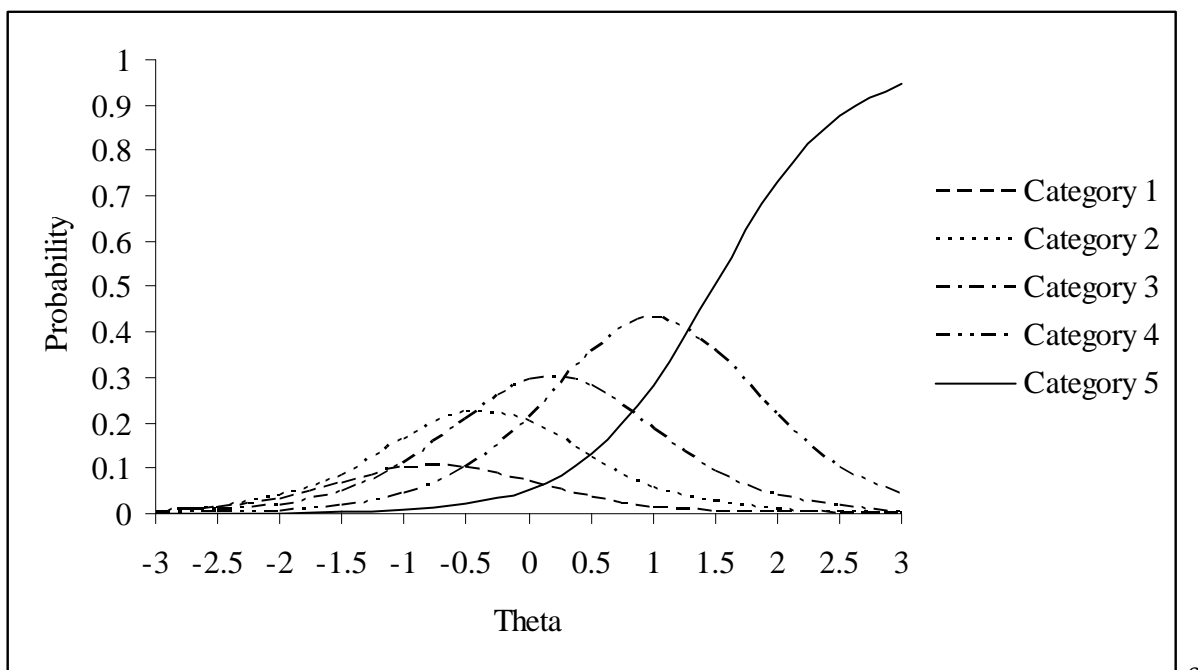
Samejima's (1969) unidimensional graded response model was fit the PTGI data using MULTILOG (Thissen, Chen, & Bock, 2003). Table 2 reports slope and threshold estimates for each item. Inspection of Table 2 shows that each item has a single slope value and each category (e.g., Strongly Disagree, Agree) has a unique threshold (difficulty) parameter. The slope parameter indicates that the model assumes that each item category is equally discriminating. The threshold parameters indicate the place on the trait continuum (post-traumatic growth) where a respondent has a 50% probability of selecting that particular category. As shown in the table, items 5, 10, and 21 were the most discriminating items. Figure 7 shows the IRFs of Item 1.

**Table 2**

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

**Figure 7**  
IRFs of PTGI Item 1



Based on previous research questioning the robustness of the PTGI across groups (Ho et al., 2004; Sheikh & Marotta, 2005), an unrestricted full-information item factor analysis using the recently developed POLYFACT program was conducted. 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, which approximate the scale's theoretical factor structure (Tedeschi & Calhoun, 1996). Table 3 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 also was 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 4 shows that the empirical factors were moderately correlated.

**Table 3***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	-0.011	0.209	<b>0.852</b>	<b>0.820</b>	0.107	0.296	-0.003	0.136	0.031	0.256
2	-0.097	0.318	<b>0.706</b>	<b>0.677</b>	0.131	0.285	0.042	0.239	0.248	0.337
3	0.339	0.453	0.125	0.341	0.024	0.304	-0.118	0.127	<b>0.551</b>	<b>0.508</b>
4	0.143	0.384	0.169	0.368	-0.049	0.256	0.150	0.313	<b>0.559</b>	<b>0.524</b>
5	<b>0.485</b>	<b>0.535</b>	0.054	0.293	0.105	0.340	-0.011	0.231	0.349	0.394
6	0.291	<b>0.431</b>	0.355	<b>0.500</b>	-0.249	0.164	0.121	0.289	0.414	0.367
7	<b>0.786</b>	<b>0.651</b>	0.262	0.377	-0.111	0.202	-0.109	0.135	0.085	0.275
8	<b>1.022</b>	<b>0.731</b>	-0.016	0.208	0.035	0.274	0.083	0.253	-0.192	0.211
9	<b>0.946</b>	<b>0.732</b>	0.113	0.173	0.138	0.358	0.075	0.245	-0.078	0.244
10	<b>0.554</b>	<b>0.594</b>	-0.114	0.235	0.074	0.310	0.000	0.260	0.477	0.480
11	<b>0.784</b>	<b>0.628</b>	-0.096	0.171	0.104	0.332	0.007	0.218	0.115	0.266
12	0.168	0.326	0.050	0.215	<b>0.908</b>	<b>0.744</b>	-0.076	0.148	-0.089	0.180
13	-0.003	0.264	0.079	0.295	<b>0.463</b>	<b>0.521</b>	-0.049	0.165	0.423	0.416
14	-0.130	0.296	0.263	0.454	0.014	0.330	-0.120	0.193	<b>0.891</b>	<b>0.576</b>
15	0.081	0.390	-0.026	0.319	-0.005	0.311	-0.107	0.161	<b>0.941</b>	<b>0.613</b>
16	0.014	0.412	-0.077	0.306	-0.129	0.226	0.195	0.342	<b>0.908</b>	<b>0.512</b>
17	0.077	0.334	0.134	0.276	0.324	0.496	<b>0.659</b>	<b>0.566</b>	-0.108	0.235
18	0.097	0.366	-0.025	0.207	-0.091	0.274	<b>0.944</b>	<b>0.796</b>	0.083	0.261
19	0.077	0.312	0.073	0.217	<b>0.950</b>	<b>0.722</b>	-0.037	0.201	-0.075	0.232
20	-0.146	0.267	-0.050	0.282	<b>0.597</b>	<b>0.623</b>	0.130	0.339	0.479	0.388
21	-0.101	0.311	-0.075	0.304	0.217	0.426	0.087	0.320	<b>0.850</b>	<b>0.585</b>

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

**Table 4***Factor Correlations*

	Factors				
	1	2	3	4	5
1	1.000				
2	0.656	1.000			
3	0.617	0.563	1.000		
4	0.634	0.770	0.632	1.000	
5	0.772	0.563	0.697	0.670	1.000

The bifactor IRT model based on Samejima's (1969) graded response IRT model was fit to the data in terms of the following model: (a) the original five factor model of the PTGI, as per Tedeschi and Calhoun (1996), and (b) the unrestricted FI item factor analysis.

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

**Table 7 (continued)**

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

Second, the FI bifactor model based on the results of the unrestricted full-information item factor analysis was fit to the data,  $\chi^2_{653} = 32,054.82$ . Table 8 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 9. 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 9 reports item thresholds, while Table 10 shows the observed and expected proportions of respondents across categories. The root mean square error value of 0.018 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 8***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 9***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 10***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
	<i>0.272</i>	<i>0.060</i>	<i>0.141</i>	<i>0.199</i>	<i>0.208</i>	<i>0.119</i>
2	0.272	0.067	0.132	0.201	0.215	0.112
	<i>0.295</i>	<i>0.085</i>	<i>0.154</i>	<i>0.198</i>	<i>0.178</i>	<i>0.089</i>
3	0.332	0.085	0.127	0.223	0.154	0.079
	<i>0.362</i>	<i>0.098</i>	<i>0.140</i>	<i>0.212</i>	<i>0.132</i>	<i>0.057</i>
4	0.297	0.040	0.107	0.205	0.215	0.136
	<i>0.318</i>	<i>0.045</i>	<i>0.124</i>	<i>0.215</i>	<i>0.197</i>	<i>0.101</i>
5	0.351	0.072	0.131	0.242	0.141	0.062
	<i>0.391</i>	<i>0.080</i>	<i>0.137</i>	<i>0.229</i>	<i>0.121</i>	<i>0.042</i>
6	0.262	0.066	0.110	0.200	0.208	0.154
	<i>0.281</i>	<i>0.077</i>	<i>0.122</i>	<i>0.206</i>	<i>0.192</i>	<i>0.123</i>
7	0.355	0.085	0.165	0.218	0.119	0.059
	<i>0.387</i>	<i>0.098</i>	<i>0.173</i>	<i>0.199</i>	<i>0.100</i>	<i>0.044</i>
8	0.427	0.091	0.146	0.192	0.087	0.056
	<i>0.470</i>	<i>0.097</i>	<i>0.145</i>	<i>0.170</i>	<i>0.075</i>	<i>0.043</i>
9	0.443	0.082	0.127	0.167	0.096	0.084
	<i>0.395</i>	<i>0.072</i>	<i>0.143</i>	<i>0.213</i>	<i>0.116</i>	<i>0.062</i>
10	0.352	0.062	0.132	0.232	0.136	0.085
	<i>0.395</i>	<i>0.072</i>	<i>0.143</i>	<i>0.213</i>	<i>0.116</i>	<i>0.062</i>
11	0.544	0.079	0.132	0.120	0.069	0.056
	<i>0.590</i>	<i>0.079</i>	<i>0.125</i>	<i>0.105</i>	<i>0.061</i>	<i>0.041</i>
12	0.241	0.079	0.122	0.310	0.180	0.069
	<i>0.254</i>	<i>0.094</i>	<i>0.140</i>	<i>0.304</i>	<i>0.157</i>	<i>0.051</i>
13	0.315	0.052	0.114	0.228	0.200	0.091
	<i>0.344</i>	<i>0.059</i>	<i>0.124</i>	<i>0.230</i>	<i>0.176</i>	<i>0.067</i>
14	0.288	0.056	0.112	0.194	0.202	0.147
	<i>0.311</i>	<i>0.068</i>	<i>0.133</i>	<i>0.197</i>	<i>0.177</i>	<i>0.114</i>
15	0.325	0.060	0.137	0.211	0.170	0.097
	<i>0.353</i>	<i>0.071</i>	<i>0.155</i>	<i>0.208</i>	<i>0.143</i>	<i>0.070</i>
16	0.286	0.072	0.085	0.185	0.201	0.171
	<i>0.311</i>	<i>0.083</i>	<i>0.094</i>	<i>0.188</i>	<i>0.187</i>	<i>0.137</i>
17	0.308	0.055	0.117	0.184	0.187	0.149
	<i>0.342</i>	<i>0.066</i>	<i>0.128</i>	<i>0.1871</i>	<i>0.165</i>	<i>0.117</i>
18	0.382	0.064	0.095	0.161	0.151	0.147
	<i>0.428</i>	<i>0.067</i>	<i>0.096</i>	<i>0.158</i>	<i>0.131</i>	<i>0.120</i>
19	0.221	0.045	0.085	0.258	0.261	0.130
	<i>0.231</i>	<i>0.052</i>	<i>0.100</i>	<i>0.279</i>	<i>0.237</i>	<i>0.102</i>

**Table 10 (continued)**

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

### Summary

The emergent use of self-report instruments in health outcomes research 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.

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 trait.

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 study demonstrate 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.

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

The real data example illustrated how exploratory and confirmatory-based factor analytic IRT procedures can be used to model health outcomes scale data. Factor analytic results of the PTGI showed that the scale's original factor structure did not provide the best fit to the data. Although emergent factors were similar to the original factors, the first factor reported in Tedeschi and Calhoun's (1996) study was less salient in the present data. 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.

As presented in this report, 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 PRO scale data.

The aim of this report 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, West, Sousa, 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, 4<sup>th</sup> 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.
- 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., & 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 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 psychometrician. *Psychometrika*, *71*, 425-440.
- Bozzette, S.A., et al. (1998). The Care of HIV-Infected Adults in the United States. *The New England Journal of Medicine*, *26*, 1897-1904.
- 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. *Psychometrika*, 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, 3<sup>rd</sup> 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: Kluwer Academic Publishers.
- Harmon, H. (1976). *Modern factor analysis* (3<sup>rd</sup> 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.
- Neumann, P. J., Goldie, S. J., & Weinstein, M. C. (2000). Preference-based measures in economic evaluation in health care. *Annual Review of Public Health*, 21, 587-611.
- 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 bifactor 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.
- 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., Boydston, 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.
- 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.
- 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* (4<sup>th</sup> 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.