Published in final edited form as:
Psychol Methods. 2011 September ; 16(3): 221–248. doi:10.1037/a0023350.

Generalized Full-Information Item Bifactor Analysis

Li Cai, **Ji Seung Yang**, and **Mark Hansen**
University of California, Los Angeles

Abstract

Full-information item bifactor analysis is an important statistical method in psychological and educational measurement. Current methods are limited to single group analysis and inflexible in the types of item response models supported. We propose a flexible multiple-group item bifactor analysis framework that supports a variety of multidimensional item response theory models for an arbitrary mixing of dichotomous, ordinal, and nominal items. The extended item bifactor model also enables the estimation of latent variable means and variances when data from more than one group are present. Generalized user-defined parameter restrictions are permitted within or across groups. We derive an efficient full-information maximum marginal likelihood estimator. Our estimation method achieves substantial computational savings by extending Gibbons and Hedeker's (1992) bifactor dimension reduction method so that the optimization of the marginal log-likelihood only requires two-dimensional integration regardless of the dimensionality of the latent variables. We use simulation studies to demonstrate the flexibility and accuracy of the proposed methods. We apply the model to study cross-country differences, including differential item functioning, using data from a large international education survey on mathematics literacy.

Keywords

hierarchical factor model; item response theory; multidimensional IRT; item factor analysis; differential item functioning

Full-information item bifactor analysis (Gibbons & Hedeker, 1992; Gibbons et al., 2007) has been increasingly recognized as an important statistical method in psychological and educational measurement. Item bifactor analysis, as a special case of confirmatory multidimensional item response theory (IRT) modeling, provides information about the dimensionality of the measurement instrument, strategies for scaling individual differences, and new approaches to computerized adaptive testing. For instance, Reise, Morizot, and Hays (2007) applied item bifactor analysis to patient reported health outcomes data and concluded that the item bifactor model provides a valuable tool for exploring dimensionality. In psychopathology research, Simms, Grös, Watson, and O'Hara (2008) found that a bifactor structure is needed for describing mood and anxiety symptoms. In the area of psychiatric services research, Gibbons et al. (2008) applied the bifactor model to the construction of item banks and computerized adaptive tests and demonstrated dramatic reductions in patient and clinician burden. In educational measurement, DeMars (2006) applied the item bifactor model to data from testlet-based assessments and found the bifactor model a practical alternative to more specialized testlet response models (e.g. Wainer, Bradlow, & Wang, 2007).

Address correspondence to: Li Cai, UCLA, Los Angeles, CA, USA 90095-1521. lcai@ucla.edu. Phone: 310.206.0583. Fax: 310.206.5830.

From a psychometric angle, the hierarchical bifactor structure (Holzinger & Swineford, 1937) has provided plausible explanations for real data under a number of different scenarios (see e.g., Jöreskog, 1969; Tucker, 1958). Some theoretical results pertaining to its applicability also exist (see e.g., Yung, McLeod, & Thissen, 1999 for a discussion of the relation between hierarchical and higher-order factor analytic models). However, we are not proponents of the bifactor structure *per se*. Our view is that the item bifactor model is one of the many possible measurement models that may be useful in practice. Further research can provide fruitful avenues of new opportunities.

From a statistical angle, the success and popularity of full-information item bifactor analysis is largely due to Gibbons and Hedeker's (1992) discovery of an efficient maximum marginal likelihood parameter estimation method. It takes advantage of the bifactor measurement structure to analytically reduce the dimensionality of the otherwise intractable integrals in the item factor analysis likelihood function. Regardless of the number of factors in the model, one only needs to numerically approximate a series of two-dimensional integrals. We refer to this general strategy as *dimension reduction*. This computational breakthrough, together with its availability in popular software programs (e.g., TESTFACT, Bock et al., 2003; BIFACTOR, Gibbons & Hedeker, 2007), have made full-information item bifactor analysis widely accessible.

We note here that because the item bifactor model is a confirmatory item factor analysis model, alternative estimation strategies exist. For instance, adaptive quadrature methods (Schilling & Bock, 2005) have already been implemented in software programs such as Mplus (Muthén & Muthén, 2008). Bayesian Markov chain Monte Carlo (MCMC) estimation methods for general confirmatory item factor analysis (see e.g., Wirth & Edwards, 2007 and the references therein), and stochastic optimization algorithms for maximum marginal likelihood estimation (Cai, 2010a; Cai, 2010b) are effective alternatives. Under appropriate circumstances, limited-information least squares estimators such as NOHARM (Fraser & McDonald, 1988) can also be applied.

We do not pursue the alternatives mentioned above for the following reasons. First, these methods share the common characteristic of not using analytical dimension reduction. When the factor pattern indeed conforms to the bifactor structure, which we assume throughout, deterministic quadrature based optimization algorithms (e.g., the EM algorithm of Bock & Aitkin, 1981) that implement dimension reduction can be substantially more efficient and stable than alternative adaptive quadrature or stochastic methods that do not actively employ dimension reduction (see e.g., Cai, 2010c, for a timing comparison). Second, duly noting the advantages of Bayesian inference, we focus on obtaining modal estimates, which, in principle, should be conducted first so that one may "begin mapping the posterior density" (Gelman, Carlin, Stern, & Rubin, 2004, p. 312). Finally, the flexibility of full-information maximum likelihood estimation becomes particularly important as the complexity of contemporary psychometric modeling surpasses the capabilities of limited-information estimators in routine use (Cai, 2010c).

Limitations of Current Methods

Successful as they are, current full-information item bifactor analysis methods are not without limitations. The original derivations in Gibbons and Hedeker (1992) relied heavily on a result specific to the bivariate normal integral, which restricted their item response model to the normal ogive model for dichotomous responses. Gibbons et al.'s (2007) recent extension to the case of graded responses continues to rely on the normal ogive argument. It turns out that the limitation to the normal ogive model is unnecessary. Using graph theory arguments, Rijmen and coworkers (Rijmen, Vansteelandt, & De Boeck, 2008; Rijmen,

2009) convincingly demonstrated that bifactor-type dimension reduction holds under more general conditions and for a wider variety of IRT models than Gibbons and Hedeker (1992) originally considered.

In practice, however, this restriction has been reinforced by software that only uses the normal ogive parameterization. For example, the unidimensional 3-parameter IRT model with an estimated lower asymptote parameter is popular in educational measurement and increasingly applied in psychological measurement (e.g. Reise & Waller, 2003). TESTFACT currently supports a multidimensional analogue of the 3-parameter model (Reckase, 1997) but does not allow the estimation of the lower asymptote as a free parameter. A recommendation (see e.g. Thissen & Wainer, 2001) that is also endorsed by the software manual is to fix the asymptote value at an estimate obtained from a prior unidimensional 3-parameter model run. This recommendation rests on the presumption that the prior estimate is unaffected by dimensionality, despite evidence to the contrary (e.g. DeMars, 2007). Estimating all parameters jointly is ideal but is not possible given current item bifactor analysis software.

Consider as another example Muraki's (1992) Generalized Partial Credit (GPC) model and its multidimensional counterpart (e.g. te Marvelde, Glas, Van Landeghem, & Van Damme, 2006). The GPC model is a popular alternative to Samejima's (1969) graded response model, particularly for constructed response items. As Thissen and Steinberg (1986) showed, it is both a special case of Bock's (1972) nominal categories model and a generalization of partial credit and rating scale models based on Rasch measurement principles (e.g. Andrich, 1978; Masters, 1982). Though Gibbons et al. (2007) implement a version of the rating scale model by restricting the thresholds of the normal ogive graded response model, it remains unclear whether GPC-type multidimensional models are possible within their framework.

It is of interest to note that current methodological research on item bifactor analysis has been confined to single-group models largely due to the complexity of the dimension reduction method. As of this writing, we are aware of only one full-information multiple-group analysis that utilizes dimension reduction (Jeon & Rijmen, 2010). However, the reality is that multiple-group IRT modeling (Bock & Zimowski, 1997) forms a cornerstone of IRT-based test linking and equating (see, e.g., Kolen & Brennan, 2004), IRT-based Differential Item Functioning (DIF) detection (see, e.g., Thissen, Steinberg, & Wainer, 1993), and IRT-based vertical/developmental scaling (see e.g., Williams, Pommerich, & Thissen, 1998), among other useful activities. As Reise et al. (2007) noted, the item bifactor model can potentially enrich all of the above areas of investigation. The key prerequisites – practical estimation methods and software implementation of flexible multiple-group item bifactor models – have yet to be developed. In particular, a multiple-group modeling framework will be most useful if it has a degree of flexibility that is comparable to structural equation modeling (e.g., the LISREL model and software, Jöreskog & Sörbom, 2001), wherein a user can impose restrictions such as fixing or equality constraints directly on the parameters, facilitating theory-driven hypothesis testing and model comparison.

Overall Objectives

In this research, we systematically generalize the bifactor model due to Gibbons and Hedeker (1992) and Gibbons et al. (2007). We aim to achieve the following objectives:

1. As a departure from the normal ogive tradition, the IRT models connecting the latent variables and observed data are multidimensional extensions of all major unidimensional logistic models, including but not limited to the 1-2-3-parameter logistic models, the logistic graded model, the GPC model, and the full nominal

model. The focus on logistic models makes any arbitrary mixings of the above models straightforward so that one may mix items with different response formats (e.g., dichotomous, ordinal, and even nominal) in a single analysis.

2. With more than one group, we allow the estimation of means and variances of all latent variables, provided that a group is chosen as the reference population. In addition, we permit generalized user-defined restrictions on all item and population mean/variance parameters. These features extend capabilities of standard multiple-group IRT models and software packages, e.g., Multilog (Thissen, 2003).
3. We preserve the computational efficiency of full-information maximum likelihood bifactor estimation by generalizing Gibbons and Hedeker's (1992) dimension reduction technique. This ensures that our item bifactor models can be estimated within a practical amount of time. There is little restriction in terms of the dimensionality of latent variables, the types of IRT models, the number of groups, the number of respondents, and the number of user-defined restrictions on parameters.
4. Finally, we illustrate the proposed methodology with simulated and real data.

The General Statistical Approach

A Bifactor-Like Structure

The standard item bifactor measurement structure has the following factor pattern

$$\begin{pmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & a_{31} & 0 \\ a_{40} & 0 & a_{42} \\ a_{50} & 0 & a_{52} \\ a_{60} & 0 & a_{62} \end{pmatrix}, \tag{1}$$

for 6 items, one general dimension and 2 specific dimensions. The a's are the item slope/discrimination parameters that are analogous to factor loadings. As can be seen, the first dimension is the general factor and the others are specific factors. The first subscript denotes the item and the second denotes the factor. As will be clear later, it is advantageous notationally to begin the index of factors from zero. We call a pattern *bifactor-like* if an item always loads on the general factor and is permitted to load on at most one specific factor, potentially with other fixing or equality constraints, e.g.,

$$\begin{pmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{11} & 0 \\ a_{30} & 0 & 0 \\ a_{40} & 0 & a_{40} \\ a_{40} & 0 & a_{40} \\ a_{60} & 0 & 0 \end{pmatrix}. \tag{2}$$

In the example above, item 3 does not load on a specific dimension and there are additional equality restrictions, as indicated by the subscripts, reducing the number of free parameters to 6. Some of these restrictions are identification conditions and some are testable constraints. For example, the equality restriction for the specific factor slopes a_{11} is an identification condition. In this case, the specific factor is defined by a pair of items, and

amounts to a residual correlation in standard confirmatory factor analysis. On the other hand, the within- and across-item equality restrictions on the slopes of items 4 and 5 may be testable. With appropriate software support for restrictions, a wide variety of interesting substantive questions can be implemented as user-defined restrictions and tested for statistical significance, e.g., are two items equally discriminating on the general dimension; are the item slopes equal across groups; do the latent variables have the same means, etc.

We may also use path diagrams to represent the relationships between items and factors. In Figure 1, we depict the bifactor-like structure in Equation 2 as a path diagram. As one can see, there are six observed items, represented as rectangles, and there are three latent factors (θ_0, θ_1, θ_2), shown as circles. What is different from a standard path diagram is the notation we use to represent the number of categories for each item and the kind of IRT model. The rectangle for item 1 contains the letter *D*, which means that we fit a dichotomous response model with a lower asymptote (i.e., a 3-parameter logistic model) to this item. Items 2 and 3 are marked with numbers 2 and 5, which denotes the use of graded response models for 2 and 5 categories, respectively. It should be noted that the 2-category graded response model is equivalent to a 2-parameter logistic model. We describe it here as a graded response model in order to empahsize that the 2-category model belongs to the larger family of models for ordinal data. We fit the GPC model to items 4 and 5 (hence the letter *G* in the rectangles) for ordinal categories. Item 4 is a 3-category item (G_3) and item 5 has 4 categories (G_4). Finally, the last item is nominal with 5 categories (N_5). The details of these IRT models will be elaborated in a later section.

The general dimension is θ_0, and the two specific factors are defined by two item doublets. The item slope/discrimination parameters are represented as single-headed arrows, with equality and fixing constraints explicitly shown. For instance, the path from θ_1 to item 3 is omitted because it is a fixed zero slope, and the slopes of items 1 and 2 on θ_1 are set equal. The latent factors are mutually orthogonal with zero means and unit variances. We will give a more in-depth discussion of the distribution of the latent factors in the section on estimation.

Some Notation

Let us suppose that there are $g = 1, \ldots, G$ independent and mutually exclusive groups, which may be defined by such variables as gender, ethnicity, education level, disease status, treatment condition, geographical region, etc. In group g, let there be $j = 1, \ldots, n_g$ items each scored in K_{jg} categories. Further, let there be $i = 1, \ldots, N_g$ independent respondents or examinees in group g. The overall sample size is $N = \sum_{g=1}^{G} N_g$. We denote the response from person i to item j in group g as y_{ijg}. Without loss of generality, we assume that y_{ijg} may either take integer values from $\{0, 1, \ldots, K_{jg} - 1\}$ or y_{ijg} may be a missing value. Let us write the $n_g \times 1$ vector of item responses from respondent i in group g as $\mathbf{y}_{ig} = (y_{i1g}, \ldots, y_{ijg}, \ldots, y_{in_g g})^t$, where the symbol ^{t} denotes the transpose of a matrix or vector.[1] The entire $N_g \times n_g$ matrix of item responses in group g is $\mathbf{Y}_g$, with its ith row given by $\mathbf{y}_{ig}^t$. Let the n_g observed items measure $S_g + 1$ latent variables, which we denote here as $\boldsymbol{\theta}_g = (\theta_{0g}, \theta_{1g}, \ldots, \theta_{S_g g})^t$. Let the mean vector of $\boldsymbol{\theta}_g$ be $(\mu_{0g}, \mu_{1g}, \ldots, \mu_{S_g g})^t$. We preserve the orthogonality requirement for the factors so that for group g, the covariance matrix of $\boldsymbol{\theta}_g$ is diagonal, i.e., $\mathrm{diag}(\sigma_{0g}^2, \sigma_{1g}^2, \ldots, \sigma_{S_g g}^2)$. Of these $Sg + 1$ factors, θ_{0g} is the general dimension, and the rest are specific.2

[1] We follow the notational convention that regular lower case symbols are scalars while symbols in bold face denote vectors or matrices.

On the Conditional Independence Assumption

An IRT model relates the conditional probabilities of the item responses $\mathbf{y}_{ig}$ to them latent variables $\boldsymbol{\theta}_g$ (see e.g. Edwards & Edelen, 2009). Model building typically starts with the conditional independence assumption (Lord & Novick, 1968), i.e., the item responses are independent given the latent variables. Conditional independence is a hallmark feature of a number of popular models in psychometrics, including exploratory factor analysis and (unidimensional) item response theory. In the unidimensional IRT literature, satisfying conditional independence amounts to unidimensionality because the unidimensional IRT model only allows a single latent factor to influence the observed item responses. Various diagnostic approaches (e.g. Braeken, Tuerlinckx, & De Boeck, 2007; Chen & Thissen, 1997; Stout, 1990) exist to check this critical assumption. Because the IRT model we consider is multidimensional, the conditional independence assumption we make is a more relaxed one in that we condition on all $S_g + 1$ latent variables simultaneously. In other words, the operating assumption is that the item responses $\mathbf{y}_{ig}$ are correlated wholly due to their being influenced by a common set of unobserved random variables $\boldsymbol{\theta}_g$. Once the influence of $\boldsymbol{\theta}_g$ is accounted for, the item responses should be independent. Indeed, proponents of the bifactor model use the specific factors to explicitly model extra residual dependence above and beyond the primary dimension.

Bifactor IRT Models

Recall that the response of individual i to item j in group g is $y_{ijg} \in \{0, 1, \ldots, K_{jg} - 1\}$. There is no requirement that the items must all have the same number of response categories, or that all the IRT models have to be the same within a group. To avoid notational clutter, we temporarily drop the subscripts, with the understanding that we are referring to a generic item j in group g. The lack of subscripts should not be taken as an indication that the parameters are equal.

A Model for Dichotomously Scored Response

When the item response is scored dichotomously (i.e., $K = 2$), we denote $y = 1$ as the correct/endorsement response and $y = 0$ as the incorrect/nonendorsement response. Dichotomous item responses are routinely encountered in educational measurement (e.g. multiple-choice test items) or psychological assessment (e.g. symptom checklists). The model we describe is a bifactor extension of the classical 3-parameter logistic model. Let the conditional probability of correct/endorsement given the general dimension θ_0 and specific factor θ_s be

$$P(y=1|\theta_0,\theta_s)=c+\frac{1-c}{1+\exp\{-[d+a_0\theta_0+a_s\theta_s]\}}, \tag{3}$$

where c is the lower asymptote ("guessing" probability), d the item intercept, a_0 the item slope on the primary factor, and a_s the item slope on specific factor s. Naturally, the conditional probability for the incorrect/nondendorsement response is $P(y = 0|\theta_0, \theta_s) = 1 - P(y = 1|\theta_0, \theta_s)$. If the item does not load on a specific factor, then a_s is zero and the conditional probability does not depend on θ_s, in which case the model reduces to the classical 3-parameter logistic model:

[2]Note that the additional g subscript permits different number of observed variables as well as different number of latent variables across the groups. This gives the data analyst yet another degree of flexibility.

$$P(y=1|\theta_0)=c+\frac{1-c}{1+\exp\{-[d+a_0\theta_0]\}}. \quad (4)$$

In the unidimensional case, the logit in Equation (4) can be reexpressed in a more convenient slope-threshold form as $d + a_0\theta_0 = a_0(\theta_0 - b)$, where $b = -d/a_0$ is the threshold (or item difficulty) parameter, indicating the point on the θ_0 scale at which the probability for correct/endorsement response is exactly .5 if $c = 0$, or .5 + .5c if c is not zero. Unfortunately, the slope-threshold form does not generalize well to truly multidimensional models, so we adopt the slope-intercept parameterization for this model and all remaining IRT models.

In qualitative terms, the item slopes reflect the strength of association between the item responses and the factors, or how indicative/discriminating the item is for the latent variables. The model effectively generates an item response surface as a function of θ_0 and θ_s. Consider the level plot (a filled contour where lighter colors indicate higher probabilities) of the correct/endorsement response surface as shown in Figure 2. The slope parameters are $a_0 = 1$ and $a_s = 1$. Thus, the item is equally discriminating along both θ_0 and θ_s. In contrast, Figure 3 presents a response surface with a markedly different shape. The slope parameters are $a_0 = 2$ and $a_s = 0.5$. In other words, this item derives most of its discrimination along the direction of the primary dimension, as reflected by a much faster rate of change of the surface along the direction of θ_0 than θ_s.

The lower asymptote for the item in Figure 2 is $c = 0.3$, and hence the height of the contour begins at 0.3 from the lower-left corner. In educational measurement, the c parameter is routinely interpreted as the guessing probability for multiple-choice exam questions. In psychological measurement, Reise and Waller (2003) argue that a non-zero lower asymptote might indicate item content ambiguity that has little to do with the underlying trait or symptom being assessed. The lower asymptote of the item in Figure 3 is equal to 0.1, which is lower than the previous item, as clearly indicated by the gray scale legend on the right.

The item intercept parameter d is negatively associated with the difficulty/location of the item (Reckase, 2009). Indeed, as shown in Equation (4), this relation is transparent as we defined the item threshold parameter as $b = -d/a_0$. For a multidimensional IRT model such as the one we are discussing, we can adopt Reckase's (2009) formula (p. 90) to convert the intercept into a location parameter (multidimensional difficulty; MDIFF) to aid interpretation

$$\text{MDIFF}=\frac{-d}{\sqrt{a_0^2+a_s^2}}. \quad (5)$$

Note that MDIFF reduces to b when one of the a's is zero. The intercept parameter is $d = -1$ for the item depicted in Figure 2, whereas the intercept of the item in Figure 3 is equal to 1. The location parameters are therefore $1/\sqrt{1^2+1^2}=.71$ and $-1/\sqrt{2^2+0.5^2}=-.49$, respectively. Thus the item in Figure 3 can be interpreted as being either "easier" (in educational measurement) or less "severe" (in psychological measurement) than the other item, pending context. We discuss additional estimation issues for the dichotomous response model in Appendix A.

A Model for Graded Response

This model is the bifactor counterpart of the logistic version of Samejima's (1969) graded response model. Muraki and Carlson (1995) describe similar normal ogive multidimensional models. The graded response model has been popular in the item analysis for psychological scales and recently has become the recommended model in patient reported outcomes measurement (Thissen, Reeve, Bjorner, & Chang, 2007). In educational measurement, the graded model has been applied in the analysis of questionnaire data (e.g. at the student or teacher level) where the items have Likert-type ordered categorical response options or when the data are judged ratings of constructed response items. Let $y \in \{0, 1, \ldots, K-1\}$ be the item response in K graded categories. The development starts from the cumulative response probabilities:

$$\begin{aligned} P(y \geq 1|\theta_0, \theta_s) &= \frac{1}{1+\exp[-(d_1 + a_0\theta_0 + a_s\theta_s)]}, \\ &\vdots \\ P(y \geq K-1|\theta_0, \theta_s) &= \frac{1}{1+\exp[-(d_{K-1} + a_0\theta_0 + a_s\theta_s)]}, \end{aligned} \tag{6}$$

where $d_1, \ldots, d_{K-1}$ are a set of $K-1$ (strictly ordered) intercepts. As before, a_0 is the item slope on the primary factor, and a_s the item slope on specific factor s. Equation (6) implies that the category response probability is a difference of two adjacent cumulative response probabilities:

$$P(y=k|\theta_0, \theta_s) = P(y \geq k|\theta_0, \theta_s) - P(y \geq k+1|\theta_0, \theta_s), \tag{7}$$

where the two boundaries are defined by $P(y = 0|\theta_0, \theta_s) = 1 - P(y \geq 1|\theta_0, \theta_s)$ and $P(y = K - 1|\theta_0, \theta_s) = P(y \geq K - 1|\theta_0, \theta_s)$. If the item does not load on specific factor s, we obtain the unidimensional graded model.

As with the dichotomous model, the graded model generates a set of response probability surfaces in terms of the primary and the specific dimensions. Barring the absence of a lower asymptote parameter, the slope and intercept parameters carry similar meaning in this model as in the dichotomous response model. Figure 4 presents the category response surfaces for a hypothetical graded item with 3 categories.

The item slope parameters are $a_0 = 1$, $a_s = 2$. This item is more discriminating along the specific factor. The level plots clearly reveal the shapes of the category response surfaces. From the lower-left corner, the lowest category's response probabilities gradually decrease to zero as both θ_0 and θ_s increase. The middle category's response probabilities first increase and then decrease, resulting in a ridge that lies in the middle of the plot. The highest category's response probabilities gradually increase to 1 as both θ_0 and θ_s increase. The overall rate of change along θ_0 is not as fast as that of θ_s, consistent with the lower slope on the primary dimension. The first intercept is $d_1 = -1$ and the second intercept is $d_2 = 1$. As with the dichotomous response model, the intercepts are negatively related to the difficulty/location/severity of the response categories. The only difference here is that an item potentially possesses more than one intercept. Finally, we mention that if the number of categories K is exactly two, the graded response model will coincide with a dichotomous model with $c = 0$.

A Generalized Partial Credit Model

This model is a bifactor extension of Muraki's (1992) unidimensional GPC model (see e.g. Yao & Schwarz, 2006), which further generalizes the rating scale model (Andrich, 1978),

and the partial credit model (Masters, 1982). The GPC model is often used in scoring constructed response items for educational assessments. The GPC model is derived from a different set of assumptions about the item response process than the graded model, but it can often generate similarly shaped item response functions as the graded response model. Because of this, it is an effective and popular alternative to the graded model in practical data analysis. Chapter 4 of Reckase's (2009) book contains excellent discussions about the multidimensional GPC and graded models.

Let $y \in \{0, 1, \ldots, K-1\}$ be the item response in K ordinal categories. Historically there are a number of equivalent ways to express the GPC model. Here, we adapt Thissen, Cai, and Bock's (2010) notation for the full multidimensional GPC model. Let

$$P(y=k|\theta_0, \theta_s)=\frac{\exp\{T_k[a_0\theta_0+a_s\theta_s]+d_k\}}{\sum_{l=0}^{K-1}\exp\{T_l[a_0\theta_0+a_s\theta_s]+d_l\}}, \quad (8)$$

be the conditional response probability for category $k = 0, \ldots, K-1$, where T_k is the so-called scoring function for category k and d_k is the category intercept. We continue to denote the item slopes as a_0 and a_s for the general and specific dimensions, respectively. Without loss of generality, the scoring function can be taken as $T_k = 0, \ldots, K-1$. They are often interpreted as the scores that would be assigned to responses by a rater. The category intercepts are themselves not directly estimable due to an indeterminacy of the multinomial logit. If we add a constant to all the d_k's, the category response probabilities are left unchanged. This is analogous to the estimablity issues in analysis of variance: only $K-1$ contrasts are identifiable for K categories. A remedy is to impose an identification restriction so that one of the d_k's (usually that of the first category) is zero. We accomplish this restriction via a reparameterization of the intercepts as described in Appendix B. As before, if the item does not load on a specific factor, the unidimensional GPC model is obtained.

Figure 5 presents the GPC category response probability surfaces in terms of the primary and the specific dimensions. This item has three categories. The surfaces behave similarly as the graded model surfaces. As θ_0 and θ_s increases, the lowest category probabilities decrease to zero, the highest category probabilities increase to one, and the middle category probabilities first increase and then decrease.

For this item, the slope parameters are $a_0 = 1$ and $a_s = 1$, so it is equally discriminating along both latent dimensions. The intercepts are $c_0 = 0$, $c_1 = 1$, and $c_2 = 2$, which are interpretable as location parameters, much as the category intercepts in the graded model. An alternative interpretation is that the intercepts control the relative frequency of endorsements in each of the categories, or equivalently stated, the height of the category surfaces, at specific points in the (θ_0, θ_s) plane. Take the origin for example. At this point, $(\theta_0, \theta_s) = (0, 0)$, and the category probabilities are entirely functions of the d_k's. For the item in Figure 5, the first category's endorsement probability is equal to $e^0/(e^0 + e^1 + e^2) = .09$ at the origin. The second category probability is $e^1/(e^0 + e^1 + e^2) = .24$, and the highest category probability is $e^2/(e^0 + e^1 + e^2) = .67$ at the origin. The calculation generalizes to any number of categories and dimensions, as long as the item parameters are known.

A Model for Nominal Response

This model is a recent extension of Bock's (1972) classical nominal categories model to multidimensional IRT (Thissen et al., 2010). The nominal model is one of the most flexible parametric IRT models and has found wide-ranging applications in psychometrics. Thissen and Steinberg (1986) used the nominal model to provide a taxonomy of a variety of

polytomous IRT models. It has been applied to data analysis involving testlets (Thissen, Steinberg, & Mooney, 1989). The nominal model has been found useful in representing the item response process (e.g., Revuelta, 2007), and accounting for individual differences in response style (e.g., Johnson & Bolt, 2010). It has also emerged as a practical tool to explore the rank order of the response categories in data analysis for patient report outcomes (see, e.g., Thissen et al., 2007).

Let $y \in \{0, 1, \ldots, K-1\}$ be the item response in K nominal (unordered) categories. Category k's response probability is defined as

$$P(y=k|\theta_0, \theta_s)=\frac{\exp\{a_k^*[a_0\theta_0+a_s\theta_s]+d_k\}}{\sum_{l=0}^{K-1}\exp\{a_l^*[a_0\theta_0+a_s\theta_s]+d_l\}}, \quad (9)$$

where d_k is the category intercept (as in Equation 8) and a_0 is the item slope on the primary factor, a_s the item slope on specific factor s. The a_k^* scoring function parameters describe the empirical "ordering" of the response categories with respect to the linear combination of latent variables: $a_0\theta_0 + a_s\theta_s$. By comparing Equations (8) and (9), one can see that the GPC model is a special case of the nominal model, with predetermined scoring functions for each category, whereas the scoring functions are estimated parameters in the nominal model.

As with the GPC model, the category intercepts are not directly estimable. A similar identification restriction must be imposed on the d_k's so that the first intercept is always zero and what are estimated are the remaining $K-1$ intercept contrasts. Due to the added complexity of estimating the "ordering" of the categories from data, two identification restrictions must be placed on the set of a_k^* parameters. Hence, we estimate $K-2$ scoring function contrasts. We present details of the reparameterization in Appendix B. Though the particular reparameterization we adopt here is not unique, it has the easily discernable feature that a_0^* is always 0 and a_{K-1}^* is always $K-1$. Regardless of the reparameterization, we emphasize that the a_k^*'s can always be interpreted as the ordering of the response categories and can be used as scoring functions. Finally we note that this model reduces to Bock's (1972) original nominal model when one of the item slopes is zero.

Figure 6 presents level plots of the category response surfaces for a nominal model with 4 categories. The item slopes are $a_0 = 1$ and $a_s = 1$. These are overall item discrimination parameters and have the same interpretation as the item slopes in the other models. The scoring function parameters for this item are $a_0^*=0$, $a_1^*=0$, $a_2^*=3$, and $a_3^*=2$. The scoring function values must be strictly ordered ($a_0^*<a_1^*<a_2^*<a_3^*$) in order for the categories to be so ordered. Clearly, the categories are not ordered for this item. Since $a_0^*=a_1^*=0$, the order of the first two categories cannot be distinguished, and the order of categories 2 and 3 are in fact reversed, with category 2 possessing a higher scoring function value than category 3. The level plots clearly reveal the reversal of categories 2 and 3, showing category 3's surface increasing and then decreasing, as if it is one of the middle categories in a set of ordered responses.

The effects of the intercepts are harder to see from the level plots. As in the GPC model, they continue to function as indicators of the relative proportions of responses in each category. For this item, the intercepts are $c_0 = 0$, $c_1 = 2$, $c_2 = 1$, and $c_3 = 2$. The curves in Figure 7 resemble a set of item characteristic functions for the standard unidimensional nominal model, but they are in fact a cross-section of the item response surfaces in Figure 6, holding θ_s constant at 0. This slice of the item response surface reveals that category 1 has a

much higher endorsement probability than category 0, though the two categories' ordering is the same. In practice, this implies that categories 0 and 1 can be collapsed into a single category. In addition, it appears from Figure 6 that the order of response categories 2 and 3 should be reversed. With these changes, an IRT model for ordinal data (e.g., the graded model) might be fitted to this item.

Putting the IRT Models Together

From Equations (3), (7), (8), and (9), we can write, generically, for respondent i's response to item j, the conditional category response probability as $P(y_{ij} = k|\theta_0, \theta_s)$, for $k = 0, \ldots, K_j - 1$. Let

$$\chi_k(y) = \begin{cases} 1, & \text{if } y=k, \\ 0, & \text{otherwise,} \end{cases}$$

be an indicator function such that $\chi_k(y) = 1$ if and only if $y = k$, for $y \in \{0, 1, 2, \ldots\}$, and $\chi_k(y) = 0$ otherwise. The conditional distribution for an observed response y_{ij} is a multinomial with trial size 1 in K_j cells, and the conditional density is

$$f(y_{ij}|\theta_0, \theta_s) = \prod_{k=0}^{K_j-1} P(y_{ij}=k|\theta_0, \theta_s)^{\chi_k(y_{ij})}. \tag{10}$$

Note that the indicator function effectively selects out the category response probability for each observed response. If y_{ij} is a missing value, then the value of the indicator function will always be 0, which implies that the conditional density $f(y_{ij}|\theta_0, \theta_s)$ is identically equal to 1. This observation will subsequently be important because the natural logarithm of 1 is 0, so a missing observation does not contribute any information to the log-likelihood. The resulting estimator thus utilizes observed information fully in the presence of of missing data.

The Distribution of Latent Factors

We make the standard bifactor analysis assumption that the general factor and the specific dimensions are jointly normally distributed and mutually orthogonal (see, e.g., Gibbons & Hedeker, 1992).3 This is also the underlying assumption of testlet response models (Wainer et al., 2007). For a single group, the location and scale of the latent variables must be set in order to achieve identification. It is typically assumed that $\boldsymbol{\theta}$ has a multivariate normal distribution with zero means and identity covariance matrix. Appendix C contains additional details.

The orthogonality and normality imply that the density function of the latent variables reduces into a product. In the case of Figure 1, the distribution of the latent factors can be written as

$$f(\theta_0, \theta_1, \theta_2) = f(\theta_0) f(\theta_1) f(\theta_2), \tag{11}$$

[3]Strictly speaking this is not necessary. Rijmen (2009) showed that the dimension reduction only requires the conditional independence of the specific factors given the general factor. Nevertheless, the orthogonality restriction is often required to identify the model.

where $f(\theta_s)=1/\sqrt{2\pi}\exp(-0.5\theta_s^2)$ is the standard univariate normal density.

For multiple-group analysis, the latent variable means and variances become estimable parameters, provided that a group is chosen as the reference, and that enough identification restrictions have been imposed on the item parameters to link the latent variable scales together. As with the item parameters, we also permit arbitrary user-defined equality or fixing restrictions on the factor means or variances, either within or between groups.

Consider the simplest situation, where there are two groups. Typically the latent variables in the reference group have zero means and an identity covariance matrix, and the location and scale of the latent variables in the other group are estimated relative to the reference group, with at least one common item's parameters set equal across the two groups to link the scales together. When there are more than two groups in an analysis, general identification rules are harder to specify, especially given the multitude of possibilities with generalized user-defined restrictions on the item parameters and/or group means and variances. As in structural equation modeling, we recommend inspecting the definiteness of the Fisher information matrix to determine the identification of a particular model.

Maximum Marginal Likelihood Estimation

The principle of maximum marginal likelihood was first applied in the IRT modeling context by Bock and Lieberman (1970). The key idea is that the individual latent variables, i.e., the θ's, are nuisance parameters (in the sense of Neyman & Scott, 1948). To obtain consistent statistical inference in the presence of nuisance parameters, a well-accepted approach is to integrate the nuisance parameters out of the likelihood function. The resulting likelihood is known as the marginal likelihood. In our context, it is solely a function of the item parameters and the group means and variances. Optimizing the marginal likelihood leads to the maximum marginal likelihood estimator, which enjoys desirable statistical properties such as asymptotic normality, asymptotic unbiasedness, and asymptotic efficiency.

Given the complexity of our generalized bifactor model, we give detailed formulas and approximations related to the maximum marginal likelihood estimator in Appendix C. Interested readers can find more discussions about computation in Cai (2010c) for a model that includes the bifactor model. In this section, we illustrate key estimation issues with the example depicted in Figure 1. That is, we only consider a single group model, $n = 6$ observed items, one general dimension and $S = 2$ specific dimensions. We attempt to give a non-technical description of the role of the generalized bifactor dimension reduction method in the derivation of efficient computational methods for full-information maximum likelihood estimation.

The Numerical Integration Problem

As a first step, we note that in the example considered here, we are only estimating the free item parameters. These item parameters include the six a parameters shown in the path diagram, as well as the c, d, and a^* parameters, depending on the type of IRT model. Without loss of generality, let us assemble all the unknown parameters together and refer to them collectively as $\boldsymbol{\beta}$. Clearly, the conditional distribution of the observed item responses in Equation (10) depends on $\boldsymbol{\beta}$. The distribution of the latent factors (Equation 11) does not depend on $\boldsymbol{\beta}$ in this example, but if any of the latent variable means and variances are free, such as in a multiple-group setting, the distribution of the latent variables would also depend on $\boldsymbol{\beta}$.

In the second step, we write the marginal distribution of the item responses. From elementary probability, we first multiply the conditional distribution of **y** given the latent variables $\boldsymbol{\theta}$ with the distribution of $\boldsymbol{\theta}$ to obtain the joint distribution. Next, we integrate $\boldsymbol{\theta}$ out of the joint distribution, and what is left is the marginal of **y**. In our example with 3 latent factors, the marginal distribution is

$$f_\beta(\mathbf{y})=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_\beta(\mathbf{y}|\theta_0,\theta_1,\theta_2) f(\theta_0,\theta_1,\theta_2) d\theta_0 d\theta_1 d\theta_2, \tag{12}$$

where the subscript $\boldsymbol{\beta}$ is added to emphasize the dependence of the conditional and marginal distributions on the unknown parameters. From Equation (11), we know that $f(\theta_0, \theta_1, \theta_2)$ is the product of three univariate normal densities. From our discussions about conditional independence in the previous section, we know that the conditional density $f_{\boldsymbol{\beta}}(\mathbf{y}|\theta_0, \theta_1, \theta_2)$ must factor into a product of conditional distributions of the individual items, $f_\beta(\mathbf{y}|\theta_0,\theta_1,\theta_2)=\prod_{j=1}^{6} f_\beta(y_j|\theta_0,\theta_1,\theta_2)$ because the model assumes that the item responses are related entirely due to their joint dependence on $\boldsymbol{\theta}$. The dependence is reflected and modeled by the IRT models, as shown in Equation (10).

In general, the integral in Equation (12) must be approximated via numerical quadrature. That is, instead of deriving an analytical answer to Equation (12), we must represent it as a sum of values evaluated at discrete points. The more points are used, the more accurate the approximation becomes. A problem with numerical quadrature is its computational complexity, which increases exponentially in the order of integration. Consider the triple integral in Equation (12), quadrature approximation takes the following form

$$f_\beta(\mathbf{y}) \doteq \sum_{q_0=1}^{Q}\sum_{q_1=1}^{Q}\sum_{q_2=1}^{Q} f_\beta(\mathbf{y}|X_{q_0},X_{q_1},X_{q_2}) W_{q_0} W_{q_1} W_{q_2}, \tag{13}$$

where the integrand is evaluated at a set of Q discrete quadrature nodes (the X_q's) for each latent factor, with weights at each node give by W_q. One can use either Gauss-Hermite quadrature or equally-spaced (rectangular) quadrature. In Gauss-Hermite quadrature, pre-computed nodes and weights are widely available from applied mathematics reference sources (e.g., Abramowitz & Stegun, 1964). With rectangular quadrature, around $Q = 20$ quadrature points equally spaced between −5 and 5 are usually necessary for a practical degree of accuracy in item factor analysis. The weights in rectangular quadrature are the (normalized) ordinates of normal densities evaluated at the corresponding quadrature nodes.

The details of the numerical integration scheme is not as important as the realization that the triple sum on the right hand side of Equation (13) implies a total of Q^3 evaluations of the integrand. If $Q = 20$ points are adopted, the total number of function evaluations is $20^3 = 8,000$. Note that in general, the direct quadrature approximation requires Q^{S+1} function evaluations. For example, if there are $S = 5$ specific factors in a bifactor model, the complexity of the direct quadrature approximation is in the order of $20^6 = 64,000,000$ function evaluations. Even if we reduce the number of quadrature points per dimension by half, it still requires 10^6, or a million function calls. This is prohibitive even for today's computers. Wirth and Edwards (2007) refer to this as the "challenge of dimensionality."

Dimension Reduction

The central thesis of the full-information item bifactor method is that by a judicial rearrangement, the integral in Equation (12) can be converted into a series of iterated integrals whose dimensionality is at most two. This facilitates the numerical integration that is required in marginal likelihood computations because the number of function evaluations is at most a constant multiple of Q^2 as opposed to Q^{S+1}. The dimension reduction method tackles the challenge of dimensionality analytically.

To see this, we need to pay special attention to the conditional distribution of item responses. While all items depend on θ_0, the first two items are the only ones that depend on θ_1, and items 4 and 5 are the only ones that depend on θ_2. For the $6 - 1$ vector of item responses $\mathbf{y} = (y_1, \ldots, y_6)^t$, conditional independence implies

$$f_\beta(\mathbf{y}|\theta_0, \theta_1, \theta_2) = f_\beta(y_1, y_2|\theta_0, \theta_1) f_\beta(y_4, y_5|\theta_0, \theta_2) f_\beta(y_3, y_6|\theta_0) \tag{14}$$

Therefore, when we seek the marginal distribution of $\mathbf{y}$, we can integrate the two specific dimensions out of the joint distribution first, and then integrate θ_0 out so that the marginal distribution of $\mathbf{y}$ is expressed as the following iterated integral:

$$f_\beta(\mathbf{y}) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f_\beta(y_1, y_2|\theta_0, \theta_1) f(\theta_1) d\theta_1 \right] \times \left[\int_{-\infty}^{+\infty} f_\beta(y_4, y_5|\theta_0, \theta_2) f(\theta_2) d\theta_2 \right] f_\beta(y_3, y_6|\theta_0) f(\theta_0) d\theta_0. \tag{15}$$

As a result of integration over the specific factors, the terms in the square brackets only depend on θ_0, which is ultimately integrated out. For the iterated integral in Equation (15), we can approximate its value to arbitrary precision as

$$f_\beta(\mathbf{y}) \doteq \sum_{q_0=1}^{Q} \left[\sum_{q_1=1}^{Q} f_\beta(y_1, y_2|X_{q_0}, X_{q_1}) W_{q_1} \right] \times \left[\sum_{q_2=1}^{Q} f_\beta(y_4, y_5|X_{q_0}, X_{q_2}) W_{q_2} \right] f_\beta(y_3, y_6|X_{q_0}) W_{q_0}, \tag{16}$$

where the number of function evaluations is $2 \times Q^2$. If Q is 20, the total number of function calls is 800, which is 1/10th of the computational complexity of the direct approximation in Equation (13).

We emphasize that it is ultimately the bifactor-like structure that allows us to reexpress the multiple integral as an iterated integral and subsequently achieve dramatic reduction in numerical integration. Equations such as (15) and (16) epitomize the power of analytical dimension reduction. Appendix C provides a derivation of the marginal likelihood under the full generality of our model.

We urge methodologically sophisticated readers to consider analytical dimension reduction when confronted with high dimensional latent variable modeling problems. Cai (2010c), Rijmen et al. (2008), and Rijmen (2009) present additional applications of dimension reduction in multidimensional IRT. Cudeck, Harring, and du Toit (2009) present another example in the context of nonlinear structural equation modeling.

Estimation

Recall that $\mathbf{y}_i$ is a vector of observed item responses from respondent i. The marginal likelihood of $\boldsymbol{\beta}$ based on $\mathbf{y}_i$ is defined as a function of the unknown parameters $\boldsymbol{\beta}$, i.e., $L(\boldsymbol{\beta}|\mathbf{y}_i) = f_{\boldsymbol{\beta}}(\mathbf{y}_i)$, treating the item responses as fixed once they are observed. Because we have

assumed the independence of respondents, for the single-group example that we have considered thus far, the marginal log-likelihood is a sum over respondents $\sum_{i=1}^{N} \log L(\beta|\mathbf{y}_i)$. For more than one independent groups, the marginal log-likelihood is a sum of the group-specific contributions

$$\log L(\beta|\mathbf{Y}_1, \ldots, \mathbf{Y}_G) = \sum_{g=1}^{G} \sum_{i=1}^{N_g} \log L(\beta|\mathbf{y}_{ig}), \tag{17}$$

where $\mathbf{Y}_g$ is an $N_g \times n_g$ matrix of item responses as defined earlier. In this case, $\boldsymbol{\beta}$ contains all free parameters in the model, across all groups.

The marginal log-likelihood contains N integrals that must be approximated using numerical quadrature, as discussed in the previous section. It can then be optimized over $\boldsymbol{\beta}$ using standard numerical techniques such as an EM algorithm due to Bock and Aitkin (1981). Let $\hat{\boldsymbol{\beta}}$ be the optimizer of log $L(\boldsymbol{\beta}|\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$. Then $\hat{\boldsymbol{\beta}}$ is the maximum marginal likelihood estimate. Cai (2010c) describes the EM algorithm in full detail.

Inference

All standard likelihood-based inferential techniques (see e.g., Pawitan, 2001) would apply in this setting. When coupled with user-defined restrictions, the log-likelihood value at the maximum likelihood estimate can be used to conduct likelihood ratio tests of nested models. For instance, the relative fit of the bifactor and the unidimensional models can be compared because the unidimensional model is nested within the bifactor model. As another example, Thissen et al.'s (1993) likelihood ratio DIF test can be directly extended to the bifactor setting, wherein the significance of DIF items is inferred from the likelihood ratio between two models: a model with item parameters constrained equal across groups and a model that relaxes the constraint.

Let Model B denote the less constrained (baseline) model. Adding constraints to Model B would produce a Model A that is nested within B. Using notation from above, let log $L_A(\hat{\boldsymbol{\beta}}_A|\mathbf{Y}_1, \ldots, \mathbf{Y}_G)$ be the maximized log-likelihood of Model A, and let log $L_B(\hat{\boldsymbol{\beta}}_B|\mathbf{Y}1, \ldots, \mathbf{Y}_G)$ be the maximized log-likelihood of Model B. Then under appropriate conditions (Haberman, 1977; see also Maydeu-Olivares & Cai, 2006), minus 2 times the difference of the two log-likelihoods

$$\chi^2_{LR} = -2[\log L_A(\widehat{\beta}_A|\mathbf{Y}_1, \ldots, \mathbf{Y}_G) - \log L_B(\widehat{\beta}_B|\mathbf{Y}_1, \ldots, \mathbf{Y}_G)] \tag{18}$$

is asymptotically distributed as a central chi-square variable under the null hypothesis that the constraints that define Model A are correct. It produces a direct test of the user-defined restrictions in the same way as nested model comparisons in structural equation modeling. The degrees-of-freedom of the chi-square variable are governed by the difference in the number of free parameters of the two models.

Furthermore, the parameter estimates and their covariance matrix can be used to construct Wald chi-square test statistics much in the same way as in regression analysis or structural equation modeling. For example, the difference in factor means or variances can be statistically tested for significance. Yet another example is the extension of Langer's (2008) multiple-group Wald DIF test to bifactor analysis, where the item parameters are directly

compared across groups, turning DIF null hypotheses into linear hypotheses that are amenable to Wald tests.

Let the null hypothesis be expressed as $H_0 : \mathbf{L}^t\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\mathbf{L}$ is a fixed matrix of contrast coefficients, and $\boldsymbol{\beta}_0$ is a vector of null hypothesized constant values for the linear combination $\mathbf{L}^t\boldsymbol{\beta}$. The null hypothesis reflects a set of linear restrictions on the parameters and the alternative hypothesis is simply the lack of such restrictions. For instance, the null hypothesis may specify that two item slope parameters are equal across two groups, in which case a potential contrast coefficient matrix is $(1, -1)^t$ and the null hypothesized value is 0. This is analogous to a two-group *t*-test or a pair-wise comparison in analysis of variance. More broadly, the following statistic

$$\chi^2_W = (\mathbf{L}^t\widehat{\beta} - \beta_0)^t [\mathbf{L}^t cov(\widehat{\beta})\mathbf{L}]^{-1} (\mathbf{L}^t\widehat{\beta} - \beta_0) \tag{19}$$

is asymptotically distributed as a chi-square variable under the null, where $cov(\hat{\boldsymbol{\beta}})$ denotes the covariance matrix of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ under the alternative hypothesis. The degrees-of-freedom are governed by the number of linear restrictions in the null hypothesis, or equivalently the rank of the contrast matrix $\mathbf{L}$. Unlike the likelihood ratio test, the success of the Wald test depends critically upon the availability of accurate covariance matrix for the parameter estimates. As in Langer (2008), we compute this covariance matrix using a Supplemented EM algorithm (Meng & Rubin, 1991). For the interested reader, Cai (2008) and Cai and Lee (2009) present additional details of the Supplemented EM algorithm in the contexts of IRT and structural equation modeling.

Finally, the asymptotic normality of the maximum likelihood estimator implies that we can construct confidence intervals for the parameter estimates, following standard recipes. For instance, the limits of a symmetric 95% univariate confidence interval could be $\hat{\beta} \pm 1.96 \times se(\hat{\beta})$, where $se(\hat{\beta})$ is the standard error estimate.

Simulations

We conduct two sets of simulations, as an empirical check of the accuracy of our proposed estimation methods and as a demonstration of some of the capabilities of our extended bifactor model. The item response data are generated with a built-in simulator in the numeric engine of IRTPRO (Cai, du Toit, & Thissen, forthcoming). The fitted models and the data generating models are the same, so the results represent parameter recovery. We use the Bock and Aitkin (1981) EM algorithm in conjunction with the bifactor dimension reduction method as implemented in the numeric engine of IRTPRO to obtain maximum likelihood estimates. For the EM algorithm, the M-step convergence criterion is 1.0×10^{-7}. The E-step cycles are terminated when the maximum absolute difference in parameter estimates drops below 1.0×10^{-3} between adjacent E-steps. The maximum number of E-steps is set at 500, and a solution is declared nonconvergent if the maximum is reached. As to quadrature, we use $Q = 21$ quadrature points equally-spaced between −5 and +5 to approximate the integrals in the marginal log-likelihood. With dimension reduction, the number of function evaluations is a constant multiple of $Q^2 = 441$ for each integral in the marginal log-likelihood. The computer we used is a workstation running Windows XP, with duo 3GHz Xeon CPUs and 3GB RAM.

Simulation One: A Two-Group Model

We generated data from an extended bifactor model in two groups ($N_1 = N_2 = 1,000$). As shown in Figure 8, group 1 has $n_1 = 16$ dichotomously scored items all fitted with the graded

response model for two categories. The specific factors are defined by $S_1 = 4$ item clusters (items 1–4, 5–8, 9–12, 13–16). In group 1, the latent variables are assumed to have zero means and unit standard deviations. In group 2, one of the item clusters (items 13–16) is not present, resulting in $n_2 = 12$ observed items and $S_2 = 3$ specific factors. The factor means and variances are freely estimated in group 2. The item parameters (intercepts and slopes) of the first 12 items are constrained to be equal across groups. Thus, the data generating model reflects measurement invariance across the two groups. The generating parameter values are listed in Table 1. These parameters are chosen so that we examine parameter recovery for a variety of plausible intercept/slope combinations.

The simulated example also shows the flexibility of our extended modeling framework. In a realistic setting, perhaps the unadministered items (items 13–16) are inapplicable to the respondents in group 2. For instance, the two groups may be defined by two countries, and the measurement instrument may contain culturally-sensitive items that are not relevant in one country. In an educational context, the groups may be students from two adjacent grade levels, and the assessment contains 12 overlapping items for vertically linking the two grades so that a developmental scale can be constructed. The non-overlapping items (items 13–16) may be inappropriate for the lower grade. As in unidimensional multiple-group IRT, the key to flexible multiple-group bifactor analysis lies in the ability to estimate factor means and variances, and the ability to impose arbitrary within/cross group restrictions. Our model has achieved that without sacrificing computational efficiency.

We ran $M = 500$ replications of the Monte Carlo simulation. In each replication, the two-group model is fitted back to the simulated data and the parameter estimates, standard errors, and computer process time (measured as CPU seconds) are recorded. For this model, all 500 replications converged. The average per-replication time is 28 CPU seconds in 113 E-cycles. [4]

We examine both bias and variability of the estimates. For a generic parameter β, let $\hat{\beta}_m$ be its estimate in the mth Monte Carlo replication. Then $\bar{\beta}=M^{-1}\sum_{m=1}^{M}\hat{\beta}_m$ is the Monte Carlo average of the point estimates. The estimated bias is defined as $\bar{\beta}-\beta$. As can be seen from Table 1, the estimated biases for all parameters are quite small (maximum absolute bias is .06), indicating good parameter recovery.

Table 2 compares the Monte Carlo averages of the estimated standard errors with the Monte Carlo standard deviations of the estimated parameters. Specifically, for a generic parameter β, let the estimated standard error from replication m be $se(\hat{\beta}_m)$. The Monte Carlo average of the estimated standard error is $\overline{se}(\beta)=M^{-1}\sum_{m=1}^{M}se(\hat{\beta}_m)$, and the Monte Carlo standard deviation is defined as $sd(\beta)=\sqrt{(M-1)^{-1}\sum_{m=1}^{M}(\hat{\beta}_m-\bar{\beta})^2}$. If the standard errors are estimated accurately, their average should agree with the Monte Carlo standard deviations of the parameter estimates. We can see that this is indeed the case (maximum discrepancy is .05).

[4]The efficiency of the bifactor dimension reduction is evident. As a comparison, for a randomly selected replication that converged in 26 seconds and 103 cycles, we also used the adaptive quadrature based EM algorithm (Schilling & Bock, 2005) implemented in IRTPRO to estimate the model. The number of adaptive Gauss-Hermite quadrature points per dimension is 5. The adaptive quadrature based EM algorithm does not implement the bifactor dimension reduction. It required 1,715 seconds in 103 E-cycles to converge to the same place as the dimension-reduced Bock and Aitkin (1981) EM algorithm for bifactor analysis.

Simulation Two: A Model for Complex Assessment

It is not uncommon in educational measurement to encounter scales or assessments made up of items with complex response formats. Thissen, Wainer, and Wang (1994) describe educational tests consisting of a multiple-choice (MC) section and a constructed response (CR) section. While both sections are intended to measure the same construct (e.g., general reading achievement), the underlying factor structure may be multidimensional. A bifactor pattern is one of the possibilities. The MC section may be composed of several clusters of items, e.g., in a passage-based reading assessment, and the CR section may exert some additional influence above and beyond the general dimension due to method effects.

In recent large-scale educational surveys, the so-called Complex Multiple-Choice (CMC) format has become popular. Take the Program for International Student Assessment (PISA) as a case in point. A typical CMC item in PISA consists of a set of tightly coupled multiple-choice questions (see e.g., Organization for Economic Co-operation and Development, 2009). In other words, a CMC item is a mini testlet, which in turn can be nested within a larger testlet potentially consisting of a number of other MC, CMC, or CR items. PISA scores a CMC item by counting the number of correct responses, and then assigns partial credit according to a pre-specified scoring rubric, akin to Thissen et al.'s (1989) nominal model approach for testlets.

This similarity brings up the potential application of our bifactor nominal response model to CMC items. Consider the simplest CMC item, which contains 2 multiple-choice questions. If we use 0 to denote the incorrect response and 1 the correct response, the pattern of responses to the two questions must be one of the following four possibilities: 00, 01, 10, and 11. Instead of collapsing the pattern of responses into a number-correct score, we can directly treat the four possibilities as a single nominal item with four potentially unordered categories. They are unordered because though the 00 pattern should in general indicate less ability or achievement of the respondent than the 11 pattern, it is unclear *a priori* which one of the 01 or 10 response patterns reflects "higher" ability/achievement. The bifactor nominal model can 1) suggest the empirical order of the categories, 2) use all available item response information, and 3) permit the item responses be influenced by more than one source of individual difference, both due to the general factor and the testlet-specific factor on which the CMC item loads.

Figure 9 shows a hypothetical test with 9 MC items, 1 CMC item, and 5 CR items. The MC items have 4 response alternatives and there is potentially a non-zero guessing probability for each item. The MC items are scored right/wrong and the dichotomous model is appropriate for these items. The CMC item consists of 2 questions and as we described earlier, we use a 4-category nominal model for this CMC item. We code the categories in the lexicographical order: 00 as category 0, 01 as category 1, 10 as category 2, and 11 as category 3. The CR items are scored in 4 ordinal categories. A popular model for these CR items is the GPC model, although as we mentioned earlier, the graded model is suitable as well. There are three factors in the model. In the MC and CMC section of the test, the first five items form into a cluster. For instance, they are potentially all based on the same reading passage or the same graph. Hence, they load on a specific factor to account for the residual influence above and beyond the general factor. The entire CR section defines another specific factor. The factors are assumed to be normally distributed with zero means and unit variances.

Table 3 shows the generating parameter values for this model. In the generating model, the CR items 11–15 are equally discriminating with respect to θ_2. We simulated data from this bifactor model and the sample size is $N = 3,000$. For each simulated data set, we fitted two models. The first model imposes equality constraints on the slopes of items 11–15 on θ_2.

The second model does not impose the equality constraints. Therefore, the first model is nested within the second model, and the likelihood ratio comparison of these two models provides a test of the equality restrictions. In this case, the likelihood ratio test statistic should be distributed as a central chi-square variable with 4 degrees-of-freedom because the 4 equality constraints are indeed correct in the data generating process. For the parameter estimates and standard errors, we report results from the second model in Tables 3 and 4.

We ran 500 replications and all replications converged. To help stabilize estimation for the MC items, we imposed the normal stochastic constraint (prior) outlined in Appendix A on the lower asymptote parameters with mean equal to −1.1 and standard deviation equal to .50. For each replication, we recorded the parameter estimates, the standard errors, the CPU time, and the maximized log-likelihood values for the two fitted models. The average per-replication runtime is 13 CPU seconds in 48 E-cycles.

Figure 10 plots the empirical quantiles of the likelihood ratio test statistic against the theoretical distribution – central chi-square with 4 degrees-of-freedom. As one can tell, the points are very close to the 45-degree reference line, showing a good distributional fit. The mean of the empirical distribution is 4.07 (the expected value should be 4) and the variance is 7.50 (the expected variance should be 8). Overall, the chi-square approximation seems to have worked well.

Table 3 presents the estimated bias of all parameters. We mention that though we actually estimate the logit of the lower asymptote parameter (as shown in Appendix A), and reparameterize the GPC/nominal intercept and nominal scoring function parameters into estimable contrasts (as shown in Appendix B), we present the item parameters in their natural form in Table 3. In other words, the Appendices are prerequisites to a full understanding of the results in Table 3.

As one can see, the item parameters for the CR items are recovered well. The slopes of the GPC model show very little bias. The nominal model has also successfully recovered the parameters for the CMC item. In particular, the true scoring function value for category 1 is $a_1^*=2.49$, which is higher than $a_2^*=2.10$, indicating that the order of the two middle categories corresponding to the 01 and 10 responses should in fact be reversed. Though the two categories share the same number-correct score, a correct response to the second one of the CMC questions indicates a higher ability/achievement than a correct response to the first.

Due to the presence of the stochastic constraint on the logit of the asymptote parameter, the estimates of guessing are all centered on the true values. However, the estimates for the slopes and intercepts of the MC items exhibit slightly larger biases than the CMC and CR items. The largest MC slope bias is .09 for item 10, but on a relative note, it is only off by 6.3 percent of the true slope (1.44). Our results are consistent with the well-known fact in unidimensional IRT that the 3-parameter model requires significantly larger N than the other item response models to achieve stable estimation.

Turning to the variability information presented in Table 4, we see that the estimated standard errors of the CR and CMC items are all closely aligned with the Monte Carlo standard deviations. Note that as we mentioned earlier, these are standard errors for the parameters that are actually estimated, i.e., the logits and the contrasts. On the other hand, while the standard errors for the MC item parameters are generally close to the Monte Carlo standard deviations, there is an interesting relationship between the difficulty/location of the item and the standard errors. Item 1 has the highest intercept, and hence it is the easiest item. With few people answering an easy item incorrectly, the data contain little information about guessing. Consequently, the average of the standard errors (.31) is not appreciably

smaller than the prior standard deviation (.50) for the logit of guessing. Note that the MC items are arranged in an descending order by item intercept (see Table 3). Hence, the items become more difficult as we move down the rows of Table 4, and the data become increasingly more informative of the guessing strategy, which results in decreased standard errors that are in closer agreement with the Monte Carlo standard deviations.

Illustrative Real Data Analysis

We analyze item responses to 15 math questions from the 2000 Program for International Student Assessment (PISA) conducted by the Organization for Economic Co-operation and Development (OECD).5 We selected students who have attempted all 15 items on Booklet 1 from 5 developed countries (sample size in parentheses): Australia (464), Germany (409), Japan (445), UK (889), US (358). There are 14 CR items and one MC item with 4 response alternatives. For the CR items, 10 are scored in 2 categories, 3 in 3 categories, and 1 in 4 categories. For the MC item, we fitted the model in Equation (3) for the dichotomously scored response. The estimation of the logit of guessing probability is stabilized by the normal stochastic constraint (prior) outlined in Appendix A. The mean for the prior is −1.1 and the prior standard deviation is increased to 1.0 to weaken the potential impact of the prior on the guessing parameter as we do not know its true value. For the CR items, we fitted the graded response model, although the GPC model can certainly be used instead.

The items belong to the following content clusters (number of items in the cluster appears in the parentheses): Cube Painting (4), Farms (2), Walking (2), Apples (3), Continent Area (1), Growing Up (3). With the exception of Continent Area, which is a singleton, each cluster defines a separate specific factor in addition to the general math literacy factor on which all items load. These specific factors reflect the residualized individual differences that are specific to the content cluster above and beyond the individual differences in general math literacy. As discussed earlier, because the specific factors for Farms and Walking only have two items each, we constrain the slopes to be equal within each factor for identification, making these two factors the equivalent of correlated residuals. Figure 11 shows the path diagram for the bifactor structure for one of the 5 countries. We believe this factor structure is reasonably appropriate for these data (see e.g., DeMars, 2006).

Invariant Items to Explore Group Differences

In this section, we assume measurement invariance. That is, we are assuming that each item functions in the same way across countries. In other words, there is no DIF. For our multiple-group bifactor model, or more generally for all parametric IRT models, the invariance assumption translates into restrictions on the item parameters. The parameters for each item are set equal across groups to indicate that their measurement properties are the same in all countries. While the PISA test developers may have conducted extensive measurement invariance analysis (Adams & Wu, 2002) to ensure the quality of the instruments, we emphasize that the absence of DIF is an assumption that should and can be checked empirically. We will relax the invariance restrictions in the second model and discuss ways to model potential DIF.

With the equality restrictions on the item parameters, we can freely estimate the group means and variances of the general factor to make cross-country comparisons of math literacy. Optionally, we allow some or all of the means/variances of the specific factors to be freely estimated. We arbitrary choose the US children as the reference group and standardize the latent factors in this group to set the location and scale.

[5]General introductions to PISA and public-use data sets can be found at http://www.pisa.oecd.org/.

As an illustration, we estimate the mean and variance of the Cube Painting factor because it has the most items of all specific factors. After all, in order for the factor means and variances to be interpretable substantively, the underlying factor should correspond to a broad enough construct, even though technically a specific factor is identified with as few as two items loading on it. Though this specific factor may not be broad enough, we note that in other settings such as quality of life research (see e.g., Gibbons et al., 2007), the specific factors often represent meaningful content subdomains. We estimate the mean of Cube Painting to illustrate what an investigator *can* do.

Table 5 presents the bifactor slope estimates and standard errors for the PISA math items. With the exception of the items for the Growing Up cluster, all slopes are strong and statistically significant. It is particularly important to note the magnitude of the specific factor slopes, as they indicate the presence of additional underlying dimensions.

Adding these specific factors has significantly improved model-fit. The unidimensional multiple-group model is nested within our extended bifactor model. Using the notation of the likelihood ratio test developed earlier (see Equation 18), the unidimensional model is Model A and the bifactor model is model B. One can obtain the unidimensional model by fixing all the specific factor slopes to 0 and not estimating the Cube Painting factor means and variances. The bifactor model has 64 free parameters, whereas the unidimensional model has 44 parameters. The likelihood ratio chi-square is $\chi^2_{LR}=570$, which is highly significant on 20 degrees-of-freedom.

Table 6 presents information about the response format, the various IRT models fitted, as well as estimates of guessing and/or intercept parameters. Note that the logit of guessing is estimated to be −2.31, which translates into a guessing probability of .09. In contrast, when we fitted a unidimensional IRT model to the data, the guessing probability is estimated to be .15. Thus, the guessing parameter *estimate* is dependent on dimensionality, and directly estimating the parameter is preferable to the TESTFACT-recommended strategy of setting guessing at a value obtained from a prior unidimensional model run.

Table 7 presents the estimated factor means and variances for each country, with the US as the reference. It can be seen that Japan has the highest mean and the smallest variance in the general dimension. Japanese students are .75 standard deviations higher than the American students in math literacy. Their achievement levels are also much more homogeneous (variance estimate of .58) than the US students (assumed variance of 1.00). Interestingly, the Japanese students as a group are not the highest achievers when we look at the context-specific skills for solving the Cube Painting exercise. In this case, the Australian students have the highest mean. Finally, it is worth noting the efficiency of the dimension reduction method because fitting this 6-dimensional model to data from all 5 countries simultaneously took only 2 minutes of CPU time on the workstation we used for simulations.

Modeling and Testing for DIF

In this section, we adapt the IRT likelihood ratio DIF (IRTLRDIF) procedure due to Thissen et al. (1993) and the updated Wald DIF test described by Langer (2008) to the case of bifactor modeling. The strategy we adopt combines the linking method in IRTLRDIF with the flexibility of the Wald test to model DIF across several groups simultaneously. To our knowledge, Jeon and Rijmen (2010) is the only reference that also adopts an IRT-based method for DIF detection in a bifactor model. Our unique contribution here is the completeness of the approach.

We would like to illustrate our proposed procedure by testing the two constructed response items in the Farms cluster for DIF. The responses to these two items are dichotomous rater

scores, and the graded response model was used. Within a group, there is one intercept, one slope on the general dimension, and one slope on the specific dimension for each item. However, the slopes on the specific dimension must be set equal across the two items within each group for identification.

As in IRTLRDIF, the first step is to link the latent variable scales. In the model assuming no DIF, all item parameters are set equal across groups. To study DIF, the cross-group equality restrictions for the Farms items must be dropped. The remaining 13 items in the other content clusters serve as the anchoring set that links the latent variable scales together. Specifically, the item intercepts, and slopes for the general factor are freely estimated in each group for the Farms items. On the other hand, the item slopes on the factor specific to the Farms cluster are set equal both within and across groups while at the same time, we freely estimate the variances for this specific factor in all groups but the US, which is the reference group and has an assumed factor variance of 1.0. Thus, though there appears to be only one slope parameter pertaining to residual dependence, we estimate 4 additional factor variances. In total, 20 new parameters are added: 8 intercepts, 8 slopes, and 4 specific factor variances. We could have chosen to set the Farms factor variances to 1 and freely estimate the slopes on the Farms factor for each group – which would lead to an equivalent model. The parameterization we adopt not only uses a unique feature of our generalized modeling framework, i.e., the ability to estimate factor variances, but also has the advantage of effectively separating out the DIF parameters that are substantively important (the intercepts and general factor slopes) from those that are nuisance (parameters for residual dependence).

The item parameters that are specific to the Farms cluster are shown in the top portions of Tables 8 and 9. Even a cursory glance reveals the fact that the intercepts and slopes are different across the countries, indicating possibly significant DIF. For instance, the highest intercept for Farms 1 is 1.06, obtained in Japan, and the lowest intercept is −.24, from the US children. Thus the item is much more difficult for the US children than their Japanese counterparts. The highest slope estimate for this item is in the UK group (2.51), which is almost twice as large as the lowest slope of 1.31 in the Germany group. Thus the discrimination of the items is also different. Table 10 presents the estimated factor means and variances. Note that the variances for the Farms factor are freely estimated relative to the US reference variance of 1.0. While the trend in the general factor means and variances remains as before, with Japan having the highest mean (.66 standard deviations above the US average) as well as the most homogeneity, the absolute magnitude of the difference between Japan and the other countries lessened somewhat in comparison to results in Table 7.

Once the latent variable scales are set, we are ready to conduct DIF tests for the items in question. For this, we employ the Wald test described earlier. For each item, there are 10 item parameters that are relevant to the Wald test, 5 general dimension slopes (one per group) and 5 intercepts (also one per group). Let us denote these item parameters as

$$\beta=\left(a_1 \quad a_2 \quad a_3 \quad a_4 \quad a_5 \quad d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5 \right),$$

where the subscripts indicate the group number. The parameters for the other items are invariant across groups and hence can be dropped from further consideration without loss of generality. As a concrete example, for Farms 1, the estimated parameters are

$$\widehat{\beta}=\begin{pmatrix} 2.11 & 1.31 & 1.38 & 2.51 & 1.90 & .76 & .03 & 1.06 & .21 & -.24 \end{pmatrix}.$$

To conduct the Wald DIF test involving multiple groups, we must choose a set of contrast coefficients that describe the group comparisons that we wish to make. In principle, any standard analysis of variance contrasts (e.g., deviation, polynomial) can be used. Here we use the so-called Helmert contrasts. Let the contrast coefficient matrix be

$$\mathbf{L}=\left(\begin{array}{cccccccc} -1 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\ -1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 0 & -1 & 0 & 2 & 0 & 0 & 0 \\ -1 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & -1 & 0 & 2 & 0 & 0 \\ 0 & -1 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right). \quad (20)$$

These contrasts have the desirable property of being mutually orthogonal, i.e., $\mathbf{L}^t\mathbf{L}$ is a diagonal matrix. The vertical bars separate the contrasts into 4 orthogonal sets. The first set contrasts the US against the average of Australia, Germany, Japan, and UK. The second set contrasts UK against the average of Australia, Germany, and Japan. The third one contrasts Japan against the average of Australia and Germany. The final set contrasts Australia and Germany. Furthermore, within each set, there are two orthogonal contrasts, one for the slope parameters and one for the intercept parameters.

The null hypothesis is of the form: $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. In other words, the null specifies that there is no DIF, so there is no difference in the item parameters across countries, and all contrasts must be zero. To test this null, we use $\chi^2_w=(\mathbf{L}^t\widehat{\beta})^t(\mathbf{L}^t cov(\widehat{\beta})\mathbf{L})^{-1}(\mathbf{L}^t\widehat{\beta})$ as the test statistic, which is the same as Equation (19) by setting $\boldsymbol{\beta}_0 = \mathbf{0}$. The degrees-of-freedom are determined by the rank of $\mathbf{L}$, which is 8 in this case. For Farms 1, $\chi^2_w=57.96$, which is highly significant on 8 degrees-of-freedom. Similarly, for Farms 2, $\chi^2_w=63.95$, again highly significant. Thus, we must reject the null hypothesis of no DIF for both items under the Farms cluster.

As in the IRTLRDIF procedure, a significant overall Wald DIF test is followed by a set of targeted DIF tests to locate the source of DIF (Langer, 2008). In other words, we would like to know whether the significant overall chi-square is caused by differences in item slopes (nonuniform DIF) or in the intercepts (uniform DIF). They resemble the step-down tests in standard analysis of variance. Given the orthogonality of our contrast coefficients, conducting step-down tests are straightforward. Table 11 presents the results of these step-down tests. In each case, the test statistic remains the Wald chi-square, with differences in the contrast coefficients. As noted above, the $\mathbf{L}$ matrix as shown in Equation (20) contains four distinct sets of orthogonal contrasts. Correspondingly, each item has 4 rows of results in Table 11.

Take the first row for Farms 1 as an example. The Total column presents a Wald chi-square for a step-down test of this particular 2 degrees-of-freedom contrast for the item slopes and intercepts. Effectively, this test is using the first two columns of $\mathbf{L}$ in Equation (20) as the

contrast coefficients. The chi-square is 19.45, which shows that the US item slope and intercept are jointly significantly different from the average slope and intercept of the other countries. This test can be further decomposed into two single degree-of-freedom tests. The *a* column reports the chi-square for slope differences alone. One can obtain this test by using the first column of **L** in Equation (20) as the contrast coefficients. There is no statistically significant slope difference between the US and the average of the other countries. The third chi-square is denoted as $d|a$, which is simply the total chi-square minus the *a* chi-square: 19.45 − .06 = 19.39. This is a direct test of intercept differences, conditional on the test of slope differences. In this case, the $d|a$ chi-square is highly significant, indicating that most of the DIF between the US and the others is due to a difference in the intercept (or difficulty) of the item. The second comparison (UK vs. the preceding countries) shows a different picture. The slope contrast results in an *a* chi-square of 9.63, which is significant. Given the significant difference in slopes, it is less clear how or whether the intercept comparisons should be interpreted. We refer the reader to the advice of Thissen et al. (1993) on the order in which specific DIF tests should be conducted (p. 88). The other rows of Table 11 can be interpreted analogously.

The DIF tests can be repeated for each item cluster, essentially "sweeping" over the item clusters. Overall and step-down tests can then be conducted for each item. This is the bifactor equivalent of the sweep procedure as implemented in the software program IRTLRDIF (Thissen, 2001), but further generalized to more than one groups. The software program IRTPRO (Cai et al., forthcoming) has implemented the Wald DIF tests as described here. We note that these tests are rather new. Simulations under the unidimensional model (Langer, 2008) show that the Wald test have well-calibrated Type I error rates and have adequate power. However, more research on their performance in the multidimensional IRT context is needed. Furthermore, we have only provided an outline of a bifactor DIF testing procedure and have not had room to discuss either the details or the ramifications of the results. We encourage future work on DIF for multidimensional IRT.

Discussion

In this research we propose a comprehensive item bifactor analysis framework with extended support for a variety of item response models. The framework can handle item responses from multiple groups, with dichotomous, ordinal, and nominal response formats. Motivated by Rijmen et al.'s (2008) results, an efficient full-information maximum marginal likelihood estimator is defined, and we show how one can extend the Gibbons and Hedeker (1992) dimension reduction technique to a much more generalized model. Arbitrary user-defined restrictions on model parameters are permitted and directly testable using standard likelihood based inferential statistics.

The simulation studies demonstrate that one can indeed reliably fit the generalized bifactor model to data. We have implemented the estimation procedures in IRTPRO (Cai et al., forthcoming). The analysis can be conducted via a mouse-driven graphical user interface. The simulations also show that the bifactor dimension reduction method results in substantial time-savings. The empirical application of the model to cross-country comparisons using real data from an OCED PISA survey serves to highlight the flexibility of the model in exploring group differences.

A generalized modeling framework opens up many opportunities previously unanticipated. Our model directly supports full-information bifactor-based multidimensional differential item functioning analyses, for more than two groups. Our model also supports bifactor-based linking/equating studies. Though Gibbons et al. (2007) only mention the computation of scale scores for the general factor, our model allows the computation of scale scores for

all the factors in the model as posterior expected values (see Cai, 2010c for detailed scoring formulas). Gibbons et al. (2008) discuss the utility of the domain-specific scores in mental health assessments.

The proposed model can be applied to explore the dimensionality of psychological and educational measurement instruments. The extension to polytomous, particularly nominal, IRT models significantly improves the generality of item bifactor analysis. The inclusion of latent variable means and variances in conjunction with the ability to impose and test user-defined constraints not only enhances the flexibility of multidimensional item response theory modeling but also offers a likelihood-based alternative to popular Bayesian item factor analytic methods (e.g., Wainer et al., 2007). The full-information estimator seamlessly handles a variety of response formats as well as data that are missing at random, providing some advantages over existing limited-information categorical factor analysis methodology. In sum, we believe that our proposed modeling framework has wide-ranging implications for data analysis and measurement modeling in psychology, education, public health, and medical research.

We conclude by noting some important limitations of our research. First, the paper is exclusively written from the perspective of IRT. Given the well-known relationship between the normal ogive IRT model and the categorical factor analysis model (Takane & de Leeuw, 1987), a large class of dichotomous and polytomous IRT models can be effectively estimated with limited-information methods using standard structural equation modeling software. In particular, a multiple-group item bifactor analysis with the normal ogive graded response model can be handled by virtually all standard software on the market today. Though limited-information estimation methods may have certain deficiencies, as mentioned in the beginning of this paper, their computational speed has made them far more attractive to data analysts. On the other hand, we wish to point out that by applying dimension reduction, full-information estimation can also be highly computationally efficient. Hence the choice between limited-information and full-information estimation should be based on the nature of the problem. Though it is our belief that full-information estimation methods are more flexible, the two approaches may be complementary.

Second, for the ease of exposition, we have restricted ourselves to normal latent variables. While the normality assumption is usually appropriate in educational measurement, it may not be for psychological assessment, particularly for clinical populations. In the context of unidimensional IRT, semi-parametric approaches to handling non-normal latent densities have been proposed (see e.g. Woods & Thissen, 2006), and in principle, can be applied to item bifactor models. This is a topic that we will pursue in future research.

Third, the other assumption we made is the orthogonality of the latent variables. This assumption is much more critical for bifactor analysis. It ensures model identification (see e.g. Rijmen, 2009) and is used to derive dimension reduction in maximum likelihood computations. However, as Rijmen (2009) noted, the assumption can be relaxed (subject to model identification) to the case of independent specific factors conditional on the general factor. In addition, recent development of exploratory and confirmatory item factor analytic methods (see, e.g., Cai, 2010a, 2010b) also provide computationally efficient alternatives under more general conditions.

Finally, we would like to emphasize the issue of model fit diagnosis. The purpose of this research is to develop a new modeling framework and we had little room for comprehensive studies of model fit. Nevertheless, it is a singularly important problem that should be addressed in any analysis of real data. Recent years have seen active research on goodness-of-fit tests for IRT models (e.g. Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-

Olivares & Joe, 2005; Orlando & Thissen, 2000), but the proposed indices have generally only been applied to unidimensional models. Future research in the area of multidimensional models would be useful.

Acknowledgments

We acknowledge financial support from the following sources: Institute of Education Sciences (R305B080016 and R305D100039), and National Institute on Drug Abuse (R01DA026943 and R01DA030466). The development of IRTPRO was supported by the National Cancer Institute in the form of a Small Business Innovative Research contract (#HHSN-2612007-00013C) awarded to Scientific Software International. The views expressed in this paper do not reflect the views and policies of the funding agencies.

We thank Kathleen Preston for helping with data analysis and simulations and are indebted to Dr. David Thissen for comments on an earlier draft. We also thank the associate editor and the reviewers for helpful suggestions. Any remaining faults are our own.

References

Abramowitz, M.; Stegun, IA. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York, NY: Dover; 1964.

Adams, R.; Wu, M. PISA 2000 technical report. Paris, France: Organization for Economic Cooperation and Development; 2002.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43:561–573.

Bock RD. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 1972; 37:29–51.

Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. 1981; 46:443–459.

Bock, RD.; Gibbons, R.; Schilling, SG.; Muraki, E.; Wilson, DT.; Wood, R. TESTFACT 4 user's guide. Chicago, IL: Scientific Software International, Inc; 2003.

Bock RD, Lieberman M. Fitting a response model for *n*dichotomously scored items. Psychometrika. 1970; 35:179–197.

Bock, RD.; Zimowski, MF. Multiple group IRT. In: van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. New York, NY: Springer-Verlag; 1997. p. 433-448.

Braeken J, Tuerlinckx F, De Boeck P. Copula functions for residual dependency. Psychometrika. 2007; 72:393–411.

Cai L. SEM of another flavour: Two new applications of the supplemented EM algorithm. British Journal of Mathematical and Statistical Psychology. 2008; 61:309–329. [PubMed: 17971266]

Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. Psychometrika. 2010a; 75:33–57.

Cai L. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics. 2010b; 35:307–335.

Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika. 2010c; 75:581–612.

Cai, L.; du Toit, SHC.; Thissen, D. IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Chicago, IL: Scientific Software International, Inc; forthcoming

Cai L, Lee T. Covariance structure model fit testing under missing data: An application of the supplemented EM algorithm. Multivariate Behavioral Research. 2009; 44:281–304.

Cai L, Maydeu-Olivares A, Coffman DL, Thissen D. Limited-information goodness-of-fit testing of item response theory models for sparse 2^ptables. British Journal of Mathematical and Statistical Psychology. 2006; 59:173–194. [PubMed: 16709285]

Chen WH, Thissen D. Local dependence indices for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997; 22:265–289.

Cudeck R, Harring JR, du Toit SHC. Marginal maximum likelihood estimation of a latent variable model with interaction. Journal of Educational and Behavioral Statistics. 2009; 34:131–144.

DeMars CE. Application of the bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement. 2006; 43:145–168.

DeMars CE. "Guessing" parameter estimates for multidimensional item response theory models. Educational and Psychological Measurement. 2007; 67:433–446.

Edwards, MC.; Edelen, MO. Special topics in item response theory. In: Millsap, R.; Maydeu-Olivares, A., editors. Handbook of quantitative methods in psychology. New York, NY: Sage; 2009. p. 178-198.

Fraser C, McDonald RP. NOHARM: Least squares item factor analysis. Multivariate Behavioral Research. 1988; 23:267–269.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian data analysis. 2. Boca Raton, FL: Chapman & Hall/CRC; 2004.

Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, et al. Full-information item bifactor analysis of graded response data. Applied Psychological Measurement. 2007; 31:4–19.

Gibbons RD, Grochocinski VJ, Weiss DJ, Bhaumik DK, Kupfer DJ, Stover A, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services. 2008; 59:361–368. [PubMed: 18378832]

Gibbons RD, Hedeker D. Full-information item bifactor analysis. Psychometrika. 1992; 57:423–436.

Gibbons, RD.; Hedeker, D. BIFACTOR [Computer software]. Chicago, IL: Center for Health Statistics, University of Illinois at Chicago; 2007.

Haberman SJ. Log-linear models and frequency tables with small expected cell counts. The Annals of Statistics. 1977; 5:1148–1169.

Holzinger KJ, Swineford F. The bi-factor method. Psychometrika. 1937; 2:41–54.

Jeon, M.; Rijmen, F. Assessing differential item functioning for testlet-based tests using the bifactor model. Paper presented at the annual meeting of the National Council on Measurement in Education; Denver, CO. 2010.

Johnson TR, Bolt DM. On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. Journal of Educational And Behavioral Statistics. 2010; 35:92–114.

Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika. 1969; 34:183–202.

Jöreskog, KG.; Sörbom, D. LISREL user's guide. Chicago, IL: Scientific Software International, Inc; 2001.

Kolen, MJ.; Brennan, RL. Test equating, scaling, and linking: Methods and practices. 2. New York, NY: Springer; 2004.

Langer, MM. Unpublished doctoral dissertation. Department of Psychology, University of North Carolina; Chapel Hill: 2008. A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation.

Lord, FM.; Novick, MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982; 47:149–174.

Maydeu-Olivares A, Cai L. A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. Multivariate Behavioral Research. 2006; 41:55–64.

Maydeu-Olivares A, Joe H. Limited and full information estimation and testing in 2^ncontingency tables: A unified framework. Journal of the American Statistical Association. 2005; 100:1009–1020.

Meng XL, Rubin DB. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association. 1991; 86:899–909.

Muraki E. A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement. 1992; 16:159–176.

Muraki E, Carlson JE. Full-information factor analysis for polytomous item responses. Applied Psychological Measurement. 1995; 19:73–90.

Muthén; Muthén. Mplus (Version 5.0) [Computer software]. Los Angeles, CA: Author; 2008.

Neyman J, Scott EL. Consistent estimates based on partially consistent observations. Econometrica. 1948; 16:1–32.

Organization for Economic Co-operation and Development.. Take the test: Sample questions from OECD's PISA assessments. Paris, France: Author; 2009.

Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24:50–64.

Pawitan, Y. In all likelihood: statistical modelling and inference using likelihood. Oxford, UK: Oxford University Press; 2001.

Reckase MD. The past and future of multidimensional item response theory. Applied Psychological Measurement. 1997; 21:25–36.

Reckase, MD. Multidimensional item response theory. New York, NY: Springer; 2009.

Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research. 2007; 16:19–31. [PubMed: 17479357]

Reise SP, Waller NG. How many IRT parameters does it take to model psychopathology items? Psychological Methods. 2003; 8:164–184. [PubMed: 12924813]

Revuelta J. The generalized logit-linear item response model for binary-designed items. Psychometrika. 2007; 73:385–405.

Rijmen, F. Educational Testing Service. 2009. Efficient full information maximum likelihood estimation for multidimensional IRT models(Tech. Rep. No. RR-09-03).

Rijmen F, Vansteelandt K, De Boeck P. Latent class models for diary method data: Parameter estimation by local computations. Psychometrika. 2008; 73:167–182. [PubMed: 20046853]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometric Monographs. 1969; 17

Schilling S, Bock RD. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. Psychometrika. 2005; 70:533–555.

Simms LJ, Grös DF, Watson D, O'Hara MW. Parsing the general and specific components of depression and anxiety with bifactor modeling. Depression And Anxiety. 2008; 25:E34–E46. [PubMed: 18027844]

Stout W. A new item response theory modeling approach with application to unidimensional assessment and ability estimation. Psychometrika. 1990; 55:293–325.

Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika. 1987; 52:393–408.

te Marvelde JM, Glas CAW, Van Landeghem G, Van Damme J. Application of multidimensional item response theory models to longitudinal data. Educational and Psychological Measurement. 2006; 66:5–34.

Thissen, D. IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. 2001. Retrieved from http://www.unc.edu/dthissen/dl.html

Thissen, D. MULTILOG 7 user's guide. Chicago, IL: Scientific Software International, Inc; 2003.

Thissen, D.; Cai, L.; Bock, RD. The nominal categories item response model. In: Nering, M.; Ostini, R., editors. Handbook of polytomous item response theory models: Developments and applications. New York, NY: Taylor & Francis; 2010. p. 43-75.

Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Quality of Life Research. 2007; 16:109–116. [PubMed: 17294284]

Thissen D, Steinberg L. A taxonomy of item response models. Psychometrika. 1986; 51:567–577.

Thissen D, Steinberg L, Mooney JA. Trace lines for testlets: A use of multiple-categorical-response models. Journal of Educational Measurement. 1989; 26:247–260.

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 67-113.

Thissen, D.; Wainer, H. Test scoring. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.

Thissen D, Wainer H, Wang XB. Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? an analysis of two tests. Journal of Educational Measurement. 1994; 31:113–123.

Tucker LR. An inter-battery method of factor analysis. Psychometrika. 1958; 23:111–136.

Wainer, H.; Bradlow, ET.; Wang, X. Testlet response theory and its applications. New York, NY: Cambridge University Press; 2007.

Williams VSL, Pommerich M, Thissen D. A comparison of developmental scales based on Thurstone methods and item response theory. Journal of Educational Measurement. 1998; 35:93–107.

Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods. 2007; 12:58–79. [PubMed: 17402812]

Woods CM, Thissen D. Item response theory with estimation of the latent population distribution using spline-based densities. Psychometrika. 2006; 71:281–301.

Yao L, Schwarz R. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. Applied Psychological Measurement. 2006; 30:469–492.

Yung YF, McLeod LD, Thissen D. On the relationship between the higher-order factor model and the hierarchical factor model. Psychometrika. 1999; 64:113–128.

Appendix A

In equation (3), the lower asymptote parameter c is bounded between 0 and 1. Stable estimation of c often requires a reparameterization $\gamma = \log[c/(1 - c)]$, so that one may estimate the unbounded logit of guessing γ instead of c. Clearly, the inverse relationship is $c = 1/[1 + \exp(-\gamma)]$. In addition, a stochastic constraint (or prior if one adopts a Bayesian intepretation) on γ is usually needed to ensure proper convergence. For instance, when analyzing items from a multiple-choice test with 4 response alternatives, one may postulate that c should be about .25 if the distractors are constructed appropriately. The logit of .25 is $\log[0.25/(1 - 0.25)] = -1.10$. In this case, a stochastic constraint (or prior) in the form of a univariate normal

$$\frac{1}{\sqrt{2\pi}\tau}\exp\left[-\frac{1}{2}\left(\frac{\gamma-\nu}{\tau}\right)^2\right]$$

with mean $\nu = -1.1$ and standard deviation $\tau = 0.5$ is a plausible choice as a penalty function (see, e.g., Thissen, 2003). The constraint penalizes γ as it moves away from -1.1. When interpreted as a prior, the prior standard deviation of .5 leaves ample room for the data to suggest where the most likely estimate of γ should be. In our implementation, we permit user-defined mean and standard deviation for the constraint function for maximal generality.

Appendix B

Let $\theta = a_0\theta_0 + a_s\theta_s$ be a linear combination of the general dimension and the specific factor s with weights given by the item slopes. It follows from Equation (8) that the category response probability for the GPC model may be written as

$$P(y=k)=\frac{\exp\{k\theta+d_k\}}{\sum_{l=0}^{K-1}\exp\{l\theta+d_l\}}.$$

Similarly, from Equation (9), the category response probability for the nominal model is

$$P(y=k)=\frac{\exp\{a_k^*\theta+d_k\}}{\sum_{l=0}^{K-1}\exp\{a_k^*\theta+d_l\}}.$$

Thissen et al. (2010) point out that for identification, the following restrictions should be in place: $a_0^*=0, a_{K-1}^*=K-1, d0=0$. This can be accomplished by reparameterizaton. Let

$$\mathbf{a}^*=\begin{pmatrix} a_0^* \\ \vdots \\ a_{K-1}^* \end{pmatrix}=\mathbf{F}\begin{pmatrix} 1 \\ \alpha \end{pmatrix}, \text{and } \mathbf{d}=\begin{pmatrix} d_0 \\ \vdots \\ d_{K-1} \end{pmatrix}=\mathbf{F}\delta.$$

The scalar parameters a_k^* and d_k are the kth elements of the K-dimensional vectors $\mathbf{a}^*$ and $\mathbf{d}$, respectively. The vector $\boldsymbol{\alpha}$ is a $(K-2)\times 1$ vector of scoring function contrasts that defines the ordering of categories, and $\boldsymbol{\delta}$ is a $(K-1)\times 1$ vector of intercept contrasts. The matrix $\mathbf{F}$ is a fixed $K\times(K-1)$ matrix of contrast coefficients that relate the a_k^*'s and d_k's to the estimable parameters in $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. By an appropriate choice of $\mathbf{F}$, the identification restrictions will be automatically satisfied. Thissen et al. (2010) propose the use of the following linear-Fourier contrast matrix

$$\mathbf{F}=\begin{pmatrix} 0 & 0 & \cdots & 0 \\ 1 & f_{2,2} & \cdots & f_{2,(K-1)} \\ 2 & f_{3,2} & \cdots & f_{3,(K-1)} \\ \vdots & \vdots & & \vdots \\ K-1 & 0 & \cdots & 0 \end{pmatrix},$$

where a typical element $f_{k,m}$ for $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, K-1$ takes its value from a Fourier sine-series:

$$f_{k,m}=\sin\left\{\frac{\pi(k-1)(m-1)}{K-1}\right\}.$$

For instance, for $K = 4$, the contrast matrix is

$$\mathbf{F}=\begin{pmatrix} 0 & 0 & 0 \\ 1 & .866 & .866 \\ 2 & .866 & -.866 \\ 3 & 0 & 0 \end{pmatrix},$$

and for $K = 5$, the contrast matrix is

$$\mathbf{F}=\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & .707 & 1 & .707 \\ 2 & 1 & 0 & -1 \\ 3 & .707 & -1 & .707 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

For the nominal model, one can verify that identification conditions are all satisfied. For the GPC model, the a_k^*'s are not relevant, and the d_k share the same identification condition as the nominal model. In fact, this derivation shows that the GPC model is a special case of the nominal model where all the $\boldsymbol{\alpha}$ contrasts are set to zero. Other contrast matrices can also be used (see, e.g., Thissen & Steinberg, 1986).

The seemingly complicated reparametrization achieves several goals. First, the linear-Fourier contrasts implement the identification restrictions. Second, this reparametrization permits one to "mix" the nominal model with other models for any arbitrary number of dimensions, as is routinely done in Multilog (Thissen, 2003) for unidimensional IRT. Third, the linear-Fourier matrix provide (partially) orthogonal bases that essentially serve to "smooth" category boundaries or define partial ordering of the categories.

Appendix C

In this technical appendix we present additional details relevant to parameter estimation. Recall that with the bifactor orthogonality restriction, in group $g = 1, \ldots, G$, the distribution of the $(S_g + 1)$-dimensional vector of latent variables $\boldsymbol{\theta}_g$ can be written as a product of $S_g + 1$ univariate normals, i.e.,

$$f_{\beta}(\theta_g) = \prod_{s=0}^{S_g} f_{\beta}(\theta_{sg} | \mu_{sg}, \sigma_{sg}), \tag{21}$$

where $f_{\boldsymbol{\beta}}(\theta_{sg}|\mu_{sg}, \sigma_{sg})$ is the density of a normal variable with mean μ_{sg} and standard deviation σ_{sg}. The means and variances are dependent on the vector of free parameters $\boldsymbol{\beta}$, and the dependence is explicitly shown in Equation (21) by the $\boldsymbol{\beta}$ subscript.

Let us denote the conditional distribution of the response pattern $\mathbf{y}_{ig}$ given $\boldsymbol{\theta}_g$ as $f_{\boldsymbol{\beta}}(\mathbf{y}_{ig}|\boldsymbol{\theta}_g)$. Similarly, the item parameters are dependent on the vector of free parameters $\boldsymbol{\beta}$. For item j, let the conditional distribution of y_{ijg} be $f_{\boldsymbol{\beta}}(y_{ijg}|\boldsymbol{\theta}_g)$. By conditional independence, we should be able to write $f_{\boldsymbol{\beta}}(\mathbf{y}_{ig}|\boldsymbol{\theta}_g)$ as a product, i.e.,

$$f_{\beta}(\mathbf{y}_{ig}|\theta_g) = \prod_{j=1}^{n_g} f_{\beta}(y_{ijg}|\theta_g). \tag{22}$$

However, recall that the bifactor-like structure implies that the response probabilities of item j depend on at most two latent variables, the general dimension θ_{0g} and (optionally) another specific factor. Thus, conditional independence in the item bifactor context can be more conveniently expressed as follows

$$f_{\beta}(\mathbf{y}_{ig}|\theta_g) = \prod_{s=0}^{S_g} \prod_{j \in \mathcal{E}_s} f_{\beta}(y_{ijg}|\theta) = \prod_{j \in \mathcal{E}_0} f_{\beta}(y_{ijg}|\theta_{0g}) \left[\prod_{s=1}^{S_g} \prod_{j \in \mathcal{E}_s} f_{\beta}(y_{ijg}|\theta_{0g}, \theta_{sg}) \right], \tag{23}$$

where $\mathcal{E}_s$ is a set of indices of items that load on specific factor $s = 1, \ldots, S_g$ and $\mathcal{E}_0$ denotes the set of items that do not load on any specific factor. Equation (23) is simply a rearrangement of the right-hand side of Equation (22). Instead of taking a product of n_g item response probabilities by the natural indexing, we break the product into $S_g + 1$ sections.

Then we immediately realize that within section $s = 1, \ldots, S_g$, the item response probabilities only depend on θ_{0g} and θ_{sg}. For instance, the bifactor structure in Equation (1) implies $S_g = 2$ sections, and the item index in section 1 is given by $\mathfrak{S}_1 = \{1, 2, 3\}$, and in section 2, $\mathfrak{S}_2 = \{4, 5, 6\}$. When there are items that load only on the primary dimension, as in the bifactor-like structure in Equation (2), the section index $\mathfrak{S}_0$ is given by $\{3, 6\}$, with $\mathfrak{S}_1 = \{1, 2\}$ and $\mathfrak{S}_2 = \{4, 5\}$.

Multiplying $f_{\boldsymbol{\beta}}(\mathbf{y}_{ig}|\boldsymbol{\theta}_g)$ by $f_{\boldsymbol{\beta}}(\boldsymbol{\theta}_g)$ and substituting in Equations (21) and (23) leads to the joint distribution of the observed and the latent variables in group g:

$$\begin{aligned} f_\beta(\mathbf{y}_{ig}, \theta_g) &= f_\beta(\mathbf{y}_{ig}|\theta_g) f_\beta(\theta_g) \\ &= \prod_{j \in \mathfrak{S}_0} f_\beta(y_{ijg}|\theta_{0g}) \left[\prod_{s=1}^{S_g} \prod_{j \in \mathfrak{S}_s} f_\beta(y_{ijg}|\theta_{0g}, \theta_{sg}) \right] \prod_{s=0}^{S_g} f_\beta(\theta_{sg}|\mu_{sg}, \sigma_{sg}) \\ &= \left[\prod_{j \in \mathfrak{S}_0} f_\beta(y_{ijg}|\theta_{0g}) f_\beta(\theta_{0g}|\mu_{0g}, \sigma_{0g}) \right] \times \left[\prod_{s=1}^{S_g} \prod_{j \in \mathfrak{S}_s} f_\beta(y_{ijg}|\theta_{0g}, \theta_{sg}) f_\beta(\theta_{sg}|\mu_{0g}, \sigma_{0g}) \right] \\ &= \left[\prod_{j \in \mathfrak{S}_0} f_\beta(y_{ijg}, \theta_{0g}) \right] \{ \prod_{s=1}^{S_g} \left[\prod_{j \in \mathfrak{S}_s} f_\beta(y_{ijg}, \theta_{0g}, \theta_{sg}) \right] \}. \end{aligned} \tag{24}$$

We now notice from the last line of Equation (24) that the joint distribution of $\mathbf{y}_{ig}$ and $\boldsymbol{\theta}_g$ conveniently factor into $S_g + 1$ terms that are mutually independent. In order to conduct maximum marginal likelihood estimation, we must integrate the latent factors out of the joint distribution. We realize from the factorization result in Equation (24) that the marginal distribution of $\mathbf{y}_{ig}$ can be written as

$$f_\beta(\mathbf{y}_{ig}) = \int_{-\infty}^{+\infty} \prod_{j \in \mathfrak{S}_0} f_\beta(y_{ijg}, \theta_{0g}) \left[\prod_{s=1}^{S_g} \int_{-\infty}^{+\infty} \prod_{j \in \mathfrak{S}_s} f_\beta(y_{ijg}, \theta_{0g}, \theta_{sg}) d\theta_{sg} \right] d\theta_{0g}. \tag{25}$$

As in Gibbons and Hedeker (1992), we may approximate the series of two-dimensional integrals in Equation (25) to any desired degree of accuracy with numerical quadrature:

$$f_\beta(\mathbf{y}_{ig}) \doteq \sum_{q_0=1}^{Q} \prod_{j \in \mathfrak{S}_0} f_\beta(y_{ijg}, X_{q_0 g}) \left[\prod_{s=1}^{S_g} \sum_{q_s=1}^{Q} \prod_{j \in \mathfrak{S}_s} f_\beta(y_{ijg}, X_{q_0 g}, X_{q_s g}) W_{q_s g} \right] W_{q_0 g}, \tag{26}$$

where the integrand functions are evaluated at a set of $Q \times Q$ quadrature nodes (the X_q's), with weights at each point give by the W_q's. The contribution to the marginal log-likelihood of $\boldsymbol{\beta}$ given $\mathbf{y}_{ig}$ is $\log L(\boldsymbol{\beta}|\mathbf{y}_{ig}) = \log f_{\boldsymbol{\beta}}(\mathbf{y}_{ig})$. When summed over the respondents and the groups, we arrive at the marginal log-likelihood given in Equation (17).

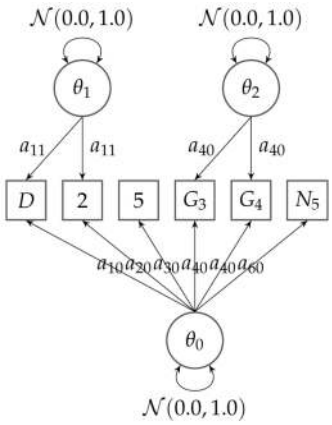

Figure 1.

Conceptual Path Diagram Showing the Bifactor-like Factor Structure in Equation (2). Key to notations for IRT models: D = Dichotomous response with lower asymptote; 2 = Graded response in 2 categories; 5 = Graded response in 5 categories; G_3 = GPC response in 3 categories; G_4 = GPC response in 4 categories; N_5 = Nominal response in 5 categories.

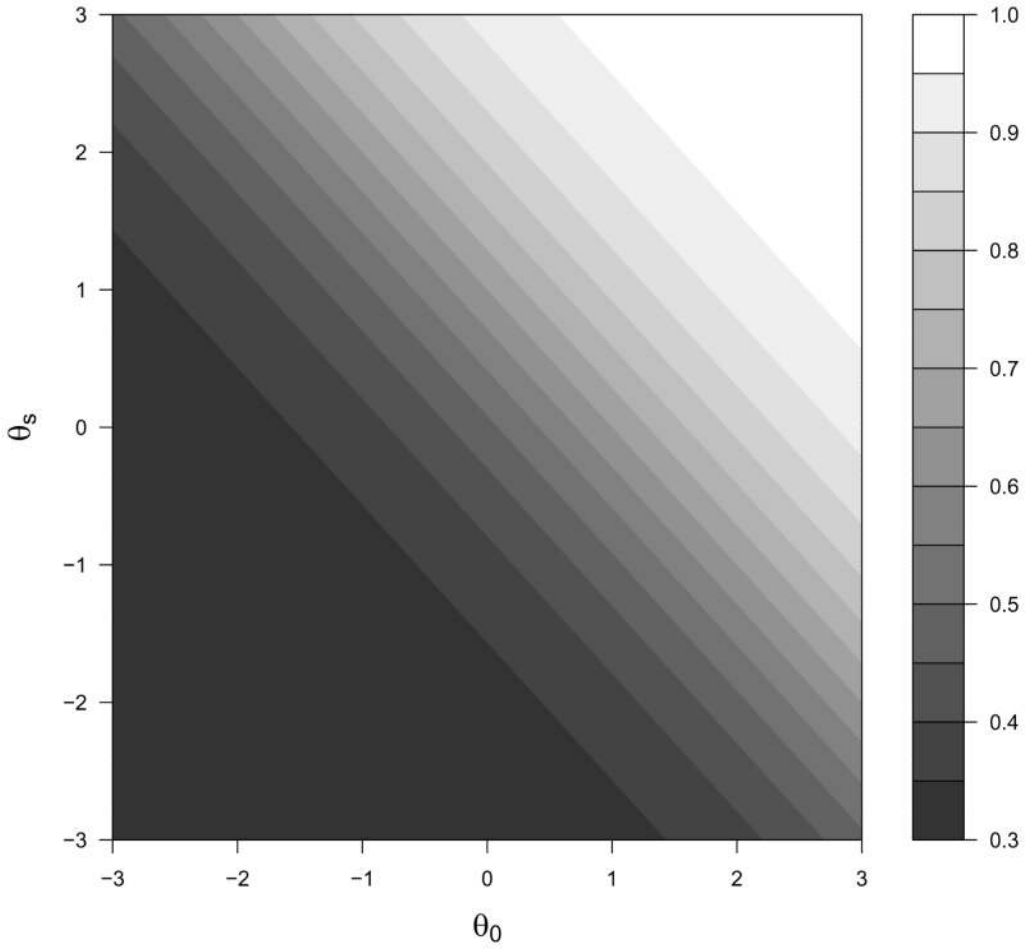

Figure 2.
Level Plot of the Item Response Surface of a Bifactor IRT Model for Dichotomous Response. The item parameters are $a_0 = 1$, $a_s = 1$, $d = -1$, $c = 0.3$.

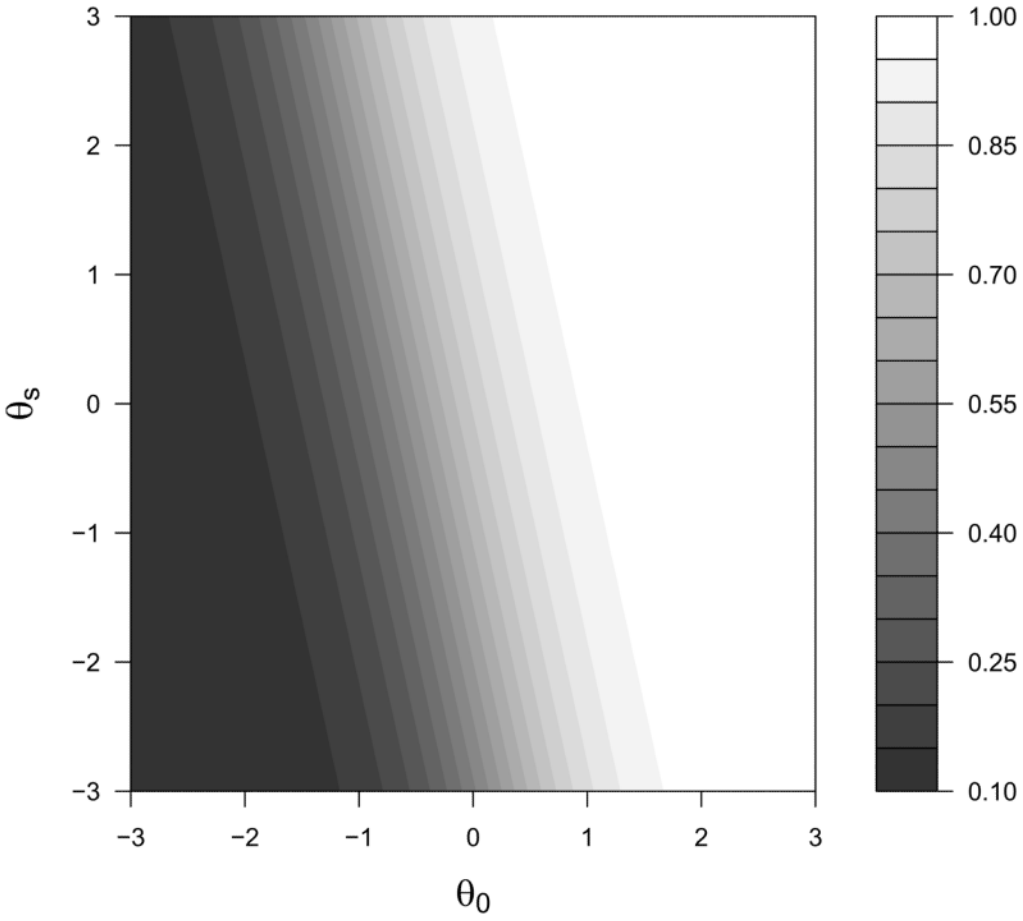

Figure 3.
Level Plot of the Item Response Surface of a Bifactor IRT Model for Dichotomous Response. The item parameters are $a_0 = 2$, $a_s = 0.5$, $d = 1$, $c = 0.1$.

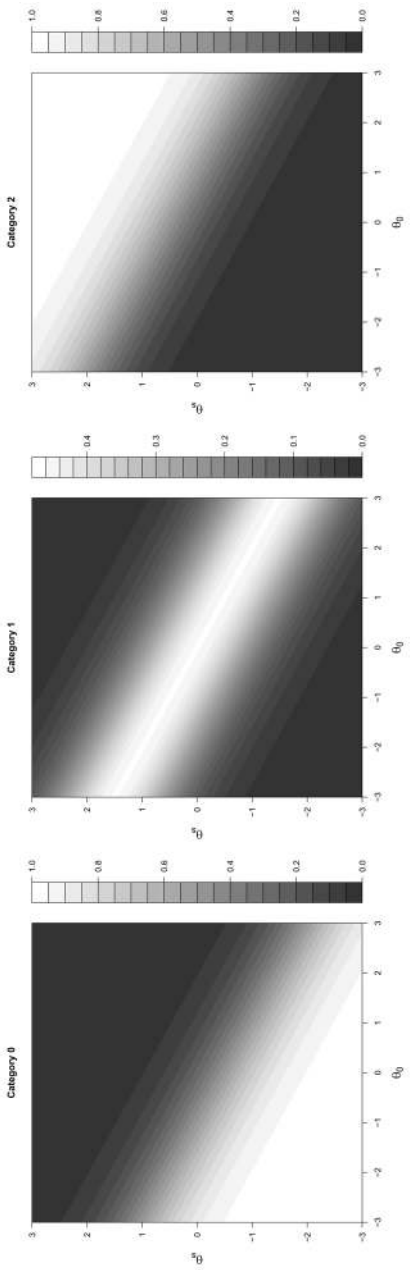

Figure 4.
Level Plots of the Category Response Surfaces of a Bifactor IRT Model for Three Graded Response Categories. The item parameters are $a_0 = 1$, $a_s = 2$, $d_1 = -1$, $d_2 = 1$.

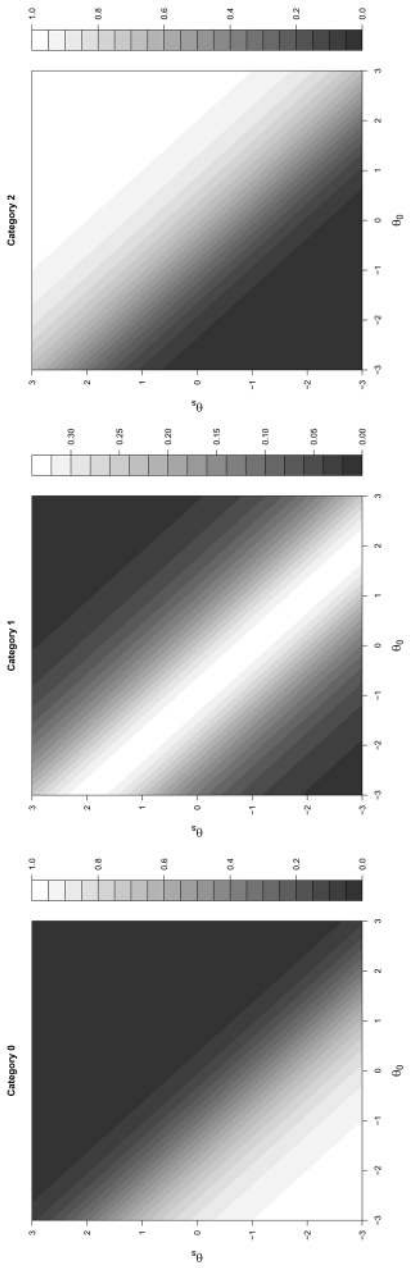

Figure 5.
Level Plots of the Category Response Surfaces of a Bifactor Generalized Partial Credit IRT Model. The item parameters are $a_0 = 1$, $a_s = 1$, $d_0 = 0$, $d_1 = 1$, and $d_2 = 2$.

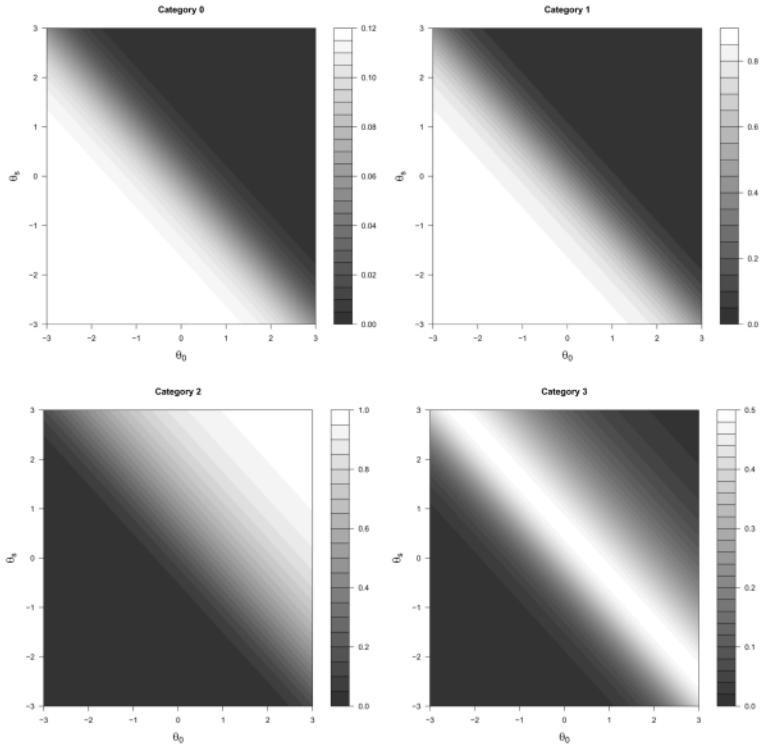

Figure 6.
Level Plots of the Category Response Surfaces of a Bifactor Nominal Categories Model.

The scoring function parameteres are $a_0^*=0, a_1^*=0, a_0^*=3$, and $a_3^*=2$. The item slopes are a_0 = 1, and a_s = 1. The intercepts are d_0 = 0, d_1 = 2, d_2 = 1, and d_3 = 2.

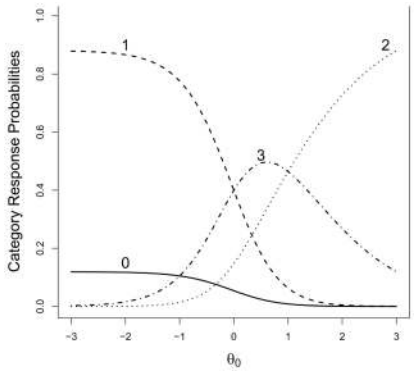

Figure 7.
Plot of a Cross-section of the Item Response Surfaces of a Bifactor Nominal Categories Model. In this plot, θ_s is held constant at 0. The item parameters are the same as Figure 6.

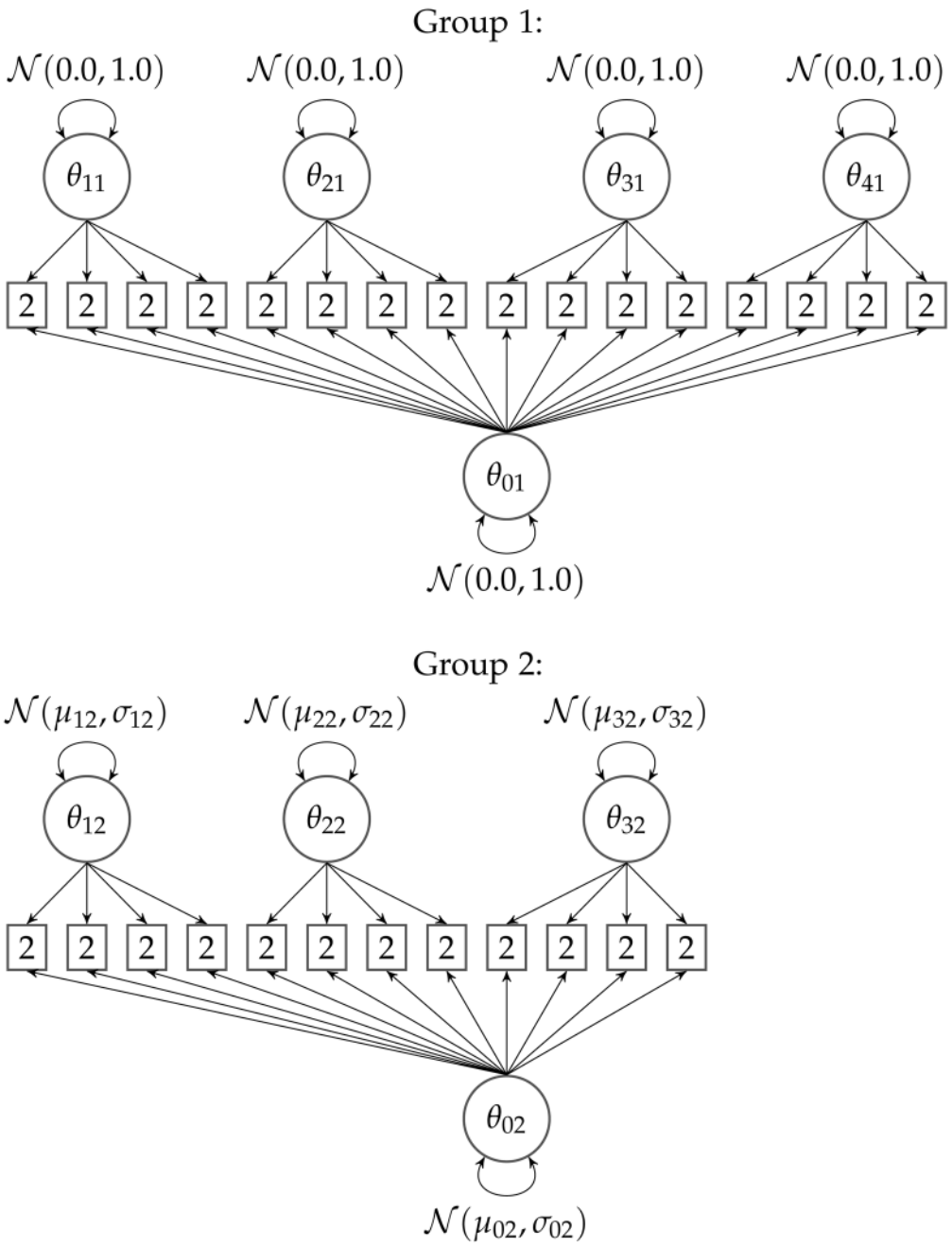

Figure 8.
Conceptual Path Diagram for a Two-Group Extended Bifactor Model. There are 16 items in group 1. These items form into 4 clusters of 4 items each. There are 12 items in group 2 and one of the clusters is not present. Though not shown directly, all item parameters are set equal across the groups to link the two groups. Group 1 is chosen as the reference with fixed means and variances for all latent variables. In group 2, the latent variable means and variances are estimated parameters.

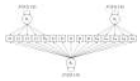

Figure 9.
Conceptual Path Diagram for an Extended Bifactor Model. There are three types of items, dichotomous, 4-category nominal, and 4-category ordinal (GPC).

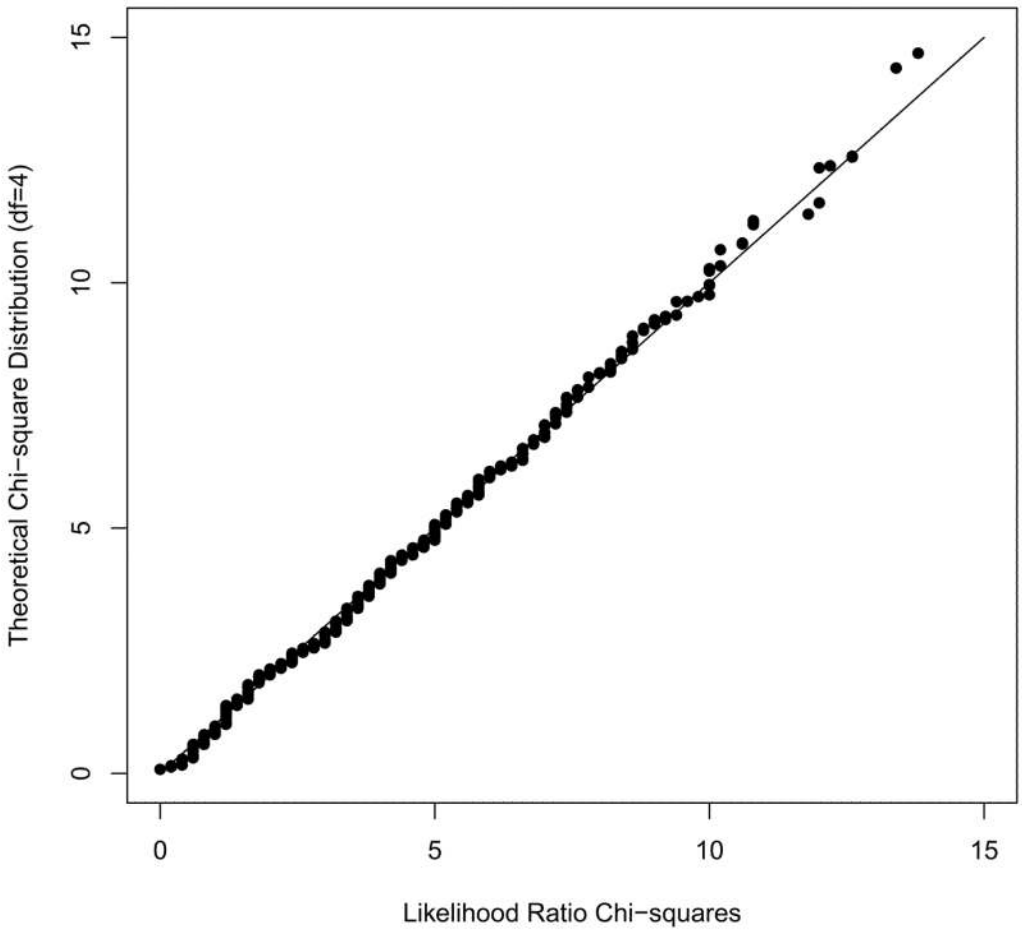

Figure 10.
Q-Q Plot of the Quantiles of the Likelihood Ratio Test Statistic for Bifactor Nested Model Comparisons Against Theoretical Quantiles of a Central Chi-square Variable with 4 Degrees-of-Freedom

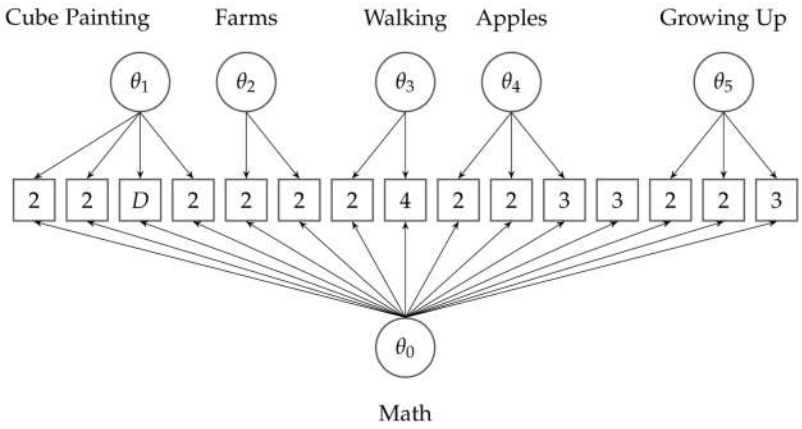

Figure 11.
Conceptual Path Diagram Showing an Extended Bifactor Model for the 15 Items on Booklet 1 of the PISA Math Assessment.

Table 1

Two Group Bifactor Model: Generating Parameter Values and Estimated Biases

		Slopes				
Items	**Intercept**	θ_0	θ_1	θ_2	θ_3	θ_4
1	1.00 (.00)	1.00 (.00)	.80 (.00)	.00 (—-)	.00 (—-)	.00 (—-)
2	.25 (−.01)	1.40 (.01)	1.50 (.06)	.00 (—-)	.00 (—-)	.00 (—-)
3	−.25 (−.01)	1.70 (.02)	1.20 (.02)	.00 (—-)	.00 (—-)	.00 (—-)
4	−1.00 (−.01)	2.00 (.03)	1.00 (.01)	.00 (—-)	.00 (—-)	.00 (—-)
5	1.00 (.01)	1.40 (.01)	.00 (—-)	1.00 (.01)	.00 (—-)	.00 (—-)
6	.25 (.01)	1.70 (.02)	.00 (—-)	0.80 (.00)	.00 (—-)	.00 (—-)
7	−.25 (.00)	2.00 (.03)	.00 (—-)	1.50 (.03)	.00 (—-)	.00 (—-)
8	−1.00 (−.02)	1.00 (.01)	.00 (—-)	1.20 (.03)	.00 (—-)	.00 (—-)
9	1.00 (.01)	1.70 (.01)	.00 (—-)	.00 (—-)	1.20 (.01)	.00 (—-)
10	.25 (.01)	2.00 (.03)	.00 (—-)	.00 (—-)	1.00 (.01)	.00 (—-)
11	−.25 (.00)	1.00 (.01)	.00 (—-)	.00 (—-)	.80 (.00)	.00 (—-)
12	−1.00 (−.02)	1.40 (.02)	.00 (—-)	.00 (—-)	1.50 (.05)	.00 (—-)
*13	1.00 (.01)	2.00 (.02)	.00 (—-)	.00 (—-)	.00 (—-)	1.50 (.01)
*14	.25 (.01)	1.00 (.03)	.00 (—-)	.00 (—-)	.00 (—-)	1.20 (.06)
*15	−.25 (.00)	1.40 (.02)	.00 (—-)	.00 (—-)	.00 (—-)	1.00 (.01)
*16	−1.00 (−.01)	1.70 (.02)	.00 (—-)	.00 (—-)	.00 (—-)	.80 (.01)
Groups						
1	Means	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
	Variances	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)
2	Means	1.00 (−.04)	−.50 (.05)	.00 (.04)	.50 (.04)	N/A
	Variances	.80 (.00)	1.20 (.02)	1.50 (−.01)	1.00 (.02)	N/A

Note.

*Items 13 to 16 are only present in group 1. For a parameter, estimated bias (in parentheses) is defined as the Monte Carlo average of the estimates minus the generating value. Dashed parameters are fixed.

Table 2

Two Group Bifactor Model: Standard Errors and Monte Carlo Standard Deviations

		Slopes				
Items	**Intercept**	θ_0	θ_1	θ_2	θ_3	θ_4
1	.09 (.09)	.09 (.09)	.12 (.12)			
2	.11 (.11)	.15 (.14)	.24 (.25)			
3	.10 (.10)	.14 (.14)	.17 (.17)			
4	.12 (.12)	.16 (.15)	.14 (.14)			
5	.10 (.09)	.12 (.11)		.15 (.16)		
6	.10 (.09)	.13 (.12)		.12 (.14)		
7	.12 (.12)	.19 (.19)		.25 (.24)		
8	.11 (.10)	.11 (.10)		.19 (.19)		
9	.11 (.11)	.14 (.14)			.19 (.19)	
10	.11 (.10)	.16 (.16)			.16 (.16)	
11	.08 (.08)	.09 (.08)			.13 (.13)	
12	.15 (.17)	.16 (.19)			.30 (.35)	
*13	.16 (.15)	.25 (.25)				.33 (.33)
*14	.10 (.10)	.14 (.13)				.27 (.27)
*15	.10 (.10)	.15 (.14)				.20 (.19)
*16	.12 (.12)	.17 (.16)				.18 (.19)
Groups						
1	Means					
	Variances					
2	Means	.11 (.10)	.16 (.15)	.15 (.14)	.16 (.14)	N/A
	Variances	.10 (.09)	.31 (.29)	.35 (.30)	.28 (.27)	N/A

Note.

* Items 13 to 16 are only present in group 1. Entries are the Monte Carlo averages of estimated standard errors and the Monte Carlo standard deviations (in parentheses) of the estimated parameters. Fixed parameters do not have standard errors.

Table 3

A Model for Complex Assessment: Generating Parameter Values and Estimated Biases

Items	c	a_0	a_1	a_2	d
1	.25(.00)	1.37(.06)	1.40(.06)	.00(—-)	−.27(−.05)
2	.25(.00)	1.42(.04)	.88(.03)	.00(—-)	−.79(−.05)
3	.25(.00)	1.57(.03)	.59(.02)	.00(—-)	−.92(−.03)
4	.25(.00)	1.24(.03)	.90(.02)	.00(—-)	−1.03(−.04)
6	.25(.00)	.83(.06)	.00(—-)	.00(—-)	−1.27(−.09)
7	.25(.00)	1.74(.01)	.00(—-)	.00(—-)	−2.05(.00)
8	.25(.00)	1.72(.02)	.00(—-)	.00(—-)	−2.36(−.03)
9	.25(.00)	1.23(.06)	.00(—-)	.00(—-)	−2.31(−.08)
10	.25(.00)	1.44(.09)	.00(—-)	.00(—-)	−2.79(−.14)

Items	a_0^*	a_1^*	a_2^*	a_3^*	a_0	a_1	a_2	d_0	d_1	d_2	d_3
5	.00(—-)	2.49(.01)	2.10(.00)	3.00(—-)	1.32(.04)	.88(.04)	.00(—-)	.00(—-)	−.91(−.04)	−3.40(−.03)	−3.27(−.06)
11	.00(—-)	1.00(—-)	2.00(—-)	3.00(—-)	2.04(.02)	.00(—-)	.95(.00)	.00(—-)	−3.87(−.03)	−6.81(−.06)	−8.04(−.10)
12	.00(—-)	1.00(—-)	2.00(—-)	3.00(—-)	2.02(.01)	.00(—-)	.95(.00)	.00(—-)	−.11(.00)	−2.01(−.02)	−2.58(−.03)
13	.00(—-)	1.00(—-)	2.00(—-)	3.00(—-)	1.35(.01)	.00(—-)	.95(.00)	.00(—-)	−.90(−.01)	−2.83(−.02)	−4.68(−.04)
14	.00(—-)	1.00(—-)	2.00(—-)	3.00(—-)	.92(.01)	.00(—-)	.95(.01)	.00(—-)	−1.20(.00)	−4.34(−.04)	−6.63(−.08)
15	.00(—-)	1.00(—-)	2.00(—-)	3.00(—-)	1.84(.01)	.00(—-)	.95(.00)	.00(—-)	.11(.00)	−.05(.00)	−.06(−.01)

Note. 1–4, 6–10 = MC items; 5 = CMC items. 11 – 15 = CR items. For a parameter, estimated bias (in parentheses) is defined as the Monte Carlo average of the estimates minus the generating value. Dashed parameters are fixed either due to restrictions or identification conditions. See Appendix B for details on identification conditions.

Table 4

A Model for Complex Assessment: Standard Errors and Monte Carlo Standard Deviations

Items			***logit*(*c*)**	**a_0**	**a_1**	**a_2**	***d***		
1			.31(.23)	.27(.25)	.32(.31)		.32(.28)		
2			.22(.19)	.22(.21)	.19(.20)		.28(.26)		
3			.19(.17)	.22(.21)	.15(.15)		.26(.24)		
4			.23(.20)	.22(.21)	.21(.20)		.32(.30)		
6			.29(.22)	.22(.21)			.41(.37)		
7			.13(.12)	.29(.28)			.38(.37)		
8			.12(.12)	.31(.32)			.44(.44)		
9			.16(.14)	.31(.29)			.52(.49)		
10			.12(.12)	.37(.37)			.62(.61)		
	Scoring Function Contrasts						Intercept Contrasts		
Items	1	2	a_0	a_1	a_2		1	2	3
5	.17(.16)	.13(.13)	.18(.18)	.19(.18)			.09(.09)	.18(.18)	.16(.17)
11			.17(.18)		.12(.12)		.21(.20)	.15(.14)	.10(.10)
12			.12(.12)		.09(.10)		.07(.07)	.10(.10)	.05(.05)
13			.08(.08)		.09(.09)		.08(.08)	.09(.10)	.05(.04)
14			.06(.06)		.10(.10)		.12(.12)	.14(.13)	.07(.07)
15			.12(.12)		.10(.11)		.05(.05)	.09(.09)	.04(.04)

Note. 1–4, 6–10 = MC items; 5 = CMC items. 11 – 15 = CR items. Entries are the Monte Carlo averages of estimated standard errors and the Monte Carlo standard deviations (in parentheses) of the *estimable*, i.e., reparameterized, parameters. For the MC items, the logit of the guessing (lower asymptote) parameter is estimated. For the CMC and CR items, 3 intercept contrasts are estimated. For the CMC item, 2 scoring function contrasts are estimated. See Appendices A and B for details. Fixed parameters do not have standard errors.

Psychol Methods. Author manuscript; available in PMC 2012 September 1.

Table 5

PISA Item Slope Estimates (Standard Errors): Measurement Invariance Assumed

Items	a_0	a_1	a_2	a_3	a_4	a_5
Cube Painting 1 (CR2)	1.21 (.09)	.93 (.14)				
Cube Painting 2 (CR2)	1.09 (.09)	.80 (.13)				
Cube Painting 3 (MC4)	2.66 (.35)	2.13 (.39)				
Cube Painting 4 (CR2)	1.33 (.10)	.89 (.16)				
Farms 1 (CR2)	2.01 (.13)		.53 (.18)			
Farms 2 (CR2)	.90 (.07)		.53 (.18)			
Walking 1 (CR2)	2.84 (.16)			1.33 (.21)		
Walking 2 (CR4)	2.72 (.13)			1.33 (.21)		
Apples 1 (CR2)	1.59 (.11)				.80 (.17)	
Apples 2 (CR2)	3.25 (.20)				.89 (.15)	
Apples 3 (CR3)	3.07 (.23)				1.69 (.37)	
Continent Area (CR3)	1.68 (.09)					
Growing Up 1 (CR2)	1.29 (.13)					1.00 (.32)
Growing Up 2 (CR2)	1.13 (.07)					.31 (.14)
Growing Up 3 (CR3)	.78 (.06)					.44 (.14)

Note. CR = Constructed Response; MC = Multiple-choice; The number of categories for each item is listed in the parentheses. Slopes on the doublets for Farms and Walking are constrained equal for identification. Fixed zero slopes are shown as blanks.

Psychol Methods. Author manuscript; available in PMC 2012 September 1.

Table 6

PISA Other Item Parameter Estimates (Standard Errors): Measurement Invariance Assumed

Items	*logit*(*c*)	Intercept 1	Intercept 2	Intercept 3
Cube Painting 1 (CR2)		.53 (.09)		
Cube Painting 2 (CR2)		−2.15 (.12)		
Cube Painting 3 (MC4)	−2.31 (.23)	1.80 (.26)		
Cube Painting 4 (CR2)		−1.41 (.11)		
Farms 1 (CR2)		.22 (.08)		
Farms 2 (CR2)		.28 (.06)		
Walking 1 (CR2)		−2.36 (.14)		
Walking 2 (CR4)		−1.86 (.12)	−4.03 (.19)	5.75 (.25)
Apples 1 (CR2)		.28 (.06)		
Apples 2 (CR2)		−3.28 (.19)		
Apples 3 (CR3)		−4.08 (.31)	−6.05 (.45)	
Continent Area (CR3)		−1.42 (.09)	−3.65 (.12)	
Growing Up 1 (CR2)		.33 (.07)		
Growing Up 2 (CR2)		−.10 (.06)		
Growing Up 3 (CR3)		2.02 (.08)	−.21 (.05)	

Note. CR = Constructed Response; MC = Multiple-choice; The number of categories for each item is listed in the parentheses.

Table 7

PISA Factor Means and Variances (Standard Errors): Measurement Invariance Assumed

	Math	Cube Painting	Farms	Walking	Apples	Growing Up
			Means			
Australia	.36 (.06)	.64 (.14)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
Germany	.20 (.06)	.07 (.14)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
Japan	.75 (.05)	.45 (.15)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
UK	.46 (.04)	.23 (.11)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
USA	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
			Variances			
Australia	1.01 (.11)	1.42 (.42)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)
Germany	1.06 (.12)	2.97 (.59)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)
Japan	.59 (.07)	1.70 (.52)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)
UK	.82 (.07)	1.34 (.36)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)
USA	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)

Note. Fixed parameters do not have standard errors.

Psychol Methods. Author manuscript; available in PMC 2012 September 1.

Psychol Methods. Author manuscript; available in PMC 2012 September 1.

Table 8

PISA Item Slope Estimates (Standard Errors): Studying DIF for the Farms Cluster

Items	a_0	a_1	a_2	a_3	a_4	a_5
Australia						
Farms 1 (CR2)	2.11 (.29)		.62 (.19)			
Farms 2 (CR2)	.39 (.13)		.62 (.19)			
Germany						
Farms 1 (CR2)	1.31 (.22)		.62 (.19)			
Farms 2 (CR2)	1.20 (.20)		.62 (.19)			
Japan						
Farms 1 (CR2)	1.38 (.25)		.62 (.19)			
Farms 2 (CR2)	.81 (.22)		.62 (.19)			
UK						
Farms 1 (CR2)	2.51 (.26)		.62 (.19)			
Farms 2 (CR2)	.66 (.10)		.62 (.19)			
US						
Farms 1 (CR2)	1.90 (.28)		.62 (.19)			
Farms 2 (CR2)	1.50 (.24)		.62 (.19)			
All Groups						
Cube Painting 1 (CR2)	1.15 (.09)	.96 (.14)				
Cube Painting 2 (CR2)	1.04 (.09)	.84 (.12)				
Cube Painting 3 (MC4)	2.45 (.28)	1.90 (.34)				
Cube Painting 4 (CR2)	1.26 (.09)	.82 (.13)				
Walking 1 (CR2)	2.71 (.16)			1.28 (.26)		
Walking 2 (CR4)	2.61 (.14)			1.28 (.26)		
Apples 1 (CR2)	1.50 (.10)				.86 (.16)	
Apples 2 (CR2)	3.05 (.19)				.94 (.15)	
Apples 3 (CR3)	2.84 (.31)				1.65 (.46)	
Continent Area (CR3)	1.61 (.09)					
Growing Up 1 (CR2)	1.22 (.16)					1.00 (.45)
Growing Up 2 (CR2)	1.11 (.07)					.29 (.13)

Items	a_0	a_1	a_2	a_3	a_4	a_5
Growing Up 3 (CR3)	.73 (.06)					.46 (.18)

Note. CR = Constructed Response; MC = Multiple-choice; The number of categories for each item is listed in the parentheses. Slopes on the doublets for Farms and Walking are constrained equal for identification. Fixed zero slopes are shown as blanks.

Table 9

PISA Other Item Parameter Estimates (Standard Errors): Studying DIF for the Farms Cluster

Items	*logit*(*c*)	Intercept 1	Intercept 2	Intercept 3
Australia				
Farms 1 (CR2)		.76 (.16)		
Farms 2 (CR2)		.35 (.11)		
Germany				
Farms 1 (CR2)		.03 (.15)		
Farms 2 (CR2)		−.43 (.15)		
Japan				
Farms 1 (CR2)		1.06 (.20)		
Farms 2 (CR2)		1.66 (.24)		
UK				
Farms 1 (CR2)		.21 (.14)		
Farms 2 (CR2)		.20 (.09)		
US				
Farms 1 (CR2)		−.24 (.15)		
Farms 2 (CR2)		.56 (.15)		
All Groups				
Cube Painting 1 (CR2)		.49 (.09)		
Cube Painting 2 (CR2)		−2.23 (.14)		
Cube Painting 3 (MC4)	−2.30 (.28)	1.65 (.22)		
Cube Painting 4 (CR2)		−1.40 (.11)		
Walking 1 (CR2)		−2.22 (.16)		
Walking 2 (CR4)		−1.74 (.13)	−3.92 (.22)	5.63 (.30)
Apples 1 (CR2)		.37 (.07)		
Apples 2 (CR2)		−3.09 (.18)		
Apples 3 (CR3)		−3.84 (.43)	−5.78 (.64)	
Continent Area (CR3)		−1.35 (.08)	−3.59 (.12)	
Growing Up 1 (CR2)		.39 (.08)		
Growing Up 2 (CR2)		−.05 (.06)		
Growing Up 3 (CR3)		2.06 (.09)	−.16 (.05)	

Note. CR = Constructed Response; MC = Multiple-choice; The number of categories for each item is listed in the parentheses.

Table 10

PISA Factor Means and Variances (Standard Errors): Studying DIF for the Farms Cluster

	Math	Cube Painting	Farms	Walking	Apples	Growing Up
			Means			
Australia	.30 (.06)	.80 (.15)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
Germany	.21 (.07)	.11 (.15)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
Japan	.66 (.06)	.68 (.15)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
UK	.45 (.05)	.32 (.12)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
USA	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)	.00 (—-)
			Variances			
Australia	1.19 (.12)	1.52 (.42)	1.04 (.93)	1.00 (—-)	1.00 (—-)	1.00 (—-)
Germany	1.17 (.13)	3.09 (.59)	2.05 (.11)	1.00 (—-)	1.00 (—-)	1.00 (—-)
Japan	.78 (.09)	1.64 (.47)	1.51 (1.54)	1.00 (—-)	1.00 (—-)	1.00 (—-)
UK	.89 (.08)	1.42 (.39)	.36 (.18)	1.00 (—-)	1.00 (—-)	1.00 (—-)
USA	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)	1.00 (—-)

Note. Fixed parameters do not have standard errors.

Psychol Methods. Author manuscript; available in PMC 2012 September 1.

Table 11

Step-down DIF Tests for the Farms Items

Contrasts													
Australia	**Germany**	**Japan**	**UK**	**US**	**Total**	***df***	***p*-value**	***a***	***df***	***p*-value**	***d|a***	***df***	***p*-value**
					Farms 1								
−1	−1	−1	−1	4	19.45	2	.000	.06	1	.810	19.39	1	.000
−1	−1	−1	3	0	13.06	2	.000	9.63	1	.002	3.43	1	.064
−1	−1	2	0	0	9.32	2	.002	1.19	1	.275	8.13	1	.004
−1	−1	0	0	0	17.07	2	.000	4.74	1	.030	12.33	1	.000
					Farms 2								
−1	−1	−1	−1	4	9.30	2	.002	9.22	1	.002	.08	1	.779
−1	−1	−1	3	0	10.06	2	.000	.95	1	.329	9.11	1	.003
−1	−1	2	0	0	43.37	2	.002	.00	1	.970	43.36	1	.000
−1	−1	0	0	0	19.88	2	.000	10.79	1	.001	9.10	1	.003