Applied Psychological Measurement
http://apm.sagepub.com/

A Bifactor Multidimensional Item Response Theory Model for Differential Item Functioning Analysis on Testlet-Based Items

Hirotaka Fukuhara and Akihito Kamata

Applied Psychological Measurement 2011 35: 604
DOI: 10.1177/0146621611428447

Applied Psychological Measurement
35(8) 604–622
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621611428447
http://apm.sagepub.com

A Bifactor Multidimensional Item Response Theory Model for Differential Item Functioning Analysis on Testlet-Based Items

Hirotaka Fukuhara[1] and Akihito Kamata[2]

Abstract

A differential item functioning (DIF) detection method for testlet-based data was proposed and evaluated in this study. The proposed DIF model is an extension of a bifactor multidimensional item response theory (MIRT) model for testlets. Unlike traditional item response theory (IRT) DIF models, the proposed model takes testlet effects into account, thus estimating DIF magnitude appropriately when a test is composed of testlets. A fully Bayesian estimation method was adopted for parameter estimation. The recovery of parameters was evaluated for the proposed DIF model. Simulation results revealed that the proposed bifactor MIRT DIF model produced better estimates of DIF magnitude and higher DIF detection rates than the traditional IRT DIF model for all simulation conditions. A real data analysis was also conducted by applying the proposed DIF model to a statewide reading assessment data set.

Keywords

differential item functioning, testlets, MCMC, MIRT

Test fairness is a major concern in standardized testing when establishing the validity of test scores. To investigate whether a test is fair for subgroups of a targeted population, an analysis of differential item functioning (DIF) is commonly conducted. It is said that DIF is present when the probability of answering an item correctly is different between a focal group and a reference group, given the same level of ability that is intended to be measured by a test.

This study proposes a DIF detection model that is based on a bifactor multidimensional item response theory (MIRT) model, which detects uniform DIF on binary response items that are nested in testlets. A testlet is defined as a set of items that are constructed and implemented together as a unit of measurement (Wainer & Keily, 1987). For example, in a reading comprehension test, a testlet is formed as a series of questions that are based on one reading passage,

[1] Pearson, San Antonio, USA

[2] University of Oregon, Eugene, USA

Corresponding author:
Hirotaka Fukuhara, 19500 Bulverde Road, San Antonio, TX 78259, USA
Email: Hiro.Fukuhara@Pearson.com

and the reading comprehension test may comprise multiple reading passages, each containing a set of questions. In this scenario, examinees' performance on questions may be affected not only by their ability in reading comprehension but also by their knowledge and/or interest in the content of the reading passage (Yen, 1993). Therefore, item responses within the same testlet may not be locally independent from each other under a unidimensional item response theory (IRT) model. Fukuhara and Kamata (2007) evaluated a DIF detection model through a confirmatory factor analysis approach (e.g., Finch, 2005) on the presence of the testlet effect or testlet dimension (i.e., local item dependence). As a result, it was demonstrated that DIF parameters and item discrimination parameters were systematically underestimated if the testlet effect was ignored. Thus, traditional DIF detection methods should not be used when the assumption of local item independence does not hold, which is likely to happen when a test is composed of testlets.

To date, two classes of DIF analysis approaches have been recognized to overcome local item dependence due to a testlet effect. The first class of approaches evaluates a set of item responses at the testlet level, and a DIF analysis is conducted at the testlet level. For example, a polytomous item response model can be applied to estimate item parameters separately for the focal group and the reference group to determine the magnitude of DIF (Thissen, Steinberg, & Mooney, 1989). In addition, Douglas, Roussos, and Stout (1996) proposed a method to investigate DIF at the testlet level, which is called differential bundle functioning, by using the SIBTEST framework (Shealy & Stout, 1993); however, DIF can be detected only at the testlet level using this approach, which we may refer to as differential testlet functioning. Accordingly, We are not able to specify which items are the source of differential testlet functioning. Furthermore, a development of a testlet is costly and time-consuming, and We would not want to eliminate an entire testlet from an item bank due to a concern with differential testlet functioning; rather, it is best to investigate which items are problematic to attempt to correct problems at the item level without discarding the entire testlet. For this reason, a method for investigating DIF at the item level, rather than a method for investigating differential testlet functioning, would be more practical and valuable in item bank development.

Another class of approaches is to detect DIF at the item level via a testlet response model. W. Wang and Wilson (2005a), for example, extended their Rasch testlet response model (W. Wang & Wilson, 2005b) for a DIF analysis. In addition, X. Wang, Bradlow, Wainer, and Muller (2008) proposed a DIF analysis procedure, based on a two-parameter testlet response model.

The proposed DIF detection model in this study belongs to the latter class of approaches; however, the proposed model is distinct from other approaches in several ways. First, the proposed model in this study detects DIF for all items simultaneously, whereas the method used by X. Wang et al. (2008) studies only one item at a time. In addition, their method estimates the DIF magnitude of a studied item by using the rest of the items in the test as anchor items, which means that the DIF magnitudes of the other items are all assumed to be zero. However, by the DIF detection model in this study, DIF magnitudes are estimated under the assumption that the average DIF magnitude is zero. Second, the proposed model includes a parameter to capture the mean ability difference between the focal and reference groups to differentiate DIF from the impact. Otherwise, We need to make a strong assumption that there is no ability difference between the focal and reference groups. For example, the DIF detection method proposed by Wang et al. (2008) does not include such a parameter. They justify the use of their method by ensuring that there is no substantial mean ability difference between the focal and reference groups; however, such an assumption is not likely to hold for many applications, and a model that parameterizes the ability difference is more desirable in many cases. Moreover, if there is only a trivial difference in mean abilities between the focal and reference groups, the inclusion of an ability difference parameter will not negatively affect the results of a DIF analysis.

The proposed DIF detection model is based on a bifactor MIRT model that is expected to capture a testlet effect. As a result, the proposed model is expected to estimate DIF magnitudes more accurately and provide more accurate information about DIF for the studied items. Moreover, when the testlet effect is trivial or nonexistent, the proposed model is reduced to a conventional IRT-based DIF detection model. Hence, it will not be disadvantageous to adopt the proposed model even if the testlet effect is absent.

Model

A Bifactor MIRT Model for Testlets With Covariates

The proposed DIF detection model is an extension of a bifactor MIRT model for testlets (see DeMars, 2006; Li, Bolt, & Fu, 2006), and the measurement part of the model is expressed by

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i\theta_j - \delta_i + \lambda_i\gamma_{d(i)j} - \beta'_i G_j, \tag{1}$$

where θ_j is the ability for person j (a random-effect associated with persons, which is the primary dimension); $\gamma_{d(i)j}$ is a testlet effect for person j on items in testlet d (a random effect associated with testlet d, which is the secondary dimension); δ_i is an item-difficulty-related parameter; a_i and λ_i are discrimination parameters for the ability and testlet effect, respectively; β'_i is the difference in item-difficulty-related parameters between groups for item i; and G_j is a group indicator for person j ($G_j = 1$ for the focal group, and $G_j = 0$ for the reference group). Ability and the testlet effect in the bifactor MIRT model for testlets are assumed to be independent and to have the standard normal distribution (DeMars, 2006; Li et al., 2006). The magnitude of a testlet effect is determined by the ratio of λ_i to a_i. The magnitude of uniform DIF for item i is captured by β'_i in Equation 1. As a result, the item-difficulty-related parameter for the focal group ($G_j = 1$) for item i is $\delta_i - \beta'_i$, whereas the item-difficulty-related parameter for the reference group ($G_j = 0$) is δ_i. Accordingly, Equation 1 can be reexpressed as

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i(\theta_j - b_i + C_i\gamma_{d(i)j} + \beta_i G_j), \tag{2}$$

where $b_i = \frac{\delta_i}{a_i}$, $C_i = \frac{\lambda_i}{a_i}$, and $\beta_i = \frac{\beta'_i}{a_i}$. Now, assuming that $\gamma_{d(i)j}$ in Equation 2 is rescaled to have a normal distribution with the mean equal to 0 and the variance equal to $\sigma^2_{\gamma_d}$ for testlet d, instead of the standard normal distribution, Equation 2 becomes

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i(\theta_j - b_i + \frac{C_i}{\sigma_{\gamma_d}}\gamma_{d(i)j} - \beta_i G_j). \tag{3}$$

Li et al. (2006) indicated that if $C_i = \sigma_{\gamma_d}$ for all items within each testlet, then Equation 3 can be simplified as

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i(\theta_j - b_i + \gamma_{d(i)j} - \beta_i G_j), \tag{4}$$

which is analogous to a two-parameter logistic (2PL) testlet response theory model (e.g., Bradlow, Wainer, & Wang, 1999). In other words, Equation 4 is a special case of Equation 1, in which ability and testlet effects are assumed to have the same item discrimination parameter (a_i). Note that although the primary and secondary dimensions share the same item discrimination parameter, this does not mean that the magnitudes of item discrimination power for the primary and secondary dimensions are the same unless the scales of the two dimensions are the same. In this study, Equation 4 was adopted as the measurement part of the proposed DIF model; therefore, the proposed DIF model is an extension of a bifactor MIRT model for testlets, assuming that $C_i = \sigma_{\gamma_d}$ holds for all items in testlet d.

It is important to note that Equation 4 does not capture the mean ability difference between the focal and reference groups, which may not be a realistic setup in many applications. For instance, when one conducts a DIF analysis between standard curriculum students and students in exceptional student education (ESE) programs, it is unreasonable to assume that they have the same mean abilities. Furthermore, if a DIF procedure is applied for an item drift study, in which the stabilities of item parameter estimates are evaluated across different test administrations of a large-scale statewide assessment, then ability levels for two adjacent administrations are expected to be different. In such cases, item performance differences need to be adjusted by introducing a parameter that captures the group mean ability difference. Accordingly, the structural model is given by

$$\theta_j = \beta_\theta G_j + \zeta_j, \tag{5}$$

where β_θ is the effect of group G_j on the ability θ_j . Thus, β_θ captures the difference in mean abilities between the focal group ($G_j = 1$) and the reference group ($G_j = 0$). Furthermore, ζ_j is the residual for person j. Then, Equations 4 and 5 can be expressed as a combined model

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i(\beta_\theta G_j + \zeta_j - b_i + \gamma_{d(i)j} - \beta_i G_j), \tag{6}$$

which is the complete form of the proposed DIF detection model in this study. Note that this is a multidimensional model because two latent factors (ζ_j and $\gamma_{d(i)j}$) predict the logit of the correct answer for each item.

2PL IRT Model With Covariates

A conventional IRT-based DIF detection model was also included in this study as a base model for evaluating the performance of the proposed bifactor MIRT DIF detection model. The authors chose the IRT-based DIF detection model as a competing model instead of a DIF detection method based on a testlet response theory model, such as those proposed by W. Wang and Wilson (2005a) and X. Wang et al. (2008), for the following reasons. The model by W. Wang and Wilson (2005a) is based on the Rasch testlet model, whereas the proposed DIF model is based on a 2PL testlet response theory model. Thus, it is obvious that the proposed model would detect DIF better than would the Rasch testlet DIF model if item discrimination varies across items. The DIF detection method by X. Wang et al. (2008) is based on a 2PL testlet response theory model, which is analogous to the proposed model in this study. However, their method uses an all-other anchor item approach (W. C. Wang, 2004), in which the scale of DIF magnitude is established by using all items on a test as anchor items, except the item that is being studied. This approach assumes that all anchor items are DIF free. However, the assumption under the proposed DIF model is that the average DIF magnitude over all test items is zero, which W. C. Wang (2004) referred to as an equal-mean-difficulty approach. Thus, each method works differently to set the scale of DIF parameters, and the two approaches do not necessarily produce the same results, depending on how their assumptions are met. In addition, because most of the DIF detection models currently implemented in practice do not account for testlet effects, it would be practically meaningful to demonstrate the benefits that can be expected from the proposed model in comparison with a comparable IRT approach that does not consider testlet effects.

The conventional IRT-based DIF detection model used by the authors was an extension of the 2PL IRT model by introducing grouping covariates for the purpose of DIF detection. This is analogous to a confirmatory factor analysis model with covariates, which is also known as the multiple indicators multiple causes (MIMIC) model, for DIF detection (e.g., Finch, 2005). This model is written as

$$\ln\left(\frac{P(y_{ji}=1)}{P(y_{ji}=0)}\right) = a_i(\beta_\theta G_j + \zeta_j - b_i - \beta_i G_j). \tag{7}$$

Equation 7 is similar to Equation 6, except a testlet effect is not in the model.

Model Parameter Estimation

In this study, parameters in both the proposed DIF model and the IRT-based DIF model were estimated with a Markov chain Monte Carlo (MCMC) method. Another computational method found to be efficient for a bifactor model, such as the proposed DIF model, is full information maximum likelihood estimation (Cai, 2010; Cai, Yang, & Hansen, 2011; Gibbons et al., 2007; Gibbons & Hedecker, 1992; Rijmen, 2009). Although the estimation for a bifactor model by full information maximum likelihood estimation is available on some off-the-shelf software, such as TESTFACT (Bock et al., 2003), it did not allow the authors to fit their proposed model. Thus, they implemented an MCMC estimation method in this study using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which is flexible enough to fit the proposed model.

A general assumption of a Bayesian approach is that all parameters are treated as random variables, whereas, in the traditional IRT perspective, only latent factors (e.g., ability and testlet effect) are treated as random variables. In addition, a Bayesian framework incorporates information about parameters of interest (namely, prior distributions of parameters) into the probability statements about the parameters. Assuming that $\boldsymbol{\omega}$ is a vector of parameters and $\mathbf{y}$ is a vector of observed data, the objective of the Bayesian approach is to make a statistical inference about $\boldsymbol{\omega}$ from the conditional probability of $\boldsymbol{\omega}$ given $\mathbf{y}$, which is called the posterior density or $P(\boldsymbol{\omega} \mid \mathbf{y})$ (Gelman, Carlin, Stern, & Rubin, 2004). The posterior density is proportional to the product of the prior distribution ($P(\boldsymbol{\omega})$) and the likelihood function ($P(\mathbf{y} \mid \boldsymbol{\omega})$) as follows:

$$P(\boldsymbol{\omega}|\mathbf{y}) \propto P(\boldsymbol{\omega})P(\mathbf{y}|\boldsymbol{\omega}). \tag{8}$$

However, a sample cannot be directly drawn from a posterior density when nonconjugate priors are used (see, for example, Wainer, Bradlow, & Wang, 2007), which was the case in this study. In such a situation, an MCMC method becomes useful (Gelman et al., 2004). A Gibbs sampling algorithm, which is a type of MCMC method, first creates subvectors of parameters, and then it draws samples from the conditional distribution for each subvector, given the remaining subvectors, and iteratively updates each parameter subvector conditioning on the other parameter subvectors. To date, a Gibbs sampler has been applied to traditional IRT models (e.g., Albert, 1992; Kim, 2001) and IRT models for testlets (Bradlow et al., 1999; Li et al., 2006; Wainer, Bradlow, & Du, 2000; X. Wang, Bradlow, & Wainer, 2002). This study also adopted a Gibbs sampling algorithm to estimate parameters for the proposed testlet-based DIF detection model and the conventional IRT-based DIF detection model. The mean of the posterior distribution for each parameter was considered as the point estimate.

Table 1. Prior Distributions of Parameters and Hyperparameters

Parameter	Prior distribution of the parameter	Hyperparameter of the parameter	Prior distribution of the hyperparameter
ζ_j	$N(0, 1)$		
a_i	$N(\mu_a, \sigma_a^2)I(0, \infty)$	μ_a	$N(0, 1000)$
		σ_a^2	$Inv - \chi^2(0.5)$
b_i	$N(\mu_b, \sigma_b^2)$	μ_b	$N(0, 1000)$
		σ_b^2	$Inv - \chi^2(0.5)$
β_i	$N(\mu_\beta, \sigma_\beta^2)$	μ_β	$N(0, 1000)$
		σ_β^2	$Inv - \chi^2(0.5)$
β_θ	$N(\mu_{\beta_\theta}, \sigma_{\beta_\theta}^2)$	μ_{β_θ}	$N(0, 1000)$
		$\sigma_{\beta_\theta}^2$	$Inv - \chi^2(0.5)$
$\gamma_{d(i)j}$	$N(0, \sigma_{\gamma_d}^2)$	$\sigma_{\gamma_d}^2$	$Inv - \chi^2(0.5)$

Prior Distributions

For the proposed bifactor MIRT DIF detection model and the conventional IRT-based DIF detection model, noninformative priors were used, so that posterior distributions of parameters were determined dominantly by observed data. The proposed priors for model parameters are summarized in Table 1. A normal distribution was used as a prior for ζ_j , $\gamma_{d(i)j}$, b_i , β_θ , and β_i in Equation 6. Note that ζ_j had the standard normal prior. However, a_i had a truncated normal prior, which included only the positive side of the normal distribution. In addition, priors for hyperparameters were specified as follows: Each of the means for the prior distributions, except ζ_j , had a normal distribution with a large variance, indicating that the prior was noninformative. Similarly, each of the variances for the prior distributions, except ζ_j , also had a noninformative prior. The prior distributions described here were analogous to those used by Bradlow et al. (1999) and Li et al. (2006).

Identifiability Problem

Two types of scale indeterminacy could occur for IRT models: additive and multiplicative identifiability problems (Befumi, Gelman, Park, & Kaplan, 2005). Assuming that a 2PL IRT model is used, an additive identifiability problem would occur in a way that allows any constant to be added to the ability and an item difficulty, while the logit of a correct answer for the item remains unchanged. However, a multiplicative indeterminacy would occur in a way that allows item discrimination parameters to be multiplied by any constant and the ability and item-difficulty parameters to be divided by the same constant, while the logit of a correct answer for the item stays the same. Befumi et al. (2005) suggested fixing the mean and standard deviation (*SD*) of ability or normalizing item parameters using the mean and *SD* of ability to identify the model.

In this study, the parameters in the proposed model (Equation 6) were treated in the following ways. First, the distribution for ζ_j was assumed to be the standard normal distribution, and the mean of $\gamma_{d(i)j}$ was fixed to be 0. Second, the parameters for the group mean ability difference and DIF magnitudes were centered around the average of the DIF magnitudes across all the items as follows:

$$\beta_{\theta}^{adj} = \beta_{\theta} - \bar{\beta}, \tag{9}$$

$$\beta_{i}^{adj} = \beta_{i} - \bar{\beta}, \tag{10}$$

where β_{θ}^{adj} and β_{i}^{adj} are adjusted parameters for the mean ability difference and DIF magnitudes, respectively, and $\bar{\beta}$ indicates the average DIF magnitude across all items. A similar approach was made for the conventional IRT-based DIF detection model (Equation 7), where the standard normal distribution was assumed for ζ_j , and Equations 9 and 10 were applied to rescale the mean ability difference and DIF magnitudes in Equation 7. The WinBUGS codes for the proposed bifactor MIRT DIF model and for the conventional IRT DIF model are extensions of Li et al. (2006) and are presented in the appendix.

Convergence Diagnosis

When parameters are estimated via an MCMC method, convergence of the parameter estimation needs to be examined. If parameter estimates do not converge, incorrect inferences about parameters of interest will result. Thus, we need to determine the number of iterations to discard (i.e., burn-in iterations) when a parameter estimation stabilizes. In addition, they need to determine the number of iterations after a burn-in period to obtain good samples of each parameter that represent the posterior distribution of the parameter. Sinharay (2004) recommended a series of convergence diagnosis methods. In this study, convergence of parameter estimation was assessed using several graphical methods. First, history plots of posterior distributions were obtained to determine when the parameter estimation became stable. Density plots were also examined to check whether posterior distributions formed smooth densities close to the anticipated shapes. In addition, autocorrelation plots that showed correlations between samples from a posterior distribution at successive time points were obtained. Generally, the higher the autocorrelation, the longer it takes to explore the entire posterior distribution, which indicates that more samples should be drawn. For the proposed model, it was decided that 15,000 samples would be drawn from each posterior distribution after 5,000 samples were discarded as burn-ins based on a preliminary analysis under one of the simulation conditions.

Simulation Study

Simulation Design

A simulation study was conducted to evaluate the proposed bifactor MIRT DIF model. Item responses were randomly generated by mimicking a testlet-based test. It was assumed that there were 42 items in a test with six testlets, each containing seven items. In each testlet, the true item difficulties for the seven items were set to −1.5, −1.0, −0.5, 0.0, 0.5, 1.0, and 1.5, respectively. Four simulation factors were considered in this simulation study: magnitude of testlet effect, magnitude of DIF, magnitude of item discrimination, and the proportion of a focal group to all examinees. The first three factors have been found to affect the estimation of DIF magnitude by an IRT-based DIF model under the violation of local item independence due to testlets (Fukuhara & Kamata, 2007). The three levels of testlet effect were 0.5, 1.0, and 2.0. The magnitude of the testlet effect was determined by the ratio of the random-effect variance of testlet effect to the random-effect variance for ability, which is a method used by other studies, such as Li et al. (2006) and Wainer et al. (2007). The true DIF magnitude (TDIF) considered in this study was either 0.5 or 0.7 in log-odds. The magnitude of DIF in this simulation was considered to be the difference in item difficulties between the focal group and the reference group. The magnitudes of item discrimination were 0.8 or 2.0, which correspond to approximately 0.4 and 0.75 biserial correlations, respectively. The same item discrimination was assigned to all items to simplify a simulation condition to determine the effect of the magnitude of item discrimination on the estimation of DIF magnitude. Finally, the number of examinees was assumed to be 1,000 for all conditions. However, the proportion of examinees in the focal group was set to either 0.5 or 0.25, so the number of examinees in the focal group (NF) was 500 (NF = 500) or 250 (NF = 250). Thus, the total number of simulation conditions resulted in $3 \times 2 \times 2 \times 2 =$ 24 conditions. The mean difference in ability levels between the two groups was set to be 0.2 in the standardized scale. In addition, this study assumed that only the fourth item in each testlet, whose item-difficulty level was 0.0, displayed nonzero DIF. Fukuhara and Kamata (2007) showed that the item-difficulty level did not affect the estimation of the DIF magnitude; thus, the difficulty level for DIF items was set to the same value to achieve more control in the simulation. It was also assumed that the DIF items in three testlets had positive DIF, whereas the DIF items in the other three testlets had negative DIF. Therefore, in each simulation condition, the average of DIF magnitudes over all items was constrained to be 0.0.

Data Generation

For each simulation condition, abilities (i.e., the primary dimension) were generated from a normal distribution. The average ability levels for the focal group and the reference group were 0.2 and 0.0, respectively. However, the variance of the ability distribution was fixed to be 1.0 for both groups. Testlet effects (i.e., random effects associated with testlets, which are the secondary dimensions) were generated from a normal distribution with the mean equal to 0.0 and the variance equal to 0.5, 1.0, or 2.0. In this study, the ability and the testlet effects were assumed to be independent of each other, which is also an assumption made under other testlet-based IRT models (Bradlow et al., 1999; DeMars, 2006; Li et al., 2006; Wainer et al., 2000; X. Wang et al., 2002). Item responses for hypothetical examinees were obtained based on the testlet response theory model (Bradlow et al., 1999; Wainer et al., 2000; X. Wang et al., 2002). This study did not consider the pseudoguessing parameter.

Model Evaluation

In each simulation condition, data generation and parameter estimation by MCMC for the bifactor MIRT DIF model as well as the IRT DIF model were replicated 100 times. Then, the proposed bifactor MIRT DIF model was evaluated in terms of bias, standard error (SE), and root mean squared error (RMSE) of the estimates of the DIF magnitude. The bifactor MIRT DIF model was also evaluated by comparing the models' bias, SE, and RMSE with those for the IRT-based model. In addition, a series of repeated measures analysis of variance (ANOVA) was conducted for bias in the DIF magnitude to determine what factors affected the estimates of the DIF magnitude. In the ANOVA for DIF items, the two DIF models were treated as a within-replication-factor, whereas magnitudes of the testlet effect, magnitudes of TDIF, and magnitudes of item discrimination were treated as between-replication-factors. In addition, testlets and the simulation conditions functioned as blocking factors. For non-DIF items, a within-factor, between-factors, and a blocking factor were the same as those for DIF items, while a series of repeated measures ANOVA were conducted separately by item-difficulty level for non-DIF items.

Moreover, the accuracy of estimating DIF magnitude by the proposed bifactor MIRT DIF model was evaluated by assessing both DIF detection rates and DIF detection error rates. The DIF detection rate was defined as the proportion of DIF items that were actually detected to be

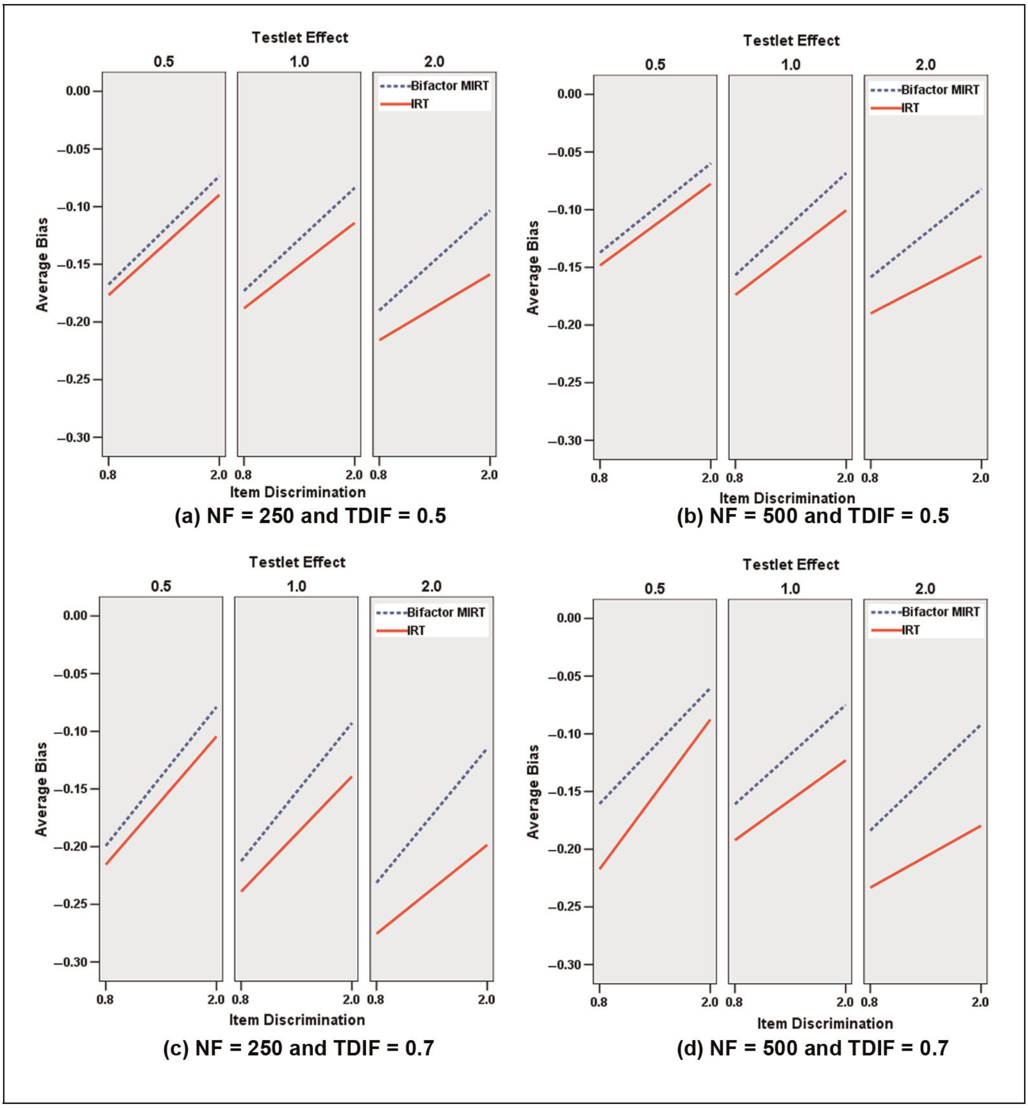
DIF items. However, the DIF detection error rate was defined as the proportion of non-DIF items that were detected as DIF items. The authors considered that an item displayed a meaningfully large DIF if the absolute value of its DIF magnitude was greater than 0.426 in the logit scale, which corresponded to 1.0 in the logistic definition of the delta scale (e.g., Monahan, McHorney, Stump, & Perkins, 2007). This value is also a cutoff effect size value for moderate DIF based on the Educational Testing Service (ETS) DIF classification (Dorans & Holland, 1993), and items with moderate or large DIF are carefully evaluated in practice. Therefore, the authors designated the absolute value of 0.426 in the logit scale as the threshold for a meaningful DIF magnitude in this study. Then, the DIF detection rate was determined by how many times DIF items displayed meaningful DIF magnitudes out of 100 replications. Similarly, the DIF detection error rate was determined by counting how many times DIF-free items had meaningfully large DIF magnitudes out of 100 replications. In practice, a hypothesis testing is also frequently used in conjunction with the evaluation of the magnitude of a DIF, such as in ETS criteria (Dorans & Holland, 1993); however, this study used only the magnitudes to make a judgment regarding meaningful DIF magnitude.

Results

Average bias in DIF magnitude over all DIF items for all simulation conditions are summarized in Panels 1a (NF = 250, TDIF = 0.5), 1b (NF = 500, TDIF = 0.5), 1c (NF = 250, TDIF = 0.7), and 1d (NF=500, TDIF = 0.7) in Figure 1. The proposed bifactor MIRT DIF model and the IRT DIF model underestimated DIF magnitude on DIF items for all simulation conditions. However, the magnitude of underestimation for the proposed DIF model was always less than it was for the IRT DIF model, given the same simulation condition. The magnitude of bias for the proposed DIF model did not change as the magnitude of the testlet effect increased, whereas a greater amount of underestimation on DIF magnitude was obtained with larger testlet effects for the IRT DIF model. The magnitude of bias was generally less with a larger item discrimination power and smaller testlet effect for the proposed DIF model. Moreover, slightly less bias was obtained when the proportion of examinees to all examinees in a focal group was 0.5. Accordingly, the bifactor MIRT DIF model produced the closest estimate of TDIF magnitude when the item discrimination was 2.0, the testlet effect was 0.5, and the proportion of a focal group to all examinees was 0.5.

Average DIF detection rates over all DIF items for all simulation conditions are shown in Panels 2a (NF = 250, TDIF = 0.5), 2b (NF = 500, TDIF = 0.5), 2c (NF = 250, TDIF = 0.7), and 2d (NF =500, TDIF = 0.7) in Figure 2. The average DIF detection rate was always higher for the bifactor model than for the IRT DIF model, and the difference in the average DIF detection rate between the two models increased as the magnitude of the testlet effect increased. In other words, the proposed bifactor MIRT DIF model had much higher DIF detection rates than the IRT DIF model with larger testlet effects. In addition, average DIF detection rates for both DIF models were higher with larger item discrimination, larger DIF magnitude, and equal group size between a focal and a reference group. Average DIF detection rates for the bifactor DIF model were close to 100% when the TDIF magnitude was 0.7, the item discrimination was 2.0, and the proportion of a focal group to whole group was 0.5.

Parameter recovery for other parameters is briefly explained here. The bifactor MIRT DIF model and IRT DIF model provided good estimates of the DIF magnitude for non-DIF items. DIF detection error rates for both models were nearly zero for all simulation conditions. One noticeable difference in the parameter recovery between the two models was that the bifactor MIRT DIF model always produced a better estimate of the ability difference between a focal



(a) NF = 250 and TDIF = 0.5

(b) NF = 500 and TDIF = 0.5

(c) NF = 250 and TDIF = 0.7

(d) NF = 500 and TDIF = 0.7

Figure 1. Average bias in DIF magnitude for TDIF = 0.5 (NF = 250 in Panel 1a and NF = 500 in Panel 1b) and TDIF = 0.7 (NF = 250 in Panel 1c and NF = 500 in Panel 1d)

Note: DIF = differential item functioning; TDIF = true DIF; NF = number of examinees in the focal group.

and a reference group than did the IRT DIF model. Because the magnitudes of bias in DIF magnitude on DIF items were also smaller for the proposed model than for the IRT DIF model, this might point to the proposed model as having the smaller bias for the difference in the average abilities between the two groups. Moreover, the bifactor MIRT DIF model estimated item discriminations better than the IRT DIF model, especially when item discriminations were higher ($a = 2.0$). In other words, the IRT DIF model largely underestimated item discriminations when item discriminations were 2.0. Finally, the bifactor MIRT DIF model estimated item difficulties well for non-DIF items. However, biases for item difficulties for DIF items were higher than

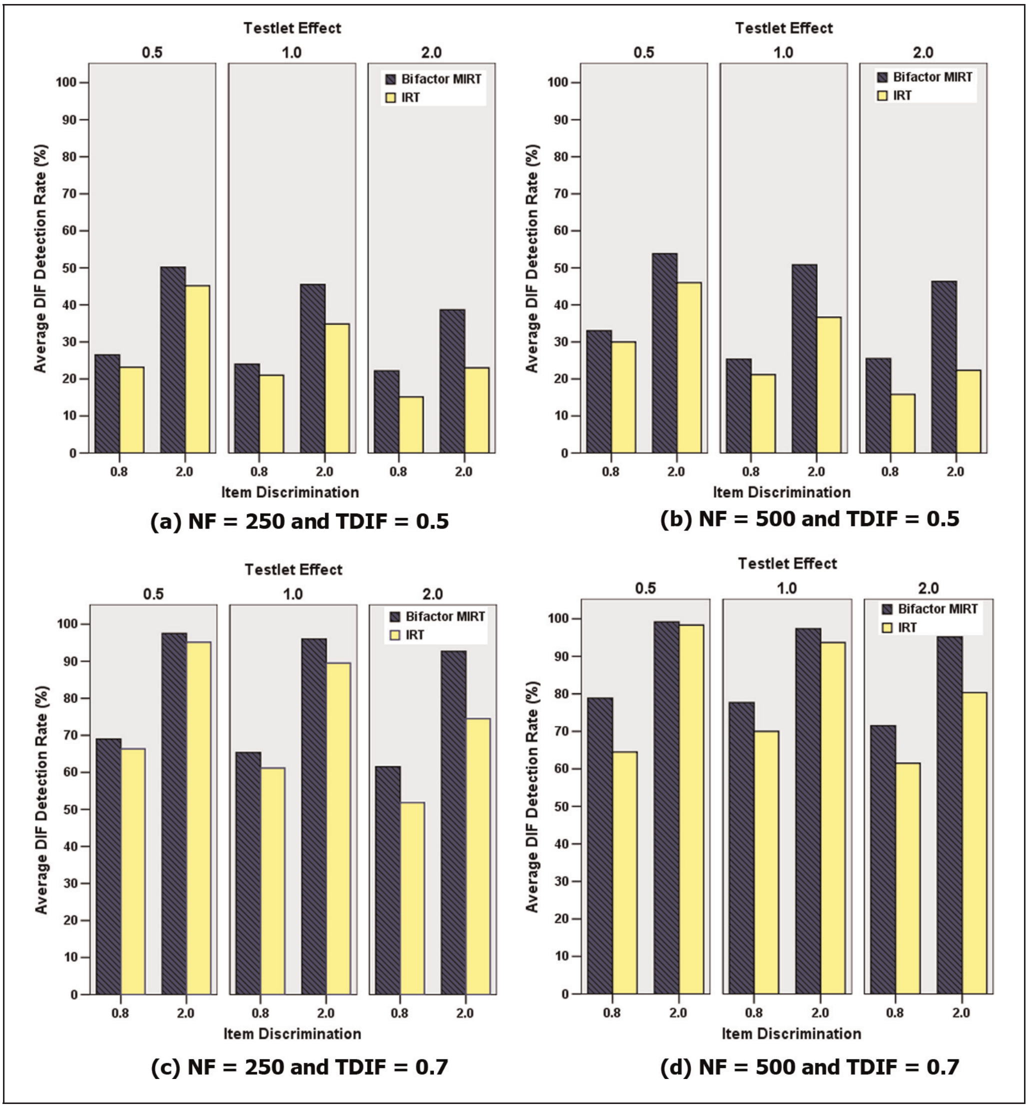


Figure 2. Average DIF detection rate for TDIF = 0.5 (NF = 250 in Panel 2a and NF = 500 in Panel 2b) and TDIF = 0.7 (NF = 250 in Panel 2c and NF = 500 in Panel 2d)

Note: DIF = differential item functioning; TDIF = true DIF; NF = number of examinees in the focal group.

those for non-DIF items. Item difficulties were overestimated for DIF items that had positive DIF, whereas they were underestimated for the DIF items that had negative DIF. The IRT DIF model did not estimate an item difficulty well when the true item difficulty was large in its absolute values (e.g., $b = -1.5, -1.0, 1.0$, or 1.5) and the magnitude of the testlet effect was large ($\sigma_\gamma^2 = 2.0$). Thus, the magnitude of bias for the bifactor MIRT DIF model was smaller than that for the IRT DIF model in those conditions. Overall, the results described above did not change for the different proportions of a focal group to the whole group. In other words, the differences in bias for the different proportions of examinees in a focal group in relation to all examinees were negligible for both DIF models. For more simulation study results on the proposed model, readers are referred to Fukuhara (2009).

Real Data Analysis

Method

The proposed bifactor MIRT DIF model was also applied to real data. A data set from the fifth-grade reading achievement test in a southeastern state in the United States was examined. The reading test consisted of 45 operational items in six testlets (reading passages), and all items were multiple-choice items and were dichotomously scored. A DIF analysis was conducted between African American students (focal group) and non-Hispanic White students (reference group). A sample of 1,000 students was randomly selected from each of the focal and the reference groups. Thus, 2,000 students were included in this analysis. DIF parameters were estimated using both the proposed bifactor MIRT DIF model (Equation 6) and the IRT-based DIF model (Equation 7). Prior to the DIF analysis, a preliminary analysis was conducted to diagnose the convergence for both models using history plots, density plots, and autocorrelation plots. As a result, the number of burn-in iterations was set to 7,000, and the number of iterations after the burn-in iterations was set to 15,000.

Results

The authors adopted the same criterion used in the simulation study with the value of 0.426 in logit to identify meaningfully large DIF items. As summarized in Table 2, the proposed bifactor MIRT DIF model flagged seven items as having meaningfully large DIF, whereas the IRT DIF model only identified two items as having meaningfully large DIF. The two items flagged by the IRT DIF model (Items 5 and 36) were also flagged as DIF items by the proposed bifactor MIRT DIF models and displayed much greater DIF magnitude than the criterion value. The other five items flagged only by the proposed bifactor MIRT DIF model (Items 11, 12, 20, 23, and 34) had DIF magnitudes that were close to the criterion value. Based on the simulation study, both the proposed bifactor MIRT DIF model and the IRT DIF model underestimated DIF magnitude, but the proposed bifactor MIRT DIF model underestimated DIF magnitude to a lesser extent than the IRT DIF model. Thus, it is likely that the IRT DIF model misclassified the five items as non-DIF items. Another finding was that both DIF models similarly estimated DIF magnitudes on items that displayed almost zero DIF (Item 2, for instance), which was consistent with the simulation study.

The magnitudes of the testlet effect were estimated to be small, ranging from 0.188 to 0.344, which were comparable with the small testlet effect in the simulation study. The simulation study revealed that even small testlet effects led to differences in DIF estimates between the proposed bifactor MIRT DIF model and the IRT DIF model. Thus, the differences in DIF estimates between the two DIF models revealed in the real data analysis were quite consistent with results from the simulation study.

Regarding the estimates of other parameters, the estimated values of item discriminations by the two models were close to each other. In the simulation study, their estimated values were close to each other when item discrimination was small (a = 0.8). On the other hand, the IRT DIF model underestimated item discrimination parameters when the magnitude of item discrimination was large. The estimated values for both models were close to each other, because the item discrimination parameters estimated by both DIF models were rather low. Thus, the real data set was similar to the data sets with low item discriminations in the simulation study

Table 2. Parameter Estimates in DIF Magnitudes and in Difference in Average Abilities Between Two Groups by the Proposed Bifactor MIRT DIF and IRT DIF Models

Parameter[a]	Bifactor MIRT DIF model	IRT DIF model
β_1^{adj}	0.296	0.206
β_2^{adj}	0.008	0.010
β_3^{adj}	−0.189	−0.156
β_4^{adj}	−0.286	−0.205
β_5^{adj}	−1.676[b]	−1.070[b]
β_6^{adj}	0.025	0.028
β_7^{adj}	0.072	0.033
β_8^{adj}	0.132	0.113
β_9^{adj}	0.316	0.209
β_{10}^{adj}	0.013	0.029
β_{11}^{adj}	0.506[b]	0.354
β_{12}^{adj}	−0.478[b]	−0.380
β_{13}^{adj}	0.076	0.089
β_{14}^{adj}	0.386	0.334
β_{15}^{adj}	0.057	0.033
β_{16}^{adj}	−0.073	−0.028
β_{17}^{adj}	0.149	0.046
β_{18}^{adj}	−0.239	−0.177
β_{19}^{adj}	−0.382	−0.296
β_{20}^{adj}	0.431[b]	0.327
β_{21}^{adj}	−0.153	−0.129
β_{22}^{adj}	−0.093	−0.084
β_{23}^{adj}	0.474[b]	0.330
β_{24}^{adj}	0.005	−0.036
β_{25}^{adj}	0.051	0.023
β_{26}^{adj}	0.054	0.064
β_{27}^{adj}	−0.044	−0.028
β_{28}^{adj}	0.288	0.193
β_{29}^{adj}	−0.009	0.000
β_{30}^{adj}	−0.199	−0.166
β_{31}^{adj}	0.051	0.051
β_{32}^{adj}	0.348	0.251
β_{33}^{adj}	0.001	−0.004
β_{34}^{adj}	−0.439[b]	−0.325
β_{35}^{adj}	−0.170	−0.136
β_{36}^{adj}	0.872[b]	0.626[b]
β_{37}^{adj}	0.060	0.040
β_{38}^{adj}	−0.088	−0.066
β_{39}^{adj}	0.008	0.010
β_{40}^{adj}	0.299	0.263
β_{41}^{adj}	0.224	0.175
β_{42}^{adj}	−0.008	−0.003
β_{43}^{adj}	−0.274	−0.231
β_{44}^{adj}	−0.077	−0.062
β_{45}^{adj}	−0.323	−0.255
β_{θ}^{adj}	−0.914	−0.879

Note: MIRT = multidimensional item response theory; IRT = item response theory; DIF = differential item functioning.

[a] Parameter estimates for the DIF magnitude and the difference in the average abilities between two groups were adjusted as described in Equations 9 and 10.

[b] DIF magnitude > 0.426

(a = 0.8). Under the simulation conditions, both models underestimated DIF magnitudes on DIF items more than those for the high item discrimination conditions (a = 2.0). Thus, both DIF models were also likely to underestimate DIF magnitudes in real data analysis. In addition, the estimated item-difficulty parameters for both DIF models were similar when the estimated values were close to zero. However, some degree of separation in the estimated item-difficulty parameters was observed for those items that had extreme item-difficulty values, such as Items 7 and 9 (approximately −3.0). These findings were also consistent with the simulation study results.

Conclusion

This study demonstrated the development and evaluation of a parametric DIF detection model for testlet-based item response data. Because testlets could cause a violation of the local item independence assumption under a unidimensional IRT model, traditional DIF detection methods based on a unidimensional IRT model may not be the best approach. Thus, a new DIF detection model was proposed by extending a bifactor MIRT model for testlets. The proposed DIF model incorporates the effect of local item dependence due to testlets with random-effect parameters for testlets.

The simulation study revealed that the bifactor MIRT DIF model produced higher DIF detection rates for DIF items than the IRT DIF model did. Moreover, the proposed DIF model based on a bifactor MIRT model estimated the DIF magnitude of DIF items more accurately than the IRT DIF model. In addition, the proposed bifactor MIRT DIF model properly estimated DIF magnitude for non-DIF items, and it had very low DIF detection error rates. The proposed bifactor MIRT DIF model also estimated other parameters relatively well in comparison with the IRT DIF model. However, the authors also found that although the magnitude of underestimation was smaller than that for the standard IRT DIF model, the proposed bifactor MIRT DIF model underestimated DIF magnitude on DIF items. They suspect that the DIF magnitude might have been underestimated for the proposed model due to an underestimation of its unadjusted DIF magnitude, because the average ability difference was estimated well. On the other hand, the comparison model (2PL IRT DIF) further underestimated DIF magnitudes because of bias in item parameters, which were caused by local item dependence in addition to the factors described above for the proposed model. Thus, the proposed model had smaller bias in estimated DIF magnitudes than the 2PL IRT DIF model for all simulation conditions.

The results from the real data analysis indicated that although the magnitudes of testlet effects were determined to be small for all testlets, the bifactor MIRT DIF model flagged more items than the IRT DIF model. This result confirmed that even small testlet effects can lead a traditional IRT DIF model to underestimate DIF magnitude and consequently misclassify some DIF items as non-DIF items. Therefore, it demonstrated the importance of applying the proposed bifactor MIRT DIF model to testlet-based data for a DIF analysis.

Standardized tests are frequently composed of testlets, especially reading comprehension tests where sets of items are associated with reading passages. In such situations, item responses may not be locally independent of each other (Wainer & Kiely, 1987). Consequently, parameter estimates by conventional IRT models may be biased, especially for item discrimination parameters (Bradlow et al., 1999). Similarly, DIF magnitude and item parameters also may not be estimated accurately under the violation of the item local independence assumption (Fukuhara & Kamata, 2007).

In addition to the application of the proposed bifactor DIF model to testlet-based tests, the model is applicable in other testing situations. As Yen (1993) argued, many situations potentially cause local item dependence. For example, a standardized science test may contain subareas, such as chemistry, biology, physics, and natural science. In such a test, subareas can be clustered and treated as if they are testlets, and the proposed bifactor DIF model can be effectively used to take the local item dependence due to subareas (e.g., reporting category) into account.

Finally, some limitations of this study and possible future research are described. First, because the proposed DIF detection model is based on a parametric approach, the size of the sample needs to be much larger than that for nonparametric DIF detection models. Parametric DIF detection models, such as the proposed DIF model in this study, may not be feasible to use in DIF studies for some ESE students, because the number of ESE students within each ESE category (e.g., visually impaired) may not be sufficiently large. Second, in the proposed bifactor MIRT DIF model, the ability and the testlet effect share the same item discrimination parameter, which is analogous to a two-parameter testlet response theory model (X. Wang et al., 2002). Li et al. (2006) concluded that the bifactor MIRT model for testlets, in which the ability and the testlet effect are assumed to have separate item discrimination parameters, is more flexible than the 2PL testlet response theory model and, thus, may fit testlet-based item response data relatively well. Therefore, if the ability and the testlet effect had two different discrimination parameters, such a model might fit better with real testlet-based data. However, if the proposed DIF model had unique item discriminations for the primary and secondary random effects, the scale of DIF in the equation would include item discrimination and get quite complicated for interpretation, because the DIF magnitude in Equation 10 was scaled so that the average DIF magnitude across all items was zero. Third, the proposed DIF model assumed only binary response items. However, many large-scale assessments include performance task items that are scored polytomously. The authors' proposed model could be extended for these types of items. Fourth, the proposed model was designed to detect only uniform DIF. Uniform DIF indicates that differential performance of an item between a focal group and reference group is uniform across all ability levels. In contrast, nonuniform DIF is present when the magnitude and/or the direction of the differential performance between two target groups depend on their ability level. A future study may extend the proposed model to detect nonuniform DIF as well. Finally, the DIF magnitude in the proposed DIF model was parameterized so that the average DIF magnitude across all items is zero. This assumption may be reasonable for many standardized tests because items in those tests are usually well developed and may have gone through an extensive item bias review process. However, if the average DIF magnitude is largely deviated from zero, which is likely in some real situations, then the estimated DIF magnitudes for studied items will be biased. However, it should be emphasized that this problem is inherent to any DIF detection model, not just to the model proposed in this study. See W. C. Wang (2004), for example, for a detailed discussion regarding this issue.

Appendix

Syntax for the Bifactor Multidimensional Item Response Theory (MIRT) Differential Item Functioning (DIF) Model With a Bayesian Estimation by WinBUGS

```
model
{
 mu.Beta1<-mean(Beta1[])
 Beta~dnorm(muBeta,sigBeta)
 mu.z<-mean(z[])
 sd.z<-sd(z[])
 for (i in 1:n){
                a[i]~dnorm(mua,siga) I(0,)
                b[i]~dnorm(mub,sigb)
                Beta1[i]~dnorm(muBeta1,sigBeta1)
                adj.Beta1[i]<-(Beta1[i]-mu.Beta1)
               }
 mua~dnorm(0,.0001)
 mub~dnorm(0,.0001)
 muBeta~dnorm(0,.0001)
 muBeta1~dnorm(0,.0001)
 siga~dchisqr(.5)
 sigb~dchisqr(.5)
 sigBeta~dchisqr(.5)
 sigBeta1~dchisqr(.5)
 for (j in 1:N){
                for (i in 1:n){
                               p[j,i]<- 1/(1+exp((-a[i]*(Beta*Group[j]+z[j]-
 b[i]+gamma[j,i]-Beta1[i]*Group[j]))))
                               U[j,i]~dbern(p[j,i])
                               gamma[j,i]<-gamtes[j,test[i]]
                              }
                z[j]~dnorm(0,1)
                for (i in 1:NT){
                                gamtes[j,i]~dnorm(0,siggam[i])
                               }
               }
 adj.Beta<-(Beta-mu.Beta1)
 for (i in 1:NT){
                 siggam[i]~dchisqr(.5)
                }

}
```

Syntax for an IRT-Based DIF Model With a Bayesian Estimation by WinBUGS

```
model
{
 mu.Beta1<-mean(Beta1[])
 mu.z<-mean(z[])
 sd.z<-sd(z[])
 for (i in 1:n){
                a[i]~dnorm(mua,siga) I(0,)
                b[i]~dnorm(mub,sigb)
                Beta1[i]~dnorm(muBeta1,sigBeta1)
                adj.Beta1[i]<-(Beta1[i]-mu.Beta1)
               }
 mua~dnorm(0,.0001)
 mub~dnorm(0,.0001)
 Beta~dnorm(muBeta,sigBeta)
 muBeta~dnorm(0,.0001)
 muBeta1~dnorm(0,.0001)
 siga~dchisqr(.5)
 sigb~dchisqr(.5)
 sigBeta~dchisqr(.5)
 sigBeta1~dchisqr(.5)
 for (j in 1:N){
                for (i in 1:n){
                               p[j,i]<- 1/(1+exp(-(a[i]*(Beta*Group[j]+z[j]-
 b[i]-Beta1[i]*Group[j]))))
                               U[j,i]~dbern(p[j,i])
                              }
                z[j] ~ dnorm(0,1)
               }
 adj.Beta<-(Beta-mu.Beta1)
}
```

Acknowledgements

The authors gratefully acknowledge the editor and anonymous reviewers for their critical and useful comments on earlier versions of this manuscript.

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 received no financial support for the research, authorship, and/or publication of this article.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Befumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171-187.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4.0 [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

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

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*, 581-612.

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

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.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*, 465-484.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278-295.

Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using a bi-factor multidimensional item response theory model: A Bayesian approach* (Unpublished doctoral dissertation). Florida State University, Tallahassee.

Fukuhara, H., & Kamata, A. (2007, November). *DIF detection in a presence of locally dependent items*. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

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

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

Kim, S. (2001). An evaluation of a Markov chain Monte Carlo method for Rasch model. *Applied Psychological Measurement*, *25*, 163-176.

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32*, 92-109.

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

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. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*, 461-488.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247-260.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: Analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *28*, 237-247.

Wang, W., & Wilson, M. (2005a). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*, 549-576.

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

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *73*, 221-261.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement*, *26*, 109-128.

Wang, X., Bradlow, E. T., Wainer, H., & Muller, E. S. (2008). A Bayesian method for studying DIF: A cautionary tale filled with surprises and delights. *Journal of Educational and Behavioral Statistics*, *22*, 363-384.

Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.