This article was downloaded by: [KU Leuven University Library]
On: 01 April 2014, At: 05:42
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Multivariate Behavioral Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmbr20

Functionally Unidimensional Item Response Models for Multivariate Binary Data

Edward H. Ip [a], Geert Molenberghs [b], Shyh-Huei Chen [c], Yuri Goegebeur [d] & Paul De Boeck [e]

[a] Department of Biostatistical Sciences, Department of Social Sciences & Health Policy, Wake Forest School of Medicine

[b] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt and K.U. Leuven

[c] Department of Biostatistical Sciences, Wake Forest School of Medicine

[d] Department of Mathematics and Computer Science, University of Southern Denmark

[e] Department of Psychology, K.U. Leuven and University of Amsterdam

Published online: 26 Jul 2013.

To cite this article: Edward H. Ip, Geert Molenberghs, Shyh-Huei Chen, Yuri Goegebeur & Paul De Boeck (2013) Functionally Unidimensional Item Response Models for Multivariate Binary Data, Multivariate Behavioral Research, 48:4, 534-562, DOI: 10.1080/00273171.2013.796281

To link to this article: http://dx.doi.org/10.1080/00273171.2013.796281

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Multivariate Behavioral Research, 48:534–562, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2013.796281

Routledge
Taylor & Francis Group

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Functionally Unidimensional Item Response Models for Multivariate Binary Data

Edward H. Ip
Department of Biostatistical Sciences, Department of Social Sciences & Health Policy
Wake Forest School of Medicine

Geert Molenberghs
Interuniversity Institute for Biostatistics and Statistical Bioinformatics
Universiteit Hasselt and K.U. Leuven

Shyh-Huei Chen
Department of Biostatistical Sciences
Wake Forest School of Medicine

Yuri Goegebeur
Department of Mathematics and Computer Science
University of Southern Denmark

Paul De Boeck
Department of Psychology
K.U. Leuven and University of Amsterdam

The problem of fitting unidimensional item response models to potentially multidimensional data has been extensively studied. The focus of this article is on response data that have a strong dimension but also contain minor nuisance dimensions. Fitting a unidimensional model to such multidimensional data is believed to result in

Correspondence concerning this article should be addressed to Edward H. Ip, Department of Biostatistical Sciences, Wake Forest School of Medicine, Medical Center Boulevard, WC23, Winston-Salem, NC 27157. E-mail: eip@wakehealth.edu

ability estimates that represent a combination of the major and minor dimensions. We conjecture that the underlying dimension for the fitted unidimensional model, which we call the functional dimension, represents a nonlinear projection. In this article we investigate 2 issues: (a) can a proposed nonlinear projection track the functional dimension well, and (b) what are the biases in the ability estimate and the associated standard error when estimating the functional dimension? To investigate the second issue, the nonlinear projection is used as an evaluative tool. An example regarding a construct of desire for physical competency is used to illustrate the functional unidimensional approach.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

To meaningfully capture and measure a construct, researchers in the psychological, social, and behavioral sciences have long strived to create instruments that are parsimonious in dimensionality while at the same time being sufficiently broad—that is, instruments that possess adequate "psychological width." However, a "cover-all-bases" approach is often prone to include in the instrument not only items that measure a targeted construct but also attributes that are related but are psychologically or cognitively distinct from the construct. Balancing the trade-off between parsimoniousness in dimensionality and scope is a delicate job. The notion of parsimonious dimensionality, particularly unidimensionality, by which a single putative construct is to be measured, is implicit in both classical testing theory and item response theory (IRT; Lord, 1980). Thurstone, for example, insisted that all measurements in psychology should only measure a single construct, and he called it a universal characteristic of all measurements (Thurstone, 1931, p. 257).

Because maintaining the psychological width of a test is so critical to psychological and educational measurement, some researchers have begun to question the utility of models with parsimonious dimensions—with the case in point being the unidimensional item response model. Based on empirical studies in education, Traub (1983), for example, went so far as to argue that "no unidimensional item response model is likely to fit educational achievement data" (p. 65). Reise, Moore, and Haviland (2010) provided examples in psychological assessment, including stress reaction and a measure of alexithymia, to demonstrate how multidimensionality could arise even when the intention of the researchers was to create unidimensional scales. Thus, a common situation in measurement is that a test has good face validity of unidimensionality (e.g., as judged by content experts) but empirically exhibits evidence of a single, strong dimension and a few smaller and perhaps statistically detectable dimensions (Embretson & Reise, 2000, p. 230; Hulin, Drasgow, & Parsons, 1983, p. 40; Drasgow & Parsons, 1983).

The intriguing issue that although our models are unidimensional our data are multidimensional has attracted enduring interest that dates back to work perhaps as early as Thurstone (1931), Humphreys (1949), and Cattell (1966). In particular, the work of Humphreys and his colleagues made a clear distinction between

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

statistical unidimensionality and psychological unidimensionality (e.g., Hulin & Humphreys, 1980; Humphreys, 1970). Two situations of multidimensional data need to be distinguished: (a) when multidimensionality can be represented with a categorical item structure and (b) when a domain is pervaded by a secondary or even a tertiary dimension in a gradual way depending on the item. In the first situation, multidimensionality is often conceptualized as subscales and methods such as the bifactor model (see Reise, 2012, for an extensive review) could be practical tools for helping practitioners in deciding whether or not the total scale score is an adequate indicator of the intended construct or whether subscores should be deployed (Reise, Bonifay, & Haviland, 2013). In the second situation, the secondary dimension could be irrelevant or even distorting; an example of this with regard to test anxiety and stereotype threat is given in Halpin, da Silva, and De Boeck (in press). The secondary dimension can also be inherent and natural to the domain for all subjects or for a subgroup. A math test that intends to tap into higher order cognitive skills would inevitably require visual and mental representation skills as well as verbal abilities. In such a case, it is often not possible to formulate "pure" items or to determine the appropriate rotational position of the true dimension. De Boeck, Cho, and Wilson (2011) described a special case of this situation in which the secondary latent dimension creates differential item functioning.

Although the first situation is common and interesting, in this article we chose to focus on the second situation because methods that involve subscales or multidimensional item response theory (MIRT) may be more appropriate for the first situation. When there is an identifiable process that affects all items (e.g., mathematical thinking and operation), but there are also other minor dimensions—whether they are substantive (e.g., visual and verbal abilities) or statistical—that may be affecting the responses beyond the effects of the dominant dimension, then we may ask what it means to fit a unidimensional model to the test data. Many researchers have forcibly argued, with support from empirical studies, that a unidimensional model would measure a *representative* dimension, which is some sort of weighted average of the dimensions, both major and minor, that are present within the test (Ackerman, 1994; Junker, 1991; Kim, 1994; Kirisci, Hsu, & Yu, 2001; Reckase, Carlson, Ackerman, & Spray, 1986; Stout,1987; 1990; Walker & Beretvas, 2003; Way, Ansley, & Forsyth, 1988; Yen, 1984; Zhang & Stout, 1999). Other authors have investigated the robustness of the unidimensional model (Junker & Stout, 1994; Kim, 1994), the detection of secondary psychological dimensions (Walker, Azen, & Schmitt, 2006), and the possibility of linearly projected unidimensional models (Ip, 2010a, 2010b; Ip & Chen, 2012; Wang, 1986; Zhang & Wang, 1998) as well as the application of such tools to practical testing problems such as the creation of short forms (Stucky, Thissen, & Edelen, 2013).

A key question in operationalizing the concept of the representative dimension is this: What does it really represent, or more precisely, what is its mathe-

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

matical construct? Perhaps a way to answer this question is to envision, from the perspective of a generative model, how a unidimensional model can be fitted to truly multidimensional response data. The fitting process can take two alternative routes: (a) the Multidimensional Modeling Route, which involves first fitting a multidimensional model (e.g., multidimensional IRT or MIRT) to the data and then somehow approximating the multidimensional model by a unidimensional IRT (UIRT), or (b) the Empirical Route, which involves directly fitting a unidimensional model (e.g., UIRT) to empirical data generated from the underlying multidimensional model without going through the intermediate step of a "provisional" MIRT. The question about the meaning of the representative dimension can then be reframed as follows: What is the closest approximation of the generative MIRT model from a one-dimensional latent space? We see later that the term *closest* needs to be defined. At this point, we continue to use the term loosely. Figure 1 shows the two routes and the data as well as two types of approximations—nonlinear and linear—for the Multidimensional Modeling Route. An alternative way to interpret the representative dimension is as a population parameter to which the large-sample (as test length increases) estimate of unidimensional ability (estimated directly from the data) converges. For the convenience of further discussion, we call this theoretical, single latent dimension of the population, if it exists, the *functional (uni)dimension* (F1D).

As far as we know, a comprehensive mathematical theory does not exist, let alone an analytic closed-form solution, for the F1D. Without a mathematical

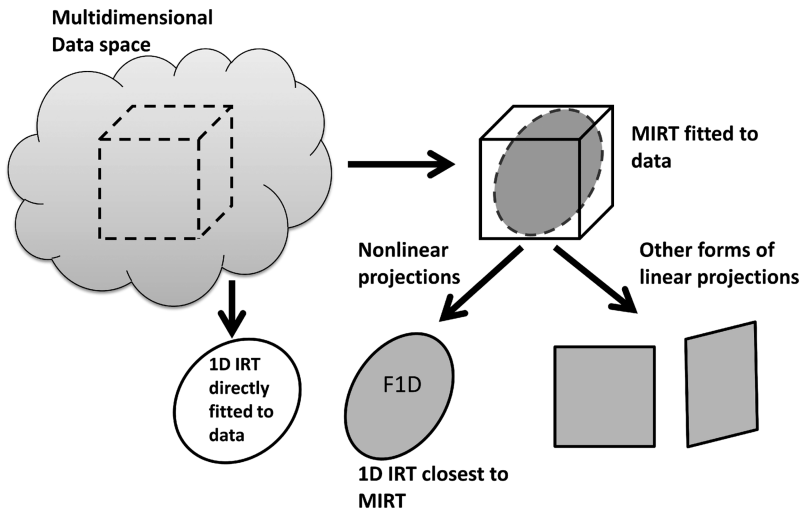


FIGURE 1 Schematic diagram showing different projection methods. IRT = item response theory; MIRT = multidimensional IRT; 1D = one-dimensional; F1D = functional unidimension.

theory describing how a unidimensional model behaves in a multidimensional data space, we have a dilemma when assessing the properties of the IRT-based estimate, namely, that there is no true value against which bias can be evaluated. To illustrate this dilemma, we use as an example a data set that was generated from a multidimensional IRT (MIRT) model (Reckase, 1997):

$$P(Y_{ij} = 1|\theta_{j1}, \theta_{j2}) = \frac{\exp(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} - d_i)}{1 + \exp(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} - d_i)}, \tag{1}$$

where Y_{ij}, $i = 1, \cdots, I$, $j = 1, \cdots, J$ is the binary response of individual j to item i; a_{i1}, a_{i2} are the discrimination parameters; $(\theta_{j1}, \theta_{j2})$ is a vector of latent traits of dimension 2; and d_i is an intercept (difficulty) parameter. We simulated response data following the specification: $a_{i1} = a_{i2} = a_i$ was sampled from a uniform distribution, *Unif*(0.2, 1.5), d_i from *Unif*(−2, 2), θ_{j1}, θ_{j2} respectively from independent univariate normal distributions $N(0, \sigma_1^2)$ and $N(0, \sigma_1^2)$ with $\sigma_1^2 = \sigma_2^2 = 0.5$, and $J = 2{,}000$. Because of the artificially imposed condition $a_{i1} = a_{i2}$, the *true* representative dimension F1D was known and given by $\theta = \theta_1 + \theta_2$, where $\theta \sim N(0, 1)$. Standard unidimensional IRT procedure was then used to first calibrate the items and subsequently compute person (ability) maximum a posteriori (MAP) estimates $\hat{\theta}$ for each individual. We evaluated the estimate $\hat{\theta}$ against the following quantities: (a) $\theta = \theta_1 + \theta_2$, the true F1D; (b) θ^*, a nonlinear approximation of the given MIRT parameters that preserves, pointwise, the item response function and serves as an approximation of F1D (procedure described later); and (c) $2\theta_1$, 2 times the first dimension. Figure 2 respectively shows, from top to bottom, the three sets of comparisons in scatterplot form. The left panel shows the results for a small number of items ($I = 10$) whereas the right panel shows the results for $I = 100$. From Figure 1 it can be seen that the IRT-based estimate $\hat{\theta}$ increasingly gets closer to the true F1D as the number of items increases. The root mean square error (RMSE; equation given later in Equation (17)) also decreases from 0.63 to 0.26. The nonlinear approximation θ^* also tends to closely mimic the behavior of the true F1D. Finally, the first coordinate $2\theta_1$, which is purposefully designated to be different from the true F1D, exhibits high levels of RMSE, regardless of the number of items.

Of course the example is not truly multidimensional; the point of the example, in which the true F1D is known for the generative MIRT model, is that it is important to distinguish between the following quantities: the true underlying single dimension, which we term F1D ($\theta = \theta_1 + \theta_2$ in the example); a numerical approximation to the F1D (θ^* in the example); and a direct UIRT estimate of F1D using response data ($\hat{\theta}$ in the example). Generally, unlike the aforementioned example, the true F1D is *not known* given the full set of parameters of an MIRT model. Therefore it is important, for the purpose of ascertaining the (mis)fitted UIRT model, to obtain an accurate numerical approximation to the F1D value.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

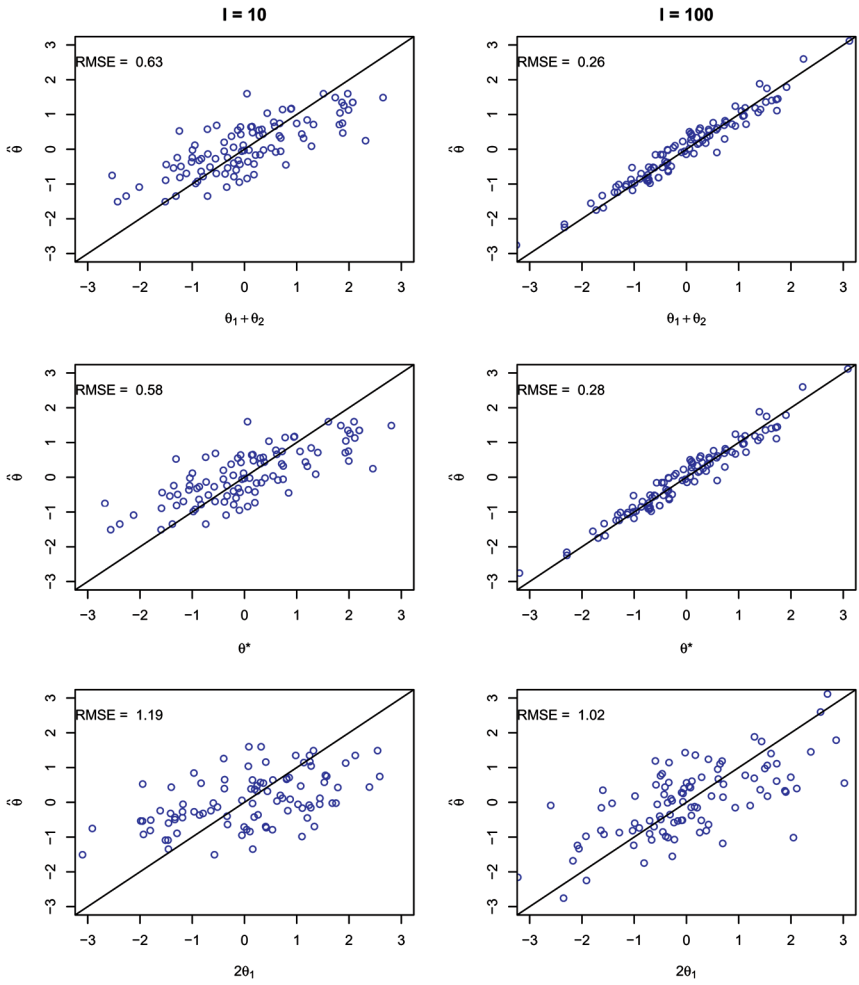


FIGURE 2 Scatterplots of a random sample of 100 simulated subjects from a larger sample of 2,000 showing item response theory (IRT)-based estimate θ versus the true functional dimensional (uppermost panels), a nonlinear approximation of the functional dimension (middle panels), and the first coordinate (bottom panels). The number of items used is $I = 10, 100$ (from left to right). RMSE = root mean square error (color figure available online).

We show that a proposed nonlinear approximation (θ^* in the aforementioned example) is both accurate and robust under a range of conditions.

In this article, our goal is to delineate the meaning of the UIRT that is directly fitted to multidimensional data. Specifically, we first explore a way to capture the true functional dimension in a multidimensional generative model. We demonstrate that that the F1D can be closely tracked by a *nonlinear approximation* of the true underlying multidimensional generative model. We then assess the performance of the UIRT in terms of bias, root mean square error (RMSE), and coverage probability of confidence interval using the nonlinear approximation of F1D as an evaluative tool.

The remainder of the article is organized as follows: the next section describes the theoretical construct F1D and procedures for finding linear and nonlinear numerical approximations. Several simulation studies for evaluating the biases in the ability point estimate and Standard Error (SE) follow. A real example is then presented to illustrate the F1D approach. The final sections contain a discussion of the strengths and limitations of the F1D approach and a brief conclusion.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

THE F1D APPROACH

The functional dimension is what a unidimensional IRT measures, with sampling error, of multidimensional response data. As the product of an intrinsically misspecified model, the functional dimension is an approximation of the multidimensional latent space onto a unidimensional latent space (see Figure 1). We distinguish between two types of approximations, which we call conformal and empirical. In conformal approximations, a given multidimensional model is approximated by a unidimensional model. Wang (1986) discussed a direct projection onto a linear combination of the coordinates that is close to the MIRT. Ip (2010a, 2010b) and Ip and Chen (2012) discussed projection onto one of the identified dimensions (e.g., the first dimension) in MIRT. For empirical approximation, on the other hand, the direction of the projection is not known a priori and is determined by the empirical response data. In the following section, we describe a conformal nonlinear approximation to the F1D and then provide some evidence of the validity of the approximation. We also compare it against a linear projection that is based on the first principal component.

Conformal Nonlinear Approximation of F1D

Broadly, the purpose of nonlinear approximation (Lee & Verleysen, 2007) is to seek a representation of points in a high-dimensional space by points in a lower dimensional space that preserves the most salient features of interest (e.g.,

distances between points). Given a multidimensional IRT model of dimension m (as a generalization of Equation (1); see Equation (3)), suppose the item parameters are denoted by $(\mathbf{a}_i, d_i)$ and the ability vector by $\boldsymbol{\theta}_j$, where $\mathbf{a}_i$ and $\boldsymbol{\theta}_j$ are both vectors of length m. Our goal is to obtain UIRT item parameters (a_i^*, d_i^*) and individual scalar-ability θ_j^* such that item response functions are preserved, pointwise:

$$P(Y_{ij} = 1|\theta_j^*) = P(Y_{ij} = 1|\boldsymbol{\theta}_j), \forall i, j. \tag{2}$$

Translating into the logistic item response model, Equation (2) is equivalent to

$$\frac{\exp(a_i^*\theta_j^* - d_i^*)}{1 + \exp(a_i^*\theta_j^* - d_i^*)} = \frac{\exp(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i)}{1 + \exp(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i)}, \forall i, j. \tag{3}$$

A naïve solution to Equation (3) is to find (a_i^*, d_i^*) and θ_j^* that satisfy the following system of equations:

$$a_i^*\theta_j^* - d_i^* = \mathbf{a}_i^T\boldsymbol{\theta}_j - d_i, \forall i, j. \tag{4}$$

However, the set of Equation (4) contains $I \times J$ equations and $2I + J$ unknowns. Thus, Equation (4) is an overdetermined system for the kind of problem we consider—that is, when $J > I$. For example, when $I = 60$ and $J = 1{,}000$, there are 60,000 equations and 1,120 unknowns. No exact solution generally exists for such an overdetermined system.

Instead of directly solving an overdetermined system, we propose to find a solution to Equation (4) by solving a relaxed system. The relaxed system is based on the observation that the system Equation (4) can be rewritten as partial sums:

$$\begin{cases} \sum_{i=1}^{n_i}(a_i^*\theta_j^* - d_i^*) = \sum_{i=1}^{n_i}(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i), n_i = 1, \ldots, I, \\ \sum_{j=1}^{n_j}(a_i^*\theta_j^* - d_i^*) = \sum_{j=1}^{n_j}(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i), n_j = 1, \ldots, J. \end{cases} \tag{5}$$

One method to reduce the number of equations is to only consider selected summation equations in Equation (5). We propose a feasible solution for the system of Equation (5) by simultaneously solving the following three sets of

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

equations, which contain exactly $2I + J$ equations.

$$\begin{cases} \sum_{i=1}^{I}(a_i^*\theta_j^* - d_i^*) = \sum_{i=1}^{I}(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i), \forall j, \\ \sum_{j=1}^{\lfloor J/2\rfloor}(a_i^*\theta_j^* - d_i^*) = \sum_{j=1}^{\lfloor J/2\rfloor}(\mathbf{a}_i^T\boldsymbol{\theta}_j - d_i), \forall i, \\ \sum_{j=\lfloor J/2\rfloor+1}^{J}(a_i^*\theta_j^* - d_i^*) = \sum_{j=\lfloor J/2\rfloor+1}^{J}(\mathbf{a}^* T_i\boldsymbol{\theta}_j - d_i), \forall i. \end{cases} \tag{6}$$

Here, $\lfloor x \rfloor$ indicates the integral part of x. Note that if (a_i^*, d_j^*) and θ_j^* are solutions of Equation (4), then they must also be solutions of Equation (6). We used the Newton method (Kelley, 2003) for solving Equation (6). In order to avoid local solutions, we employed multiple runs of the Newton method, each with randomly generated initial values, and then averaged the results from the multiple runs to form a solution. The Newton method was accomplished through the BBsolve program (Barzilai & Borwein, 1988), which implements a low-memory and computationally efficient method for solving large systems of nonlinear equations and optimizing nonlinear objective functions (Varadhan & Gilbert, 2009).

Conformal Linear Approximation

The idea of a conformal linear approximation is to produce a set of values $a_i^\alpha, d_i^\alpha, \theta_j^\alpha)$ through linear projection onto a direction $\boldsymbol{\alpha}$ such that the response function is close to a given MIRT model specified by the set of parameters $(\mathbf{a}_i, d_i, \boldsymbol{\theta}_j)$. The item parameters for the unidimensional model were obtained through the following linear projection procedure: given the MIRT and the ability vector $(\mathbf{a}_i, d_i, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta} \sim N(0, \Sigma)$, and for a *given* constant vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \geq 0$ with $\boldsymbol{\alpha}^T\Sigma\boldsymbol{\alpha} = 1$, the item parameters of a linearly projected unidimensional IRT model (a_i^α, d_i^α) were computed using the following equations (Wang, 1986):

$$\begin{aligned} a_i^\alpha &= (1 + \sigma_i^2)^{-\frac{1}{2}}\mathbf{a}_i^T\Sigma\boldsymbol{\alpha}, \\ d_i^\alpha &= (1 + \sigma_i^2)^{-\frac{1}{2}}d_i, \end{aligned} \tag{7}$$

where

$$\sigma_i^2 = \mathbf{a}_i^T\Sigma\mathbf{a}_i - (\mathbf{a}_i^T\Sigma\boldsymbol{\alpha})^2. \tag{8}$$

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Comparison of Approximations

Three sets of analysis respectively derived from the following generative models (approximation to be evaluated) follow: (a) unidimensional IRT as generative model (nonlinear approximation), (b) two-dimensional IRT as generative model (nonlinear approximation), and (c) two-dimensional IRT as generative model (linear projection onto the first principal component of the a parameters). Item response data were generated respectively using a two-parameter logistic (2-PL) unidimensional model (see Equation (3)) and the MIRT model in Equation (1). The commonly used IRT software MULTILOG (Thissen, 1991) was used to estimate abilities. Specifically, the maximum a posteriori (MAP) estimate was used. To compare the properties of the IRT estimates and the approximations, we generated item parameters for $I = 240$ items and then randomly selected 10, 40, 80, and finally all 240 items to create several experimental conditions, the idea being that if the nonlinear approximation to the F1D is accurate, then it should mimic large-sample IRT estimates of ability (as the number of items increases) regardless of the dimension of the underlying generative model. To avoid instability in item parameter estimation, we used $J = 2{,}000$ individuals for all reported experiments.

Unidimensional IRT as a generative model: Nonlinear approximation. For the unidimensional IRT, the following specifications were used to generate items and item responses: $a_i \sim Unif(1, 4)$, $d_i \sim Unif(-5, 5)$, and $\theta_j \sim N(0, 1)$.

Figure 3 shows scatterplots of the IRT estimate versus the true (unidimensional) ability (left panels) and the scatterplots of the nonlinear approximation to the latent ability versus the true ability (right panels). The corresponding RMSE is indicated in an insert within the scatterplot. Two observations can be made from Figure 3. First, the IRT-based estimate functions as expected—showing increasing improvement toward the true ability as the number of items increases; second, the nonlinear approximation to the latent ability closely tracks the true IRT ability without noticeable bias. The nonlinear approximation also appears to be rather robust and does not vary as a function of the number of items.

Two-dimensional IRT as a generative model: Nonlinear approximation. The dimension represented by a_{i1} in Equation (1) was designated as the dominant dimension, whereas the second dimension was considered a nuisance dimension, which was assigned two levels of conditions: low (L) and high (H). The high and low levels, which depend on the range of the values for a_{i2}, are an indication of the strength of the nuisance dimension. Specifications of the generative MIRT were as follows:

$$\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}) \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right), d_i \sim Unif(-5, 5),$$

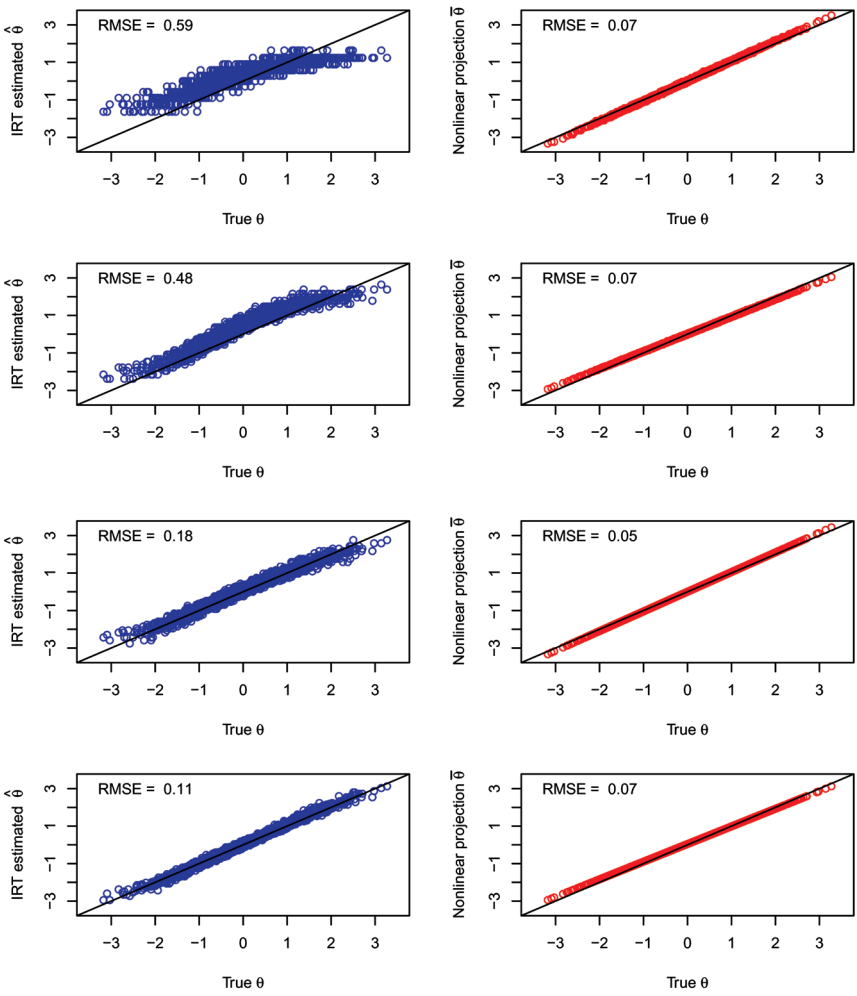


FIGURE 3 Scatterplots of the true (unidimensional) ability versus the item response theory (IRT) estimate (left panels) and the true ability versus the nonlinear approximation (right panels). The number of items used is $I = 10, 40, 80, 240$ (from top to bottom). RMSE = root mean square error (color figure available online).

and

$$\mathbf{a}_i = (a_{i1}, a_{i2}) \text{ with } a_{i1} \sim Unif(1,4), a_{i2} \sim \begin{cases} Unif(0.5,1) & \text{L,} \\ Unif(1,2) & \text{H.} \end{cases}$$

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

TABLE 1
RMSEs of Nonlinear and Linear Projections as a Function of Number of Items

	Nonlinear Projection				*Linear Projection*
No. of Items	*10*	*40*	*80*	*240*	*240*
Low a_{i2}	0.88	0.35	0.19	0.15	0.22
High a_{i2}	0.39	0.27	0.17	0.22	0.45

Note. RMSE = root mean square error.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Table 1 shows the RMSEs of the IRT estimate and the nonlinearly projected approximation to F1D for $I = 10, 40, 80$, and 240 across the L and H conditions. The IRT estimate with a large number of items ($I = 240$) is now used as a proxy for the F1D. The behavior of the nonlinear approximation tracks the large-sample behavior (as I increases) of the IRT estimate, suggesting that the nonlinear approximation captures the underlying single-dimensional latent trait quite well. An inspection of the scatterplots of IRT estimates versus the nonlinear approximation (not shown) also reveals no evidence of bias.

Two-dimensional IRT as generative model: Linear projection. We further evaluated the performance of the linearly projected approximation of F1D using the same set of conditions for the two-dimensional MIRT experiment. In both the L and H conditions, the maximum number of items ($I = 240$) were used. The RMSEs are 0.22 (0.45) for the low (high) levels, which are considerably higher than those derived from the nonlinear approximation (respectively 0.15 and 0.22). The variation of linear projections as a function of estimated ability also appears to be larger than that of the nonlinear approximation.

Based on these results, we made the decision to use the nonlinear approximation to F1D as the "true" F1D dimension in all of our subsequent simulation experiments. The aforementioned procedure made use of a customized R program for IRT data generation, the R function prcomp for principal component analysis, and an open-source R package BB (specifically BBsolve) for solving the relaxed system of equations in Equation (6).

SIMULATION EXPERIMENTS

The purpose of the simulation experiments is to evaluate the performance of UIRT when the multivariate binary response data contain multiple dimensions.

We used the nonlinear approximation as our "benchmark" for evaluating bias, RMSE, and 95% coverage of confidence intervals.

Design

We conducted four sets of Monte Carlo experiments using two-dimensional and one-dimensional stimulated response data, respectively. In Simulation Experiment 1, all response data Y_{ij}, $i = 1, \cdots, I$, and $j = 1, \cdots, J$ were generated from the two-dimensional MIRT model Equation (1). By fixing the variances of both θ_1 and θ_2 to 1.0, the model was identified and the strength of each dimension was controlled by the magnitude of the respective discrimination parameters, a_1 and a_2. Here, the first dimension θ_1 is designated as the dominant trait, whereas the second dimension θ_2 is assumed to be a nuisance factor. In the design, the strength of the first dimension, as indicated by the value of a_1 (assumed positive), was significantly higher than that of the second, which is indicated by the value of a_2 (also assumed positive).

A factorial design was used for generating Monte Carlo simulated data sets in the first experiment. We list the factors and their respective levels as follows: (a) number of items I, four levels at 15, 25, 35, and 45 items; (b) number of individuals J, three levels at 500, 1,000, and 2,000 persons; (c) relative strength of the minor dimension a_2, two levels as low (L), $Unif(0.5, 1)$, and high (H), $Unif(1, 2)$, with a_1 sampled from $Unif(1, 4)$ at both levels; and (d) the correlation ρ between the two dimensions, five levels at $\rho = -0.6, -0.3, 0.0, 0.3$, and 0.6. For all of the specified conditions, we replicated the experiment 10 times using the same setting in order to obtain an estimate of the variability of the performance metrics. In all, $(4 \times 3 \times 2 \times 5) \times 10 = 1{,}200$ data sets were generated and analyzed in the first experiment.

In Simulation Experiment 2, a five-dimensional MIRT (Reckase, 2009, p. 153) with realistic item parameter values was used to generate data. The parameters are shown in Table 2. We list the factors and their respective levels as follows: (a) number of items I, three levels at 10, 20, and 30 items; (b) number of individuals J, three levels at 500, 1,000, and 2,000 persons. For all conditions, $\boldsymbol{\theta} \sim N(0, \Sigma)$ where Σ is a 5×5 identity matrix.

In Simulation Experiment 3, unidimensional response data were generated using the specification described in the first experiment except that there was no condition for a_2 and ρ. This experiment was intended to "stress test" the approximations in situations in which the true parameter is known with certainty.

The fourth experiment, which was suggested by a reviewer, aimed to compare the scores and standard errors from the F1D approach with the bifactor approach. A five-dimension bifactor model with parameter values extracted from Gibbons and Hedeker (1992, p. 430, Table 2) was used for response data generation

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

TABLE 2
Item Parameters (Reckase, 2009, p. 153) for Generating Five-Dimensional Response Data for Simulation Experiment 2

Item	a_1	a_2	a_3	a_4	a_5	d
1	0.75	0.03	0.14	0.06	0.07	−0.18
2	0.46	0.01	0.07	0.02	0.04	0.19
3	0.86	0.01	0.40	0.02	0.20	0.47
4	1.01	0.01	0.05	0.02	0.03	0.43
5	0.55	0.02	0.15	0.04	0.08	0.44
6	1.35	0.01	0.54	0.02	0.27	0.58
7	1.38	0.09	0.47	0.18	0.24	1.04
8	0.85	0.04	0.26	0.08	0.13	−0.64
9	1.01	0.01	0.20	0.02	0.10	−0.01
10	0.92	0.01	0.30	0.02	0.15	−0.09
11	0.80	0.24	0.00	0.12	0.00	−0.81
12	1.19	0.19	0.00	0.10	0.00	0.19
13	0.71	0.09	0.06	0.05	0.12	−0.45
14	2.14	0.33	0.02	0.17	0.04	1.84
15	0.86	0.05	0.03	0.03	0.06	−0.41
16	0.93	0.15	0.02	0.08	0.04	−0.30
17	1.36	0.29	0.03	0.15	0.06	−0.18
18	0.90	0.22	0.00	0.11	0.00	−0.51
19	0.73	0.47	0.00	0.24	0.00	−1.13
20	0.64	0.09	0.01	0.05	0.02	−0.02
21	0.97	0.31	0.00	0.16	0.00	−0.62
22	0.50	0.18	0.00	0.09	0.00	0.20
23	1.11	0.41	0.20	0.21	0.40	0.37
24	1.73	0.15	0.03	0.08	0.06	1.76
25	0.67	0.15	0.00	0.08	0.00	0.24
26	1.24	0.29	0.02	0.15	0.04	−0.49
27	1.49	0.13	0.00	0.07	0.00	0.34
28	0.48	0.05	0.00	0.03	0.00	−0.29
29	0.46	0.21	0.01	0.11	0.02	−0.01
30	1.12	0.18	0.09	0.09	0.18	−0.03

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

($I = 20$; $J = 1{,}000$). Besides scores comparisons (procedure described later), estimated item parameters were compared across the three models.

All four simulation experiments were implemented through R for IRT and MIRT data generation and F1D estimation. Particularly, the R package ltm (Rizopoulos, 2006) was adopted for IRT estimation. In all of our reported analyses, the MAP estimate was used for reporting the IRT-based estimate.

The metrics for assessing performance of an estimation procedure were bias, the RMSE, and coverage probability of the 95% confidence intervals, which are

respectively defined as follows:

$$\text{Bias} = \hat{\theta}_j - \theta_j^*, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j^*)^2}, \tag{10}$$

where $\hat{\theta}$ is the unidimensional MAP estimate based upon simulated data, θ^* is the nonlinear approximation of F1D, and

$$\text{Coverage Probability} = \frac{\#\text{ respondents whose } \theta^* \text{ fall within 95\% confidence limits}}{\#\text{ total respondents}}, \tag{11}$$

where the 95% confidence limits were centered at $\hat{\theta}$ and were of length equal to 1.96× SE in either direction. If the SE calculation is correct, then one expects the value θ^* to fall within the confidence limits with a probability approximately equal to the coverage probability defined in Equation (11).

Results for Experiment 1: Two-Dimensional Response Data

Assessing bias in ability point estimates. For the assessment of bias and RMSE, we only report results using $J = 2{,}000$. The results for smaller sample sizes are consistent with those for $J = 2{,}000$ and are not shown here due to space limitations. To visually inspect bias, we obtain the RMSEs and scatterplots (a random sample of 200 simulated subjects from a larger sample of 2,000) of the nonlinearly projected approximation θ^* versus the IRT estimate for each of the conditions described earlier. Figure 4 shows selected scatterplots of UIRT estimates versus the benchmark value θ^* at low levels of the minor dimension. Individual RMSE values are reported in an insert within each respective scatterplot. Note that reporting the averaged bias could be misleading because bias in either direction could cancel out. Both visualization through scatterplot and reporting RMSE would provide a more overall accurate picture of bias. Because of space limitation, we only show two scatterplots that represent the most extreme RMSEs as well as two that have average values of RMSE. Three observations can be made from Figure 4. First, there is no evidence of bias in the ability estimate. Second, the RMSEs due to using the F1D approach for estimating the multidimensional response data are consistent across the various simulation conditions (range 0.24, 0.39). Finally, there does not appear to be

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

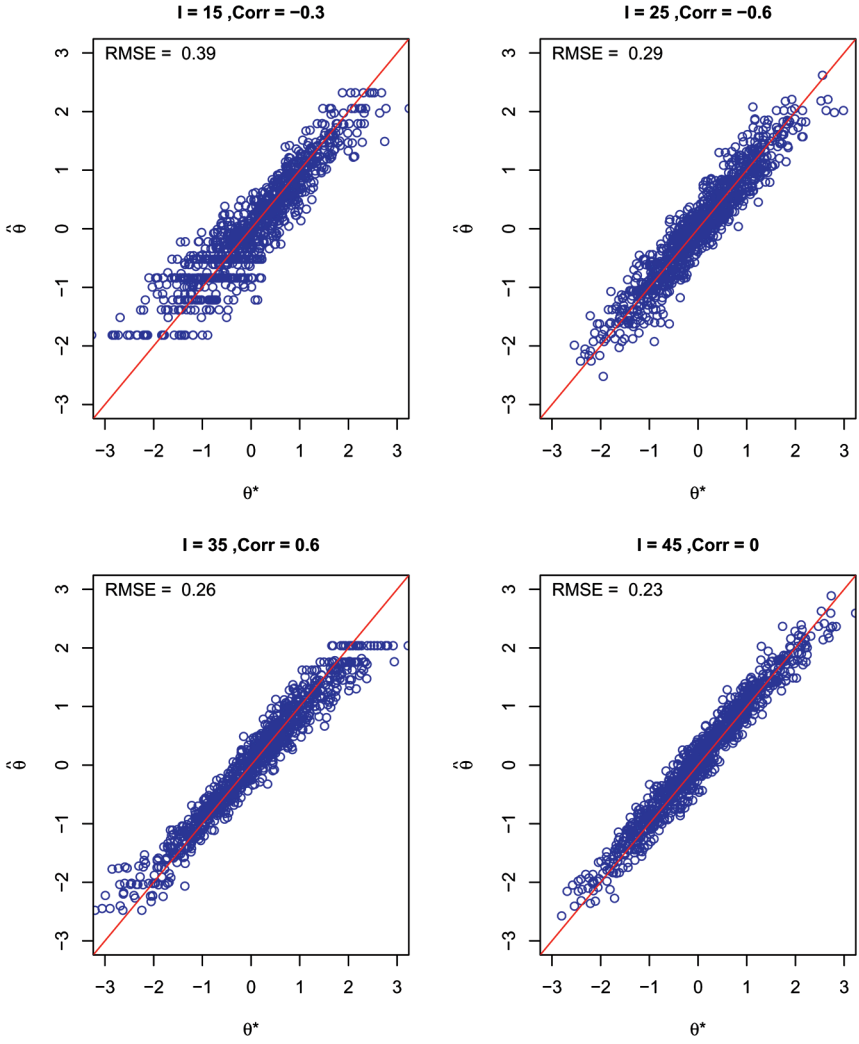


FIGURE 4 Selected scatterplots of the nonlinear approximation θ* versus the item response theory (IRT) estimate θ at low level of the minor dimension. RMSE = root mean square error (color figure available online).

a noticeable difference in the pattern of RMSE compared with those reported in Table 1. The graph (not shown) for a high level of a_2 (Condition H) looks similar to that for a low level of a_2 (Condition L). The range of RMSE for a Condition H is (0.24, 0.34). The results collectively suggest that, in terms of bias, the relative strength of the minor dimension a_2 does not substantially affect the UIRT estimate $\hat{\theta}$.

Assessing bias in SE. The results of the simulation experiments for UIRT 95% coverage probabilities across the stated conditions for low and high levels of the minor dimension are reported in Table 3. The average value of coverage probability across 10 replications of the same condition is reported within each cell. The standard deviation, which was estimated using the empirical standard deviation across the 10 replications, is reported parenthetically.

The conventional IRT (CSE) estimate from the UIRT for confidence limits produces surprisingly accurate coverage probabilities when the number of items is small and the strength of the nuisance dimension is low. However, the coverage drops with an increasing number of items, suggesting a possible underestimation of SE. The effect is more pronounced for high levels of the minor dimension, suggesting that CSE tends to underestimate SE when the strength of the minor dimension increases. The averaged numbers of undercoverage across all the cells (i.e., across all 1,200 replications) is 80.7. An unbiased estimate of SE should be approximately 50%. Using this criterion, the CSE is judged to contain confidence intervals that are too short.

Results for Experiment 2: Five-Dimensional Response Data

Table 4 shows the coverage of the 95% confidence intervals for the 5-D data set across the stated conditions. The CSE continues to perform quite well. One explanation for this phenomenon is that the minor dimensions in the 5-dimensional (5D) data are not strong (see Table 2) and resemble Condition L in Experiment 1. Inspection of the scatterplots (not shown) also did not reveal any evidence of bias.

Results for Experiment 3: One-Dimensional Response Data

Table 5 shows the coverage of the 95% confidence intervals for truly unidimensional response data. With the model now correctly specified, the "stress test" shows that CSE performs very well in terms of its coverage probability. As expected, there also is no evidence of bias in the point estimate. The result for assessing bias of the point estimate is not shown to save space.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

TABLE 3
Coverage of the 95% Confidence Interval in Simulation Experiment 1: Low (L) and High (H) Levels of Minor Dimension, Averaged Over 10 Replications

		15		25		35		45	
J	I Correlation	L	H	L	H	L	H	L	H
500	−0.6	94.66	95.12	94.94	89.18	95.20	94.46	90.74	90.92
		(1.03)	(1.37)	(1.35)	(0.69)	(1.53)	(1.43)	(3.96)	(1.64)
	−0.3	93.32	94.28	95.82	95.30	94.64	91.38	93.32	93.08
		(0.95)	(0.77)	(0.67)	(1.54)	(0.88)	(0.92)	(2.70)	(1.31)
	0	94.82	89.06	93.06	88.98	92.52	89.82	91.42	86.90
		(1.07)	(0.86)	(0.79)	(2.51)	(1.96)	(3.77)	(2.95)	(5.01)
	0.3	94.50	92.84	93.88	91.30	94.20	86.40	89.86	89.26
		(1.00)	(1.13)	(1.66)	(3.19)	(0.86)	(1.74)	(3.91)	(6.11)
	0.6	90.74	92.78	90.82	92.10	92.54	90.12	92.62	88.34
		(1.65)	(1.23)	(4.22)	(2.53)	(3.71)	(2.72)	(3.88)	(3.39)
1,000	−0.6	93.81	97.02	95.03	94.62	89.96	94.49	94.49	94.47
		(0.58)	(0.96)	(0.49)	(0.74)	(1.21)	(0.60)	(0.91)	(1.01)
	−0.3	94.60	93.34	94.20	92.80	87.73	94.54	93.50	88.63
		(0.81)	(0.78)	(0.68)	(0.35)	(1.54)	(1.12)	(2.20)	(1.66)
	0	95.95	87.22	93.23	86.34	91.31	92.07	89.11	91.09
		(0.55)	(1.83)	(0.99)	(3.71)	(1.58)	(1.53)	(1.72)	(5.63)
	0.3	96.17	95.67	94.80	87.38	90.32	93.62	89.51	92.07
		(0.49)	(0.67)	(0.83)	(1.81)	(2.56)	(2.94)	(2.81)	(2.81)
	0.6	93.43	95.86	94.41	92.82	92.41	89.31	91.17	87.43
		(0.90)	(1.04)	(1.02)	(1.21)	(1.65)	(1.41)	(2.12)	(2.84)
2,000	−0.6	97.30	96.98	96.00	95.36	92.92	92.14	95.53	90.89
		(0.26)	(0.37)	(0.31)	(0.33)	(0.87)	(0.77)	(0.48)	(0.98)
	−0.3	94.79	93.85	95.48	94.44	89.73	88.05	87.68	90.76
		(0.40)	(0.57)	(0.36)	(0.48)	(0.79)	(1.40)	(1.78)	(0.70)
	0	96.95	94.74	92.79	95.87	85.98	91.84	94.55	93.41
		(0.41)	(0.66)	(0.67)	(0.54)	(0.92)	(0.66)	(0.91)	(1.58)
	0.3	96.64	94.99	92.11	95.15	94.35	91.71	90.58	91.78
		(0.54)	(0.60)	(1.43)	(0.47)	(0.68)	(0.66)	(1.27)	(1.20)
	0.6	91.65	95.41	87.33	91.05	93.23	87.87	93.60	91.56
		(0.62)	(0.26)	(1.30)	(1.11)	(0.69)	(1.39)	(0.70)	(2.41)

Note. Standard deviation across replications is shown in parentheses.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Results for Experiment 4: Bifactor Response Data

Table 6 shows the factor loadings of the bifactor model and the estimated item parameters from the UIRT and the nonlinear approximation. The coverage of the CSE-based 95% confidence intervals is 93.6%. The RMSE is 0.57.

TABLE 4
Coverage of 95% Confidence Interval in Simulation Experiment 2: Five-Dimensional Response Data, Averaged Over 10 Replications

	$J = 500$	$J = 1{,}000$	$J = 2{,}000$
$I = 10$	96.22 (0.67)	99.27 (0.18)	98.53 (0.31)
$I = 20$	98.24 (0.60)	97.26 (0.54)	98.26 (0.31)
$I = 30$	89.36 (1.33)	88.51 (0.93)	97.61 (0.37)

Note. Standard deviation across replications is shown in parentheses.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

DESIRE FOR PHYSICAL COMPETENCE DATA EXAMPLE

The F1D approach was illustrated using a small psychological data set, which was collected from a study of older adults' *desire* for physical competence. The study aimed to examine the role that motives and expectations play in the process of physical disablement. The specific psychological construct of motive was measured by an instrument, the Desire for Physical Competence (DPC) survey, which contained 16 items. The development and validation of the DPC survey were reported in Rejeski, Ip, Katula, and White (2006). An example of an item is "What is your *level of desire* for having the ability to do heavy work in the house or yard?" The original item responses were each measured on a 5-point Likert scale (*No desire whatsoever* to *Very strong desire*). Because of the sparseness for some categories, we followed Rejeski et al. and dichotomized the response categories that corresponded to low levels of desire (No, Low, and Moderate desire) and high levels of desire (Strong and Very Strong desire).

Data from a total of 157 older adults (mean age = 73.3, range = 60–95) were available for analysis. We first applied factor analysis to the original 5-point response data using varimax rotation. Figure 5 shows the scree plot of the first 10 factors. The first two factors explained approximately 95% of the variance.

TABLE 5
Coverage of the 95% Confidence Interval in Simulation Experiment 3: One-Dimensional Response Data, Averaged Over 10 Replications

	$J = 500$	$J = 1{,}000$	$J = 2{,}000$
$I = 15$	94.10 (1.22)	94.63 (1.02)	94.11 (0.60)
$I = 25$	94.38 (1.82)	93.56 (0.73)	95.34 (0.33)
$I = 35$	94.02 (1.43)	93.77 (1.98)	95.34 (0.44)
$I = 45$	92.12 (3.75)	95.25 (1.05)	94.59 (0.97)

Note. Standard deviation across replications is shown in parentheses.

TABLE 6
Comparison of Item Parameters From the Generative Bifactor Model (Gibbons & Hedeker, 1992, p. 430, Table 2), the Unidimensional Item Response Theory (UIRT) Model, and Nonlinear Approximation

		Item Parameters			*UIRT Estimates*		*Nonlinear Approximation*	
Item	*Group*	*Difficulty*	*General*	*Specific*	*d*	*a*	*d*	*a*
1	1	−0.465	1.058	0.261	−0.409	1.037	−0.465	1.259
2	1	−0.690	0.408	0.202	−0.663	0.462	−0.693	0.574
3	1	−0.781	0.868	0.901	−0.691	1.249	−0.801	1.625
4	1	−0.254	1.181	0.599	−0.305	1.215	−0.263	1.671
5	1	−0.097	1.034	0.715	0.046	1.135	−0.110	1.627
6	2	−1.109	0.582	0.518	−0.923	0.617	−1.130	0.516
7	2	−1.170	0.733	0.612	−0.934	0.644	−1.195	0.655
8	2	−0.236	0.149	−0.032	−0.213	0.061	−0.233	0.148
9	2	−0.491	0.392	0.739	−0.515	0.320	−0.523	0.309
10	2	−0.206	0.452	0.952	−0.196	0.523	−0.249	0.346
11	3	−0.122	0.953	0.278	−0.046	0.825	−0.123	0.711
12	3	−0.735	0.464	0.240	−0.676	0.518	−0.738	0.263
13	3	−0.533	0.814	0.383	−0.463	0.769	−0.538	0.491
14	3	−0.141	1.221	0.399	−0.084	0.890	−0.143	0.877
15	3	0.051	1.007	0.532	0.102	1.022	0.044	0.562
16	4	0.603	0.199	0.463	0.543	0.232	0.604	0.324
17	4	0.373	0.761	1.645	0.272	0.498	0.376	1.204
18	4	0.347	0.520	1.237	0.276	0.689	0.348	0.854
19	4	1.121	0.722	1.047	1.060	0.783	1.125	0.998
20	4	1.144	0.104	0.053	0.965	0.060	1.144	0.116

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Also, there are two eigenvalues that are greater than 1, and the "elbow" of the plot is at two factors. In summary, the scree plot shows evidence of the presence of a dominant dimension and a second, minor component. These data thus seem to fit the characterization of F1D.

We examined the goodness-of-fit indices of several models: UIRT; MIRT of two, three, and four dimensions; and the bifactor model. The bifactor model was created using the factor loadings from the 2-D MIRT model. Dimensionality analysis was conducted using an R-based MIRT program (Chalmers, 2012). Table 7 shows several commonly used goodness-of-fit indices for the fitted models. Root mean square error of approximation (RMSEA) is smallest for the 1-D model, but other indices seem to suggest that either a bifactor model or a 2-D model is better. Examining the factor loadings of the various models (not shown) suggested that there was a strong factor and one or more factors of which the interpretations were not immediate.

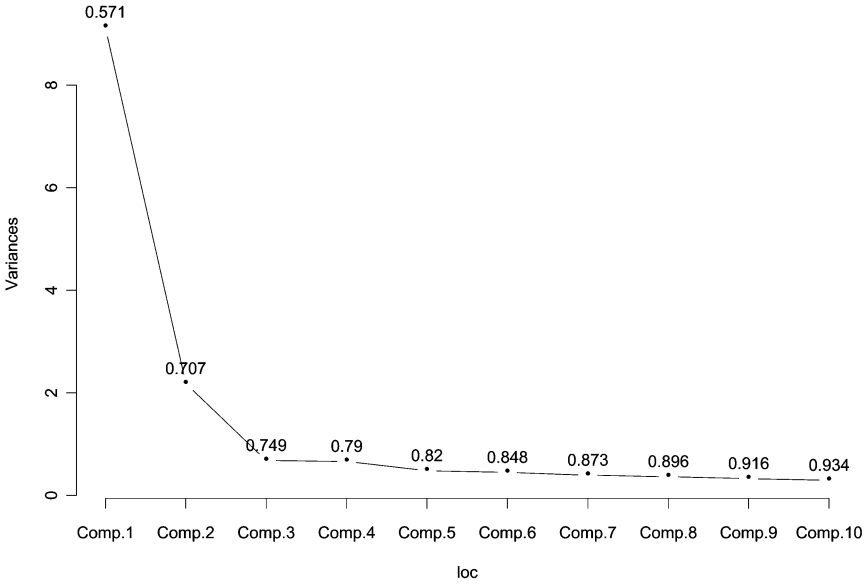


FIGURE 5 Scree plot for desire for physical competence items from factor analysis.

Examining the items in the 2-D model revealed that those that loaded high on the second factor were items that tended to involve more difficult tasks than those loaded high on the first factor, but otherwise neither set of items exhibited shared common attributes within the set. If anything, all of the highly loaded items on the first (dominant) dimension tended to be items that were related to functions for the so-called Instrumental Activities of Daily Living. On the

TABLE 7
Goodness-of-Fit Indices for Assessing Dimensionality of the DPC Example

Index	1-D IRT	2-D MIRT	3-D MIRT	4-D MIRT	Bifactor
AIC	1,975.473	1,926.368	1,928.536	1,973.885	1,912.106
BIC	2,073.273	2,070.011	2,114.967	2,200.047	2,049.637
G^2	563.75	484.65	458.82	478.17	474.39
TLI	0.591	0.554	0.432	0.139	0.581
RMSEA	0.206	0.215	0.243	0.299	0.208

Note. 1-D IRT = one-dimensional item response theory; MIRT = multidimensional item response theory; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; G^2 = likelihood ratio test statistic; TLI = Tucker Lewis index; RMSEA = root mean square error of approximation.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

other hand, the second factor appeared to involve desire for more advanced and difficult activities that may not be considered essential for day-to-day living for this age group (e.g., jogging a short distance). The preliminary analysis suggested that the DPC items appeared to function coherently and were oriented toward measuring desire for engaging in physical activities—despite the presence of a detectable second dimension that reflects functioning that is more advanced but not necessarily critical for everyday activities from an older adult's perspective.

Unlike the simulation experiments, the true values of the latent construct were not available, making the evaluation of performance of UIRT difficult. For the purpose of comparison, we compare the CSE and a robust version of SE estimate (RSE), which was used for enhancing UIRT SE estimates (Ip & Chen, 2012). The RSE is based on a so-called sandwich estimate that has been used in the statistical literature for robustifying SE estimates against possible deviations from parametric statistical assumptions such as sample independence (Huber, 1964; White, 1980). The scatterplot in Figure 6 shows a comparison between RSE and CSE. Each dot in the graph represents an individual or multiple individuals with the same SE. Figure 6 shows that the RSE estimate is generally larger than the UIRT-based CSE. Furthermore, Figure 6 suggests that the difference in SE tends to be larger for larger values of SE, which typically occur at the low and high ends of the ability scale. Because of the limitation of using real data, it is not clear if CSE is biased here. Given that the CSE is applied to a misspecified model and that there is evidence that the robust estimate of SE is higher, it is quite possible that the CSE does underestimate the SE.

In summary, the application highlights the challenges many researchers have faced when statistical tests did not immediately point to a unidimensional model. The simple F1D approach is defensible when the minor dimensions appear to be mild and can be explained by idiosyncratic features in the items.

DISCUSSION

We start with the issue of researchers "trying to measure one thing while simultaneously measuring diverse aspects of this same thing" (Reise et al., 2010). There are different scenarios that would entail different strategies and solutions. For example, subscale scores or second- or higher order models can be used if the test is judged to be truly multidimensional and has a specific factor structure. Another situation is when there exist minor dimensions that may arise from content development and that are difficult to separate as a subscale. In this article, we asked the following question: What does a unidimensional model mean if it is fitted to multidimensional response data? We called the theoretical underlying dimension, to which the large-sample estimate approaches, the functional dimension. We first argued that a nonlinear approximation satisfactorily tracks

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

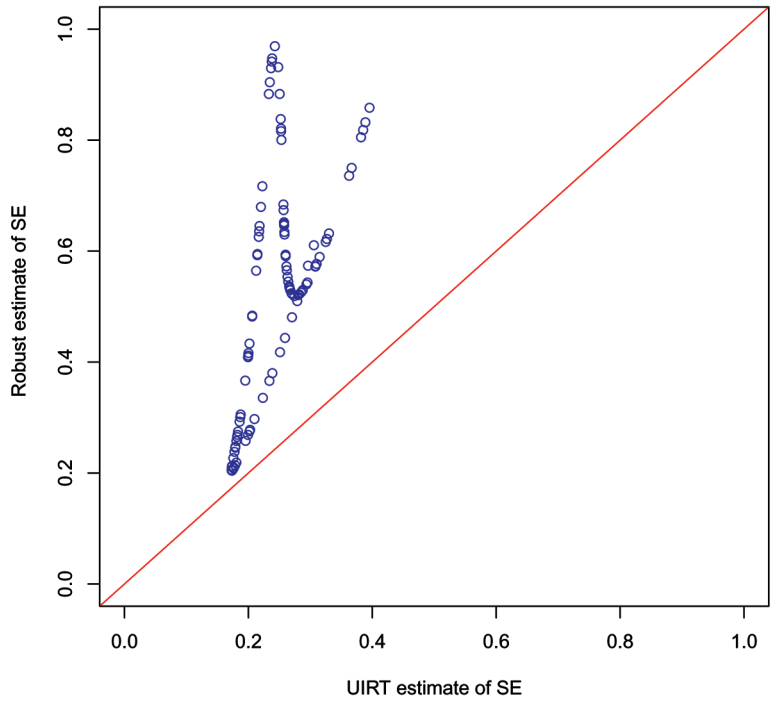


FIGURE 6 Standard error (SE) estimates from standard item response theory (IRT) and robust method for desire for physical competence responses. UIRT = unidimensional item response theory (color figure available online).

the functional dimension. Subsequently, using the nonlinear approximation as an evaluative tool, the performance of the UIRT estimate was assessed using Monte Carlo experiments.

There are two findings from this study regarding the functional dimension. First, the functional dimension appears to be better approximated by the nonlinear, unidimensional approximation of the multidimensional latent space than are its linear counterparts. Here, we are referring to the latent space specified by the respective parametric IRT and MIRT models, not the response space. It is worth emphasizing again that the F1D model is an inherently misspecified model in the presence of multidimensional data, and thus there really is no "true" ability from a generative model perspective. The nonlinear approximation appears to serve as a promising proxy for "true" ability. This observation is, of course, only based on results from simulated data and one version of nonlinear approximation that we proposed. A tremendous amount of work has been focused on linear projection, such as the form $\theta_1 \sin\alpha + \theta_2 \cos\alpha$ (Kim, 1994), but our preliminary results lean

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

toward the assertion that the closest approximation might well be nonlinear in nature. Further theoretical studies such as those using the geometric approach (Zhang, 2007), as well as the work from the weak, empirical indistinguishability between a multidimensional IRT model and a locally dependent unidimensional item response model (Ip, 2010a), could shed light on conformal approximations and projections. On the other hand, a general analytical solution for finding lower dimensional IRT representation for multidimensional response data might be possible, although the prospect seems unlikely (Junker, 1994).

The second finding from this article is that the IRT estimate is quite robust in estimating the F1D, especially in the point estimate. This finding agrees with results from previous studies (Junker & Stout, 1994; Kirisci et al., 2001). For assessing SE when strong minor dimensions exist in the data, conventional UIRT methods have a tendency to underestimate SE, or, alternatively, overestimate the accuracy of the point estimate. Fortunately, the extent to which CSE is underestimated is generally quite mild when the nuisance dimension is weak. The conventional IRT method performs well in this situation even when the dimension of response data is high, as demonstrated in the second and fourth simulation experiments. One explanation is that the IRT-based estimate essentially captures the preponderance of the variance, and there is not much noise (correlated or uncorrelated) left in the remaining structure to produce severely biased confidence limits.

A practical implication of this article is the utility of the IRT models, as supported by evidence of the robustness of UIRT for measuring the underlying trait of a test, or the functional dimension, in the presence of multidimensional response data. Admittedly, we have attempted to avoid providing explicit summative criteria for recommending when to use the F1D approach as opposed to a multidimensional IRT model. Reise et al. (2013) provided guidelines that included the use of measures such as Omega and OmegaH (McDonald, 1999) in determining when to use a summative score and when to start using subscales. The OmegaH measure, for example, is an indication of the extent to which a total scale score is interpretable as an indicator of the target construct of interest (Reise et al., 2013). As observed in Gustafsson and Aberg-Bengtsson (2010), the longer the test is, the better the total scale score indicates a single source of common variance. In this sense, the UIRT ability estimate derived from multidimensional data is analogous to the sample total-scale score of the same multidimensional data, whereas the F1D is analogous to the limiting case (as test length goes to infinity) for the "single source of common variance."

The F1D is conceptually distinct from the general factor in a bifactor model. It is also not the same as a composite score as measured by a particular set of items; nor is it the same as the marginalized trace line that could be formed by integrating out nuisance dimensions (Ip, 2010a, 2010b; Ip & Chen, 2012; Stucky et al., 2013). Although all four entities could be referred to as a one-number

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

summary of what can be derived from a test for an individual, mathematically they need to be carefully distinguished. To illustrate their differences, take a generative bifactor model as an example. The general factor in a bifactor model represents a common dimension *simultaneously specified* with other domain-specific dimensions, the marginalized trace line represents the mean effect of the general factor averaged over the domain-specific factors, the weighted sum of the bifactor scores represents a measure of the composite abilities, and the F1D represents the single dimension that is closest to the generative bifactor model in the latent space.

Besides directly fitting UIRT to the data, there exist alternative strategies for using unidimensional IRT models to handle potentially multidimensional response data. Luecht and Miller (1992), for example, adopted a multiple representative-dimension approach and used a two-stage procedure to first identify item clusters in a multidimensional space and then applied unidimensional IRT separately to each cluster. Wainer et al. (2001) discussed the use of collateral information to improve the accuracy of unidimensional estimates in the presence of nuisance dimensions. More recently, de la Torre (2009) discussed the use of hierarchical unidimensional IRT models for domain-based tests. Kahraman and Thompson (2011) suggested two-stage methods for handling multidimensional data in which the first stage used pure unidimensional items and the second stage individually calibrated suspected multidimensional items using the scale established in the first stage as one of the dimensions. These strategies are all useful for different measurement circumstances. The F1D approach described in this article could well be incorporated into some of these strategies—for example, deploying the F1D method to item clusters identified in the Luecht and Miller approach.

In structural equation modeling (SEM), there is an emerging trend to use multidimensional measurement models for commonly used or validated measures in accounting for potential multidimensionality in the data. It can be argued that in many such cases the instruments (validated or not) have been constructed as unidimensional tools, but secondary or tertiary dimensions get unintentionally included. As pointed out by Reise et al. (2013), the added complexity of using multidimensional measurement models may not result in practical gains in precision. In such cases, the F1D approach may be an appropriate way to derive single summary scores for latent constructs present in the network of SEM variables.

The current study has limitations. First, we did not attend to bias with respect to item parameter estimation when nuisance dimensions are present. This is clearly an important issue (Ansley & Forsyth, 1985; Harrison, 1986; Reckase & McKinley, 1991), but given the scope of this article we chose to focus on the issue of bias for the ability estimate, which we believe to be a less well-researched issue in measurement applications. It should be noted here that there is a substantial literature on bias in confidence limits that arises from

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

the viewpoint of local item dependency (Ip, Wang, De Boeck, & Meulders, 2004; Scott & Ip, 2002; Wainer, Bradlow, & Wang, 2007). Second, we have limited our discussion to multivariate binary responses. Extension of the F1D approach for polytomous items could be a fruitful area for future research work. Third, although the nonlinear approximation is pointwise close to the generative model, it may not be close to the F1D in general because θ is not a point but follows a distribution. However, the empirical evaluation of all the comparison models used the same points, and thus the comparison results should be quite robust. Another limitation is the lack of tests of two or higher dimensions in the simulation study, which limits the generalizability of the Monte Carlo findings.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

CONCLUSION

An emerging trend in the field of measurement is the proliferation of complex models. The unidimensional IRT, however, remains one of the mainstays of psychometric models for measurement. From an F1D perspective, the IRT model parameters and the derived scores, even in the presence of multidimensional data, are easy to interpret (Gustafsson & Aberg-Bengtsson, 2010) and, with the availability of user-friendly IRT software tools, easy to obtain. To extend the power of the IRT, it is therefore of importance to understand the mathematical and psychometric properties of the functional dimension and its estimates. This article offers a preliminary look at using the UIRT as a tool to estimate the functional dimension—one that we conceptualize as a latent variable that ostensibly explains a common source of variance between items. Questions such as general parameter invariance, robustness of estimates to deviation from unidimensionality, decision-support approaches for when to use what model (e.g., UIRT vs. MIRT), and the relationships between F1D and other models are all important and challenging issues that require further research.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255–278.

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.

Barzilai, J., & Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis, 8,* 141–148.

Cattell, R. B. (1966). Psychological theory and scientific method. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 1–18). Chicago, IL: Rand McNally.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modelling of latent DIF. *Applied Psychological Measurement, 35,* 583–603.

de la Torre, J. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33,* 620–639.

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

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

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

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association.

Halpin, P. F., da Silva, C., & De Boeck, P. (in press). A confirmatory factor analysis approach to test anxiety. *Structural Equation Modeling: A Multidisciplinary Journal*.

Harrison, D. A. (1986). Robustness of parameter estimation to violations to the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91–115.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35,* 73–101.

Hulin, C. L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Hulin, C. L., & Humphreys, L. G. (1980). Foundations of test theory. In E. B. Williams (Ed.), *Construct validity in psychological measurement* (pp. 5–10). Princeton, NJ: ETS.

Humphreys, L. G. (1949). Test purity. *American Psychologist, 3,* 245.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneburg (Ed.), *Current problems and techniques in multivariate psychology: Proceeding of a conference honoring Paul Horst* (pp. 22–32). Seattle: University of Washington Press.

Ip, E. H. (2010a). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63,* 395–415.

Ip, E. H. (2010b). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34,* 467–482.

Ip, E. H., & Chen, S.-H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement, 36,* 581–601.

Ip, E. H., Wang, Y., De Boeck, P., & Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika, 69,* 191–216.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255–278.

Junker, B. W. (1994, November). *Inference, robustness, and essential unidimensionality.* Talk given at the 10th annual Workshop on Item Response Theory Modeling, University of Twente, Enschede, The Netherlands.

Junker, B. W., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 31–61). Ottawa, Canada: University of Ottawa.

Kahraman, N., & Thompson, T. (2011). Relating unidimensional IRT parameters to a multidimensional response space: A review of two alternative projection IRT models for subscale scores. *Journal of Educational Measurement, 48,* 146–164.

Kelley, C. T. (2003). *Solving nonlinear equations with Newton's method*. Philadelphia, PA: SIAM.

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data* (Doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign).

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146–162.

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction.* New York, NY: Springer.

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

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279–293.

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

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 271–286). New York, NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY, and Heidelberg, Germany: Springer.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data.* Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario, Canada.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 14,* 361–373.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personal Assessment, 95,* 129–140.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personal Assessment, 96,* 544–559.

Rejeski, J., Ip, E. H., Katula, J., & White, L. (2006). Older adults' desire for physical competence. *Medicine and Science in Sports and Exercise, 38,* 100–105.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17,* 1–25.

Scott, S., & Ip, E. H. (2002). Empirical Bayes and item clustering effects in latent variable hierarchical models: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association, 97,* 409–419.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 586–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293–326.

Stucky, B. D., Thissen, D., & Edelen, M. O. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement, 37,* 41–57.

Thissen, D. (1991). *Multilog user's guide*. Lincolnwood, IL: Scientific Software International.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26,* 249–269.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 57–70). Vancouver, Canada: Educational Research Institute of British Columbia.

Varadhan, R., & Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software, 32,* 1–26.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its application*. New York, NY: Cambridge University Press.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, R. B., Rosa, K., Nelson, L., ... Thissen, D. (2001). Augmented scores: "Borrowing strengths" to compare scores based on small numbers of items. In H. Wainer & D. Thissen (Eds.), *Test scoring* (pp. 342–386). Mahwah, NJ: Erlbaum.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality. *Educational and Psychological Measurement, 66,* 721–738.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255–275.

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional response data.* Paper presented at the ONR Contractors Conference, Gatlinburg, TN.

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817–838.

Yen, W. M. (1984). Effects of local dependency on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72,* 69–91.

Zhang, J., & Stout, W. F. (1999). Conditional covariance structure of generalized multidimensional compensatory items with applications. *Psychometrika, 64,* 129–152.

Zhang, J., & Wang, M. (1998, April). *Relating reported scores to latent traits in a multidimensional test.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Downloaded by [KU Leuven University Library] at 05:42 01 April 2014