Applied Psychological Measurement
2016, Vol. 40(8) 573–591
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616664304
apm.sagepub.com
SAGE

After Differential Item Functioning Is Detected: IRT Item Calibration and Scoring in the Presence of DIF

Sun-Joo Cho[1], Youngsuk Suh[2], and Woo-yeol Lee[1]

Abstract

Researchers are commonly interested in group comparisons such as comparisons of group means, called impact, or comparisons of individual scores across groups. A meaningful comparison can be made between the groups when there is no differential item functioning (DIF) or differential test functioning (DTF). During the past three decades, much progress has been made in detecting DIF and DTF. However, little research has been conducted on what researchers can do after such detection. This study presents and evaluates a confirmatory multigroup multidimensional item response model to obtain the purified item parameter estimates, person scores, and impact estimates on the primary dimension, controlling for the secondary dimension due to DIF. In addition, the item response model approach was compared with current practices of DIF treatment such as deleting and ignoring DIF items and using multigroup item response models through simulation studies. The authors suggested guidelines for DIF treatment based on the simulation study results.

Keywords

differential item functioning, impact, item response theory, scoring, measurement invariance

Introduction

Many psychological studies involve group comparisons such as cultural, ethnic, gender, or treatment group comparisons. The group comparisons can be made between group means or individual scores across groups. To make meaningful group comparisons, the measurement of a construct or set of constructs is assumed to be equivalent across the groups, which is called measurement invariance (e.g., Meredith & Millsap, 1992). Measurement invariance implies that the distribution of the test score, conditional on a given value of the construct or the latent variable, is invariant across groups. If the measurement invariance does not meet at the test level or

[1]Vanderbilt University, Nashville, TN, USA
[2]Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

Corresponding Author:
Sun-Joo Cho, Peabody College, Vanderbilt University, Peabody Hobbs 213A, 230 Appleton Place, Nashville, TN 37203, USA.
Email: sj.cho@vanderbilt.edu

at the item level, the test or the item is said to present differential test functioning (DTF) or differential item functioning (DIF), respectively, in the item response theory (IRT) approach.

When a test is intended to be unidimensional, DTF or DIF can be viewed as the consequence of one or more dimensions not explained by the *primary* dimension to be measured and the failure to account for the *secondary* dimension can result in DTF or DIF (e.g., Ackerman, 1992; Bolt & Stout, 1996; Shealy & Stout, 1993). Specifically, Shealy and Stout (1993) pointed out that DTF occurs when there are secondary dimension(s) and the reference and focal groups do not have an equal number of secondary dimension(s) (i.e., the two groups have different means).

Procedures for detecting DIF items are by now well established in psychometric research using item response models (see Millsap & Everson, 1993, for an overview of DIF detection methods). However, little attention has been paid to how one can treat DIF items for valid group comparisons. In the item development stage, a larger number of items than is needed is created, and the detected DIF items can be revised or removed. When researchers cannot get involved in item development and have to use developed tests for their own research purposes, it may be a rare case in which they can revise detected DIF items and then recollect data with the revised items. Furthermore, deleting DIF items in the developed tests may result in lowering test reliability and content validity.

Lord (1980) cited a solution (suggested by Gary Marco) to purify a test by deleting DIF items and then scoring only based on non-DIF items. When a large portion of items in a test (e.g., above 50%) are detected as DIF items, using a separate scale for each group is often recommended (Bolt, Hare, Vitale, & Newman, 2004). However, when a portion of test items (e.g., less than 30% of test items) is known as DIF, an IRT purification method can be used to estimate the item parameters, person scores, and group mean difference (called *impact*[1] hereafter) on the primary dimension. De Boeck, Cho, and Wilson (2011) presented a secondary dimension modeling approach to obtain purified item parameter estimates using a confirmatory mixture multidimensional item response model when DIF items are known and groups of interest are unknown (i.e., latent classes or mixtures).

The purpose of this article is to present an IRT purification method for item calibration and scoring in the presence of DIF after a subset of items is detected as DIF items and to compare the performance of the method with that of other current DIF item treatments. The importance of DIF item treatment may differ depending on the purposes of the test used (e.g., Borsboom, 2006). The authors of the current study consider the use of test scores to detect impact and individual differences in the construct being measured. Unlike in De Boeck et al. (2011), this article focuses on manifest groups such as gender and ethnicity (instead of latent classes) using *confirmatory multigroup multidimensional item response model* and evaluates the secondary dimension modeling approach to obtain purified IRT item parameter estimates and person scores via a simulation study. The multigroup multidimensional item response model has been used in DIF detection contexts (e.g., Oshima, Raju, & Flowers, 1997). Novel presentation of the model in the current study is to specify a dimension structure to model a secondary dimension due to DIF. The specified methods can easily be extended to include a number of primary and secondary dimensions; however, for the sake of simplicity, this article focuses on the two dimensional models, one primary dimension and one secondary dimension.

The article is organized as follows. First, survey results about the practices of DIF treatment are presented. Subsequently, the purification method for item calibration and scoring in the presence of DIF items using a confirmatory multigroup multidimensional item response model are described. In addition, other DIF item treatments are described using item response models, based on the survey results of current practice for DIF items. Next, a simulation study was conducted to evaluate the proposed method with a comparison with current practices to treat DIF items. Finally, the article concludes with a summary and discussion.

Examples of DIF Item Treatment

To report how researchers treat DIF items in practice, 27 articles published in five American Psychological Association journals were reviewed. For review results and details, see Table 1 in Online Appendix A. It was observed that there are five distinct practices to deal with DIF items: (a) delete DIF items (30%), (b) no further action (33%; i.e., a specific DIF treatment was not mentioned), (c) ignore DIF items (26%; i.e., all items including DIF items were calibrated), (d) calibrate items for each group (7%; i.e., multigroup analysis), and (e) model DIF (4%). There was one article, Nye and Drasgow (2011), which showed modeling DIF approach. However, they did not model a secondary dimension separate from a primary dimension implying that the group difference and individual scores in their model may not be meaningful for group comparisons.

Below, the modeling DIF approach is presented with a two-parameter confirmatory multigroup multidimensional item response model. For the comparison with the modeling DIF approach, two-parameter unidimensional item response models for deleting DIF, ignoring DIF, and multigroup analysis were shown in Online Appendix B. Multigroup approach (Bock & Zimowski, 1997) allows for separate item parameter estimates for each group regarding DIF items, but through non-DIF items connects the estimates between groups to a common latent metric.

Modeling DIF Using a Multigroup Multidimensional Item Response Model

In the modeling DIF approach, it is assumed that DIF items are known after DIF detection methods are used. In addition, it is assumed that there are shifts with the DIF magnitudes on item parameters for the items suspected of DIF for a focal group. A secondary dimension is modeled to explain individual differences in endorsement probabilities for the focal group and DIF items (see a section of ''DIF items and multidimensionality'' for details in Online Appendix C). The logic of the method is to estimate item parameters, person scores, and impact on a primary dimension from all persons and items, and controlling for the secondary dimension from persons in the focal group and for DIF items. That is, the reference group has one dimension (θ_{1j}), as an interaction between all items and persons in the reference group. The focal group has two dimensions where the first dimension (θ_{1j}) as a primary dimension is applied for all items as in the reference group and the second dimension (θ_{2j}) as the secondary dimension is from an interaction between DIF items and persons in the focal groups. This structure of dimensionality between the two groups can be imposed with a design matrix to map an item to a dimension, specified as q_{ig} for an item i and a group g. Below, the structure is specified using an equation.

A confirmatory multiple-group multidimensional item response model can be described as follows:

$$\text{logit}\left[P\left(y_{jig}=1|\theta_{1jg},\theta_{2jg}\right)\right]=\alpha_{1ig}\cdot q_{1i.g}\theta_{1jg}+\alpha_{2ig}\cdot q_{2i.g}\theta_{2jg}-\beta_{ig}, \tag{1}$$

where α_{1ig} is a group-specific item discrimination for a primary dimension, α_{2ig} is a group-specific item discrimination for a secondary dimension, β_{ig} is a group-specific item location (e.g., item difficulty), θ_{1jg} is a group-specific primary dimension to be measured, θ_{2jg} is a group-specific secondary dimension, $q_{1i.g}$ is a group-specific element of a design matrix to map an item to the primary dimension, and $q_{2i.g}$ is a group-specific element of a design matrix to map an item to the secondary dimension. Figure 2 in Online Appendix D depicts the modeling DIF approach, Equation 1.

For the reference group ($g = 1$), $q_{1i.1}$ for the primary dimension and $q_{2i.1}$ for secondary dimension can be specified as follows for a subset of DIF items and a subset of non-DIF items:

$$\begin{array}{lc} & \begin{array}{cc} q_{1.1} & q_{2.1} \end{array} \\ \begin{array}{l} \mathit{nonDIF} \\ \mathit{nonDIF} \\ \vdots \\ \mathit{nonDIF} \\ \mathit{nonDIF} \\ \mathit{DIF} \\ \mathit{DIF} \\ \vdots \\ \mathit{DIF} \\ \mathit{DIF} \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \end{array}$$

For the focal group ($g = 2$), $q_{1i.2}$ for the primary dimension and $q_{2i.2}$ for the secondary dimension can be specified as follows for a subset of DIF items and a subset of non-DIF items:

$$\begin{array}{lc} & \begin{array}{cc} q_{1.2} & q_{2.2} \end{array} \\ \begin{array}{l} \mathit{nonDIF} \\ \mathit{nonDIF} \\ \vdots \\ \mathit{nonDIF} \\ \mathit{nonDIF} \\ \mathit{DIF} \\ \mathit{DIF} \\ \vdots \\ \mathit{DIF} \\ \mathit{DIF} \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array}$$

In the reference group, all items load on only the primary dimension. In the focal group, all items load on the primary dimension, and only DIF items load on the secondary dimension.

To identify the model, the following constraints are imposed: $\theta_{j11} \sim N(0, 1)$, $\theta_{j22} \sim N(0, 1)$, and $\mathrm{cov}(\theta_{j11}, \theta_{j22}) = 0$. In addition to the three constraints to identify the model, more constraints are required to reflect the structure specified in the matrices above. First, the equality constraint between the reference and focal groups is imposed on item discriminations loaded on the primary dimension and on item locations (i.e., $\alpha_{1i1} = \alpha_{1i2}$ and $\beta_{i1} = \beta_{i2}$ for all items). This indicates that there is the same (purified) primary dimension, θ_{1j}, across the two groups. Second, in the reference group, the mean and variance of the secondary dimension are 0, respectively, and the covariance between the primary dimension and the secondary dimension is 0 because the secondary dimension is modeled only for the focal group. Third, item discriminations for non-DIF items on the secondary dimension are set to 0 in the focal group.

Each item discrimination parameter for the primary dimension (α_{1ig} with the equality constraint) can be a purified parameter. Because the DIF magnitude for each item is explained by the secondary dimension weighted by α_{2ig} (i.e., $\alpha_{2ig} \cdot q_{2i.g}\theta_{2jg}$), an item location parameter, and β_i, a purified parameter. A score for θ_{1j} is a purified score. The mean and variance of the primary dimension in the focal group (i.e., $\theta_{j12} \sim N(\mu, \sigma^2)$) can be estimated because of $\alpha_{1i1} = \alpha_{1i2}$ and $\beta_{i1} = \beta_{i2}$. The mean, μ, is an impact on the primary dimension, which can be used for the meaningful group comparison.

Comparisons Among Four DIF Treatment Practices

Online Appendix B shows the summary of the four DIF treatment approaches, as specified in the earlier section. An example was provided in Online Appendix E to illustrate the four DIF treatment approaches, deleting, ignoring, multigroup, and modeling. In this section, the advantages and disadvantages of each DIF treatment approach in terms of estimating the item parameters for all items (α, β), impact (μ), and person scores (θ) from a primary dimension are explained.

In the deleting DIF items, only non-DIF items can be calibrated. This results in lowering test reliability and content validity, especially when the DIF magnitude is high and the number of DIF is large. The impact parameter cannot be estimated simultaneously with the item parameters, unless it is calculated based on the person scores as a subsequent analysis.

In DIF study literature, the degree of DIF is mainly characterized with respect to DIF magnitudes and the number of DIF items (e.g., Kim & Cohen, 1992; Oshima et al., 1997). Thus, ignoring DIF approach may not be problematic when the DIF magnitude is low and the number of DIF items is small. However, in the presence of non-ignorable DIF (e.g., high DIF magnitudes or the large number of DIF items), item parameter estimates and person scores can be biased. As in the deleting DIF approach, the impact parameter cannot be estimated simultaneously with the item parameters.

In the multigroup DIF approach, because item parameters without DIF magnitudes are estimated only with the reference group for DIF items, the standard errors of item parameter estimates for DIF items can be larger than those with the reference and the focal groups (as in the multigroup DIF approach with two-step and in the modeling DIF approach). In the presence of DIF, the impact and person scores from the multigroup DIF approach with one step are from different dimensions between the two groups. As explained earlier, the reference group has the primary dimension, and the focal group has the primary and the secondary dimension. Thus, in the multigroup DIF approach, the impact is not meaningful, and the person scores cannot be compared on the same scale. These limitations in the one-step approach can be overcome with a multigroup DIF approach with two steps, where the first step is to have the same item parameter estimates between the two groups (using item parameter estimates from the reference group) and the second step is to obtain the impact and person scores. However, this requires the additional step (compared with the modeling DIF approach). In addition, the uncertainty of item parameter estimates can be ignored in estimating the impact because item parameters are considered known parameters in the second step. Thus, the standard error of the impact estimate can be smaller than that of the impact estimate (compared with the modeling DIF approach).

In the modeling DIF approach, comparable item parameter estimates and person scores between the reference and focal groups are obtained using all items, by controlling for the secondary dimension due to the DIF items. In this regard, the modeling DIF approach does not hurt content validity, which is not the case for the deleting DIF approach. Furthermore, because item parameters on the primary dimension are estimated with the equality constraints on the item parameters between the reference group and the focal group, the standard errors of the item parameter estimates can be smaller (compared with the multigroup DIF approach with one step). The item parameters and impact parameter on the primary dimension are estimated simultaneously, such that the uncertainty of the item parameter estimates can be incorporated in the estimation of the impact parameter. However, the number of parameters is larger than the multigroup approach (with two steps) because of the simultaneous modeling for the primary and secondary dimensions. Accordingly, the sampling variability in the modeling DIF approach can be larger than that in the multigroup DIF approach, especially when the number of DIF items increases (because the number of item discriminations of the secondary dimension increases).

Simulation Study

The main interests in the simulation study are the following two questions, assuming the presence of DIF (a) Does the modeling approach perform well in explaining the secondary dimension due to DIF to have purified IRT item parameter estimates and person scores? (b) What are the consequences of deleting and ignoring DIF items (as the two common current practices according to survey results presented in Online Appendix A) in item parameter estimates and person scores? To answer these two questions, DIF and impact were generated based on the two-group (two-parameter) item response model as a population data-generating model (a special case of the multigroup analysis [Equation 2 in Online Appendix B] when some portions of the items are non-DIF items). The model was chosen over the confirmatory multigroup multidimensional item response model (Equation 1) as a population data-generating model because the interest is in how the secondary dimension modeling approach can perform to obtain purified IRT item parameter estimates and person scores, not in parameter recovery of the model.

For the first research question in the simulation study, item parameters and person scores in the population model were compared with the item parameter estimates and the predicted person scores from the primary dimension in the modeling DIF approach using the confirmatory multigroup multidimensional item response model. In comparison with the modeling DIF approach, multigroup DIF approaches with one step (for item parameters) and two steps (for impact and person scores) were fit to the same generated datasets. For the second question, the models for deleting and ignoring DIF practices were fit to the same generated datasets.

Simulation Designs

As the focus of this article was on investigating the effect of DIF items on item parameter estimates and person scores, varying conditions were considered for different patterns of DIF. The simulation conditions include the number of DIF items (10%, 30%, or 50%), magnitude of DIF (low or high), and type of DIF (uniform or nonuniform). In addition, the sample size design (a balanced group design or an unbalanced group design) was considered a varying condition that affects IRT item parameter estimation. As shown in Table 1 in the Online Appendix, item response theory–likelihood ratio–differential item functioning (IRT-LR-DIF) is the most commonly used IRT DIF detection method. Woods (2008) reported a literature review regarding the number of persons and items in 16 papers in which IRT-LR-DIF was applied). The mean number of examinees in the reference group was 1,081, and the mean number of items (excluding a few outliers) was 20. Furthermore, the 20-item test was also chosen in other measurement invariance studies (e.g., Clark & LaHuis, 2012; Finch, 2005; Flowers, Oshima, & Raju, 1999; Meade & Bauer, 2007). For these reasons, fixed conditions were chosen in this study for the two groups (i.e., the reference group and the focal group), 20 items and 2,000 persons in total, as used in Woods. Fully crossed conditions defined by the four varying conditions resulted in 24 (=3×2×2×2) conditions. Five hundred replications were simulated for each of the 24 conditions.

All DIF items were introduced against the focal group. Namely, first, the item parameters for the reference group were generated and then the item parameters for the focal group were manipulated by introducing DIF magnitudes in designated DIF items.

A latent variable for the reference group, θ_{j1}, was generated following $\theta_{j1} \sim N(0, 1)$ (for $j = 1, \ldots, 1000$) and a latent variable for the focal group, θ_{j2}, was generated following $\theta_{j2} \sim N(-0.5, 1)$ (for $j = 1001, \ldots, 2000$). A value of -0.5 was the impact, indicating the mean of the focal group was 0.5 lower than that of the reference group. The generated person scores for all persons ($j = 1, \ldots, 2000$) were from the primary dimension (θ_j) and the impact was on the primary dimension, although the item responses generated with DIF magnitudes ($\delta_i^{(\alpha)}$ and $\delta_i^{(\beta)}$) cannot be explained fully by the primary dimension. Accordingly, the generated person scores and impact were compared with the person scores and impact estimate in the modeling DIF approach. However, as explained earlier, each group had its own dimensionality in the presence of DIF items in the multigroup DIF approach so that person scores from the different groups cannot be compared on the same scale and the impact was not meaningful. Thus, the person scores from the two-step approach in the multigroup DIF approach were compared with the generated person scores. To investigate the consequences of deleting and ignoring DIF items, the generated person scores were compared with the person scores from the deleting and ignoring DIF approaches.

For the reference group, item discriminations were generated from a log-normal distribution with a mean of 0 and a variance of .25 used as a prior distribution in the BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996). Item locations were generated from a standard normal distribution. Table 5 in Online Appendix F presents the generated item parameters for 10% number of the DIF items, nonuniform DIF, and high magnitude to illustrate the generation of DIF conditions to be explained in the following. Item parameters for the reference group in the population data-generating model (α_i and β_i) were those without DIF magnitudes, and these item parameters were compared with purified item parameter estimates from the modeling DIF approach. They were also compared with item parameter estimates from a reference group in the multigroup DIF approach. In addition, the population item parameters for the reference group were compared with the item parameter estimates in the deleting DIF treatment for non-DIF items and in the ignoring DIF treatment for all items. Below, each simulation condition is described in more detail.

Number of DIF items. The 10%, 30%, and 50% DIF items (two items, six items, and 10 items, respectively) were considered the number of DIF items. In the DIF study literature, 30% DIF is considered a large number of DIF (e.g., Oshima et al., 1997). As indicated in the introduction, the purification method may not be recommended over having a separate scale for each group when there is a large portion of DIF items. Reise, Widaman, and Pugh (1993) noted that partial measurement invariance may hold when less than half of the items had significant modification indices (MIs) for factor loadings of the common factor model. The 50% DIF item condition was included to investigate the relative performance of the different DIF treatment in the presence of larger DIF items. The first 18 items, 14 items, and 10 items in Table 5 of Online Appendix F were used as for anchor items (i.e., non-DIF items) for 10%, 30%, and 50%, respectively.

DIF magnitudes and type of DIF items. The DIF items were simulated under each of the four DIF conditions: low and high levels of uniform DIF and low and high levels of nonuniform DIF conditions. The DIF magnitudes were chosen to coincide with other DIF studies (e.g., Oshima et al., 1997; Suh & Bolt, 2011).

For the uniform DIF type, the item location parameters for DIF items increased by 0.5 for the focal group, thus making these items harder for the focal group. The 0.5 difference in the item location represents a low level of uniform DIF magnitude.[2] A high level of DIF magnitude was simulated by introducing a 1.0 difference in the item location parameter.

For the nonuniform DIF type, a low level of nonuniform DIF was introduced at a shift level of 0.3 in the item discrimination parameter, such that the item discrimination parameters for the focal group were set 0.3 lower than for the reference group. For a low level of nonuniform DIF, the item-location parameter(s) for the focal group decreased by 0.5 representing DIF in location. A high level of nonuniform DIF condition was simulated by decreasing a 1.0 difference in location and decreasing a 0.6 difference in discrimination.

Two scale-level effect sizes, signed test difference in the sample (*STDS*) and unsigned test difference in the sample (*UTDS;* Meade, 2010), were calculated to show how much DIF exists in the designed DIF conditions regarding the numbers, types, and magnitudes of DIF. The patterns of scale-level effect sizes can differ, depending on the manipulation of the DIF patterns. Table 6 in Online Appendix G presents two scale-level DIF effect size measures, the *STDS* and the *UTDS*; the values are on the total score scale (the two measures can range from 0 to 20) using one simulated data set for each simulation condition.

Sample size design. Balanced and unbalanced designs were considered. For the balanced design, 1,000 persons were assigned to each group. According to Woods's (2008) literature review of the applications of IRT-LR-DIF, the mean ratio of the mean ratio of the number of examinees in the reference group to the number of examinees in the focal group was 3. Thus, for the unbalanced design, 1,500 persons were assigned to a reference group and 500 persons were assigned to a focal group.

Evaluation Measures

Two accuracy measures were considered to compare the item parameter estimates and person scores across four different DIF approaches: bias and root mean square error (RMSE). The bias is given by $(\hat{\mu} - \mu)$ for an impact estimate as an example and implies the accuracy of the parameter estimates. The RMSE was computed using $\sqrt{\sum_{r=1}^{R}(\hat{\mu} - \mu)^2/R}$, where r indicates the rth replication from a converged solution ($r = 1, \ldots, R$). When a parameter estimate is unbiased, the RMSE quantifies precision (i.e., the variance of the estimator). For biased parameter estimates, the RMSE combines the bias and the precision into the overall accuracy. The average bias and RMSE across items were reported for item parameters (α, β), and the average bias and RMSE across persons were presented for person scores (θ). In addition, to show the differences in item parameter estimate precision in the multigroup and modeling DIF approaches, the ratio of the standard error (*SE*) of the item parameter estimate from the multigroup DIF approach (SE_{MG}) to that of the modeling DIF approach (SE_M; i.e., relative efficiency) was calculated. Then, the average ratio across items was considered. Furthermore, IRT reliability (Green, Bock, Humphreys, Linn, & Reckase, 1984) for person scores was compared across four DIF approaches mainly to show to what extent reliability can be affected by each approach.

Analysis

Mplus 7.11 (Muthén & Muthén, 1998-2014) was used to fit four models with marginal maximum likelihood estimation (ESTIMATOR=MLR in Mplus). For multigroup and modeling approaches, the KNOWNCLASS option for TYPE=MIXTURE was used in Mplus. Prediction errors of the person scores are not available for multigroup and modeling approaches with the MLR estimator. Thus, ESTIMATOR=BAYES (with default priors and hyperpriors) was used to obtain the prediction errors, and 100 imputations were used in Mplus. The posterior median of the person scores was used to calculate the IRT reliability when the posterior distribution for the portion of the person scores was not symmetric.

Results

Due to the page limit, the hypotheses for the simulation study are reported in Online Appendix H. No convergence problems were encountered in any replication for all four approaches in the balanced design. However, in the unbalanced design, one replication had a convergence problem in the modeling DIF approach. The replication was excluded from the analyses of the results.

As expected from the research hypotheses in Online Appendix H, the patterns in the results were similar between the balanced and unbalanced designs for DIF effects, even though there were different magnitudes of bias and RMSE due to the different number of persons in the reference and focal groups in the designs. Regarding the magnitudes of RMSE in the balanced and unbalanced designs, the expected results were found for the multigroup and modeling DIF approaches (except for the impact estimates in the two-step multigroup DIF approach, the impact estimates were not much different between the balanced and unbalanced designs). However, unexpectedly, in the deleting and ignoring DIF approaches, the RMSEs were smaller in the unbalanced design than those in the balanced design for the item location estimates and the person scores. Further investigation showed that these unexpected results were from the different location shift when the impact was ignored in the deleting and ignoring DIF approaches. Specifically, the mean true person scores across all persons was 0.264 in the balanced design, whereas it was 0.137 in the unbalanced design. The smaller shift in the unbalanced design than in the balanced design resulted in a smaller bias for the item location estimates and the person scores.

Because this study focused on DIF item treatment comparisons, the differing DIF effects results are reported for the balanced design below. Results for the unbalanced design are reported in Online Appendix I.

Item parameters. Tables 1 and 2 report the average bias, RMSE, and ratio across the items for the item discrimination estimates and the item location estimates, respectively. The item parameter results for the deleting DIF approach were not comparable with the other three approaches because only non-DIF items were used for calibration. However, bias and RMSE are reported in Tables 1 and 2 to interpret them within the deleting DIF approach. The results of the multigroup approach were from the item parameter estimates with the one-step approach, and the results for the item discrimination parameters of the modeling DIF approach were based on the primary dimension.

In the deleting DIF approach, bias for the item discrimination (α) estimates was similar across DIF conditions, and RMSE increased mainly with the increasing number of DIF items. Bias and RMSE for the item-location parameter (β) estimates increased mainly with the increasing number of DIF items.

For the other three approaches, the following patterns were found for the item discrimination parameter (α). First, in terms of bias, the multigroup and modeling DIF approaches had similar values across the simulation conditions and produced smaller bias than the ignoring DIF approach (except one condition, nonuniform type, low magnitude, and 30% DIF item). Second, RMSE for the multigroup DIF approach was smaller than that of the modeling DIF approach,

Table 1. Average Bias, RMSE, and Ratio Across Items for Item Discrimination Estimates.

		Uniform				Nonuniform			
No. of DIF items	Magnitude	Deleting	Ignoring	Multigroup	Modeling	Deleting	Ignoring	Multigroup	Modeling
A. Bias									
10%	Low	0.032	0.037	0.000	0.001	0.032	0.007	0.000	0.005
	High	0.030	0.037	−0.004	−0.002	0.032	0.011	0.000	0.003
30%	Low	0.031	0.047	−0.004	−0.002	0.031	0.000	−0.003	0.001
	High	0.033	0.062	0.000	0.002	0.032	−0.026	0.000	0.010
50%	Low	0.032	0.060	−0.003	−0.001	0.036	−0.018	0.000	−0.001
	High	0.032	0.092	−0.002	−0.003	0.035	−0.059	0.000	0.004
B. RMSE									
10%	Low	0.090	0.092	0.083	0.089	0.089	0.102	0.085	0.095
	High	0.089	0.092	0.082	0.093	0.090	0.099	0.086	0.091
30%	Low	0.096	0.098	0.083	0.099	0.096	0.093	0.092	0.092
	High	0.098	0.116	0.084	0.128	0.095	0.108	0.092	0.101
50%	Low	0.106	0.108	0.083	0.108	0.109	0.096	0.099	0.092
	High	0.106	0.153	0.084	0.177	0.107	0.113	0.100	0.100
C. Ratio									
10%	Low	—	—	—	0.995	—	—	—	1.017
	High	—	—	—	0.996	—	—	—	1.012
30%	Low	—	—	—	0.999	—	—	—	1.083
	High	—	—	—	1.010	—	—	—	1.061
50%	Low	—	—	—	1.002	—	—	—	1.143
	High	—	—	—	1.038	—	—	—	1.127
Aggregated bias									
No. of DIF items	10%	0.032	0.023	−0.001	0.002				
	30%	0.032	0.021	−0.002	0.002				
	50%	0.034	0.019	−0.001	0.000				
Magnitude	Low	0.032	0.022	−0.002	0.001				
	High	0.032	0.020	−0.001	0.002				
Type	Uniform	0.032	0.056	−0.002	−0.001				
	Nonuniform	0.033	−0.014	−0.001	0.004				
Aggregated RMSE									
No. of DIF items	10%	0.090	0.096	0.084	0.092				
	30%	0.096	0.104	0.088	0.105				
	50%	0.107	0.118	0.092	0.119				
Magnitude	Low	0.098	0.098	0.088	0.096				
	High	0.098	0.114	0.088	0.115				
Type	Uniform	0.098	0.110	0.083	0.116				
	Nonuniform	0.098	0.102	0.092	0.095				
Aggregated ratio									
No. of DIF items	10%	—	—	—	1.005				
	30%	—	—	—	1.038				
	50%	—	—	—	1.078				
Magnitude	Low	—	—	—	1.040				
	High	—	—	—	1.041				
Type	Uniform	—	—	—	1.007				
	Nonuniform	—	—	—	1.074				

Note. RMSE = root mean square error; DIF = differential item functioning; - = not applicable.

Table 2. Average Bias, RMSE, and Ratio Across Items for Item Location Estimates.

		Uniform				Nonuniform			
No. of DIF items	Magnitude	Deleting	Ignoring	Multigroup	Modeling	Deleting	Ignoring	Multigroup	Modeling
A. Bias									
10%	Low	0.298	0.327	0.020	0.020	0.299	0.267	0.021	0.024
	High	0.299	0.350	0.022	0.020	0.298	0.340	0.020	0.019
30%	Low	0.313	0.378	0.020	0.026	0.313	0.359	0.020	0.026
	High	0.313	0.448	0.020	0.025	0.314	0.412	0.020	0.028
50%	Low	0.344	0.427	0.020	0.033	0.343	0.395	0.019	0.037
	High	0.343	0.549	0.021	0.034	0.345	0.486	0.019	0.049
B. RMSE									
10%	Low	0.306	0.334	0.070	0.079	0.306	0.289	0.070	0.081
	High	0.306	0.357	0.070	0.101	0.306	0.346	0.070	0.089
30%	Low	0.321	0.384	0.073	0.112	0.321	0.365	0.073	0.108
	High	0.321	0.454	0.073	0.181	0.321	0.418	0.073	0.172
50%	Low	0.350	0.432	0.076	0.134	0.350	0.400	0.077	0.132
	High	0.350	0.553	0.077	0.219	0.350	0.491	0.078	0.226
C. Ratio									
10%	Low	—	—	—	1.010	—	—	—	1.011
	High	—	—	—	1.003	—	—	—	1.006
30%	Low	—	—	—	1.048	—	—	—	1.061
	High	—	—	—	1.027	—	—	—	1.041
50%	Low	—	—	—	1.081	—	—	—	1.105
	High	—	—	—	1.057	—	—	—	1.078
Aggregated bias									
No. of DIF items	10%	0.299	0.321	0.021	0.021				
	30%	0.313	0.399	0.020	0.026				
	50%	0.344	0.464	0.020	0.038				
Magnitude	Low	0.318	0.359	0.020	0.028				
	High	0.319	0.431	0.020	0.029				
Type	Uniform	0.318	0.413	0.021	0.026				
	Nonuniform	0.319	0.377	0.020	0.031				
Aggregated RMSE									
No. of DIF items	10%	0.306	0.332	0.070	0.088				
	30%	0.321	0.405	0.073	0.143				
	50%	0.350	0.469	0.077	0.178				
Magnitude	Low	0.326	0.367	0.073	0.108				
	High	0.326	0.437	0.074	0.165				
Type	Uniform	0.326	0.419	0.073	0.138				
	Nonuniform	0.326	0.385	0.074	0.135				
Aggregated ratio									
No. of DIF items	10%	—	—	—	1.008				
	30%	—	—	—	1.044				
	50%	—	—	—	1.080				
Magnitude	Low	—	—	—	1.053				
	High	—	—	—	1.035				
Type	Uniform	—	—	—	1.038				
	Nonuniform	—	—	—	1.050				

Note. RMSE = root mean square error; DIF = differential item functioning; - = not applicable.

which was the expected result because of the larger number of parameters in the modeling DIF approach. In this approach, RMSE can be larger, with an increasing number of DIF items and high magnitudes in uniform and nonuniform DIF types. However, this pattern was found only in the nonuniform DIF type in the multigroup DIF approach. Third, the average ratio of SE_{MG} to SE_M was higher than 1.0 for all conditions, except three simulation conditions in the uniform DIF type, in which the standard errors of the item discrimination estimates were smaller in the modeling DIF approach than in the multigroup DIF approach. For all simulation conditions except the three conditions, larger differences in the ratio were found as the number of DIF items increased especially with the nonuniform DIF type.

For the item-location parameter (β), the following patterns were observed in the ignoring, multigroup, and modeling DIF approaches. First, compared with the item discrimination parameter, larger differences in bias and RMSE for the item location were found between the multigroup approach and the ignoring approach and between the modeling approach and the ignoring approach. Second, bias and RMSE for the multigroup DIF approach were similar across all DIF simulation conditions, which were expected because it is a population data-generating model. However, bias and RMSE for the modeling DIF approach increased as the number of DIF items and the DIF magnitudes increased. Third, bias was similar between the multigroup and modeling DIF approaches when there were 10% and 30% DIF items. However, the bias for the multigroup DIF approach was smaller than the bias for the modeling DIF approach where there were 50% DIF items. Fourth, the RMSE for the multigroup DIF approach was smaller than the RMSE for the modeling DIF approach. The difference between the two approaches became larger as the number of DIF items increased. Fifth, according to the average ratio of SE_{MG} to SE_M, the standard errors of the item location estimates were smaller in the modeling DIF approach than in the multigroup DIF approach for all simulation conditions, as expected.

Person scores and IRT reliability. Table 3 presents the average bias and RMSE across persons for person scores (θ) and IRT reliability for each DIF treatment approach. In the table, the results of the multigroup DIF approach were based on the two-step approach. When only non-DIF items were used for scoring in the deleting DIF approach, the bias of the person scores did not change across all levels of DIF conditions, whereas the RMSE for the person scores was influenced by the number of DIF items. Among the other three DIF treatments, the following patterns were found. First, bias and RMSE for the ignoring DIF approach were larger than the multigroup and modeling DIF approaches. Although there were some patterns, there were small differences between the two approaches. Second, overall, slightly larger bias and RMSE (except two conditions) were found in the modeling DIF approach than in the multigroup DIF approach for the uniform condition. However, the opposite pattern was found for the nonuniform condition except three conditions in which the same RMSE was found between the multigroup and modeling DIF approaches. Third, bias and RMSE in the multigroup and modeling approaches increased as the number of DIF items and the DIF magnitudes increased (except conditions with 10% and nonuniform DIF type). As expected, IRT reliability was mainly influenced by the number of DIF items in the ignoring DIF approach. IRT reliability was similar between the multigroup DIF treatment (with two steps) and the modeling DIF approach, and was not affected by the simulation conditions.

Impact and variances of person scores. The impact (μ) and variance (σ^2) of person scores for the focal group can be estimated in the multigroup DIF approach (with two steps) and the modeling DIF approach. Results are reported in Tables 4 and 5 for impact and variance estimates, respectively. The following patterns were found for the impact estimates. First, overall, larger bias and RMSE were found in the multigroup DIF approach than the modeling DIF approach (except one condition in RMSE, 10% DIF, low magnitude, and uniform DIF type). Second, for all conditions, the impact was underestimated in the multigroup DIF approach (except one condition, 10% DIF, low magnitude, and nonuniform type in which there was no cancelation of DIF across items and persons), whereas it was slightly overestimated in the modeling DIF approach. Third,

Table 3. Average Bias and RMSE Across Persons for Person Scores and IRT Reliability.

		Uniform				Nonuniform			
No. of DIF items	Magnitude	Deleting	Ignoring	Multigroup	Modeling	Deleting	Ignoring	Multigroup	Modeling
A. Bias									
10%	Low	0.264	0.261	0.003	0.001	0.264	0.263	0.054	0.040
	High	0.264	0.264	−0.015	−0.029	0.264	0.264	−0.006	−0.005
30%	Low	0.263	0.264	−0.035	−0.038	0.263	0.264	−0.019	−0.015
	High	0.264	0.264	−0.082	−0.095	0.264	0.264	−0.055	−0.046
50%	Low	0.264	0.264	−0.067	−0.071	0.264	0.264	−0.036	−0.034
	High	0.264	0.265	−0.149	−0.170	0.264	0.265	−0.092	−0.085
B. RMSE									
10%	Low	0.499	0.482	0.413	0.414	0.498	0.487	0.416	0.415
	High	0.499	0.482	0.415	0.418	0.499	0.486	0.419	0.419
30%	Low	0.523	0.481	0.420	0.422	0.523	0.489	0.424	0.424
	High	0.524	0.485	0.437	0.445	0.523	0.499	0.445	0.441
50%	Low	0.553	0.482	0.431	0.434	0.553	0.495	0.435	0.435
	High	0.553	0.491	0.474	0.489	0.553	0.516	0.474	0.470
C. Reliability									
10%	Low	0.773	0.798	0.780	0.786	0.773	0.792	0.786	0.786
	High	0.772	0.798	0.780	0.785	0.773	0.792	0.787	0.786
30%	Low	0.729	0.800	0.781	0.786	0.729	0.789	0.785	0.786
	High	0.730	0.802	0.782	0.786	0.730	0.782	0.782	0.787
50%	Low	0.669	0.802	0.784	0.787	0.670	0.786	0.784	0.788
	High	0.669	0.807	0.784	0.784	0.669	0.775	0.781	0.789
Aggregated bias									
No. of DIF items	10%	0.264	0.263	0.009	0.002				
	30%	0.264	0.264	−0.048	−0.049				
	50%	0.264	0.265	−0.086	−0.090				
Magnitude	Low	0.264	0.263	−0.017	−0.020				
	High	0.264	0.264	−0.067	−0.072				
Type	Uniform	0.264	0.264	−0.058	−0.067				
	Nonuniform	0.264	0.264	−0.026	−0.024				
Aggregated RMSE									
No. of DIF items	10%	0.499	0.484	0.416	0.417				
	30%	0.523	0.489	0.432	0.433				
	50%	0.553	0.496	0.454	0.457				
Magnitude	Low	0.525	0.486	0.423	0.424				
	High	0.525	0.493	0.444	0.447				
Type	Uniform	0.525	0.484	0.432	0.437				
	Nonuniform	0.525	0.495	0.436	0.434				
Aggregated reliability									
No. of DIF items	10%	0.760	0.795	0.783	0.786				
	30%	0.730	0.793	0.783	0.786				
	50%	0.669	0.793	0.783	0.787				
Magnitude	Low	0.724	0.795	0.783	0.787				
	High	0.715	0.793	0.783	0.786				
Type	Uniform	0.715	0.801	0.782	0.786				
	Nonuniform	0.724	0.786	0.784	0.787				

Note. Results for multigroup approach were based on the two-step approach to compare results between the reference and focal groups; bias and RMSE for the multigroup approach were based on maximum likelihood estimates for the comparison with deleting and ignoring approaches; reliability for the multigroup (with the two-step) and modeling approaches were calculated based on Bayes estimation. RMSE = root mean square error; IRT = item response theory; DIF = differential item functioning.

Table 4. Bias, RMSE, and Ratio for Impact Estimates.

		Uniform		Nonuniform	
No. of DIF items	Magnitude	Multigroup	Modeling	Multigroup	Modeling
A. Bias					
10%	Low	−0.021	0.016	0.080	0.015
	High	−0.057	0.017	−0.040	0.016
30%	Low	−0.096	0.012	−0.063	0.013
	High	−0.191	0.017	−0.138	0.015
50%	Low	−0.162	0.013	−0.100	0.015
	High	−0.326	0.017	−0.212	0.016
B. RMSE					
10%	Low	0.026	0.028	0.042	0.028
	High	0.059	0.029	0.081	0.028
30%	Low	0.097	0.029	0.065	0.029
	High	0.191	0.030	0.139	0.029
50%	Low	0.163	0.030	0.101	0.032
	High	0.327	0.031	0.212	0.032
Aggregated bias					
No. of DIF items	10%	−0.010	0.016		
	30%	−0.122	0.014		
	50%	−0.200	0.015		
Magnitude	Low	−0.060	0.014		
	High	−0.161	0.016		
Type	Uniform	−0.142	0.015		
	Nonuniform	−0.079	0.015		
Aggregated RMSE					
No. of DIF items	10%	0.052	0.028		
	30%	0.123	0.029		
	50%	0.201	0.031		
Magnitude	Low	0.082	0.029		
	High	0.168	0.030		
Type	Uniform	0.144	0.030		
	Nonuniform	0.107	0.030		

Note. Impact for the multigroup approach was estimated with the two-step approach. RMSE = root mean square error; DIF = differential item functioning.

bias and RMSE increased when the number of DIF items and the magnitudes of the DIF items increased in the multigroup DIF approach. This pattern was also found for bias in the modeling DIF approach, but the degree of the effects of the simulation factors was not as large as in the multigroup DIF approach. RMSE was mainly influenced by the number of DIF items in the modeling DIF approach.

The following patterns were observed regarding the variance of person scores. First, bias in the multigroup DIF approach was larger than bias in the modeling DIF approach across all conditions. However, RMSE was smaller in the multigroup DIF approach than that of the modeling DIF approach in the uniform DIF type (except one condition, 30% DIF, high magnitude, uniform DIF type), whereas the opposite pattern was found in the nonuniform DIF. Second, as shown in Tables 4 for impact estimates, bias and RMSE in the multigroup DIF approach increased as the number of DIF items and the magnitudes of the DIF items increased. This pattern was also observed for bias in the modeling approach, but the degree of the effects of the simulation factors was not as large as in the multigroup approach.

Table 5. Bias, RMSE, and Ratio for Variance Estimates of Person Scores.

		Uniform		Nonuniform	
No. of DIF items	Magnitude	Multigroup	Modeling	Multigroup	Modeling
A. Bias					
10%	Low	−0.017	−0.005	−0.081	−0.003
	High	−0.029	0.003	−0.121	−0.003
30%	Low	−0.014	0.001	−0.162	−0.001
	High	−0.055	−0.007	−0.315	−0.008
50%	Low	0.018	−0.004	−0.233	−0.001
	High	0.024	−0.007	−0.432	0.001
B. RMSE					
10%	Low	0.043	0.055	0.088	0.059
	High	0.047	0.058	0.126	0.059
30%	Low	0.041	0.056	0.165	0.066
	High	0.067	0.060	0.316	0.066
50%	Low	0.044	0.058	0.235	0.075
	High	0.048	0.058	0.433	0.076
Aggregated bias					
No. of DIF items	10%	−0.062	−0.002		
	30%	−0.137	−0.004		
	50%	−0.156	−0.003		
Magnitude	Low	−0.082	−0.002		
	High	−0.155	−0.004		
Type	Uniform	−0.012	−0.003		
	Nonuniform	−0.224	−0.003		
Aggregated RMSE					
No. of DIF items	10%	0.076	0.058		
	30%	0.147	0.062		
	50%	0.190	0.067		
Magnitude	Low	0.103	0.062		
	High	0.173	0.063		
Type	Uniform	0.048	0.058		
	Nonuniform	0.227	0.067		

Note. Variance for the multigroup approach was estimated with the two-step approach. RMSE = root mean square error; DIF = differential item functioning.

Discussion

The purpose of this study was to present the modeling DIF approach and to evaluate it by comparing its performance with that of other DIF treatments such as deleting, ignoring, and multigroup (with one step for item parameters and two steps for impact and person scores) approaches. Overall, the simulation results were consistent with the hypothesized ones, with few exceptions as noted earlier. The following general patterns were found in the simulation study. First, the multigroup and modeling DIF approaches outperformed the deleting and ignoring approaches for item parameter estimates and person scores. Second, overall, the multigroup approach with two steps works well, compared with the modeling DIF approach. Third, the modeling DIF approach can be a viable method to treat DIF items for most DIF conditions, except for the larger number of DIF items (e.g., 50%). Below, guidelines in choosing one DIF treatment method over another based on the simulation results are provided.

Parameters and Information of Interest

Given the mixed results for the multigroup and modeling DIF approaches, researchers can choose one of the methods depending on the parameters of interest (i.e., item parameters, person scores, impact, and variance of the person scores) and information they need (i.e., accuracy [quantified with bias], overall accuracy [quantified with RMSE], or precision [quantified with standard error]).

The simulation results showed that the multigroup DIF approach with one step can provide better overall accuracy for item parameter estimates than the modeling DIF approach in most DIF conditions. For the nonuniform DIF type, the overall accuracy for item discrimination parameter estimates in the modeling DIF approach can be similar to that in the ignoring DIF approach. The overall accuracy of the item location estimates in the modeling DIF approach was similar to that of the multigroup approach with one step only when there are 10% and 30% DIF items. The modeling DIF and the multigroup DIF with a two-step approach for item parameter estimates are recommended when the number of DIF items is not large (e.g., less than 30%).

However, the standard error of the item parameter estimates in the modeling DIF approach can be smaller than that of the multigroup DIF approach, because item parameters are estimated based on all persons of the data in the modeling DIF approach, whereas they are estimated from persons in the reference group. The item parameter estimates from the modeling DIF approach can be more precise than those from the multigroup DIF approach, especially when there are high DIF magnitudes and a large number of DIF items (e.g., 50%). If there are a larger number of persons in the reference group, the precision can be relatively good for the multigroup approach with a two-step approach. Unless there are many more persons in the reference group, the modeling DIF approach is preferred to the multigroup DIF approach when researchers need to use the standard errors of the item parameter estimates such as creating an item bank and implementing IRT equating.

For the person scores, there were small differences between the multigroup DIF approach and the modeling DIF approach based on the simulation results. However, there was the pattern that the multigroup DIF approach can be slightly better than the modeling DIF approach for the uniform DIF type. On the contrary, the modeling DIF approach can be better than the multigroup DIF approach for the nonuniform DIF type. Thus, one of the approaches can be chosen, depending on the DIF type in DIF analyses. Regarding impact and variance, the modeling DIF approach performed better than the multigroup DIF approach in terms of accuracy (bias) and overall accuracy (RMSE).

Multigroup Analysis Versus Modeling DIF

The specification of the multigroup and modeling approaches was based on the assumption that one of the two groups can be set as a reference group and another is a focal group. It is more critical to justify the reference group selection in the multigroup approach than in the modeling approach. When one of the two groups is arbitrarily set as the reference group, results (i.e., purified item parameters, person scores, impact, and variance of the person scores) can change in the multigroup approach because of the two-step nature for scoring and impact estimation in the multigroup approach. In contrast, the results from the modeling approach will not change in the case of having an arbitrarily chosen reference group. Thus, when a reference group cannot be clearly justified, using modeling approach is recommended over multigroup approach.

DIF Effect Sizes and Test Score Uses

DIF effects can be characterized differently, depending on the number of DIF items, the magnitude of the DIF, and the type of DIF. The DIF effect sizes can be calculated at the item and at the test levels to quantify the effects of these factors on the total score scale or latent variable scale. The DIF effect sizes can be a useful guideline in deciding a DIF treatment method.

In interpreting DIF effect sizes, it is important to decide whether the interpretation is made for the individual score level or the whole population level (e.g., Borsboom, 2006). As an example of the whole population level, when there is a large portion of nonuniform DIF items, cancelation is allowed between a reference group and a focal group. In this case, it is expected that differential effects between ignoring DIF and multigroup analysis or between ignoring DIF and modeling DIF will be small in impact estimate. When researchers care about the impact only, ignoring DIF would not cause bias in the presence of a full cancelation (i.e., $|STDS| - UTDS = 0$), even though the test scores are composed of DIF items. In addition, the ignoring DIF approach can be a better approach than the deleting DIF approach when researchers are interested in the impact. Deleting DIF items can result in bias for the impact because it can disturb the cancelation at the test level. However, the cancelation occurs at the population distribution score level, rather than at the individual score level (Borsboom, 2006). Thus, ignoring DIF can result in invalid score comparisons for individuals.

In fact, it may be challenging to provide a general guideline for interpreting DIF effect sizes because the interpretation of the magnitude of DIF effect size may vary across test score uses. For example, the scale-level DIF effect size (ranged from 0 to 10), $STDS = 2$, can be seriously taken for high-stakes tests, whereas it can be an ignorable effect size in low-stakes tests. It was not the aim of this study to provide an absolute guideline for interpreting the DIF effect size. It is hoped that the simulation study results can facilitate discussions of interpreting DIF effect sizes, regarding various DIF conditions and DIF treatments.

This study has several limitations. First, as in other simulation studies, the simulation conditions employed in the study design were limited, including the 20-item test and -0.5 impact, because the authors' main interest in the simulation study was to evaluate their proposed method and compare it with other methods under various DIF conditions. Investigating the four different approaches in more extensive simulation studies including various testing conditions is needed to provide more general guidelines for item calibration and scoring practices in the presence of DIF items.

Second, the multigroup and modeling approaches were based on the assumptions that there are two categories of items (i.e., DIF items and non-DIF items), and the DIF items were flagged with the right criterion. Thus, the performance of these approaches can differ depending on the quality of the DIF detection. It has been known that power and Type I error of DIF item detection vary across DIF detection methods (e.g., Bolt, 2002). Thus, it may be common to have different categorizations of DIF items (i.e., DIF items vs. non-DIF items) depending on the DIF detection method, which can be a potential problem with using the multigroup and modeling approaches. One possible strategy for dealing with this problem is to show the sensitivity of the item parameter estimates, person scores, and impact results to different DIF categorizations with different DIF detection methods. When consistent results occur among the different DIF detection methods, the results can finally be reported.

Third, the model specification for the modeling DIF approach is limited to one primary dimension and one secondary dimension for binary responses. It would be worthwhile to extend the modeling DIF approach for more complex DIF patterns such as more than one primary and secondary dimensions and/or for polytomous responses.

The current study presented and evaluated IRT purification methods for item calibration and scoring in the presence of DIF after a subset of items was detected as DIF items. In addition, the method was compared with current DIF treatment methods, deleting DIF, ignoring DIF, and multigroup approaches. It is hoped that this article improves the practice of dealing with DIF items and leads to further discussions and studies on the treatment of DIF items in evaluating measures.

Acknowledgment

The authors are grateful to Dr. Isabel Gauthier (Vanderbilt University) for making the data available for application and to Dr. Sonya Sterba (Vanderbilt University) for comments on an earlier draft.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Angoff (1993) noted that *impact* can be used to refer to ‘‘the true or unassailable difference between the groups’’ and ‘‘an artifactual difference brought about by the use of inappropriate and irrelevant (DIF) items’’ (p. 18). In this article, the authors used the term *impact* to indicate ‘‘the true or unassailable difference between the groups.’’
2. Low does not mean an absolute size of differential item functioning (DIF) such as small size of DIF. Low may represent a medium size of DIF in practice. Likewise, low and high nonuniform DIF should be interpreted relatively, not absolutely.

Supplemental Material

The online appendices are available at http://apm.sagepub.com/supplemental

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67-91.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Bock, D. R., & Zimowski, M. F. (1997). The multiple group IRT. In W. J. van der Linden & R. K. Hambleton. (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer-Verlag.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113-141.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist-revised. *Psychological Assessment*, *16*, 155-168.

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

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Clark, P. C., & LaHuis, D. M. (2012). An examination of power and type I errors for two differential item functioning indices using the graded response model. *Organizational Research Methods*, *15*, 229-246.

De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*, 583-603.

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.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT Framework. *Applied Psychological Measurement*, *23*, 309-326.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347-360.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection DIF. *Journal of Educational Measurement*, *29*, 51-66.

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

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*, 728-743.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*, 611-635.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Pscyhometrika*, *57*, 289-311.

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

Muthén, L. K., & Muthén, B. O. (1998-2014). Mplus [Computer program]. Los Angeles, CA: Muthén & Muthén.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966-980.

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.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*, 188-205.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*, *32*, 511-526.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software.