Journal of Educational and Behavioral Statistics
2013, Vol. 38, No. 1, pp. 32–60
DOI: 10.3102/1076998611432173

Modeling Differential Item Functioning Using a Generalization of the Multiple-Group Bifactor Model

Minjeong Jeon
University of California, Berkeley

Frank Rijmen
Educational Testing Service

Sophia Rabe-Hesketh
University of California, Berkeley and Institute of Education, University of London

The authors present a generalization of the multiple-group bifactor model that extends the classical bifactor model for categorical outcomes by relaxing the typical assumption of independence of the specific dimensions. In addition to the means and variances of all dimensions, the correlations among the specific dimensions are allowed to differ between groups. By including group-specific difficulty parameters, the model can be used to assess differential item functioning (DIF) for testlet-based tests. The model encompasses various item response models for polytomous data by allowing for different link functions, and it includes testlet and second-order models as special cases. Importantly, by assuming that the testlet dimensions are conditionally independent given the general dimension, the authors show, using a graphical model framework, that the integration over all latent variables can be carried out through a sequence of computations in two-dimensional subspaces, making full-information maximum likelihood estimation feasible for high-dimensional problems and large datasets. The importance of relaxing the orthogonality assumption and allowing for a different covariance structure of the dimensions for each group is demonstrated in the context of the assessment of DIF. Through a simulation study, it is shown that ignoring between-group differences in the structure of the multivariate latent space can result in substantially biased estimates of DIF.

Keywords: *multiple-group bifactor model, conditional independence, differential item functioning, bifactor model, testlets, graphical model, EM algorithm*

Introduction

The bifactor model is an item response theory (IRT) model that accommodates conditional dependencies between items grouped in item clusters by including "specific" dimensions for the clusters, in addition to the "general" dimension.

For testlet-based tests, item clusters correspond to testlets (Bradlow, Wainer, & Wang, 1999). If there are K item clusters, the bifactor model has one general or primary dimension (or factor) and K specific dimensions. Each item i has a nonzero loading or discrimination parameter α_{ig} on the primary dimension and a second loading (α_{ik}, $k = 1, 2, ..., K$) on one of the K specific dimensions. For instance, for 4 items with $K = 2$, the bifactor pattern matrix can be written as

$$\begin{pmatrix} \alpha_{1g} & \alpha_{11} & 0 \\ \alpha_{2g} & \alpha_{21} & 0 \\ \alpha_{3g} & 0 & \alpha_{32} \\ \alpha_{4g} & 0 & \alpha_{42} \end{pmatrix}.$$

Here, Items 1 through 4 all load on the primary dimension with loadings in the first column (α_{1g}–α_{4g}). Items 1 and 2 load on the first specific dimension (loadings α_{11} and α_{21}), while Items 3 and 4 load on the second specific dimension (loadings α_{32} and α_{42}). The primary dimension usually stands for the latent variable of central interest, whereas the K specific dimensions account for additional dependence between items in the same item clusters. For instance, in a reading comprehension test, the primary dimension describes the target skill and additional factors describe content area knowledge within a passage. Given the primary dimension, items in different clusters are conditionally independent, whereas items in the same clusters are conditionally dependent (Gibbons & Hedeker, 1992).

The bifactor structure was first introduced by Holzinger and Swineford (1937) as a special case of confirmatory factor analysis for continuous responses. The model has been extended to item-level analysis for binary data by Gibbons and Hedeker (1992) and for polytomous data by Gibbons et al. (2007). An important contribution of Gibbons and Hedeker (1992) and Gibbons et al. (2007) was the use of full-information maximum likelihood estimation (MLE). By taking advantage of the independence of the dimensions and nonoverlapping item clusters, the dimensionality of the integration can be reduced from $(K + 1)$ to two.

The bifactor model has been found useful in psychological and educational measurement. For instance, Reise, Morizot, and Hays (2007) suggested the bifactor model for checking the dimensionality of measurement instruments. Gibbons et al. (2008) applied the bifactor model to construct item banks and computerized adaptive tests. DeMars (2006) applied the bifactor model to testlet-based tests as an alternative to testlet models. Testlet models (Bradlow et al., 1999; Wainer, Bradlow, and Wang, 2007) are typically used to accommodate local item dependence among groups of items (the testlets) having a common stimulus. In this case, the secondary dimensions are often treated as nuisance dimensions. The bifactor model, on the other hand, tends to be applied when item clusters correspond to different domains so that the secondary dimensions have substantive interpretations. Technically, either model is applicable to tests with item clusters (see, e.g., DeMars, 2006; Li, Bolt, & Fu, 2006a; Rijmen, 2009, 2010). In fact, it

has been shown that the bifactor model is a generalization of the testlet model which is equivalent to the second-order model (De la Torre & Song, 2009; Li et al., 2006a; Rijmen, 2009, 2010). There are other approaches that take into account local dependence among items in item response models (e.g., Braeken, Tuerlinckx, & De Boeck, 2007; Ip, 2002, 2010).

In this article, we present a generalization of the multiple-group bifactor model for assessing differential item functioning (DIF) for testlet-based tests. Unlike any previous work on multiple-group bifactor or testlet models, our model relaxes the assumption of complete independence of all (specific and general) dimensions. In our approach, it is only assumed that the specific dimensions are conditionally independent, given the general dimension. Using a graphical model framework, we show that the assumption of conditional independence suffices to reduce the dimensionality of the integration from $(K + 1)$ to two.

Our model allows the means and variances of all dimensions and the correlations among the dimensions to differ between groups. Li et al. (2006a) and Cai, Yang, and Hansen (2011) proposed similar multiple-group testlet and bifactor models, respectively, but assumed independence of the latent variables. For the bifactor model, Fukuhara and Kamata (2011) allowed only the mean of the general dimension to differ between groups, whereas Wang and Wilson (2005) and Wang, Bradlow, Wainer, and Muller (2008) did not allow for any group differences in the latent variable distributions.

Allowing for a different distribution of ability is especially important when it comes to the assessment of DIF. In unidimensional IRT approaches to DIF assessment, it is generally acknowledged that at least the mean of the ability distribution should be allowed to differ between groups to allow for "impact" of group on ability when testing group-by-item interactions. Ainsworth (2007) demonstrated that ignoring group differences in the means of the secondary dimensions in the multiple-group bifactor model can lead to detection of DIF when none exists. In this article, we investigate whether, in addition to the means of the specific dimensions, it is important to allow the mean of the general dimension, the variances of all dimensions, and the correlations among the specific dimensions to differ between groups. We perform a simulation study that shows that our general approach is more appropriate for assessing DIF when group differences in all of these aspects of the latent variable distribution are present. For example, misspecifying the model by assuming uncorrelated dimensions can lead to biased DIF estimation.

Using the graphical model framework, we derive an efficient method for marginal MLE that does not require high-dimensional numerical integration. This method is generally faster than other ML methods and applicable to various settings without regard to the number of dimensions, items and examinees, and the types of item responses. An alternative estimation approach would be the use of Markov chain Monte Carlo techniques. However, these techniques tend to be slow, and it is difficult to monitor convergence to a stationary distribution and

to specify vague priors for variance parameters (e.g., Browne & Draper, 2006; Natarajan & Kass, 2000). Contemporary Bayesian approaches to multiple-group bifactor models do not appear to be feasible for very large problems as discussed by Wang et al. (2008) and Sinharay and Dorans (2010). We demonstrate the utility of our efficient approach by applying it to a realistically large dataset from the Progress in International Reading Literacy Study (PIRLS) study with 10 testlets, 126 binary and polytomous items, and over 5,000 examinees.

The rest of this article is organized as follows. We first describe the proposed multiple-group bifactor model and the estimation method. We then discuss methods for DIF assessment for testlet-based tests. In the next section, the performance of our model for assessing DIF will be compared with alternative models in a simulation study. An empirical study follows to illustrate the use of the proposed method for testing DIF in a large-scale assessment. We end with some final remarks.

Proposed Model

In this section, we describe the proposed multiple-group bifactor model and how it can be used for the assessment of DIF. We begin with the multiple-group unidimensional model and then introduce the multiple-group bifactor model.

Multiple-Group Unidimensional DIF Model

Suppose there are H groups to be compared. DIF can be investigated using the two-parameter logistic (2PL) multiple-group item response model

$$\text{logit}(\Pr(y_{j(h)i} = 1|\theta_{j(h)})) = \alpha_i(\theta_{j(h)}\sigma_h + \mu_h) + \beta_i + \delta_{ih}, \quad (1)$$

where $y_{j(h)i}$ is the response by person j in group h ($h = 1, \ldots, H$) to item i, and α_i and $-\beta_i$ are the item discrimination and difficulty parameters, respectively. $\theta_{j(h)}$ is a latent variable that follows a standard normal distribution, $\theta_{j(h)} \sim N(0, 1)$. Hence, the ability $\theta^*_{j(h)}$ for person j in group h is $\theta^*_{j(h)} = \theta_{j(h)}\sigma_h + \mu_h$, with group-specific mean μ_h and standard deviation σ_h.

We call the first group, $h = 1$, the reference group. To identify the model, the reference group has a standard normal distribution with zero mean and unit standard deviation, $\mu_1 = 0$ and $\sigma_1 = 1$. The parameter δ_{ih} represents uniform DIF. In the reference group, $\delta_{i1} = 0$, so that δ_{ih} represents the difference in item difficulty between the reference group and group h for item i. For the anchor items, which do not show DIF, $\delta_{ih} = 0$ for all h.

Multiple-Group Bifactor DIF Model

The multiple-group bifactor DIF model is a multidimensional extension of model (1). Suppose there are K item clusters. The multiple-group bifactor DIF model is then formulated as

$$\text{logit}(\Pr(y_{j(h)i(k)} = 1|\theta^*_{j(h)g}, \theta^*_{j(h)k})) = \alpha_{ig}\theta^*_{j(h)g} + \alpha_{ik}\theta^*_{j(h)k} + \beta_i + \delta_{ih}, \quad (2)$$

where $y_{j(h)i(k)}$ is the response by person j in group h to item i in testlet k, α_{ig} and α_{ik} are the item discrimination parameters on the general and the kth testlet-specific dimension, respectively, $-\beta_i$ represents the item difficulty parameter for item i (in the reference group if the item has DIF) and δ_{ih} represents DIF as before. The general and specific factors can be written as

$$\theta^*_{j(h)g} = \theta_{j(h)g}\sigma_{gh} + \mu_{gh}, \quad \theta^*_{j(h)k} = \theta_{j(h)k}\sigma_{kh} + \mu_{kh}, \quad (3)$$

where $\boldsymbol{\theta}_{j(h)} = (\theta_{j(h)g}, \theta_{j(h)1}, ..., \theta_{j(h)K})$ of dimension $(K + 1)$ follows a multivariate standard normal distribution, $\boldsymbol{\theta}_{j(h)} \sim N(\mathbf{0}, \mathbf{I})$. Hence, μ_{gh} and σ_{gh} are the group-specific mean and standard deviation of the general dimension, and μ_{kh} and σ_{kh} are the group-specific means and standard deviations of the specific dimensions. To identify the model, the means and standard deviations in the reference group for all dimensions are set to zero and one, respectively, $\mu_{g1} = 0$, $\mu_{k1} = 0$, $\sigma_{g1} = 1$, and $\sigma_{k1} = 1$ for all k. In addition, the set of anchor items should contain at least one item from each testlet.

When the mean μ_{kh} of the kth testlet dimension in group h is not zero, it is said that the kth testlet functions differently for that group than for the reference group and that we have differential testlet (or bundle) functioning (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Shealy & Stout, 1993). The model is similar to the multiple-group bifactor model that is being developed independently by Cai et al. (2011). It contains the multiple-group testlet model by Li et al. (2006a) as a special case in which, within each testlet, the item discrimination parameters on the specific dimension are proportional to the discrimination parameters on the general dimension.

Model (3) incorporates the common assumption that all dimensions are independent in the bifactor model. Gibbons and Hedeker (1992) relied on this assumption in their proof that the dimensionality of the integration can be reduced from $(K + 1)$ to two. Rijmen (2009) showed that this assumption can be relaxed. As we will show in the section on estimation, it is sufficient to assume conditional independence of the testlet-specific dimensions, given the general dimension. However, a model defined in that way needs K additional restrictions because the model is invariant under rotation of the latent variables (Rijmen, 2009). For the single-group model with normally distributed latent variables, the model can be identified by setting the correlations to zero. For the multiple-group model, it is sufficient to set the correlations to zero in the reference group only; for all other groups, the correlations between general and each of the specific dimensions can be estimated in addition to group-specific means and variances for the $(K + 1)$ dimensions.

We now propose a multiple-group bifactor model that relaxes the assumption of independent dimensions to the assumption that the specific dimensions are

conditionally independent, given the general dimension. The idea is to allow the specific dimensions to depend on the standardized general dimension θ_{gh}. Using a Cholesky decomposition, the latent variables in Equation (2) are modeled as

$$\theta^*_{j(h)g} = \theta_{j(h)g} c_{ggh} + \mu_{gh}, \quad \theta^*_{j(h)k} = \theta_{j(h)g} c_{gkh} + \theta_{j(h)k} c_{kkh} + \mu_{kh}, \tag{4}$$

where c_{ggh}, c_{kkh}, and c_{gkh} are the elements of the Cholesky decomposition $\boldsymbol{C}_h$ of the variance–covariance matrix ($\boldsymbol{\Sigma}_h$) for $\boldsymbol{\theta}^*_{j(h)} = (\theta^*_{j(h)g}, \theta^*_{j(h)1}, ..., \theta^*_{j(h)K})$. $\boldsymbol{C}_h$ is a lower triangular matrix of dimension $(K+1) \times (K+1)$ such that $\boldsymbol{\Sigma}_h = \boldsymbol{C}_h \boldsymbol{C}'_h$, and therefore $\boldsymbol{\theta}^*_{j(h)} = \boldsymbol{C}_h \boldsymbol{\theta}_{j(h)} + \boldsymbol{\mu}_h$.

If $\theta^*_{j(h)1}, ..., \theta^*_{j(h)K}$ are conditionally independent of each other given $\theta^*_{j(h)g}$, $\mathbf{C}_h$ has nonzero off-diagonal elements only in the first column (see, e.g., Sun & Sun, 2005). Then,

$$\boldsymbol{C}_h = \begin{bmatrix} c_{ggh} & 0 & 0 & \dots & 0 \\ c_{g1h} & c_{11h} & 0 & \dots & 0 \\ c_{g2h} & 0 & c_{22h} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{gKh} & 0 & 0 & \dots & c_{KKh} \end{bmatrix}. \tag{5}$$

From Equation (5), the covariance matrix $\boldsymbol{\Sigma}_h$ of $\boldsymbol{\theta}^*_{j(h)}$ can be reconstructed as

$$\boldsymbol{\Sigma}_h = \begin{bmatrix} c^2_{ggh} & c_{ggh}c_{g1h} & c_{ggh}c_{g2h} & \dots & c_{ggh}c_{gKh} \\ c_{g1h}c_{ggh} & c^2_{g1h} + c^2_{11h} & c_{g1h}c_{g2h} & \dots & c_{g1h}c_{gKh} \\ c_{g2h}c_{ggh} & c_{g2h}c_{g1h} & c^2_{g2h} + c^2_{22h} & \dots & c_{g2h}c_{gKh} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{gKh}c_{ggh} & c_{gKh}c_{g1h} & c_{gKh}c_{g2h} & \dots & c^2_{gKh} + c^2_{KKh} \end{bmatrix}. \tag{6}$$

The diagonal elements of $\boldsymbol{\Sigma}_h$ represent the variances, and the off-diagonal elements represent the covariances.

The model is identified by setting the means in the reference group for all dimensions to zero, all variances to one, and all covariances between dimensions to zero, $\mu_{g1} = 0$, $\mu_{k1} = 0$ for all k, $c_{gg1} = 1$, $c_{kk1} = 1$ for all k, and $c_{km1} = 0$ for all pairs k and m with $k \neq m$. The model can be extended for polytomous items in a straightforward way by choosing a suitable link function (Fahrmeir & Tutz, 2001). For instance, if we use a partial credit model (Masters, 1982) with an adjacent category logit link, DIF can then be investigated by allowing step difficulties for each category to differ between groups.

Returning to binary responses, δ_{ih} represents the DIF effect for item i, the difference in item difficulty between the reference group and group h after allowing for distributional differences in all latent trait dimensions. DIF can be assessed by

either a Wald test or likelihood ratio (LR) test of the null hypothesis that $\delta_{ih} = 0$. Both tests are asymptotically equivalent. The Wald test is based on a quadratic approximation to the log-likelihood function. When the log-likelihood is nonquadratic, the discrepancy between the LR and Wald test increases with the distance between the ML estimate of a parameter and its value under the null hypothesis (e.g., Fahrmeir & Tutz, 2001).

Alternative Models

As a comparison to the general approach in Equation (4), we present several models that progressively incorporate stronger assumptions on the equivalence of the ability distributions across groups. All models make the same assumption regarding the conditional response probabilities $\Pr(y_{j(h)i(k)} = 1|\theta^*_{j(h)g}, \theta^*_{j(h)k})$ as in Equation (2). The first alternative is the model with independent dimensions as formulated in Equation (3). This model can be obtained from the multiple-group bifactor DIF model of Equation (4) by setting all c_{gkh} in Equation (4) to zero (then, c_{ggh} in Equation (4) corresponds to σ_{gh} in Equation (3) and c_{kkh} to σ_{kh}). We call this model the multiple-group bifactor "independence" model. From now on, Model (4) will be referred to as the "correlation" model.

The second alternative model takes into account group differences in the general dimension only. This model can be written as

$$\theta^*_{j(h)g} = \theta_{j(h)g}\sigma_{gh} + \mu_{gh}, \quad \theta^*_{j(h)k} = \theta_{j(h)k}. \tag{7}$$

We call Model (7) the multiple-group bifactor "main" model. Finally, we can think of a model assuming no distributional differences in any dimensions,

$$\theta^*_{j(h)g} = \theta_{j(h)g}, \quad \theta^*_{j(h)k} = \theta_{j(h)k}. \tag{8}$$

We call Model (8) the multiple-group bifactor "no impact" model. We also consider two variant models allowing for mean differences between groups in Models (2) and (7) but not for differences in the variances, $\sigma_{gh} = 1$, $\sigma_{gk} = 1$ for all k. We call these models the multiple-group "independence mean" model and "main mean" model, respectively. Table 1 summarizes the six bifactor models and their assumptions.

Estimation

In general, marginal MLE of multidimensional models is computationally very intensive because of the high-dimensional integration over the latent variables. For instance, if Gauss–Hermite quadrature is used to approximate the marginal likelihood function over the latent variables, a $(K + 1)$ dimensional IRT model would require $q^{(K+1)}$ point quadrature with q quadrature points per dimension. The computation becomes intractable for even a small number of quadrature

TABLE 1
Summary of the Six Multiple-Group Bifactor Models According to the Assumptions That Are Relaxed Compared With the "No Impact" Model

Model		Corr.	General Dimension		Specific Dimensions	
Label	Name	$c_{gkh} \neq 0$	$\mu_{gh} \neq \mu_{g1}$	$\sigma_{gh} \neq \sigma_{g1}$	$\mu_{kh} \neq \mu_{k1}$	$\sigma_{kh} \neq \sigma_{k1}$
1	No impact ()					
2m	Main mean (1)		√			
2	Main (1,2m)		√	√		
3m	Independence mean (1,2m)		√		√	
3	Independence (1,2m,2,3m)		√	√	√	√
4	Correlation (1,2m,2,3m,3)	√	√	√	√	√

Note: $\neq$ means that the model permits inequality; under the model name, labels for models nested in the model are given in parentheses.

points per dimension; for example, for five dimensions with eight quadrature points, it requires 32,768 ($= 8^5$) function evaluations.

Fortunately, the conditional independence relations that are assumed in the bifactor model can be exploited for MLE. This dimension reduction technique was first described by Gibbons and Hedeker (1992) and used to estimate a single-group bifactor model under the conditions of normally and independently distributed latent variables and for the probit link. The limiting conditions were due to the fact that these authors relied on properties of the multivariate normal distribution. The multiple-group bifactor model by Cai et al. (2011) was based on Gibbons and Hedeker's (1992) formulation.

Glas, Wainer, and Bradlow (2000) derived a marginal ML algorithm for a single-group logistic (3PL) testlet model, which was based on a similar dimension reduction technique to Gibbons and Hedeker (1992). Li, Bolt, and Fu (2006b) presented a multiple-group testlet model based on Glas et al.'s formulation.

All mentioned approaches rely on the assumption that the general and specific dimensions are independent of one another. Using a graphical model framework, Rijmen (2009) showed that the assumption of independently distributed latent variables can be relaxed to conditional independence of the specific dimensions, given the general dimension. In addition, one does not have to rely on properties of the normal distribution, so that the result remains valid under any link function other than the probit function, and for latent variables that are not normally distributed. For the estimation of the multiple-group bifactor model proposed in this article, we use the efficient full-information ML method described in Rijmen (2009).

The method relies on a modification of the expectation–maximization (EM) algorithm. In the E-step of the modified EM algorithm, the graphical model framework is used to perform computations on subsets of variables that are conditionally independent (e.g., Lauritzen, 1995). The M-step proceeds in the same way as the traditional EM algorithm to update parameter estimates. A detailed

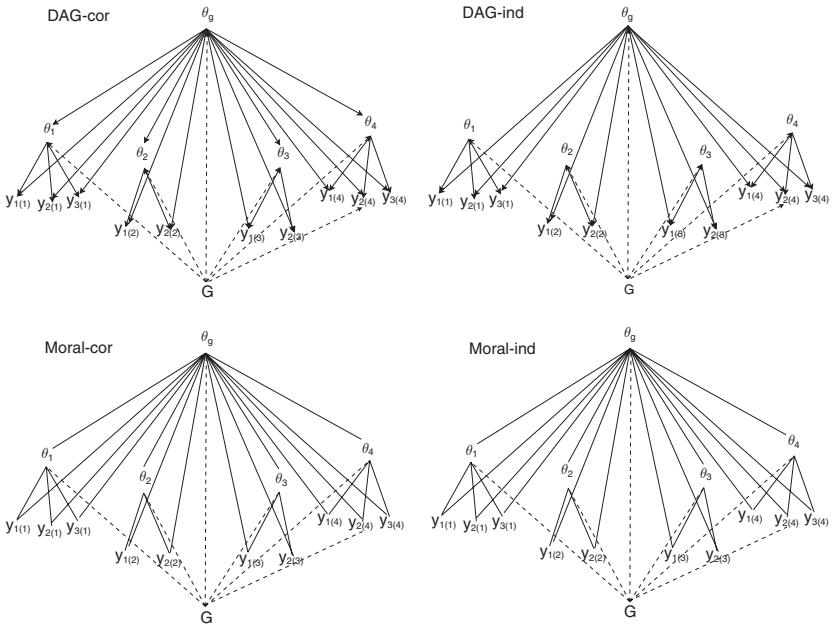

FIGURE 1. *Directed acyclic graphs (DAG) and moral graphs (Moral) for the multiple-group bifactor "correlation" and "independence" models for person* j *in group* h *with differential item functioning (DIF) for* $y_{2(4)}$. *The upper panel shows DAGs and the lower panel shows the moral graphs for the correlation model (left) and the independence model (right). Group effects are shown as dashed lines.*

account of the EM algorithm within a graphical model framework is given in Rijmen, Vansteelandt, and De Boeck (2008). Thus a brief description below will suffice for the purpose of this article.

In the first step, a directed acyclic graph (DAG) is created for the statistical model. The directed graph consists of nodes (variables) and arcs (directed edges; Hunter, Battersby, and Whitehead, 1986) to represent the conditional dependence/independence relations implied in the statistical model (Rijmen et al., 2008). The DAG is then transformed to an undirected graph, called the moral graph. The moral graph can be obtained by adding an undirected edge between all nodes with a common child that are not joined yet and dropping directions from all edges.

Figure 1 illustrates directed graphs and moral graphs for the multiple-group correlation and independence models for person *j* in group *h* with DIF for Item 2 in Testlet 4.

In each graph, θ_g represents the general dimension for person *j*, θ_1 to θ_4 represent the four specific dimensions, and $y_{i(k)}$ represents the observed response to the *i*th item in the *k*th specific dimension. The variable *G* represents the known

group membership for person j and has arrows to all latent variables; in addition, G has an arrow to the response $y_{2(4)}$ for the item that has DIF.

In the DAG for the correlation model, there are no direct links between θ_1 and θ_4, but they are connected via θ_g, implying conditional independence of θ_1 to θ_4 given θ_g. The same logic can be applied to the item responses. The item responses in $\boldsymbol{y}_k$ are conditionally independent given θ_k, and responses to items associated with different specific dimensions are conditionally independent given θ_g. Notice that in the DAG for the independence model, there are no direct edges from θ_g to θ_1 to θ_4, which implies independence among θ_g and θ_1 to θ_4.

Most importantly, when transforming the DAG into a moral graph, the same moral graph is obtained for both the correlation and the independence model. The reason is that, for the "independence model," edges are added between θ_g and each of the θ_k because θ_g has children in common with each of them. This implies that the more general correlation model will be estimated with the same computational efficiency as the independence model so that the independence assumption can be relaxed without extra computational cost.

The moral graph is further transformed into a so-called junction tree. In a junction tree, nodes correspond to sets of variables. The variables in one node are conditionally independent of the variables in another node, given the variables that are in the intersection of both nodes. For the bifactor model, there are K nodes in the junction tree. Each node consists of the general dimension θ_g, the specific dimension θ_k, and the item responses $\boldsymbol{y}_k$ pertaining to testlet k. In the E-step of the modified EM algorithm, the conditional independence relations between the latent variables are maximally exploited by carrying out a sequence of local computations defined on the sets of variables that constitute the nodes in the junction tree. It is important to realize that the modified E-step does not differ from a traditional E-step in what is computed, but only in how it is computed. As for any other EM-algorithm, the expected complete data log-likelihood is computed in the modified E-step, conditional on the observed responses. However, rather than computing the complete data log-likelihood for any combination of the $(K + 1)$ latent variables and subsequently marginalizing over the joint latent space, the expected complete data log-likelihood can be obtained through a sequence of computations that involve only two latent variables at a time. See Appendix B of Rijmen (2009) for details.

The EM algorithm described in this section was implemented in the Bayesian Networks with Logistic Regression Nodes (BNL) MATBLAB toolbox (Rijmen, 2006). The BNL toolbox can be obtained by contacting the second author.

The EM algorithm does not provide the observed or expected information matrix evaluated at the parameter estimates as a by-product. Hence, standard errors of the parameter estimates are not automatically obtained. Several procedures have been developed to obtain standard errors when using the EM algorithm. An overview can be found in, for example, McLachlan and Krishnan

(1997). In BNL, the observed information matrix evaluated at the ML estimates is approximated by the empirical information matrix (Meilijson, 1989). The empirical information matrix is obtained as the sum, over cases, of the outer product of the individual contributions to the gradient of the log-likelihood function.

In order to check for local maxima, a model was run 10 times from relatively diffuse starting values and its log-likelihood and parameter estimates were the same across the multiple runs. We also compared BNL with the ML software PROC NLMIXED in SAS (Wolfinger, 1999) and gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2005) in Stata, and with the Bayesian software WinBUGS (Spiegelhalter, Thomas, Best, & Gilks, 1996) using a small simulated dataset with 12 items, 4 testlets, and 1,000 examinees. The differences in the parameter estimates were mostly negligible between software. In terms of computation time, however, BNL was much faster than other software; for example, for the proposed correlation model, BNL took 20 minutes, SAS took one and a half days, and WinBUGS took one and a half hours for 3,000 iterations and three chains. For the main model, gllamm took 7 hours and BNL 4 minutes. For a simulated dataset with 10,000 examinees, 40 items, and 4 testlets, WinBUGS took nearly 2 days to run 1,000 iterations (three chains), whereas BNL took one and a half hours with 20 quadrature points. Detailed results on this comparison and the code and dataset can be found in the supplementary material provided on the JEBS website (http://jeb.sagepub.com/).

DIF Assessment for Testlet-Based Tests

As we have seen in the previous section, DIF can be modeled as the interaction effect between an item and group membership in an IRT model (e.g., Swaminathan & Rogers, 1990, Swanson, Clauser, Case, Nungester, & Featherman, 2002; Van den Noortgate & De Boeck, 2005). The items for which no interactions are included are called anchor items. If the anchor items include items that show DIF, this will lead to biased parameter estimates (including the DIF effect) for the studied item. An iterative item purification procedure is therefore recommended to ensure that the anchor set does not contain items that show DIF (see e.g., Holland & Thayer, 1988).

For modeling DIF, it is crucial to properly take into account the overall group differences, referred to as ''impact'' (see e.g., Millsap & Everson, 1993). Otherwise, the DIF effect may merely reflect group differences in the ability distributions, and one would erroneously conclude that an item shows DIF and hence needs further investigation or should even be dropped from the assessment altogether. More seriously, some items that do show DIF may not be flagged and inferences may be biased and thus unfair for certain subgroups.

How to take into account overall differences in the ability distribution depends on the statistical model being in place. For unidimensional IRT models, group differences in the ability distribution are typically taken into account by allowing for a different mean and variance for each group. A similar approach can be

followed for multidimensional IRT models including the bifactor model; since the latent space is now multidimensional, group differences in the multidimensional space must be properly taken into account for DIF assessment (for an in-depth discussion, see e.g., Meredith, 1993).

Recently, Wang et al. (2008) presented a DIF assessment method using a Bayesian testlet model. They allowed the studied item's parameters to differ between two examinee groups, using the rest of the test as the anchor, similarly to the LR approach by Thissen, Steinberg, and Wainer (1988). Unfortunately, Wang et al. (2008) assumed that the ability distributions are the same for the focal and reference groups. Similarly, Fukuhara and Kamata (2011) extended the bifactor model for DIF assessment and estimated it using WinBUGS but allowed only the mean of the general dimension to differ between groups. Furthermore, the methods by Wang et al. (2008) and Fukuhara and Kamata (2011) do not appear to be feasible for practical DIF assessment. For instance, in Wang et al. (2008), assessing DIF in a dataset on 903 examinees took about 3 hours, and Sinharay and Dorans (2010) pointed out that it would take a few days to apply their procedure to a test such as the SAT with nearly half a million students.

In an ML framework, Li et al. (2006b) and Cai et al. (2011) presented the multiple-group testlet and bifactor models, respectively. Ainsworth (2007) presented a similar multiple-group bifactor model viewed as a multiple indicator multiple cause model. However, these models assumed independence of the latent variables. The model we propose encompasses these approaches but is more general in that it allows for any correlation structure among the latent variables that is consistent the conditional independence of the specific dimensions given the general dimension.

Another model-based approach to DIF assessment is the Simultaneous Item Bias Test (SIBTEST; Shealy, 1989; Shealy & Stout, 1991, 1993). The underlying idea of this multidimensional IRT approach is that there are one or more nuisance dimensions that act as a source of bias with regard to inferences concerning the intended dimensions. However, the performance of SIBTEST has not been established yet for testlet-based tests. Lee, Cohen, and Toro (2009) investigated SIBTEST and Poly-SIBTEST for DIF detection in testlets-based tests, and they found that Type I error rates of both tests increase as the sample size increases and the number of items decreases.

Alternatives to model-based approaches are the Mantel–Haenszel (MH) method (Holland & Thayer, 1988) and area measures. The MH method is based on stratifying the sample according to measures of ability (often observed total scores), and testing for associations between item responses and group membership within ability strata. However, for multidimensional tests, the MH method requires that strata be defined by multivariate matching which could be cumbersome and arbitrary (Clauser, Nungester, Mazor, & Ripkey, 1996; Mazor, Kanjee, & Clauser, 1995). Further, this approach involves heavy data requirements because of the need to cross the levels of all variables that go into the match (Dorans & Holland, 1993). For this reason, the logistic regression procedure, regressing item responses

on the measures of ability and group, was recommended for DIF analysis for multivariate tests (e.g., Clauser et al., 1996; Mazor et al., 1995). In addition, Wang et al. (2008) showed that the MH method appeared unstable for items with extreme difficulty values. Others have also found that the MH method can be misleading when the true model is a two-parameter model (Bolt & Stout, 1996; DeMars, 2009; Meredith, 1993; Roussos & Stout, 1996).

Area measures can also be used for DIF detection for multidimensional tests (e.g., Raju, 1988). The differential functioning of items and tests (DFIT) procedure of Oshima, Raju, and Flowers (1997) would be an example. This method requires prior estimation of the item response functions (IRFs) for focal and reference groups. The measure of DIF is then a weighted average of squared differences between the estimated focal and reference group IRFs for the studied item. However, as far as we are aware, the performance of the DFIT method for the bifactor model has not been examined yet.

Simulation Study

A simulation study was carried out to examine the performance of the proposed multiple-group bifactor DIF model and to assess the effects of misspecification of the distribution of the latent variables on the estimation of DIF.

Design

We considered a 40-item test composed of four testlets with 10 items within each testlet. We generated data for 10,000 test takers, evenly divided into a focal group and a reference group. There were five conditions corresponding to different specifications of the latent variable distributions. For each of these five specifications, two different DIF sizes for δ_{ih} were used for Item 37 (7th item of the fourth testlet): 0.2 for small DIF and 0.5 for medium–large DIF. In all models, there were 40 difficulty parameters, 80 discrimination parameters, and 1 DIF parameter. We assumed a multivariate normal distribution for the ability vector, with zero means and an identity covariance matrix in the reference group. In the focal group, we denote the five-dimensional mean vector $\boldsymbol{\mu}_f$ and the 5×5 covariance matrix $\boldsymbol{\Sigma}_f$. The five conditions were (using the names and labels of Table 1):

No impact (1): No distributional differences in any dimension (121 parameters), $\boldsymbol{\mu}_f = \mathbf{0}$ and $\boldsymbol{\Sigma}_f = \mathbf{I}$.

Main (2): Differences in the mean and standard deviation of the general dimension only (123 parameters), $\boldsymbol{\mu}_f = (0.5, 0, 0, 0, 0)'$ and $\boldsymbol{\Sigma}_f$ has diagonal elements $(1.2^2, 1.0, 1.0, 1.0, 1.0)$.

Independence (3): Differences in the means and standard deviations of all dimensions, with dimensions that are independent of each other (131 parameters), $\boldsymbol{\mu}_f = (0.5, 0.2, 0.5, 0.7, -1.0)'$ and $\boldsymbol{\Sigma}_f$ has diagonal elements $(1.2^2, 0.5^2, 0.7^2, 0.9^2, 1.2^2)$.

Correlation, low (4): Low correlations between the general dimension and the specific dimensions (135 parameters), $\boldsymbol{\mu}_f = (0.5, 0.2, 0.5, 0.7, -1.0)'$ and $\boldsymbol{\Sigma}_f$, diagonal elements $(1.2^2, 0.5^2, 0.7^2, 0.9^2, 1.2^2)$, and correlations (0, 0.2, 0.3, 0.4) between the general dimension and the four specific dimensions.

Correlations, high (4): Moderate to high correlations between the general dimension and the specific dimensions (135 parameters), $\boldsymbol{\mu}_f = (0.5, 0.2, 0.5, 0.7, -1.0)'$ and $\boldsymbol{\Sigma}_f$, had diagonal elements $(1.2^2, 0.5^2, 0.7^2, 0.9^2, 1.2^2)$, and correlations (0, 0.5, 0.6, 0.7) between the general dimension and the four specific dimensions.

Table 2 lists the true values for the item discrimination parameters in the general and the four testlet-specific dimensions (α_g, α_1, ..., α_4) and the true values for minus the item difficulty parameters (β_i).

All six models previously discussed and shown in Table 1 were fitted to the datasets generated under the first four conditions. For the "correlation, high" model, only the three more complex models (independence mean, independence, and correlation) were fitted. The fitted models correspond directly to the generating models, except that two additional models were fitted: the main mean model (2m) with 122 parameters and the independence mean model (3m) with 126 parameters, which are constant-variance versions of the main model (2) and independence model (3), respectively. A given fitted model is misspecified if, looking this model up in Table 1, the generating model is not given in parentheses under the model name. All models were fitted using Gaussian quadrature with 20 quadrature points per dimension. We used a convergence criterion of 10^{-5} for absolute differences in parameter estimates and relative changes in the value of the log-likelihood function. The algorithm stopped when both criteria were reached.

Results

We first consider the error (estimated minus true values) for the DIF parameter estimates under each condition. The boxplots in Figures 2 and 3 summarize the results for small and medium–large DIF, respectively.

Each panel corresponds to a fitted model, in order of increasing complexity of the models, and each boxplot within a panel corresponds to one of four or five data generating conditions. When the fitted model is misspecified (i.e., the generating model is not nested in the fitted model), the boxplot is shaded.

As expected, the correlation model performs very well in all conditions, showing less error than the misspecified constrained models. Interestingly, for a given generated model, the estimates from the correlation model do not tend to be more variable than those from the more constrained correctly specified models, suggesting that the correlation model could be used as the default model. Assuming equal testlet means (first row of figures) resulted in severe underestimation of the DIF effect when the data were generated from the independence and correlation models because in these models, the testlet mean is -1 for the focal group (and 0 for the reference group) for the testlet that includes the DIF item. In all

TABLE 2
True Values for Item Parameters

	α_g	α_1	α_2	α_3	α_4	β
Item 1	1.5	1.0	0	0	0	−1.0
Item 2	0.7	0.5	0	0	0	−0.5
Item 3	1.2	1.5	0	0	0	0
Item 4	2.0	0.2	0	0	0	0.5
Item 5	2.0	0.8	0	0	0	1.0
Item 6	1.5	1.2	0	0	0	−1.0
Item 7	1.2	1.5	0	0	0	−0.5
Item 8	2.0	0.2	0	0	0	0
Item 9	2.0	0.8	0	0	0	0.5
Item 10	0.7	0.5	0	0	0	1.0
Item 11	1.5	0	1.0	0	0	−1.0
Item 12	2.0	0	0.2	0	0	−0.5
Item 13	0.8	0	2.0	0	0	0
Item 14	0.7	0	0.5	0	0	0.5
Item 15	1.2	0	1.5	0	0	1.0
Item 16	1.5	0	1.2	0	0	−1.0
Item 17	2.0	0	0.8	0	0	−0.5
Item 18	0.7	0	0.5	0	0	0
Item 19	1.2	0	1.5	0	0	0.5
Item 20	2.0	0	0.2	0	0	1.0
Item 21	1.5	0	0	1.0	0	−1.0
Item 22	0.7	0	0	0.5	0	−0.5
Item 23	1.2	0	0	1.5	0	0
Item 24	2.0	0	0	0.2	0	0.5
Item 25	2.0	0	0	0.8	0	1.0
Item 26	1.5	0	0	1.2	0	−1.0
Item 27	1.2	0	0	1.5	0	−0.5
Item 28	2.0	0	0	0.2	0	0
Item 29	2.0	0	0	0.8	0	0.5
Item 30	0.7	0	0	0.5	0	1.0
Item 31	1.5	0	0	0	1.0	−1.0
Item 32	2.0	0	0	0	0.2	−0.5
Item 33	0.8	0	0	0	2.0	0
Item 34	0.7	0	0	0	0.5	0.5
Item 35	1.2	0	0	0	1.5	1.0
Item 36	1.5	0	0	0	1.2	−1.0
Item 37	2.0	0	0	0	0.8	−0.5
Item 38	0.7	0	0	0	0.5	0
Item 39	1.2	0	0	0	1.5	0.5
Item 40	2.0	0	0	0	0.2	1.0

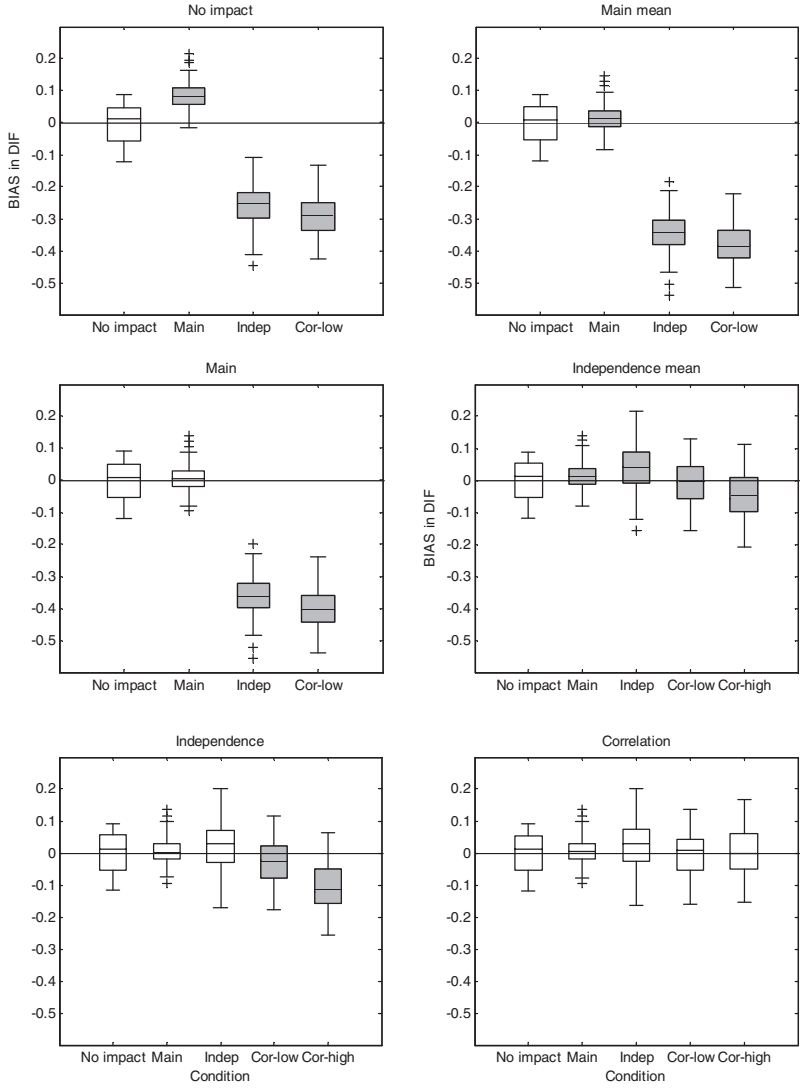

FIGURE 2. *Boxplots of error* ($\hat{\delta}_{ih} - \delta_{ih}$) *in the differential item functioning (DIF) estimates when DIF = 0.2. Panels correspond to fitted models and generating models are shown on the* x*-axes. For misspecified models, boxplots are shaded gray.*

conditions, the correctly specified models perform well. Although misspecified, the independence mean and independence models show good recovery under the "correlation, low" condition but produce downward bias under the "correlation, high" condition. These results are consistent across the two DIF conditions.

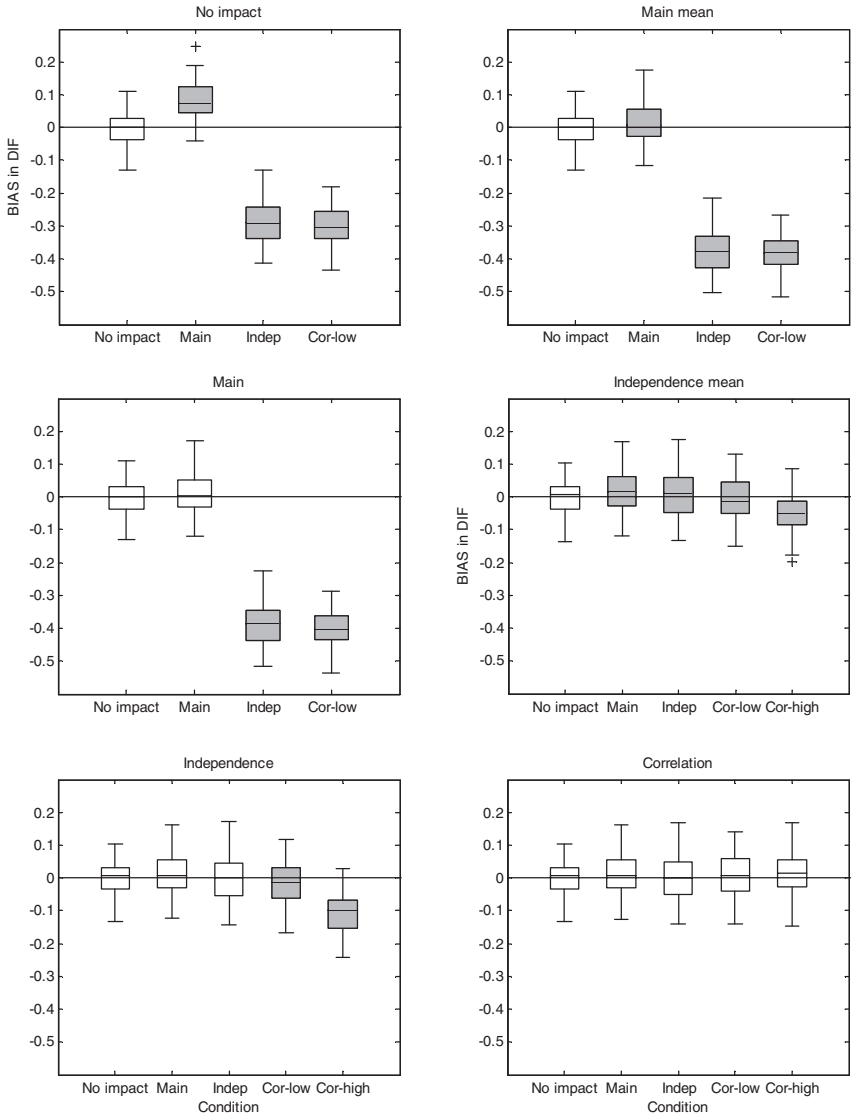

FIGURE 3. *Boxplots of error ($\hat{\delta}_{ih} - \delta_{ih}$) in differential item functioning (DIF) estimates when DIF = 0.5. Panels correspond to fitted models and generating models are shown on the* x*-axes. For misspecified models, boxplots are shaded gray.*

As advocated by Skrondal (2000), we performed a statistical analysis of the simulation results, separately for small and medium–large DIF. For each true model for the two DIF conditions, we treated the differences between the

estimated and true DIF parameter for the 50 simulated datasets and three or six fitted models as the response variable and used repeated-measures multivariate analysis of variance (MANOVA) to test the difference in the mean error (i.e., the bias) between fitted models. The overall F test (based on Wilk's λ) was significant at the 1% level for each condition except for the no impact condition (where all models are correctly specified). For all conditions except the no impact condition, we used contrasts based on the repeated measures MANOVA to test the difference in bias between each fitted model and the true model. When the model was not misspecified, we never found a significant difference at the 1% level in either DIF condition (except for the correlation model fitted to data from the independence model under the high DIF condition). When the model was misspecified, we always found a significant difference in both DIF conditions.

Now we discuss bias of the estimates for the other model parameters. To summarize the results for a large number of item parameters, we averaged the absolute values of the estimated bias across items sharing the same true item difficulty (five different values), and similarly for the discrimination parameters for the general dimension (five different values) and specific dimensions (eight different values). Figure 4 shows the results when DIF $= 0.2$.

Each panel represents a fitted model, from the no impact to the independence models. The correlation model is not presented because its bias is negligible in all data-generating conditions. The rows represent the item discrimination parameter for the general dimension (α_g), the item discrimination parameter for the testlet-specific dimensions (α_k), and the item easiness parameter (β). In each panel, the average absolute bias estimates are plotted against the true values. Different line patterns represent different data-generating conditions.

In general, all fitted models show some degree of bias if the data were generated from a more complex model, except for the difficulty parameters that do not appear to be affected by misspecification of the variances and correlations. Estimates show negligible bias if the data were generated from the same or a less complex model. The pattern of bias differs somewhat between parameters. For α_k, the average bias tends to increase as the true value increases. In contrast, the average bias for β is relatively constant. For α_g, we notice that there are some peaks in the absolute bias for α_g. These peaks were observed where the true value for α_g is 0.8. There are 2 items (Items 13 and 33) that have 0.8 as true value. Item 33 shows the biggest bias and Item 13 shows the fifth biggest bias among 40 items. The peaks are especially high when the data were generated from the correlation, high models (in the fourth and fifth columns). This shows that ignoring the correlation structure of the latent variables can result in a particularly large bias for α_g. For brevity, we do not present results for the means, standard deviations, and correlations.

We performed a sequence of LR tests to select among the fitted models for a given simulation condition. To obtain a sequence of nested models, we considered only (from least to most constrained) the correlation, independence, main,

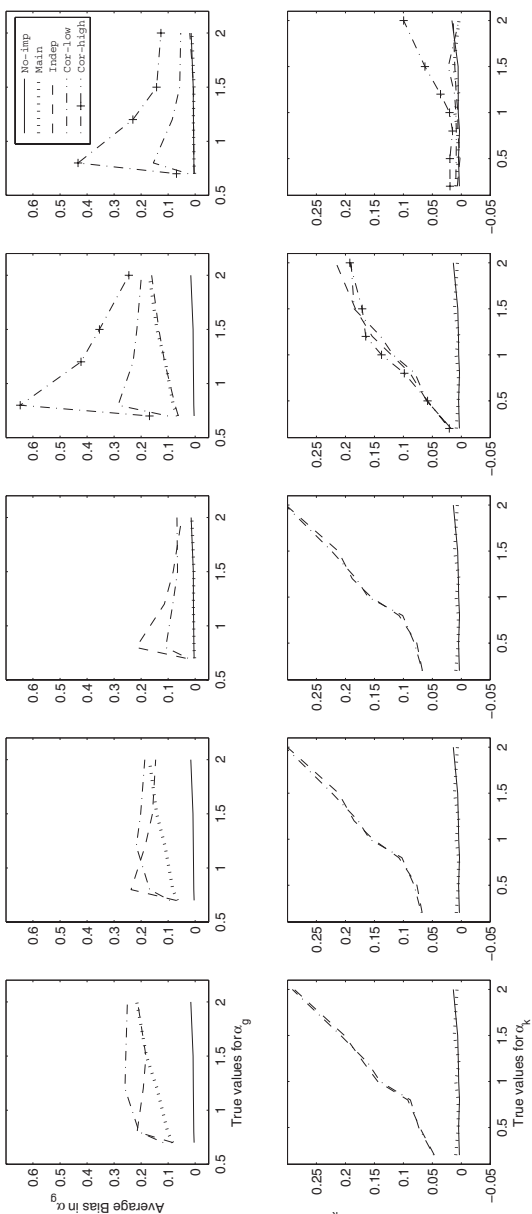

FIGURE 4. *The average of the absolute values of the bias averaged over the items that have the same parameter values for* α_g, α_k, *and* β. *Panels correspond to fitted models and different lines represent results under different data-generating conditions. In each plot, the* x*-axis represents the true parameter value and the* y*-axis represents absolute bias.*

and no impact models. The test began by comparing the least constrained (correlation) model with the next, more constrained model in the sequence (independence model); if the test was not rejected, the test proceeded to compare the independence with the main model and finally the main with the no impact model. At the 5% level of significance and when DIF = 0.2, the true model was chosen 90%, 92%, 98%, and 100% of the time under the no impact, main, independence, and correlation models, respectively. A similar pattern was observed under the other DIF condition.

We now present the root mean squared error (RMSE) of the parameter estimates. For simplicity, we computed the average RMSE, as the square root of the average MSE over the 40 items. Figure 5 shows the results when DIF = 0.2.

Each panel corresponds to a fitted model, whereas the generating models are shown on the *x*-axes. For each generating model, three points are displayed, from left to right for α_g, α_k, and β. The symbols are shaded whenever the fitted model is misspecified. This figure shows that the RMSE tended to increase with increasing model misspecification and this tendency is observed for all conditions. The maximum RMSE decreases as the complexity of the fitted model increases (0.58, 0.47, 0.43, 0.25, 0.18, and 0.09 for the no impact, main mean, main, independence mean, independence, and correlation in order). This result indicates that the increase in variance due to fitting a more complex model is outweighed by the increase in bias due to fitting a model that is more constrained than the true model. Interestingly, as for the DIF parameter, the correlation model does not produce appreciably more variable estimates for the more constrained generating models than the respective true models.

Type I Error Rate for DIF Detection

We assessed the Type I error rate of the LR test for DIF detection for the correlation, high model. We generated 200 datasets using the same parameter values as in the simulation study described earlier, except that the DIF parameter was set to zero. We then fitted the correlation model with and without a DIF parameter for Item 37, yielding 13 LR statistics greater than the critical value at the 5% level for a chi-square distribution with 1 degree of freedom. The estimated Type I error rate is therefore 0.065 (=13/200) and this does not differ significantly from the nominal level of 0.05 (*p* value from the binomial test is 0.33). The exact 95% confidence interval for the Type I error is $[0.035, 0.109]$.

Empirical Study

The multiple-group bifactor model can be applied to large-scale educational assessments such as the National Assessment of Educational Progress, the International Adult Literacy Study, Trends in Mathematics and Science Study, the Programme for International Student Assessment, and the PIRLS. These

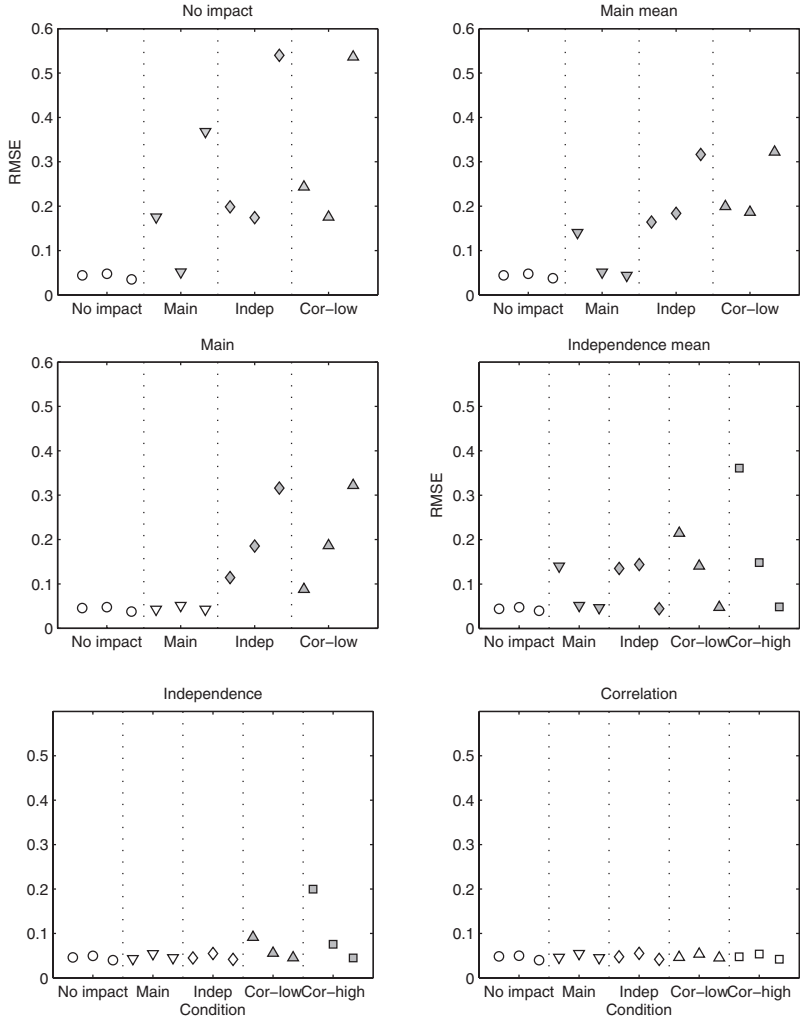

FIGURE 5. *Root mean squared error (RMSE) for the item parameters for the six fitted models, averaged over the 40 items. Panels correspond to fitted models and data-generating models are shown on the* x*-axis. The parameters are from left to right, the item discrimination parameter for the general dimension (α_g), the item discrimination parameter for the testlet-specific dimensions (α_k), and the item easiness parameter (β). For misspecified models, the symbols are shaded gray.*

assessments consist of a large number of testlets, and DIF is of great concern when reporting results for subgroups defined by gender, country, or other variables. In this article, we applied our model to the PIRLS data.

Data and Method

The 2006 U.S. PIRLS data were used to examine gender DIF. There were 5,187 students in total: 2,582 female students (the reference group) and 2,605 male students (the focal group). The assessment consists of 10 passages (testlets), five of which have a literary purpose and five of which have an informational purpose. Each passage is accompanied by 11 to 14 test items. About half of the items are multiple-choice items; the other half are short constructed-response items. There are 126 items in total, 92 binary items and 34 polytomous items (28 items with three categories and 6 items with four categories). The passages and items are distributed across 13 test booklets, and each student responded to the items of one booklet.

We first fitted the correlation model and investigated the covariance structure of the multiple latent traits (one general dimension plus 10 testlet dimensions). We then performed a series of LR tests to determine the most appropriate multiple-group bifactor model for this dataset. Following the procedure that was described in the simulation study, we started with the most general correlation model and simplified it until the reduced model was rejected. Ten quadrature points were used to fit these models. For testing statistical significance of the DIF effect, we used the same kind of LR tests as in the simulation study. Since this required fitting 126 models, we used five quadrature points (for the first 10 binary items, discrepancies between 5 and 10 quadrature points were minor).

Results

Figure 6 shows the estimated means, standard deviations, and correlations for the male students (the focal group) for the 10 testlet-specific dimensions in addition to the general dimension. The horizontal lines at zero represent the value for the female students (the reference group).

The estimated means, standard deviations, and correlations differ somewhat between males and females and across dimensions, with estimated means ranging from -1.06 to 0.5, standard deviations from 0.81 to 1.81, and correlations from -0.20 to 0.30.

The LR test indicates that the correlation model does not improve upon the independence model (LR $= 16.00$, $df = 10$, $p = .10$). However, the independence model fits significantly better than the independence mean model (LR $= 36.80$, $df = 11$, $p < .001$) and the main model (LR $= 126.4$, $df = 22$, $p < .001$). These test results suggest that the appropriate multiple-group bifactor model for the PIRLS data is the independence model, incorporating different means and standard deviations for all dimensions for the focal group, but not allowing for correlations among specific dimensions in the focal group. Therefore, the independence model was adopted to assess DIF. The total number of parameters in the independence model is 437 including 125 for $\boldsymbol{\alpha}_g$, 125 for $\boldsymbol{\alpha}_k$, 165 for $\boldsymbol{\beta}$, 11 for $\boldsymbol{\mu}_f$ and, 11 for the diagonal elements of $\boldsymbol{\Sigma}_f$.

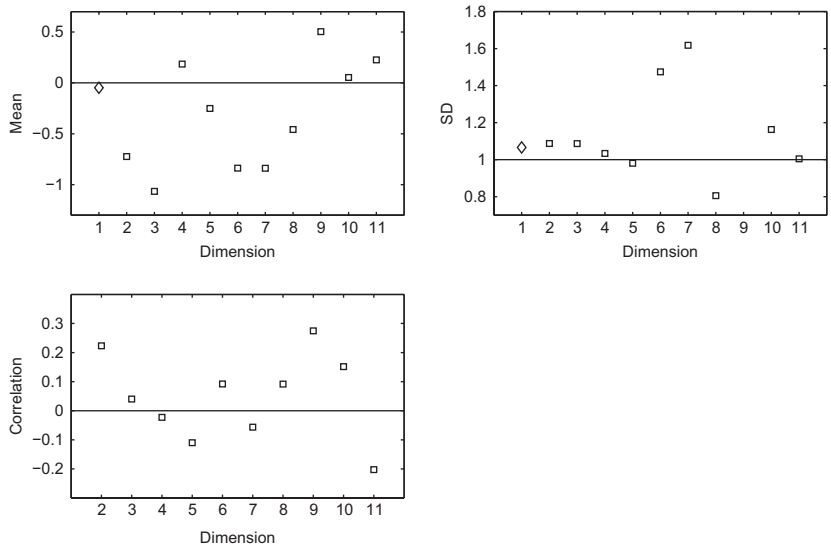

FIGURE 6. *Mean (μ), standard deviation (σ), and correlation (ρ) estimates for the Progress in International Reading Literacy Study (PIRLS) 2006 data. The* y*-axis represents the estimates and the* x*-axis represents the general dimension (Dimension 1) and 10 testlet-specific dimensions (Dimensions 2–11) in order.*

DIF was assessed separately for each item, treating the remaining items as anchor items. (Although it was not applied in this study, item purification is highly recommended in practice to find anchor items for detecting DIF for a studied item; see, e.g., Rogers & Swaminathan, 1993; Zumbo, 1999). Figure 7 presents estimates of the DIF parameter for the 91 binary items and 34 polytomous items. The DIF estimates range from −3.0 to 1.5 across binary and polytomous items.

Adjusting the significance level using the Bonferroni correction for multiple comparisons, $\alpha = .0004 (= .05/126)$, one binary item and one polytomous item were found to have statistically significant DIF parameters. In practice, it is recommended that the effect size be examined along with the p values especially when the sample size is large. For instance, Penfield's (2007) classification scheme for DIF effect sizes can be applied to interpret the DIF effect δ_{ih} in Equation (2). The cut-point for negligible DIF (Class A) is 0.426 for the absolute value on the logit scale. The cut-point for moderate DIF (Class B) is 0.638. Items with a DIF parameter greater than 0.638 are referred as large DIF (Class C) items. In Figure 7, three binary items and one polytomous item were identified as large DIF items.

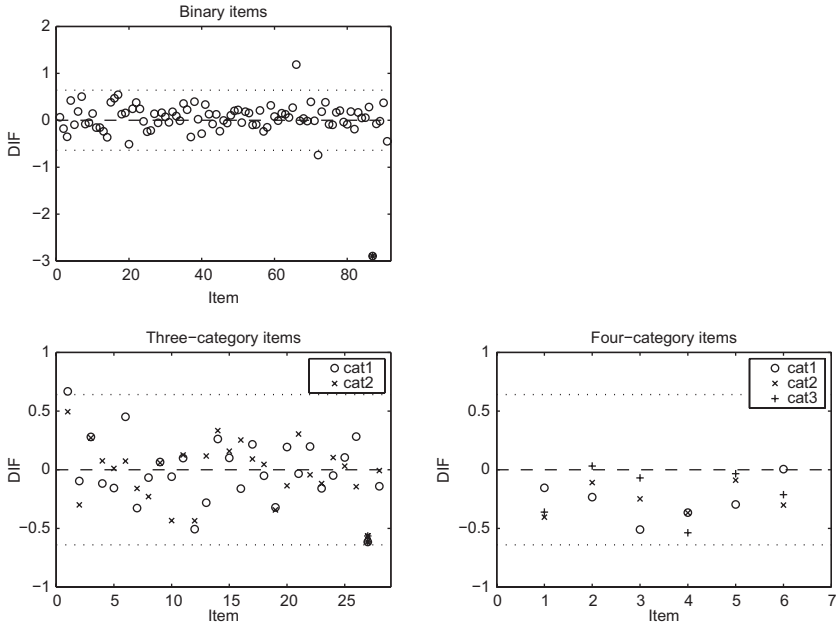

FIGURE 7. *Differential item functioning (DIF) estimates for the Progress in International Reading Literacy Study (PIRLS) 2006 data (○) for binary items and ○, ×, and + for the two or three categories of polytomous items). *represents a significant DIF effect (δ_i) for item* i *at the .04% level. The dotted horizontal lines are at 0 and ± 0.638, the cutoff between moderate and large DIF.*

Final Remarks

We presented a generalization of the multiple-group bifactor model for assessing DIF for testlet-based tests. The proposed model has four major features: First, it takes into account group differences in the multidimensional latent space. Second, it relaxes the typical assumption that all dimensions are independent to the assumption that the specific dimensions are conditionally independent given the general dimension. Third, the proposed method is flexible and can be applied to various measurement models including testlet and second-order models for binary and polytomous responses. Fourth, the model can be estimated efficiently using a full-information ML method for realistically large problems with many items, testlets, and examinees. Our extensive simulation study shows that ignoring group differences as well as ignoring the correlation structure of the multivariate latent space can result in biased item parameter estimates. In particular, DIF estimates may be substantially biased.

We assumed uniform DIF throughout this article. Nonuniform DIF could be taken into account with an additional parameter, analogous to models for nonuniform DIF in the logistic regression approach (Swaminathan & Rogers, 1990).

The proposed multiple-group bifactor model can also be used to model multidimensionality that is construct-driven. Then, the general factor accounts for the common variance in the constructs being assessed, and the specific factors represent the residual dependencies between items that are assessing the same construct/subconstruct. When the constructs are highly correlated, a testlet or second-order model (which are formally equivalent) may be preferred to the bifactor model.

A few nonmodel-based approaches have been developed for assessing DIF for multidimensional tests (e.g., DFIT). As far as we are currently aware, however, there has been no formal investigation to evaluate the application of such methods to the bifactor models. Future studies will be required to examine their performance on bifactor structures and compare them with the model-based approach that we proposed here.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D110027 to Educational Testing Service. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Ainsworth, A. (2007). *Dimensionality and invariance: Assessing differential item functioning using bifactor multiple indicator multiple cause models*. Unpublished Ph.D. dissertation, University of California, Los Angeles.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, *23*, 67–95.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, *72*, 393–411.

Browne, W., & Draper, D. (2006). A comparison of Bayesian and likelihood methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514.

Cai, L., Yang, J. S., & Hansen, M. P. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*, 221–248.

Clauser, B. E., Nungester, R., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement*, *33*, 202–214.

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

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

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correlation. *Journal of Educational and Behavioral Statistics*, *34*, 149–170.

Dorans, N. J., & Holland, P. H. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models*. New York, NY: Springer.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*, 604–622.

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

Gibbons, R. D., Grochocinski, V. J., Weiss, D. J., Bhaumik, D. K., Kupfer, D. J., & Stover, A. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*, 361–368.

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

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, *20*, 26–36.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston, MA: Kluwer-Nijhoff.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–146). Hillsdale, NJ: Lawrence Erlbaum.

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

Hunter, M., Battersby, R., & Whitehead, M. (1986). Relationships between psychological symptoms, somatic complaints and menopausal status. *Maturitas*, *8*, 217–228.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*, 367–386.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidi-mensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395–416.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, *19*, 191–201.

Lee, Y.-S., Cohen, A., & Toro, M. (2009). Examining Type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review*, *10*, 365–375.

Li, Y., Bolt, D. M., & Fu, J. (2006a). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Li, Y., Bolt, D. M., & Fu, J. (2006b). *A multiple-group testlet model and its application to DIF assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*, 131–144.

McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, NY: Wiley.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B*, *51*, 127–138.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Natarajan, R., & Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, *95*, 227–237.

Oshima, T. C, Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement*, *34*, 253–272.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement In Education*, *20*, 335–355.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31.

Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression* (Technical Report). Vrije Universiteit Medical Center, Amsterdam.

Rijmen, F. (2009). *An efficient EM algorithm for multidimensional IRT models: Full information maximum likelihood estimation in limited time* (ETS Research Report RR0903). Princeton, NJ: Educational Testing Service.

Rijmen, F. (2010). Formal relations and an empirical comparison between the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, *47*, 361–372.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika*, *73*, 167–182.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement*, *33*, 215–230.

Shealy, R. T. (1989). *An item response theory-based statistical procedure for detecting concurrent internal bias in ability tests*. Unpublished Ph.D dissertation, University of Illinois, Urbana-Champaign.

Shealy, R. T., & Stout, W. F. (1991). *An item response theory model for test bias*. Washington, DC: Office of Naval Research.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Sinharay, S., & Dorans, N. J. (2010). Two simple approaches to overcome a problem with the Mantel-Haenszel statistic: Comments on Wang, Bradlow, Wainer, and Muller. *Journal of Educational and Behavioral Statistics*, *35*, 474–488.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*, 137–167.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS 0.5 Bayesian Analysis using Gibbs Sampling. Manual (version II)*. Cambridge, England: MRC-Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml

Sun, X., & Sun, D. (2005). Estimation of the Cholesky decomposition of the covariance matrix for a conditional independent normal model. *Statistics and Probability Letters*, *73*, 1–12.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53–75.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464.

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

Wang, W., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Wang, X., Bradlow, E., Wainer, H., & Muller, E. (2008). A Bayesian method for studying DIF: A cautionary tale filled with surprises and delights. *Journal of Educational and Behavioral Statistics*, *33*, 363–384.

Wolfinger, R. D. (1999). Fitting non-linear mixed models with the new NLMIXED procedure (Technical Report). Cary, NC: SAS Institute.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differential item functioning: Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Authors

MINJEONG JEON is a Graduate Student at the Graduate School of Education, University of California, Berkeley, CA 94720; mjj@berkeley.edu. Her primary research interests include item response modeling, multilevel modeling, and research methods.

FRANK RIJMEN is a Principal Research Scientist at the Educational Testing Service, Princeton, NJ 08541; frijmen@ets.org. His primary research interests include item response modeling, latent class models, and graphical models.

SOPHIA RABE-HESKETH is a Professor at the Graduate School of Education and Graduate Group in Biostatistics, University of California, Berkeley, CA 94720, and at the Institute of Education, University of London; sophiarh@berkeley.edu. Her primary research interests include multilevel and latent variable modeling.

Manuscript received October 12, 2010
Revision received May 29, 2011
Accepted August 10, 2011