Jin and Kang *Large-scale Assess Educ (2016) 4:18*
DOI 10.1186/s40536-016-0033-3

Large-scale Assessments in Education

METHODOLOGY Open Access

Comparing DIF methods for data with dual dependency

Ying Jin[1*] and Minsoo Kang[2]

*Correspondence:
ying.jin@mtsu.edu
[1] Department of Psychology, Middle Tennessee State University, Jones Hall, 308, Murfreesboro, TN 37130, USA
Full list of author information is available at the end of the article

Abstract

Background: The current study compared four differential item functioning (DIF) methods to examine their performancesin terms of accounting for dual dependency (i.e., person and item clustering effects) simultaneously by a simulation study, which is not sufficiently studied under the current DIF literature. The four methods compared are logistic regression accounting neither person nor item clustering effect, hierarchical logistic regression accounting for person clustering effect, the testlet model accounting for the item clustering effect, and the multilevel testlet model accounting for both person and item clustering effects. The secondary goal of the current study was to evaluate the trade-off between simple models and complex models for the accuracy of DIF detection. An empirical example analyzing the 2011 TIMSS Mathematics data was also included to demonstrate the differential performances of the four DIF methods. A number of DIF analyses have been done on the TIMSS data, and rarely had these analyses accounted for the dual dependence of the data.

Results: Results indicated the complex models did not outperform simple models under certain conditions, especially when DIF parameters were considered in addition to significance tests.

Conclusions: Results of the current study could provide supporting evidence for applied researchers in selecting the appropriate DIF methods under various conditions.

Keywords: Multilevel, Testlet, TIMSS

Background

During the past few decades, there have been many studies conducted to evaluate the comparative performance of differential item functioning (DIF) methods under various conditions. These conditions, for example, include small and unbalanced sample size between groups (Woods 2009), short tests (Paek and Wilson 2011), various levels of DIF contamination (Finch 2005), multilevel data (French and Finch 2010), violation of the normality assumption of latent traits (Woods 2011), and violation of the unidimensionality assumption (Lee et al. 2009). Among these conditions, violation of the local independence assumption has gained more attention recently, especially for large-scale assessments where local independence assumption is often violated. For example, the Trends in International Mathematics and Science Study (TIMSS) collected data from more than 60 countries worldwide in year 2011. Data collected from such an assessment, which consist of subdomains of a specific subject (e.g., algebra in the mathematics

achievement test, or biology in the Science achievement test), are multilevel in nature because the primary sampling units are schools instead of individual students from each country.

The dependency of such data has two sources, person clustering effect due to the sampling strategy (e.g., individual students from the same school are dependent) and item clustering effect due to the format of the assessment (e.g., items within the same subdomain are dependent). Previous studies, however, have investigated person and item clustering effects on the comparative performance of several DIF methods, separately (e.g., French and Finch 2013; Wang and Wilson 2005).

For the current study, the primary goal is to compare four DIF methods to examine their performance in terms of accounting for dual dependency (i.e., person clustering effect and item clustering effect, Jiao and Zhang 2015) simultaneously using a simulation study, which is not sufficiently studied under the current DIF literature. An empirical example analyzing the 2011 TIMSS Mathematics data is also included to demonstrate the differential performance of the DIF methods. A number of DIF analyses have been done on the TIMSS data, and rarely had these analyses accounted for the dual dependence of the data (e.g., Innabi and Dodeen 2006; Klieme and Baumert 2001; Wu and Ercikan 2006).

Results of the current study are expected to supplement the current DIF literature when data are dually dependent in terms of both simulation and empirical studies. In the following sections, dual dependency in the DIF literature and the four DIF methods will be briefly reviewed. The review will focus on the effect of dual dependency on the comparative performance of DIF methods in terms of significance tests (e.g., type I error rate). Additionally, we will evaluate the trade-off between simple and complex DIF methods for the accuracy of DIF detection when data is dually dependent. Related previous research will also be reviewed.

Item clustering effect

An item clustering effect is often observed in achievement assessments where testlets are included, and the items within the same testlet are not locally independent due to the shared content of the testlet. A typical example is several items clustering within the same reading passage. Students' reading achievements are typically evaluated by the target ability as well as a secondary ability to understand the content of the passage. For example, passages in a reading achievement test may contain sports-related content, where the target ability is reading skills and the secondary ability is understanding what the content said about sports.

When IRT-based DIF methods are used, inaccurate DIF detection results might occur when the unidimensionality assumption of IRT models is violated due to the item clustering effect (Fukuhara and Kamata 2011). In addition, the performance of non-parametric DIF methods can also be adversely affected by the item clustering effect. Lee et al. (2009) study found out that the SIBTEST method (Shealy and Stout 1993) was conservative in terms of type I error rate unless the DIF size was large (e.g., DIF size = 1 indicating the mean ability between the reference and focal groups differ by one standard deviation under the scale of standard normal distribution).

In order to account for the item clustering effect on DIF analysis, several DIF methods have been developed. Wainer et al. (1991) developed a polytomous approach to detecting

DIF at the testlet level, such that the responses of dichotomous items within the same testlet were added up to form a polytomous item for each testlet. This approach detects DIF at the testlet level. Researchers who are interested in DIF analysis at the item level might find this approach less feasible. To detect DIF at the item level, Wang and Wilson (2005) developed a Rasch testlet model by including a random testlet effect to account for the item clustering effect, and a DIF parameter for DIF detection. Their testlet model can be extended to 2-parameter and 3-parameter IRT testlet models for DIF detection by including discrimination and guessing parameters.

Another DIF method was to employ the bifactor model to account for the item clustering effect (Cai et al. 2011; Jeon et al. 2013). Each item was loaded on the primary factor (i.e., target ability) and the secondary factor (i.e., secondary ability measured by the content of the testlet) to account for the item clustering effect. A DIF parameter was included in the bifactor model for DIF detection, and the Wald test or the likelihood ratio test was used for significance tests. Fukuhara and Kamata (2011) detected DIF under the bifactor model framework by including a covariate (i.e., the grouping variable) instead of a DIF parameter. The regression coefficient of the covariate was considered as the effect size estimate of DIF. These DIF methods have been demonstrated to be efficient in terms of both significance tests and recovery of DIF parameter estimates. These methods, however, only focused on the item clustering effect in DIF analysis.

Person clustering effect

Concurrently, DIF analyses accounting for the person clustering effect have also been investigated by researchers. Hierarchical logistic regression (HLR) is a natural choice for DIF detection in terms of accounting for the person clustering effect because of its feasibility of incorporating person dependency within clusters by a higher level regression analysis. Previous studies have examined the comparative performance between HLR and other standard DIF methods without accounting for the person clustering effect (e.g., logistic regression or Mantel–Haenszel test, French and Finch 2010, 2013). Results of these studies showed that HLR outperformed other DIF methods in terms of significance tests as the level of person dependency increased under certain conditions.

Jin et al. (2014) further found out that logistic regression (LR) performed equivalently as HLR when the covariate (i.e., total score) can explain most of the between cluster variance under the Rasch model, or when there was not much variance between discrimination parameters under the 2PL model. When type I error can be reasonably controlled under these conditions, applied researchers might prefer using the simple model (i.e., LR) for its ease of implementation and interpretation. A number of previous studies conducting DIF analysis on large-scale assessments ignored person clustering effect (e.g., Babiar 2011; Choi et al. 2015; Hauger and Sireci 2008; Innabi and Dodeen 2006; Mahoney 2008; Mesic 2012; Ockey 2007; Oliveri et al. 2014; Sandilands et al. 2013). Therefore, evaluating the trade-off between complex versus simple modeling of DIF may provide supporting evidence for the findings of these studies.

Jiao et al. (2012) developed a four-level multilevel testlet IRT model to account for the dual dependency. Their study showed that the four-level model was accurate in parameter recovery, but was less efficient due to the complexity of the model (i.e., large standard errors). Although their study is not intended for DIF detection, it provides evidence

that there is a trade-off between choosing the complex model for a slight improvement on parameter recovery but lower efficiency and the simple model for less accuracy but higher efficiency, which is similar to the concept of "the curse of dimensionality" in cluster analysis (James et al. 2013). In addition, analyzing complex models is not time-efficient. For example, when an achievement assessment contains 4 testlets, it requires five dimensions of integrations over the latent variables for the computation of the likelihood function, one dimension for the general factor and four dimensions for the secondary factors.

Although algorithms (e.g., bifactor dimension reduction, Cai et al. 2011; Gibbons and Hedeker 1992) have been proposed to reduce the number of integrations, some mainstream software do not have them implemented. In the study of Jeon et al., they compared the time spent on analyzing their proposed bifactor model using four different software, including Bayesian Networks with Logistic Regression Nodes (BNL) MATLAB toolbox (Rijmen 2006) with the dimension reduction algorithm implemented, PROC NLMIXED in SAS (Wolfinger 1999), gllamm (Rabe-Hesketh et al. 2005) in Stata, and WinBUGS (Spiegelhalter et al. 1996). The time spent ranged from 20 min (BNL) to more than a day (SAS) analyzing a simulated dataset with 12 items, 4 testlets, and 1000 examinees. Time-related issues can be of concern, especially for simulation studies, where a large number of replications needed to be analyzed to assess the performances of statistical methods.

In addition, current software, with the dimension reduction algorithm implemented to reduce the analysis time, cannot analyze multilevel models (e.g., TESTFACT, Bock et al. 2003; BIFACTOR, Gibbons and Hedeker 2007). It is difficult for researchers to be time-efficient, and to detect DIF via a model-based approach similar as the four level testlet model in Jiao et al. at the same time.

For applied researchers, it might be of particular interest to see the comparative performance between the complex and simple models for DIF detection using the mainstream software, which can model item and person clustering effects simultaneously. Therefore, the secondary goal of the current study is to evaluate the trade-off between simple models (e.g., models ignoring the dual dependency or accounting for partial dependency) and complex models (e.g., models accounting for dual dependency) for the accuracy of DIF detection. The evaluation of the trade-off can help researchers in selecting the appropriate DIF method in empirical settings when there is dual dependency in their data.

The four evaluated DIF methods

The current study focuses on detecting uniform DIF under the Rasch model, meaning that the difference between groups are constant across the entire domain of the latent variable and there is no discrimination difference between items. Due to the complexity of certain DIF methods included in this study, we chose the Rasch model to improve the efficiency of the simulation study because the Rasch model estimates fewer parameters than other models (e.g., 2-parameter IRT model). The four DIF methods included in the current study are LR ignoring the dual dependency, HLR accounting for the person clustering effect, the testlet model accounting for the item clustering effect, and the multilevel testlet model accounting for the dual dependency.

The LR model is

$$\eta_i = \beta_0 + \beta_1 G_i + \beta_2 X_i, \tag{1}$$

where $\eta_i = \ln\left(\frac{P(Y_i=1|X_i,G_i)}{P(Y_i=0|X_i,G_i)}\right)$, the logit of correct response for person i (i.e., $Y_i = 1$). G_i is the grouping variable. Significance test of the regression coefficient β_1 in Eq. (1) is used to determine the presence of uniform DIF, and the magnitude of β_1 is DIF size. X_i is the covariate (i.e., the total score) to match the latent trait between groups.

The HLR model is

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{10} G_{ij} + \beta_{20} X_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \end{aligned} \tag{2}$$

where $\eta_{ij} = \ln\left(\frac{P(Y_i=1|X_{ij},G_{ij},W_j)}{P(Y_i=0|X_{ij},G_{ij},W_j)}\right)$ for person i and cluster j, X_{ij} is the person level covariate (i.e., the total score), and the random components $u_{0j} \sim N(0, \tau^2)$. Significance tests of the regression coefficients β_{10} and γ_{01} are used to determine the presence of DIF, and the magnitude of β_{10} and γ_{01} are used as estimates of DIF size of the grouping variables G_{ij} and W_j at within-cluster (e.g., gender) and between-cluster level (e.g., country), respectively. The current study focuses on the grouping variable at the cluster level, which is consistent with the empirical example introduced later.

The testlet model is

$$\eta_{ik} = \theta_i - b_k + \gamma_{d(k)i} - \beta_k G_i \tag{3}$$

where $\eta_{ik} = \ln\left(\frac{P(Y_i=1|\theta_i,b_k,\gamma_{d(k)i},G_i)}{P(Y_i=0|\theta_i,b_k,\gamma_{d(k)i},G_i)}\right)$ for item k in testlet d for person i, θ_i is the latent trait for person i, b_k is the item difficulty parameter, $\gamma_{d(k)i}$ is the testlet effect, and β_k is the regression coefficient of the person level grouping variable used to determine the magnitude of DIF. The testlet model can be considered as the bifactor Multiple Indicators and Multiple Causes (MIMIC) model. The MIMIC model has been shown to be an effective DIF method detecting uniform DIF (Finch 2005; Woods 2009). In the MIMIC model, each item is regressed on the target latent trait and the grouping variable, and the target latent trait is regressed on the grouping variable to control for the mean difference of the target latent trait between groups. The presence of DIF is determined by the significance test of the regression coefficient of the grouping variable on each item. The bifactor MIMIC model adds a testlet factor, and each item is regressed on both target latent trait and the testlet factor.

The multilevel testlet model is

$$\begin{aligned} &\eta_{ijk} = \theta_{ij} - b_k + \gamma_{d(k)ij} \\ &\text{Level 1:} \\ &\theta_{ij} = \beta_{0j} + \beta_{10} G_{ij} + e_{ij} \\ &\gamma_{d(k)ij} = \pi_{0j} + \pi_{10} G_{ij} + \varsigma_{ij} \\ &\text{Level 2:} \\ &\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ &\pi_{0j} = \kappa_{00} + \kappa_{01} W_j + \zeta_{0j} \end{aligned} \tag{4}$$

where $\eta_i = \ln\left(\frac{P(Y_i=1|\theta_{ij},b_k,\gamma_{d(k)ij},G_{ij},W_j)}{P(Y_i=0|\theta_{ij},b_k,\gamma_{d(k)ij},G_{ij},W_j)}\right)$ for item k in testlet d for person i in cluster j, θ_{ij} is the latent trait for person i in cluster j, $\gamma_{d(k)ij}$ is the testlet effect in cluster j, e_{ij} and ς_{ij} are the level one residual variances of the target latent ability and the testlet factor, u_{0j}

and ζ_{0j} are the level two residual variances of the intercepts of the target latent ability and the testlet factor. Regression coefficients π_{10} and κ_{01} are the effects of the grouping variables on the testlet factor. Regression coefficients β_{10} and γ_{01} are used to determine the magnitude of DIF of the grouping variables G_{ij} and W_j at within-cluster and between-cluster level, respectively. The multilevel testlet model assumes that the person and item clustering effects are independent of each other. The multilevel testlet model can be extended to 2 parameter testlet model by including discrimination parameters for dichotomous items, and to multilevel testlet partial credit models by including step difficulty parameters for polytomous items (Jiao and Zhang 2015). The multilevel testlet model can also be considered as the multilevel bifactor MIMIC model where each item is regressed on the target latent trait, the testlet factor, and the grouping variables. Such a model can be analyzed using both IRT (e.g., IRTPRO) and structural equation modeling software (e.g., Mplus).

Methods

The current study manipulated seven factors to reflect various conditions in practical settings. The factors are impact (i.e., mean ability difference between groups: 2 levels), person clustering effect (3 levels), item clustering effect (3 levels), testlet contamination (2 levels), DIF contamination (2 levels), item difficulty (3 levels), and DIF methods (4 levels). Levels of each factors are fully crossed to create 864 conditions, and each condition is replicated 100 times.

Factors that were not manipulated are sample size, number of clusters, test length, and number of testlets. The sample size was 1500 for both reference and focal groups, with 30 people within each cluster. The selection of sample-size related conditions was consistent with large-scale assessment settings where sample size is at least in thousands. Some large-scale assessments employ rotated booklet design, meaning that each item is answered by a subset of the entire sample. Although the total sample size of large-scale assessments maybe large, the actual sample size for DIF analysis is less than the total sample size because DIF analysis is an item-by-item approach. The current study is particularly interested in small number of items within each testlet. The test length is set to 10 items with 5 items in each testlet, and is relatively consistent with the empirical example introduced later. The number of testlets is set to 2 for the purpose of computation efficiency.

Item responses in Eq. (4) were generated by manipulating different levels of impact, item, and person clustering effects in both θ_{ij} and $\gamma_{d(k)ij}$, and item difficulty parameters b_k. Latent ability of the reference group was generated from $N(0, 1)$ and latent ability of the focal group was generated from $N(0, 1)$ and $N(-1, 0)$ to make the two levels of the impact factor. One standard deviation in latent ability distribution between the reference and the focal groups is commonly observed in previous simulation studies as well as in empirical settings (e.g., Finch 2005; Oort 1998). For example, 2011 TIMSS 8th grade mathematics scores of participating countries have standard deviations from −1.7 to 1.1 from the scale center point. Asian countries with top scale scores, on average, have a 0.98 standard deviation away from the center point, and the United States' scale scores have a 0.1 standard deviation away from the center point (Mullis et al. 2012). Applied

researchers who are interested in the evaluation of Asian mathematics curriculum adoption might find the results of the current study beneficial to their research.

The person clustering effect in θ_{ij} had three levels $N(0, 0)$, $N(0, 0.25)$, and $N(0, 1)$; and the item clustering effect in $\gamma_{d(k)ij}$ had the same three levels: $N(0, 0)$, $N(0, 0.25)$, and $N(0, 1)$. The $N(0, 0)$ conditions were treated as baseline conditions where there is neither person nor item clustering effect, and the $N(0, 0.25)$ and $N(0, 1)$ conditions were considered as small-to-medium and medium-to-large person and item clustering effects, respectively (Jiao and Zhang 2015). The reference or focal group latent ability, person clustering effects in θ_{ij}, and item clustering effects in $\gamma_{d(k)ij}$ were additive and mutually exclusive. Item difficulty parameter b_k was within the range of $(-1, 1)$ and randomly assigned to each item. Item difficulty parameters were not generated outside the range of $(-1, 1)$ to avoid sparse cells, which might cause non-converged or extreme solutions, especially when the most complex model is fitted to the data (Bandalos 2006).

We considered two types of contamination factors in this study: testlet contamination and DIF contamination. Two levels of testlet contamination were manipulated by either generating item clustering effect in the second testlet, or not generating item clustering effect in the second testlet. Two levels of DIF contamination were manipulated by either using 3 additional DIF-present items (i.e., 30 % DIF contamination) throughout the test, or using no DIF-present items other than the studied items throughout the test. The studied items were generated to be DIF-free or DIF-present for the computation of type I error and power, respectively. Three studied items were included in the first testlet, representing items with low ($b_k = -1$), medium ($b_k = 0$), and high ($b_k = 1$) difficulty parameters. Purified total scores were used as the matching variable (i.e., sum of item scores other than the 3 studied items) to avoid the confounding effect due to DIF contamination conditions.

Selections of levels within the manipulated factors were based on two principles. First, we chose levels to closely link to the empirical data analyzed in the later section. For example, items from the first booklet in TIMSS 2011 Mathematics test were analyzed as a demonstration of the differential performance of the four DIF methods. The average number of items within each testlet was 5.25 (please see the detailed description in the empirical study section), so five items within each testlet were generated. Second, levels within some factors were adopted from previous simulation studies. For example, the levels within the item and person clustering effect factors were adopted from the four-level model in Jiao et al. simulation study.

The four DIF methods: LR, HLR, the testlet model, and the multilevel testlet model were analyzed using Mplus 7.2 (Muthén and Muthén 2014). Full-information maximum likelihood estimation method was used to estimate model parameters. LR estimated 9 parameters as in Eq. (1): $3\beta_1$ for the 3 studied item, $3\beta_2$ for the purified total score (e.g., sum of DIF-free items), and 3 threshold parameters (e.g., parameters estimated under the latent response variable formulation for categorical variables, Muthén and Asparouhov 2002) for the 3 studied items. HLR estimated 12 parameters as in Eq. (2): $3\beta_{20}$ for the purified total score at the within-cluster level, $3\gamma_{01}$ for the 3 studied items at the between-cluster level, 3 threshold parameters, and 3 residual variances for the 3 studied items.

The testlet model estimated 36 parameters: 9 factor loadings of the target ability, 4 factor loadings for the first testlet factor, 4 factor loadings for the second testlet factor, 1 regression coefficient of the grouping variable on the target ability, 2 regression coefficients of the grouping variable on the 2 testlet factors, 3 regression coefficients of the grouping variable on the 3 studied items, 10 threshold parameters for all items, 1 residual variance of the target ability, and 2 residual variances of the 2 testlet factors. The multilevel testlet model estimated 56 parameters: at the within-cluster level, 17 factor loadings of the target ability and 2 testlet factors, 1 variance of the target ability and 2 variances of 2 testlet factors; at the between-cluster level, 17 factor loadings of the target ability and 2 testlet factors, 1 regression coefficient of the grouping variable on the target ability, 2 regression coefficients of the grouping variable on the 2 testlet factors, 3 regression coefficients of the grouping variable on the 3 studied items, 10 threshold parameters for all items, 1 residual variance of the target ability, and 2 residual variances of the 2 testlet factors.

The performance of each DIF method was evaluated by type I error rate, power, bias, and mean square error (MSE). Type I error rate was computed as the percentage of falsely identified DIF-present items out of the 100 replications. Power was computed as the percentage of correctly identified DIF-present items out of the 100 replications. The medium DIF size of 0.5 (i.e., the difference of item difficulty parameters of the studied items between the reference and focal groups is 0.5) was used to compute power. Bias and MSE of the DIF parameter (i.e., regression coefficient of the grouping variable of the four DIF methods) were computed as in Eqs. (5) and (6):

$$Bias = E(\hat{coef}) - coeff \tag{5}$$

$$MSE = Bias^2 + Var(\hat{coeff}) \tag{6}$$

where $\hat{coef}$ is the estimated DIF parameter and *coeff* is the true DIF parameter. We performed two sets of analysis of variance (ANOVA) on Bias and MSE. Significance tests (*F*-test) at alpha level of 0.05 were used to determine main effects and higher-order interaction effects of the manipulated factors. Effect size estimates were used to determine the magnitude of the effects of the manipulated factors on the comparative performance of the four DIF methods. Effect sizes were reported using $f \cong \sqrt{\eta^2/(1-\eta^2)}$ as in Cohen (1969). Cutoffs of small, medium, and large effect sizes are 0.10, 0.25, and 0.40, respectively.

Results

Type I error rate

Figures 1, 2, 3 and 4 present type I error rate of the four DIF methods across different levels of person and item clustering effects, impact, testlet and DIF contamination when the studied item's difficulty is low. Similar patterns are observed when the studied item's difficulty is medium or high. Their figures are not presented here, but are available upon request. Figures 1 and 2 show that under the condition of no impact and no DIF contamination, all four DIF methods perform equivalently in terms of controlling type I error

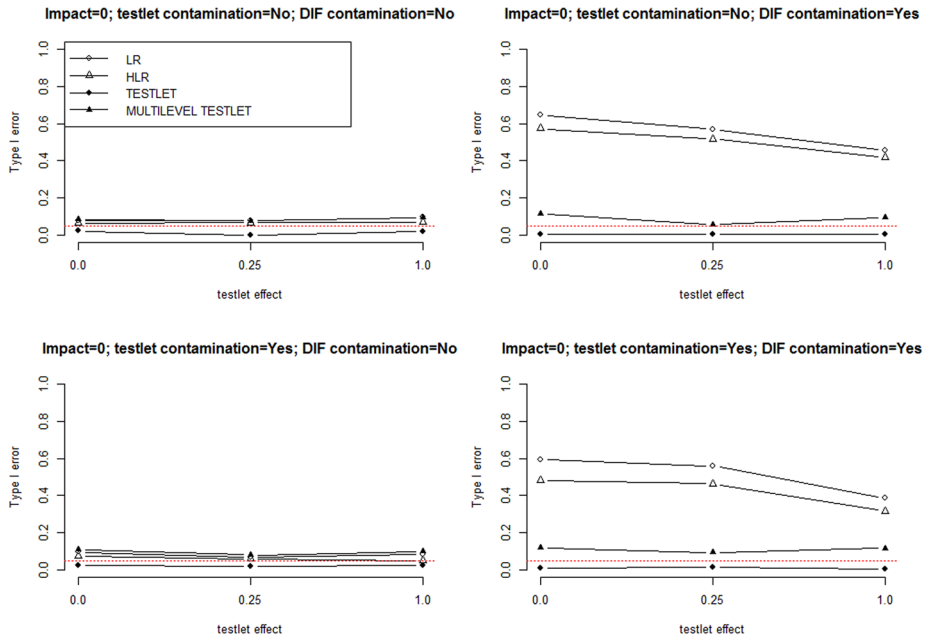

Fig. 1 Effects of testlet contamination and DIF contamination at each level of item clustering effect when there is no impact between groups. The *dotted line* is the theoretical type I error rate of 0.05

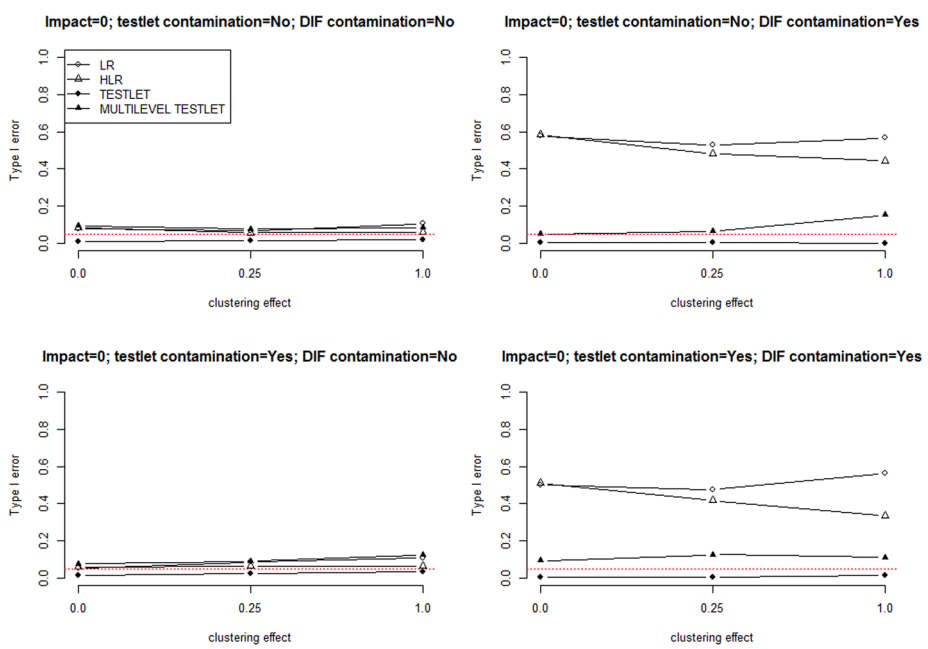

Fig. 2 Effects of testlet contamination and DIF contamination at each level of person clustering effect when there is no impact between groups. The *dotted line* is the theoretical type I error rate of 0.05

rate at the nominal level, regardless of the levels of item and person clustering effects and testlet contamination. The testlet model and the multilevel testlet model, however, outperform LR and HLR when there is DIF contamination.

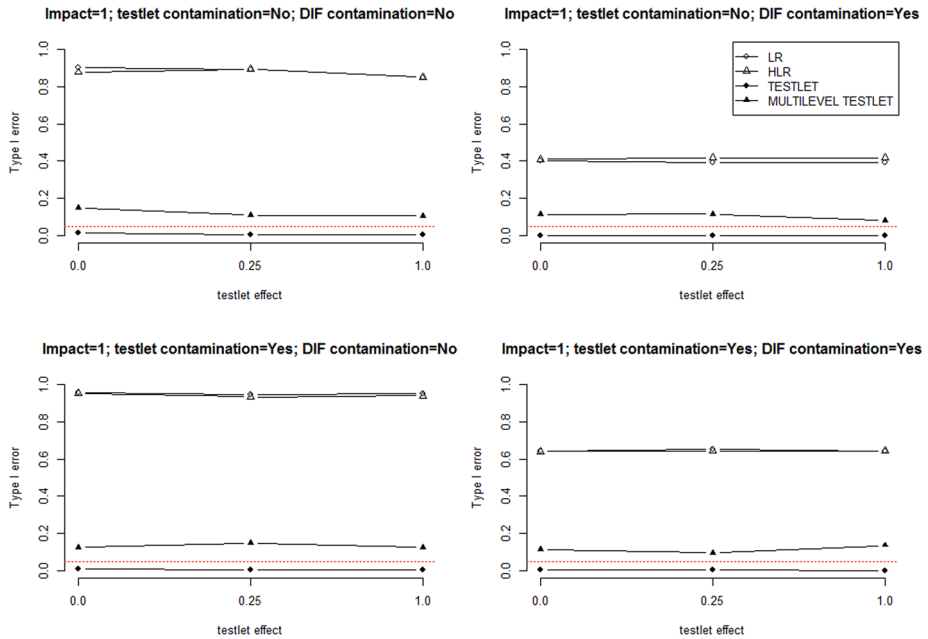

Fig. 3 Effects of testlet contamination and DIF contamination at each level of item clustering effect when there is impact between groups. The *dotted line* is the theoretical type I error rate of 0.05

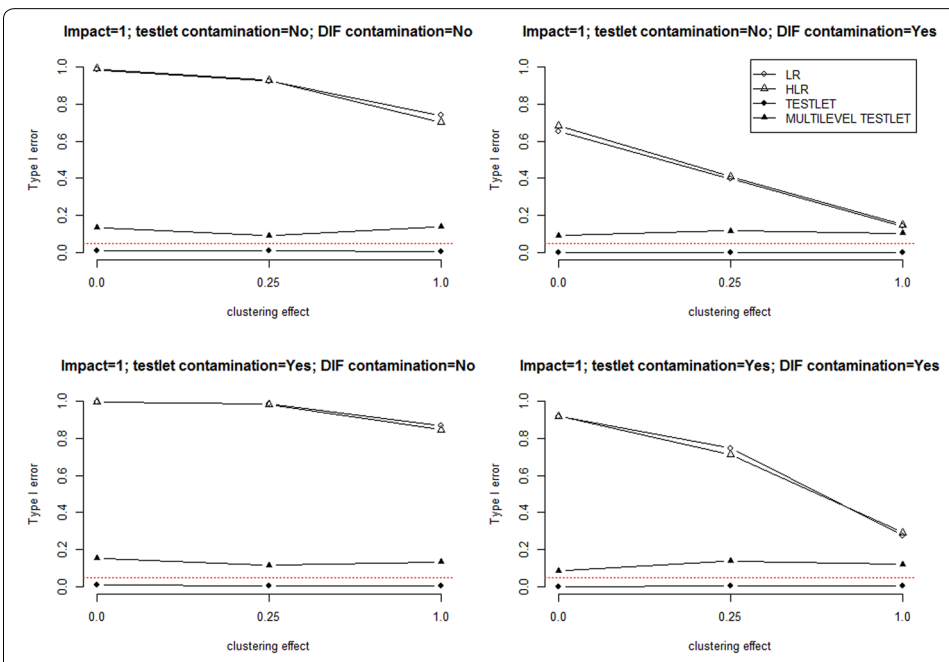

Fig. 4 Effects of testlet contamination and DIF contamination at each level of person clustering effect when there is impact between groups. The *dotted line* is the theoretical type I error rate of 0.05

Figures 3 and 4 show that under the condition of the presence of impact, the testlet model and the multilevel testlet model outperform LR and HLR regardless levels of item and person clustering effects, testlet contamination, and DIF contamination. Based

on the comparison of Figs. 3 and 4, the item clustering effect seems to have negligible effect on the performance of the four DIF methods (i.e., flat lines across levels of testlet effects), whereas the person clustering effect has no impact on the testlet model and the multilevel testlet model, but has an effect on LR and HLR. In summary, the most important factors affecting type I error rate of the four DIF methods are impact and DIF contamination. When there is no impact and no DIF contamination, the four DIF methods perform equally well, when there is impact and DIF contamination, the testlet model, and the multilevel testlet model outperform LR and HLR. In terms of the comparison of the four DIF methods across all levels of other factors, the testlet model is a little conservative (i.e., type I error is slightly below 0.05), and the multilevel testlet model is a little liberal (i.e., type I error is slightly above 0.05). HLR outperform LR under most of the conditions, but the advantage is small, the average difference of type I error rate between HLR and LR is 0.02 across all conditions.

Power

Power of HLR and LR is exceptionally high due to the excessive inflation of type I error rate of LR and HLR under most conditions. Power of HLR and LR, therefore, is not compared to the power of the testlet model and the multilevel testlet model. The testlet model and the multilevel testlet model perform equivalently across all conditions in terms of DIF detection rate. The average difference of power between the two models is 0.07. For both models, their equivalent performance are consistent regardless of person and item clustering effects, and testlet contamination. As compared to the power when there is DIF contamination, power of both models is consistently higher when there is no DIF contamination. The average power of the testlet model and the multilevel testlet model is 0.61 and 0.43 when there is no DIF contamination. When there is DIF contamination, the average power of the testlet model and the multilevel testlet model is 0.04 and 0.08, respectively, which are extremely low. Impact also has an effect on power. The average power of the testlet model and the multilevel testlet model is 0.35 and 0.30 when there is no impact. When there is impact, the average power of the testlet model and the multilevel testlet model is 0.30 and 0.21, respectively. The lower power under impact conditions is confounded by the DIF contamination conditions. In general, the effect of DIF contamination is larger than the effect of impact on power for both models: the average difference of power is 0.46 between the DIF contaminations conditions, whereas the average difference of power is 0.07 between the impact conditions. At last, similar patterns are observed among levels of item difficulty under the previously discussed conditions.

Bias

Most of the main effects and two-way interaction effects of the manipulated factors are statistically significant. Impact, DIF contamination, and item difficulty have small effect sizes ($f = 0.10, f = 0.14$, and $f = 0.16$, respectively), DIF method has medium effect size ($f = 0.38$). The two-way interaction of impact and DIF method has small effect size ($f = 0.16$), and the two-way interaction of item difficulty and DIF method has medium effect size ($f = 0.28$). Effect sizes of the rest of the factors, including higher order interaction, are negligible (i.e., $f < 0.10$). When there is no impact between groups, LR and HLR

outperform the testlet and multilevel testlet models, and their average bias are −0.06, −0.06, 0.21, and 0.38, respectively. When there is impact between groups, the testlet model outperform both LR and HLR, and their average bias are 0.36, −0.42, and −0.44, respectively. The multilevel testlet model perform the best when the item difficulty is low and medium with small average bias (−0.02 and −0.04), and perform the worst when the item difficulty is high with large average bias (0.92), which can be explained by the sparse cells due to high difficulty, resulting in extreme solutions. In summary, LR and HLR outperform the testlet and the multilevel testlet models when there is no impact between groups. When there is impact between groups, the multilevel testlet model is most accurate in estimating the DIF parameter with small and medium item difficulty parameters of the studied item. Generally speaking, LR and HLR underestimate the DIF parameter, whereas the testlet and the multilevel testlet models overestimate the DIF parameters under most conditions. The person and item clustering effects have negligible effects on bias of the four DIF methods, and can be explained by that point estimates are relatively robust against the violation of the independence assumption (Raudenbush and Bryk 2002).

MSE

Among the statistically significant main and interaction effects, the only factor having meaningful effect size is the DIF method with a medium effect size ($f = 0.33$), the effects of the other factors and their interactions are negligible (i.e., $f < 0.10$). Figures 5, 6, 7 and 8 present *MSE* of the four DIF methods across different levels of person and item clustering effects, impact, testlet and DIF contamination when the studied item's difficulty is low. Similar patterns are observed when the studied item's difficulty is medium or high.

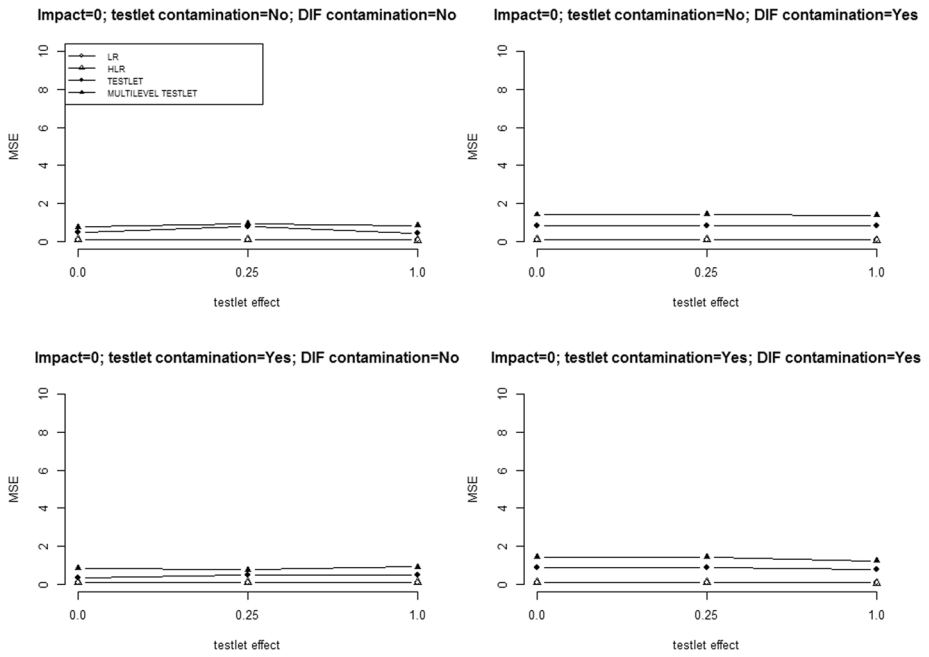

Fig. 5 Effects of testlet contamination and DIF contamination at each level of item clustering effect when there is no impact between groups

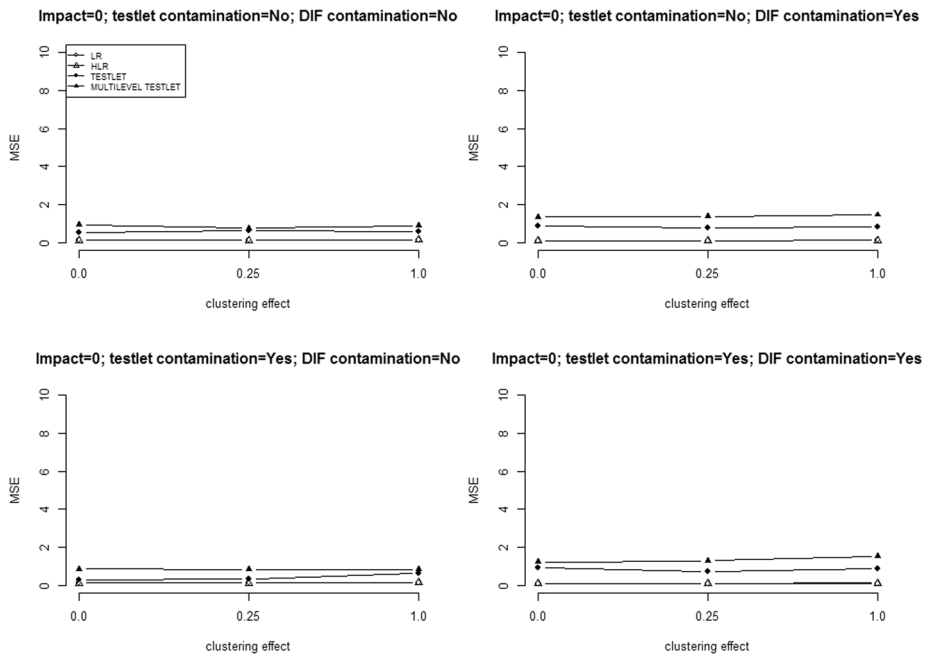

Fig. 6 Effects of testlet contamination and DIF contamination at each level of person clustering effect when there is no impact between groups

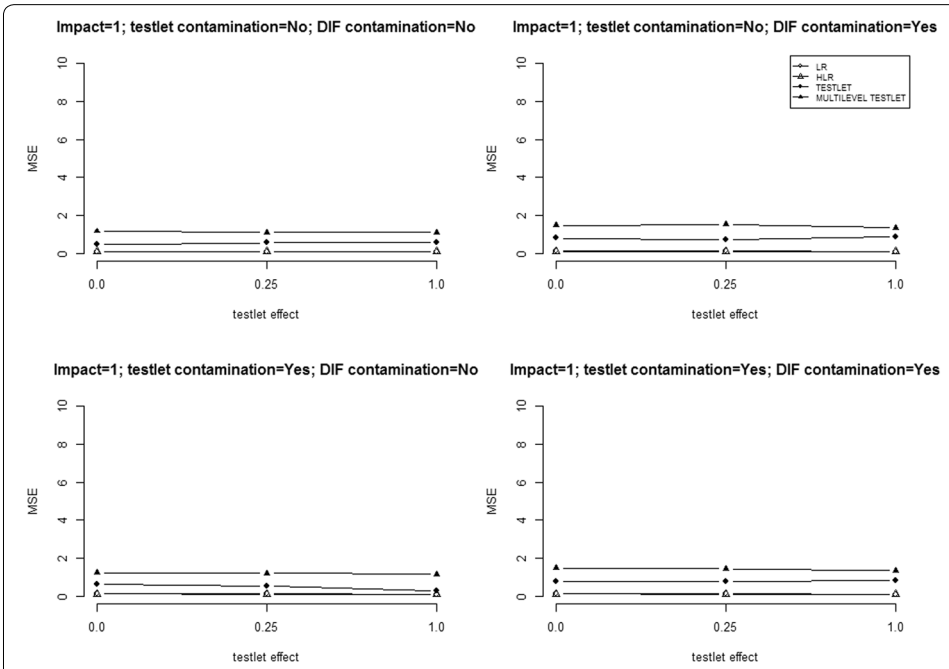

Fig. 7 Effects of testlet contamination and DIF contamination at each level of item clustering effect when there is impact between groups

As shown in the figures, the complex models show more variation in DIF parameter estimation than the simple models under all condition, and their *MSE* are in order as follows: the multilevel testlet model (1.271), the testlet model (0.675), HLR (0.112), and

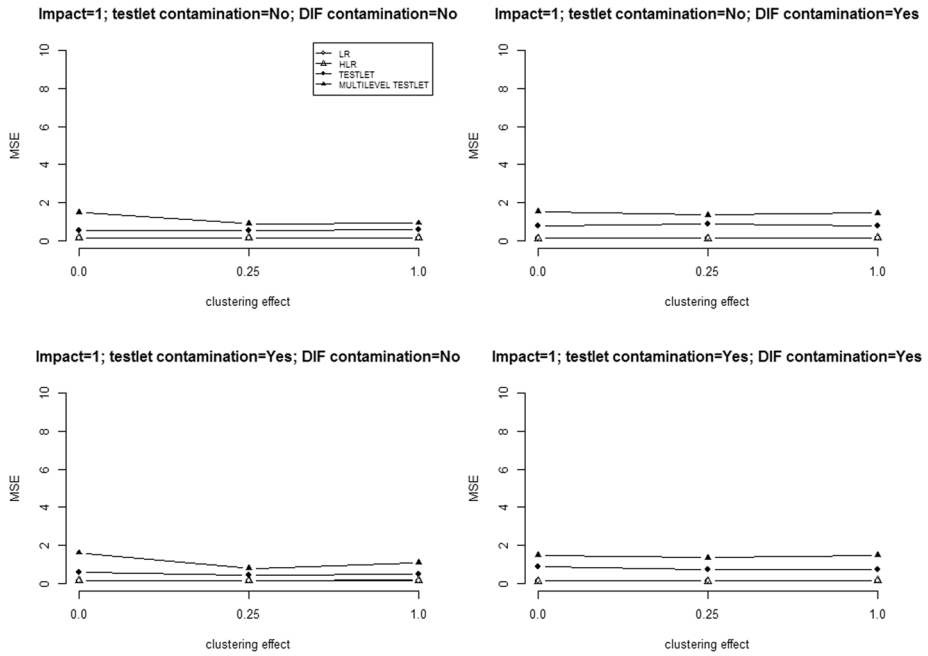

Fig. 8 Effects of testlet contamination and DIF contamination at each level of person clustering effect when there is impact between groups

LR (0.110). This can be explained by the larger number of parameters estimated in the complex models (e.g., 56 parameters are estimated from the multilevel testlet model and 9 parameter are estimated from the LR model).

An empirical example

The 2011 TIMSS Mathematics data were analyzed to show the differential performances of the four DIF methods. For demonstration purposes, only items in the first out of the 14 booklets were evaluated for DIF. There are 21 dichotomous items in the first booklet, four subdomains are measured by those items: number (8 items), data and chance (4 items), algebra (5 items), geometry (4 items). Data from two pairs of countries were selected to evaluate country level DIF. United States (average math achievement score of 509 with *SD* of 2.6) and Finland (average math achievement score of 514 with *SD* of 2.5) were selected to reflect the no impact condition in the simulation study. There were 1032 students from 613 schools from both countries taking items from the first booklet. Korea (average math achievement score of 613 with *SD* of 2.9) and New Zealand (average math achievement score of 488 with *SD* of 5.5) were selected to reflect the condition when there was impact between countries. There were 755 students from 302 schools from both countries in this dataset. DIF analyses were done for each pair of countries using the four DIF methods.

Table 1 presents the DIF analysis for the United States and Finland data. LR and HLR are consistent in terms of both significance tests and DIF parameter estimates. Both methods flag the same 14 items as DIF-present items. DIF-present items are flagged by significance tests described in Eqs. (1), (2), (3), and (4) at alpha level of 0.05. The estimated DIF parameters are from −0.002 to 0.004, and are almost identical for both

Table 1 DIF parameter estimates for USA and Finland data

	LR	HLR	Testlet	Multilevel testlet
Number				
Item 1	*0.001*	*0.001*	0.000	*0.001*
Item 2	*0.003*	*0.004*	*0.002*	*0.004*
Item 3	*−0.002*	*−0.002*	−0.001	*−0.002*
Item 4	*0.001*	*0.001*	0.000	*0.002*
Item 5	*0.001*	*0.001*	−0.002	*0.002*
Item 6	*−0.002*	*−0.002*	−0.001	*−0.001*
Item 7	*0.001*	*0.001*	0.001	*0.002*
Item 8	0.000	0.000	0.000	0.000
Geometry				
Item 1	*−0.001*	*−0.001*	*0.000*	−0.001
Item 2	*−0.001*	*−0.001*	−0.001	−0.001
Item 3	0.000	0.000	0.000	0.000
Item 4	−0.001	−0.001	0.000	0.000
Data and chance				
Item 1	*−0.001*	*−0.001*	0.000	−0.001
Item 2	0.000	0.000	0.000	0.000
Item 3	*−0.001*	*−0.001*	0.000	−0.001
Item 4	*−0.001*	*−0.001*	0.000	−0.001
Algebra				
Item 1	0.000	0.000	0.000	0.000
Item 2	0.001	0.001	0.000	0.001
Item 3	*0.001*	*0.001*	0.001	*0.001*
Item 4	0.000	0.000	0.000	0.000
Item 5	*0.002*	*0.002*	0.001	*0.002*

Italised DIF parameter estimates are significant at 5 %

methods. Because the DIF parameter estimates (i.e., DIF effect size) are so small, indicating that there is negligible DIF between groups, the flagged 14 out of 21 items by both LR and HLR can be considered false positives, which provides evidence of inflated type I error rate of LR and HLR, which is consistent with the results of the simulation study. The testlet model flags 2 out of 21 items as DIF-present items. Given the negligible estimated DIF parameters, the testlet model outperforms both LR and HLR in terms of controlling type I error rate. The multilevel testlet model flags 9 items as DIF-present items. With negligible DIF parameter estimates of these 9 items, the multilevel testlet model cannot control type I error rate as well as the testlet model, which is consistent with Figs. 1 and 2, where the multilevel testlet model almost always exhibits slight type I error inflation as compared to the testlet model. In summary, LR and HLR flag the most number of items as DIF-present items, and the testlet model flags the least number of DIF-present items by significance tests. When evaluating DIF based on DIF parameter estimates, all four methods perform equivalently. All 21 items exhibit negligible DIF, indicating low power of the methods. Simpler models that do not account for item or person clustering effects perform equivalently as the complex models.

Table 2 presents the DIF analysis for the Korea and New Zealand data. Similar to the United States and Finland DIF analysis, LR and HLR flag the same number of DIF-present items (16 out of 21 items), and their DIF parameter estimates are almost identical.

Table 2 DIF parameter estimates for Korea and New Zealand data

	LR	HLR	Testlet	Multilevel testlet
Number				
Item 1	*0.014*	*0.014*	0.005	*0.012*
Item 2	*−0.006*	*−0.007*	−0.002	*−0.007*
Item 3	0.000	−0.001	−0.002	*−0.004*
Item 4	*−0.006*	*−0.006*	0.000	*−0.010*
Item 5	*0.006*	*0.007*	0.002	*0.008*
Item 6	*0.005*	*0.005*	0.002	0.003
Item 7	*−0.012*	*−0.012*	−0.005	*−0.014*
Item 8	*−0.007*	*−0.007*	−0.004	−0.003
Geometry				
Item 1	*−0.005*	*−0.006*	−0.011	*−0.013*
Item 2	*−0.010*	*−0.011*	−0.007	*−0.019*
Item 3	0.002	0.002	0.002	*−0.005*
Item 4	*0.007*	*0.008*	0.001	−0.003
Data and chance				
Item 1	0.000	0.000	−0.001	0.000
Item 2	*0.017*	*0.017*	0.005	*0.009*
Item 3	*−0.011*	*−0.012*	−0.007	*−0.017*
Item 4	*0.003*	*0.040*	0.000	0.002
Algebra				
Item 1	*0.006*	*0.007*	0.002	0.000
Item 2	*0.004*	*0.004*	0.001	*−0.004*
Item 3	0.003	0.003	0.002	−0.004
Item 4	0.000	0.000	−0.003	−0.003
Item 5	*−0.004*	*−0.005*	0.000	*−0.009*

Italised DIF parameter estimates are significant at 5 %

The DIF parameter estimates are from −0.012 to 0.017. These considerably small DIF parameter estimates provide evidence of inflated type I error of both LR and HLR. The testlet model, on the other hand, flags no items as DIF-present items, which is consistent with the simulation study that the testlet model is conservative under the conditions when there is impact between countries. The DIF parameter estimates are from −0.011 to 0.005. The multilevel testlet model flags 13 out of 21 items as DIF-present items. The DIF parameter estimates are from −0.019 to 0.012. Given the small size of DIF parameter estimates and large number of flagged DIF items, the multilevel testlet model does not control type I error rate as well as the testlet model. This might be explained by that the multilevel testlet model is the most complex model, and estimation problems can occur, such as unconverged or extreme solutions. In order to achieve converged and meaningful solutions in this study, the estimation of country difference on math achievement was left out, which was contradictory to the data, where Korea and New Zealand differed on their math achievement. Failure to account for impact between groups can lead to inflated type I error due to the confounding effect of impact (Finch 2005).

In addition, the DIF parameter estimates of LR, HLR, and the testlet model are more consistent than the DIF parameter estimates of the multilevel testlet model, meaning that the difference of DIF parameter estimates between the multilevel testlet model and the other three methods are larger than the differences between LR, HLR, and the testlet

model. The reason might be explained by that TIMSS uses matrix sampling, where students from the same school and the same country receive different booklets of the test, so the number of students within each school is limited and the person clustering effect is not necessarily large. For the Korea and New Zealand data, the intraclass correlation (ICC) of the test scores is 0.47, which is considered to be a relatively large person clustering effect. DIF analysis, however, is an item-by-item approach, ICC of the individual item score is not as large as ICC of the test score. ICC of some items can be as small as 0.008, and the average ICC of items in the first booklet is 0.176. With some items having extremely small ICC (e.g., 0.008), the multilevel testlet model is likely to overfit the data, leading to inconsistent parameter estimates.

Discussion

The primary goal of the current study was to examine the comparative performance of the four DIF methods when the data exhibited dual dependency due to item and person clustering effects. The multilevel testlet model accounting for dual dependency exhibited slight inflation of type I error across all simulated conditions, but had low power, especially when there was DIF contamination and impact between groups. When there was impact between groups, the multilevel testlet model was most accurate in terms of estimating DIF parameter among the four methods when the studied item's difficulty parameter was low or medium, but was the least efficient method due to its complexity. The testlet model accounting for the item clustering effect was slightly conservative in terms of type I error rate across all conditions, and the power was even lower than the multilevel testlet model under certain conditions. The testlet model cannot estimate DIF parameter as accurately as the multilevel testlet model under most conditions, but was slightly more efficient than the multilevel testlet model. HLR accounting for the person clustering effect can control type I error rate as well as the other methods when there was no impact between groups, but exhibited serious type I error inflation when there was impact between groups. HLR can accurately estimate DIF parameter when there was no impact between groups. LR ignoring both item and person clustering effect exhibited slightly higher type I error inflation and bias than HLR, and LR was the most efficient method due to its simplicity. In general, LR performed relatively equivalent as HLR under most conditions. Based on the results of this study, applied researchers should use caution and prior knowledge about their data when choosing an appropriate DIF method. For example, the multilevel testlet model and the testlet model are not appropriate DIF methods due to the low power of these two methods, especially when the researchers have a strong indication that DIF might exist (e.g., cultural DIF when the grouping variable is country).

The secondary goal of the current study was to evaluate the trade-off between simple models and complex models for the accuracy of DIF detection. Based on the results of the simulation study and the empirical example analysis, simple models suffered from type I error inflation due to the failure in accounting for the dependency, and complex models suffered from extremely low power under certain conditions, possibly due to the overfitting of the data. Given the inadequate significance test, researchers can always look into DIF parameter estimates for more information. When there was no impact between groups, simple models estimated DIF parameters more accurately than

complex models. When there was impact between groups, complex models had lower bias under certain conditions, but were not as efficient as simple models in estimating DIF parameters. When considering both significance tests and DIF parameter estimates, simple models may be preferred when the researchers have prior knowledge that the groups do not differ on their mean abilities (i.e., no impact between groups). When no prior knowledge is available on impact, complex models might be preferred, but with cautions such that DIF-present items may not be successfully detected due to low power, or DIF parameter estimates can be extreme values to achieve model identification.

The current study intended to simulate conditions that were commonly observed in practical settings. These conditions, however, were limited for its results to be generalized. Most limitations were related to computation efficiency. First, the current study generated 2 testlets, which was considered a small number of testlets as compared to other studies examining testlet effect (e.g., Jiao et al. 2012). Part of the reason was to be time-efficient. With 2 testlets, the dimensions of numerical integration for a multilevel testlet model was 6, 3 dimensions at the within-cluster level and 3 dimensions at the between-cluster level. The average time spent on analyzing each condition was about 10 h, and the current study generated 216 conditions for the multilevel testlet model. Increasing the number of testlets with this large number of conditions would be a great challenge. Second, each condition was replicated 100 times, which was relatively small in simulation studies. With small number of replications, the construction of the sampling distribution of parameter estimates might not be sufficient, leading to biased point estimates. Increasing the number of replications also increased the analysis time significantly for complex models. Future studies can take these two factors into account once software are updated with features such as dimension reduction algorithm to save analysis time greatly.

Future studies should also be conducted to address limitations of the current simulation study because some conditions were not closely linked to empirical conditions in large-scale assessments (e.g., TIMSS). For instance, item difficulty parameters were generated within the range of $(-1, 1)$ to reduce the number of non-converged or extreme solutions. Future studies should include easier or more difficult test items like large-scale assessments do to evaluate the comparative performance of the four DIF methods under a broader range of difficulty levels.

The current study assumed orthogonality of the primary factor and the secondary factors (i.e., testlets). Previous studies indicated that the bifactor model for DIF detection for testlet-based tests was more flexible and accurate when the orthogonality assumption was relaxed (Jeon et al. 2013). The testlet model in the current study was essentially a bifactor model with a DIF parameter included as a covariate. When allowing correlations between the primary factor and secondary factors, the complex models may be more appropriate than simple models because they can incorporate the correlations explicitly. Future studies may investigate this advantage of the testlet and the multilevel testlet model more extensively by relaxing the orthogonality assumption when data exhibit dual dependency.

Based on the results of the empirical example, the four DIF methods performed differently in terms of both significance tests and parameter estimates. Applied researchers should be aware of the difference when they conduct DIF analysis on data with dual

dependency. Selecting inappropriate DIF methods by ignoring dual dependency will lead to costly decisions, such as removing item with negligible DIF from the test, or declaring group difference confounded by falsely identified DIF-present items. Researchers should also be aware of the challenge when applying the complex model accounting for the dual dependency, such as constraining parameters to achieve model identification due to the complexity of the model. To facilitate the selection of the most appropriate DIF method, it would be better for researchers to be familiar with their data. With prior knowledge of group difference (i.e., impact), item difficulty, and the magnitude of dependency (i.e., item or person clustering effect), researchers will be more likely to select the most appropriate DIF method based on the results of the current study.

The current study focused on identifying DIF-present items rather than exploring source of DIF when data were dually dependent. Admittedly, it is very important to understand why DIF happens, given that DIF-present items are accurately detected. Previous studies utilized both significance tests and effect size estimates to explore source of DIF. For example, Wu and Ercikan (2006) studied DIF due to cultural difference among countries (e.g., extra lesson hours after school) by the decreased magnitude of DIF effect size and decreased number of DIF-present items after including the cultural factor as an extra matching variable. To insure meaningful exploration of DIF, items suspicious of DIF need to be identified correctly first, especially when data have complex structure (e.g., dual dependency).

Authors' contributions

YJ carried out the study and drafted the manuscript. Both authors read and approved the final manuscript.

Author details

[1] Department of Psychology, Middle Tennessee State University, Jones Hall, 308, Murfreesboro, TN 37130, USA. [2] Department of Health and Human Performance, Middle Tennessee State University, Murphy Center #128, Murfreesboro, TN 37132, USA.

Competing interests

The authors declare that they have no competing interests.

Received: 4 December 2015 Accepted: 19 September 2016
Published online: 29 September 2016

References

Babiar, T. C. (2011). Exploring differential item functioning (DIF) with the Rasch model: A comparison of gender differences on eighth grade science items in the United States and Spain. *Journal of Applied Measurement, 12*(2), 144–164.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In R. C. Serlin, G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385–462). Greenwich: Information Age.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4 user's guide*. Chicago: Scientific Software International.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221–248.

Choi, Y.-J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS, 2007 mathematics test. *International Journal of Testing, 15*, 239–253.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego: Academic Press.

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*(4), 278–295.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299–317.

French, B. F., & Finch, W. H. (2013). Extensions of Mantel-Haenszel for multilevel DIF Detection. *Educational and Psychological Measurement, 73*(4), 648–671.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*(8), 604–622.

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

Gibbons, R. D., & Hedeker, D. (2007). *Bifactor [Computer software]*. Chicago: Center for Health Statistics, University of Illinois at Chicago.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*, 237–250.
Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. *School Science and Mathematics, 106*(8), 328–337.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38*(1), 32–60.
Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82–100.
Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68*, 65–83.
Jin, Y., Myers, N. D., & Ahn, S. (2014). Complex versus simple modeling for DIF detection when the intraclass correlation coefficient (ρ) of the studied item is less than the ρ of the total score. *Educational and Psychological Measurement, 74*(1), 163–190.
Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education*, *16*, 385–402.
Lee, Y.-S., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review, 10*, 365–375.
Mahoney, K. (2008). Linguistic influence on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing, 8*, 14–33.
Mesic, V. (2012). Identifying country-specific cultures of physics education: A differential item functioning approach. *International Journal of Science Education, 34*(16), 2483–2500.
Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.
Muthén, B. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Mplus Web Notes: No. 4, Version 5. 2002.
Muthén, L. K., & Muthén, B. O. (2014). Mplus: Statistical analysis with latent variables (version 7.2) [Computer software]. Los Angeles, CA: Author.
Ockey, G. J. (2007). Investigating the validity of math word problems for English language learns with DIF. *Language Assessment Quarterly, 4*(2), 149–164.
Oliveri, M. E., Ercikan, K., Zumbo, B. D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—comparing a latent class to a manifest DIF approach. *International Journal of Testing, 14*, 265–287.
Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107–124.
Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023–1046.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*, 301–323.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.
Rijmen, F. (2006). *BNL: A Matlab toolbox for Bayesian networks with logistic regression (Technical Report)*. Amsterdam: Vrije Universiteit Medical Center.
Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13*, 152–174.
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.
Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). BUGS 0.5 Bayesian Analysis using Gibbs Sampling. Manual (version II). Cambridge, MRC-Biostatistics Unit.
Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197–219.
Wang, W.-C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement, 65*(4), 549–576.
Wolfinger, R. D. (1999). *Fitting non-linear mixed models with the new NLMIXED procedure (Technical Report)*. Cary: SAS Institute.
Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1–27.
Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the mantel and GMH tests, and IRT-LR-DIF when the latent distribution is non normal for both groups. *Applied Psychological Measurement, 35*(2), 145–164.
Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing*, *6*(3), 287–300.