Macrothink Institute™

Journal of Educational Issues
ISSN 2377-2263
2016, Vol. 2, No. 2

DIF Analysis with Multilevel Data: A Simulation Study Using the Latent Variable Approach

Ying Jin (Corresponding author)

Department of Psychology, Middle Tennessee State University

Jones Hall 308, Murfreesboro, TN, 37130, USA

E-mail: ying.jin@mtsu.edu

Hershel Eason

Department of Psychology, Middle Tennessee State University

Jones Hall 308, Murfreesboro, TN, 37130, USA

E-mail: hje2b@mtmail.mtsu.edu

Received: September 20, 2016 Accepted: November 21, 2016
Published: December 2, 2016

doi:10.5296/jei.v2i2.10045 URL: http://dx.doi.org/10.5296/jei.v2i2.10045

Abstract

The effects of mean ability difference (MAD) and short tests on the performance of various DIF methods have been studied extensively in previous simulation studies. Their effects, however, have not been studied under multilevel data structure. MAD was frequently observed in large-scale cross-country comparison studies where the primary sampling units were more likely to be clusters (*e.g.*, schools). With short tests, regular DIF methods under MAD-present conditions might suffer from inflated type I error rate due to low reliability of test scores, which would adversely impact the matching ability of the covariate (*i.e.*, the total score) in DIF analysis. The current study compared the performance of three DIF methods: logistic regression (LR), hierarchical logistic regression (HLR) taking multilevel structure into account, and hierarchical logistic regression with latent covariate (HLR-LC) taking multilevel structure into account as well as accounting for low reliability and MAD. The results indicated that HLR-LC outperformed both LR and HLR under most simulated conditions, especially under the MAD-present conditions when tests were short. Practical implications of the implementation of HLR-LC were also discussed.

Keywords: DIF, Multilevel, Reliability

1. Introduction

Measurement invariance (MI) is one of the most important aspects of item validity (Millsap, 2011). Examination of MI detects unfavorable effects of secondary factors due to group membership on the performance of the studied item measuring the target ability. A variety of differential item functioning (DIF) methods have been developed to test for MI. Recently, the performance of several DIF methods have been examined and compared under the multilevel data structure, where the local independence assumption was violated (French & Finch, 2010, 2013; Jin, Myers, & Ahn, 2014). Results of these studies have shown that standard DIF methods that did not account for multilevel data structure may or may not perform equivalently as modified DIF methods that accounted for multilevel data structure under specific conditions. These simulation studies, however, left out some important factors that are observed rather frequently in practice. The current study included these factors to provide a more complete literature on the performance of DIF methods under the multilevel data structure.

It is well known that applying standard statistical tests to multilevel data inflates type I error rate due to the violation of the independence assumption (Raudenbush & Bryk, 2002). In DIF analysis, the standard statistical test logistic regression (LR) and Mantel-Haenszel test (MH) have been shown to be not as effective as hierarchical logistic regression (HLR) in controlling type I error rate because LR and MH do not model random effects due to clusters (French & Finch, 2010, 2013). Jin et al. (2014) extended French and Finch's study by showing that HLR outperformed LR when the intraclass correlation (ρ) was medium to large (*e.g.*, $\rho > 0.25$), but LR performed equally well as HLR when ρ was small to medium (*e.g.*, $\rho < 0.25$). Their findings were based on the sufficiency of the covariate (*i.e.*, the total score), which, by convention, is called the matching variable in DIF analysis. As Jin et al. demonstrated, when the covariate was a sufficient estimate of the latent ability (θ), it can be an accurate estimate of θ, as well as being able to maintain the multilevel structure of θ (*i.e.*, ρ of the covariate is close to ρ of θ). Therefore, most between-cluster variance can be explained by the covariate. LR, under such a circumstance, can perform equivalently as HLR because little between-cluster variance was left for multilevel modeling in HLR.

The sufficiency of the covariate can be compromised under certain conditions, where the covariate cannot maintain the multilevel structure of θ, meaning that ρ of θ cannot be closely estimated by ρ of the covariate. One of these conditions is low reliability of the covariate. Under multilevel data structure, ρ can be interpreted as the correlation between individuals within the same cluster (Bock, 1989). The estimated correlation, however, can be attenuated by low reliability of scores (Crocker & Algina, 1986; Murphy & Davidshofer, 1988; Zimmerman & Williams, 1977), leading to underestimated ρ. One of the factors causing low reliability is test length. Previous studies have shown that reliability of a short test was generally lower than that of a long test (*e.g.*, Cronbach, 1951; Tavakol & Dennick, 2011). In DIF analysis, test length was a commonly manipulated factor in simulation studies, and results of these studies have shown that inflated type I error rate was more frequently observed for short tests (*e.g.*, Fidalgo, Mellenbergh, & Muniz, 2000). With short tests, especially for standard DIF methods, results of multilevel DIF analysis might be inaccurate

due to low reliability of the covariate as it cannot maintain the multilevel structure of θ. One of the current study's goals, therefore, is to investigate the effect of short tests, which has not been examined in previous multilevel DIF analysis studies, on the performance of both standard DIF method (*i.e.*, LR) and DIF method accounting for multilevel data structure (*i.e.*, HLR).

Another consequence of compromised sufficiency of the covariate is that the matching ability of the covariate can be adversely affected, resulting in inaccurate DIF analysis. For some DIF methods (*e.g.*, LR), the function of including the total score as the covariate in DIF analysis is to match θ between focal and reference groups (*i.e.*, population of interest and mainstream population) to make sure that the true difference in θ is not contaminated by DIF (Osterlind & Everson, 2009). In order to match θ accurately, the covariate needs to be an accurate estimate of θ reflecting the true difference between groups, as well as being able to maintain the multilevel structure of θ for each group. Lower reliability (*e.g.*, due to relatively shorter test) could attenuate the estimated effect of the true mean difference between groups (*e.g.*, non-significant regression coefficient of the covariate in LR, leading to failure in controlling mean difference between groups). A factor normally referred to as mean ability difference (MAD) was often included in previous simulation studies to investigate the matching ability of the covariate. Previous studies have shown that inflated type I error rate was observed more frequently when there was MAD between groups (*e.g.*, Finch, 2005).

The effect of MAD has not been investigated in multilevel DIF analysis. Situations where MAD is present are not uncommon in large-scale international achievement assessments. For example, the primary sampling units of Trends in International Mathematics and Science Study (TIMSS) were schools within each country (Liu, Wu, & Zumbo, 2006). TIMSS scores can be quite different from country to country due to different curriculum designs or education systems (Hagiwara & Matsubara, 2012). It is very important for the educators to understand the distinctions between the true difference in TIMSS scores across countries and DIF caused by secondary factors (*e.g.*, cultural characteristics, Wu & Ercikan, 2006) before any high-stakes education policy is made (*e.g.*, adoption of foreign education or testing system). The second goal of the current study, therefore, is to examine the effect of MAD on the performance of both LR and HLR under multilevel data structure.

2. Logistic Regression and Hierarchical Logistic Regression

Swaminathan and Rogers (1990) had adopted LR for detecting both uniform and nonuniform DIF. Uniform DIF indicates consistent change between reference and focal groups across θ, whereas nonuniform DIF indicates inconsistent change. The current study focused on uniform DIF detection to be consistent with previous studies for comparison purposes. The LR model is written as,

$$\eta_i = \gamma_0 + \gamma_1 G_i + \gamma_2 X_i, \quad (1)$$

Where, $\eta_i = g(P(Y_i = 1 | X, G))$ for person i, and g is the logit link. $G_i = 1$ for focal group and $G_i = 0$ for reference group. The significance test of the regression coefficient γ_1 in Equation (1) is used to determine the presence of uniform DIF. X_i is the covariate (*i.e.*, the total score) to

match θ between groups.

When data were sampled from clusters, HLR was recommended for DIF analysis by researchers to account for dependency between person level scores (French & Finch, 2010, 2013; Jin et al., 2014), especially when ρ was medium to large. The HLR model is written as,

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{10} X_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} G_j + \mu_{0j} \end{aligned} \tag{2}$$

Where, $G_j = 1$ for focal group and $G_j = 0$ for reference group, X_{ij} is the person level covariate (*i.e.*, the total score), and the random components $\mu_j \sim N(0, \tau^2_{y|x})$. The regression coefficient γ_{01} is used to determine the presence of DIF in terms of significance test. The current study focused on the grouping variable being at the cluster level (*e.g.*, country, Klieme & Baumert, 2001) instead of person level (*e.g.*, gender, Taylor & Lee, 2012) because previous studies have shown that type I error rate can be well maintained at the nominal level when the grouping variable was at the person level, but was inflated when the grouping variable was at the cluster level (*e.g.*, French & Finch, 2010).

For both LR and HLR, the covariate was included at the person level to match θ between groups. According to Jin et al. (2014), LR and HLR performed equivalently when ρ was small to medium with no MAD, given that the person level covariate was a sufficient estimate of θ for both reference and focal groups. The authors of the current study hypothesized that with short tests and the presence of MAD, both DIF methods, especially LR, will be more likely to fail to control type I error rate at the nominal level, because the presence of MAD between groups might be falsely classified as DIF, especially when the reliability of the covariate is low due to fewer items in a test.

3. Hierarchical Logistic Regression with Latent Covariates (HLR-LC)

One solution to account for low reliability of the covariate is to use the latent variable approach, which has been demonstrated by researchers as advantageous when the covariate was not reliable (Asparouhov & Muthén, 2006; Bollen, 1989; Ludtke et al., 2008). By adopting the latent covariate approach, the authors of the current study proposed HLR with latent covariates (HLR-LC) for multilevel DIF detection to account for low reliability as well as the presence of MAD at the cluster level. HLR-LC is modeled as,

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{wj} \theta_{ijw} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} G_j + \gamma_{0b} \theta_{ijb} + \mu_{0j} \end{aligned} \tag{3}$$

Where, the observed person level covariate X_{ij} in Equation (2) is decomposed into two unobserved latent components: the within-cluster latent covariate (θ_{ijw}) and the between-cluster latent covariate (θ_{ijb}). HLR-LC was expected to control type I error rate more effectively than HLR and LR in three aspects. First, the multilevel data structure can be explicitly modeled by HLR to account for the violation of the independence assumption. Second, measurement error due to low reliability can be modeled by decomposing the covariate into two latent components at both within- and between-cluster levels. Third, in addition to the grouping variable (G_j), θ_{ijb} in HLR-LC is included to model MAD at the

cluster level.

4. Hypotheses for the Current Study

When MAD is present with short tests, the following hypotheses regarding the performance of the three DIF methods, LR, HLR, and HLR-LC are: (1) LR may not perform equivalently as HLR even when ρ is small to medium as shown in previous studies; (2) HLR-LC would outperform LR and HLR with respect to significance tests. Results of the current study are expected to expand the existing literature on multilevel DIF analysis in two ways. First, the current study attempts to investigate the effects of short tests and MAD on the accuracy of DIF detection of LR and HLR, which have not been studied in previous multilevel DIF analysis. Second, the current study also provides supporting evidence for HLR-LC as an effective DIF method when tests are short and MAD is present under multilevel data structure.

5. Methods

Six factors were manipulated in the current Monte Carlo study to investigate the comparative performance of LR, HLR, and HLR-LC. Levels within each factor were selected to replicate conditions generated in previous studies (*e.g.*, studies with longer tests and no MAD) as close as possible for comparison purposes. Test length (2 levels), mean ability difference (2 levels), intraclass correlation (4 levels), DIF size (2 levels), number of clusters (3 levels), and sample size within each cluster (3 levels) were crossed to create 288 conditions. Each condition was replicated 1000 times. Item responses were generated by incorporating the manipulated factors under the Rasch model using the R package for statistical computing (R core Team, 2013). Item parameters were obtained from a subset of item parameters in Narayanan and Swaminathan (1996). The item difficulty parameters ranged from -2.7 to 1.68. Uniform DIF (*i.e.*, difference in difficulty parameters) reflecting different DIF size was generated to be consistent with previous studies.

5.1 Manipulated Factors

Tests with 5 and 10 items were generated to represent short tests as compared to previous studies with long tests (*e.g.*, 20-item test, Jin et al., 2014). The number of items selected was within the range of short tests defined in previous research (*e.g.*, Woods, 2009). In practice, tests, especially tests containing several testlets measuring different sub-domains, normally contain fewer items. Because of the unidimensionality assumption, DIF analyses were often conducted on each sub-domain instead of on the entire test (Jiao et al., 2012). Within the 5- or 10-item test, one studied item was generated to be a DIF-free or DIF-present item when computing type I error rate and power, respectively. The rest of the items on the tests were generated to be DIF-free because DIF contamination was not of primary interest and was not manipulated in the current study.

For the MAD-absent condition, θ of both reference and focal groups was generated from $N\sim(0, 1)$. For the MAD-present condition, θ of the reference group was generated from $N\sim(0, 1)$, whereas θ of the focal group was generated from $N\sim(-1, 1)$. The one standard deviation difference in θ between the reference and focal groups was commonly observed in previous

simulation studies (*e.g.*, Finch, 2005; Oort, 1998).

Four levels of ρ were simulated to cover a wide range of ρ observed frequently in empirical studies (Hedges, 2007; Hedges & Hedberg, 2007; Liu et al., 2006). The four levels are 0.1, 0.2, 0.3, and 0.4. Levels of ρ were generated by adding different magnitudes of random effects due to clusters to θ values thus generating item responses. Two levels of DIF size were generated to represent commonly observed small to medium DIF size (0.3) and medium to large DIF size (0.6) for the purpose of computing power (DeMars, 2009; Jodoin & Gierl, 2001).

Two factors were manipulated to generate different levels of sample size. Three levels of number of clusters were generated to reflect small (30), medium (50), and large (100) number of clusters. Three levels of sample size within each cluster were generated to reflect small (10), medium (30), and large (50) sample size within each cluster. The sample size parameters were set close to parameters of previous multilevel DIF analysis studies for comparison purposes.

5.2 Evaluation

The analyses of LR, HLR, and HLR-LC were conducted using Mplus 7.1 (L. K. Muthén & B. O. Muthén, 2013). Maximum likelihood estimation was implemented for all three methods. Type I error rate and power were reported as outcomes to evaluate the comparative performance of these three methods, where Type I error rate was calculated as the ratio of the number of DIF-free items falsely identified as DIF-present items to the number of replications. Power was calculated as the ratio of the number of DIF-present items correctly identified as DIF-present items to the number of replications.

6. Results

6.1 Type I Error Rate

Table 1 provides type I error rate and its standard deviation (SD) aggregated across conditions for each level of the manipulated factors. Across most of these conditions, HLR-LC outperformed both HLR and LR in terms of controlling type I error rate at the nominal level of 0.05. Conditions where HLR-LC exhibited some level of inflated type I error rate were shorter tests (*i.e.*, 5-item test), small ρ (*i.e.*, 0.1), large number of clusters (*i.e.*, 100), and small samples within each cluster (*i.e.*, 10). Among all the manipulated factors, MAD appeared to be the most influential factor because type I error rate was much more stable under the MAD-absent conditions (*i.e.*, SD: 0.008-0.037) than under the MAD-present conditions (SD: 0.077-0.264). In addition, the differences in SDs between MAD-present and MAD-absent conditions were the largest as compared to the SD differences between levels of the other factors. Test length, although not as influential as MAD, appeared to have a relatively larger effect on the performance of the three DIF methods than ρ and sample size.

Table 1. Averaged type I error rate across conditions (standard deviations aggregated across conditions)

Factors	Levels	LR	HLR	HLR.LC
Test length	5	0.474 (0.394)	0.436 (0.408)	0.085 (0.079)
	10	0.331 (0.323)	0.311 (0.323)	0.051 (0.024)
Mean ability	no	0.089 (0.037)	0.055 (0.009)	0.041 (0.008)
	yes	0.716 (0.264)	0.692 (0.27)	0.095 (0.077)
Intraclass	0.1	0.431 (0.404)	0.415 (0.407)	0.101 (0.101)
	0.2	0.415 (0.382)	0.392 (0.388)	0.068 (0.049)
	0.3	0.397 (0.36)	0.363 (0.365)	0.054 (0.028)
	0.4	0.368 (0.327)	0.324 (0.333)	0.049 (0.016)
Number of	30	0.33 (0.308)	0.294 (0.302)	0.049 (0.026)
	50	0.401 (0.362)	0.371 (0.366)	0.064 (0.047)
	100	0.477 (0.414)	0.455 (0.428)	0.091 (0.087)
Sample size	10	0.271 (0.266)	0.253 (0.262)	0.081 (0.077)
	30	0.436 (0.38)	0.405 (0.386)	0.066 (0.057)
	50	0.501 (0.405)	0.462 (0.424)	0.057 (0.044)

To better understand the effects of MAD and test length at each level of the other manipulated factors, type I error rate was plotted for both conditions of MAD across levels of test length, ρ, number of clusters, and sample size within each cluster. Figure 1 shows that under the MAD-absent condition, HLR and HLR-LC were able to control type I error rate consistently across levels of ρ even with shorter tests (*i.e.*, 5-item test). LR, on the other hand, exhibited inflated type I error rate as ρ increased and the magnitude of inflation was greater when the test was shorter. Under the MAD-present condition, HLR-LC outperformed both HLR and LR across levels of ρ. HLR and LR exhibited greater type I error inflation, especially with shorter tests. In summary, test length seemed to have no effect on HLR and HLR-LC under the MAD-absent condition, but had an effect on LR and HLR under the MAD-present condition. HLR-LC was able to control type I error rate under the MAD-present condition, and was slightly conservative under the MAD-absent condition across levels of ρ. The inflation of type I error rate of LR, however, was greater as test length decreased and as ρ increased.

Figures 2 and 3 show the effects of MAD and test length with sample size factors. Under the MAD-absent condition, the number of clusters seemed to have no effect on the relative performance of LR, HLR, and HLR-LC. Under the MAD-present condition, as the number of clusters increased, type I error rate of HLR and LR also increased. With relatively longer tests

(*i.e.*, 10-item test), HLR-LC was able to maintain type I error rate at different levels of number of clusters. On the other hand, sample size within each cluster had no effect on HLR and HLR-LC, but had an effect on LR under the MAD-absent condition. As sample size within each cluster increased, type I error of LR also increased with a greater magnitude of inflation with shorter tests. Under the MAD-present condition, as sample size within each cluster increased, type I error rate of HLR and LR also increased. With relatively longer tests (*i.e.*, 10-item test), HLR-LC was able to maintain type I error rate at different levels of sample size within each cluster. In summary, the two sample size related factors had greater effects on LR and HLR than on HLR-LC.

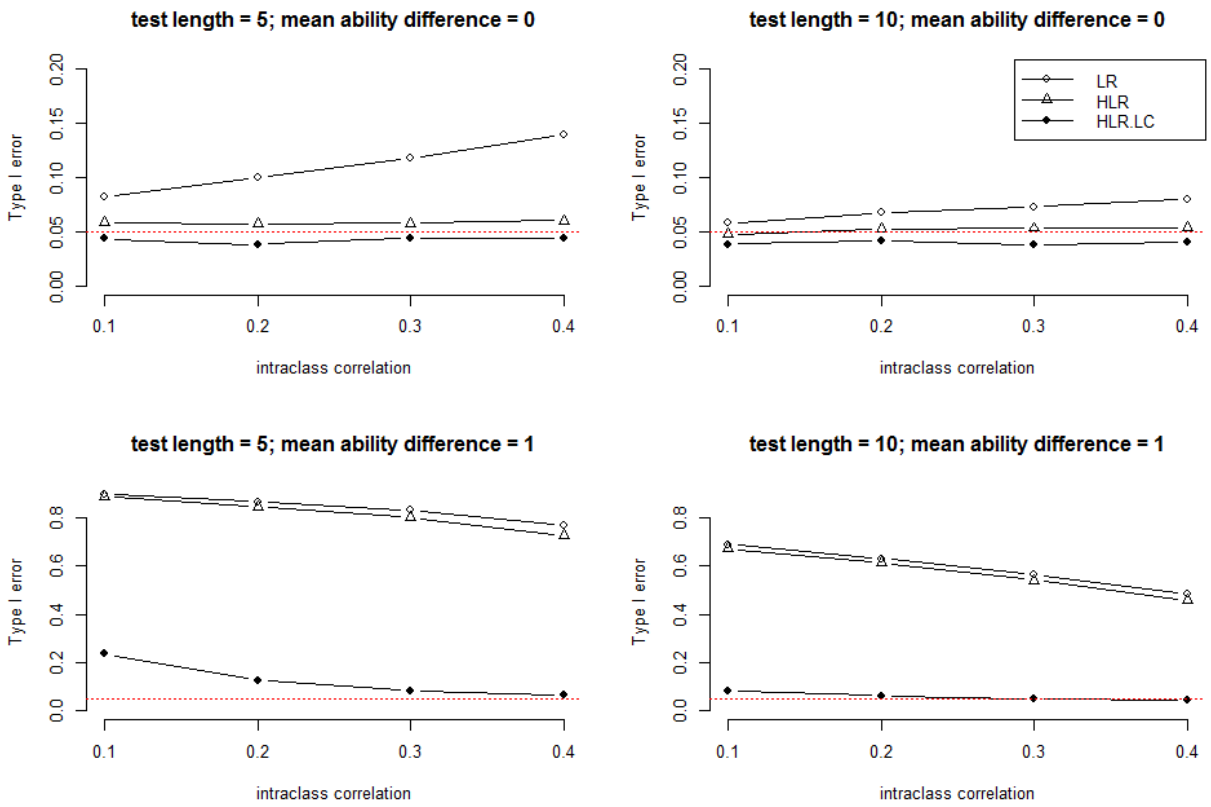

Figure 1. Effects of test length and MAD at each level of intraclass correlation

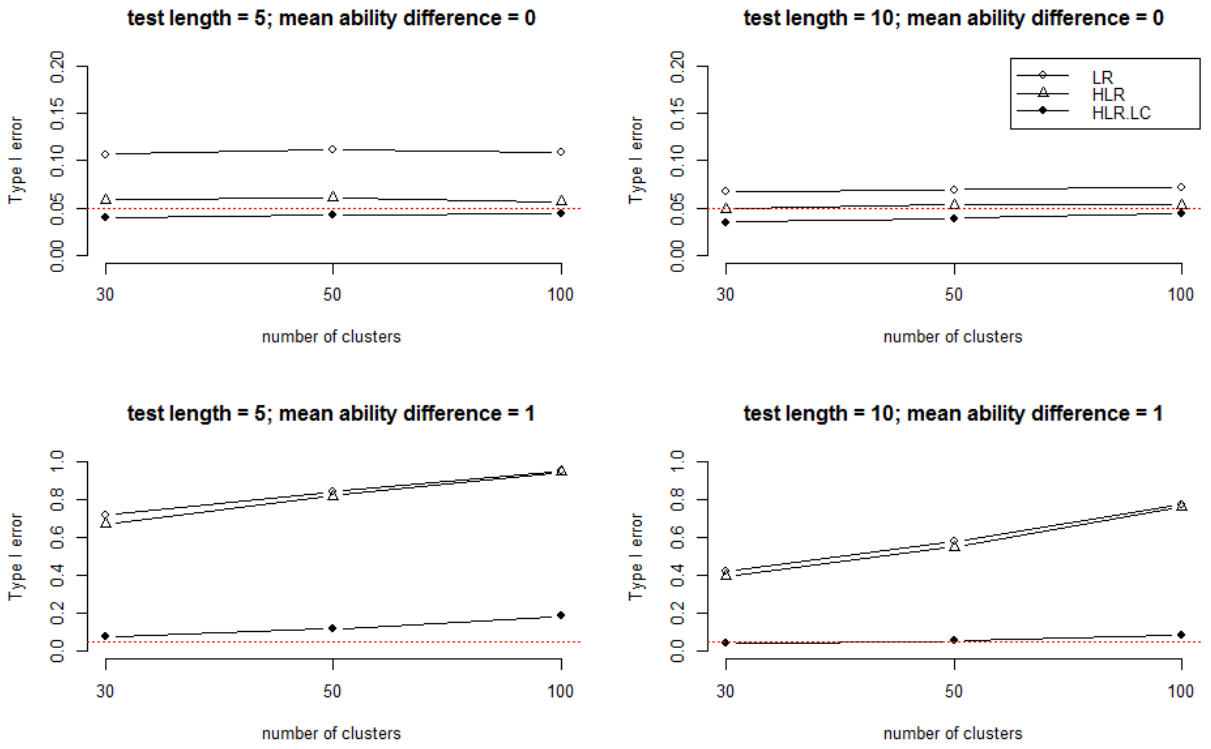

Figure 2. Effects of test length and MAD at each level of number of clusters

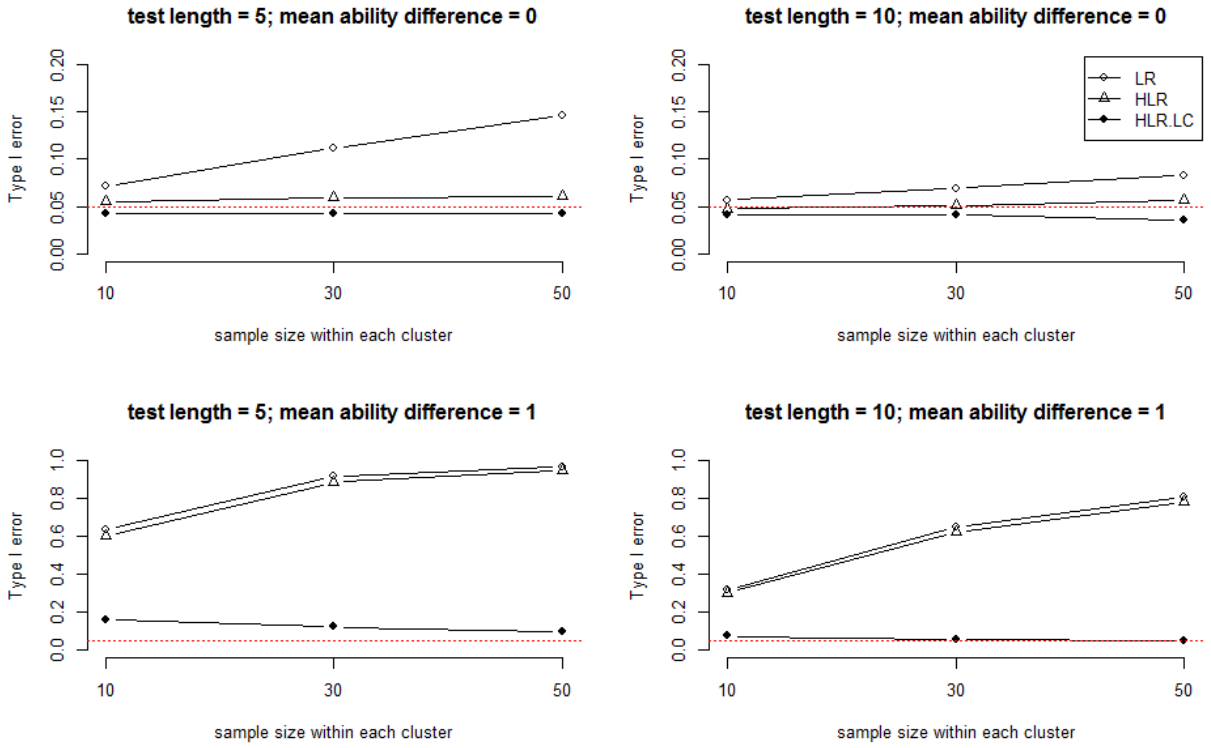

Figure 3. Effects of test length and MAD at each level of sample size within each cluster

6.2 Power

Table 2 provides power and its standard deviation (SD) aggregated across conditions for each level of the manipulated factors. For both LR and HLR, power exceeded the acceptable level of 0.8 across all levels of the manipulated factors, which was not meaningful because of inflated type I error rate, especially for LR. HLR-LC appeared to be powerful under most conditions except when levels of sample size related factors were small (*i.e.*, 30 clusters and 10 individuals within each cluster). The effects of most manipulated factors on the performance of the three DIF methods were consistent: as DIF size, test length, number of clusters, and sample size within each cluster increased, power also increased for all three DIF methods; as ρ increased, power of the three DIF methods decreased. As for the effect of MAD, power of LR and HLR was greater under the MAD-present condition than under the MAD-absent condition, which again might not be trustworthy because of the greater magnitude of type I error inflation under the MAD-present conditions for both LR and HLR. Power of HLR-LC under both MAD-absent and MAD-present conditions was quite similar.

Table 2. Averaged power across conditions (standard deviations aggregated across conditions)

Factors	Levels	LR	HLR	HLR.LC
DIF size	0.3	0.901 (0.170)	0.878 (0.189)	0.797 (0.218)
	0.6	0.992 (0.027)	0.989 (0.035)	0.975 (0.056)
Test length	5	0.945 (0.133)	0.926 (0.157)	0.887 (0.180)
	10	0.948 (0.127)	0.941 (0.136)	0.885 (0.184)
Mean ability difference	no	0.905 (0.169)	0.881 (0.188)	0.890 (0.193)
	yes	0.988 (0.044)	0.986 (0.049)	0.882 (0.170)
Intraclass correlation	0.1	0.956 (0.121)	0.951 (0.130)	0.890 (0.178)
	0.2	0.949 (0.129)	0.939 (0.141)	0.888 (0.180)
	0.3	0.944 (0.132)	0.929 (0.149)	0.887 (0.181)
	0.4	0.937 (0.139)	0.914 (0.165)	0.880 (0.192)
Number of clusters	30	0.900 (0.176)	0.876 (0.197)	0.781 (0.229)
	50	0.953 (0.118)	0.940 (0.134)	0.901 (0.158)
	100	0.987 (0.044)	0.984 (0.052)	0.976 (0.056)
Sample size within each cluster	10	0.871 (0.194)	0.856 (0.209)	0.764 (0.238)
	30	0.976 (0.062)	0.963 (0.093)	0.926 (0.123)
	50	0.992 (0.026)	0.981 (0.061)	0.968 (0.068)

To examine the effects of MAD and test length at each level of the other manipulated factors, plots similar to Figures 1, 2, and 3 were created for power of LR, HLR, and HLR-LC at each level of DIF size. Figure 4 shows that ρ had not effect at each level of MAD and test length on power for all three DIF methods. Figure 5 shows that as number of cluster increased, power of all three DIF methods also increased. Figure 6 shows similar pattern as in Figure 5, where power of all three DIF methods increased as sample size within each cluster increased. Figures 4, 5, and 6 were created for DIF size = 0.3. When DIF size = 0.6, the patterns of the effects of MAD and test length at each level of ρ and sample size related factors were consistent except that power for DIF = 0.6 were higher than power for DIF size = 0.3. Figures for DIF size = 0.6 were not included but are available upon request.

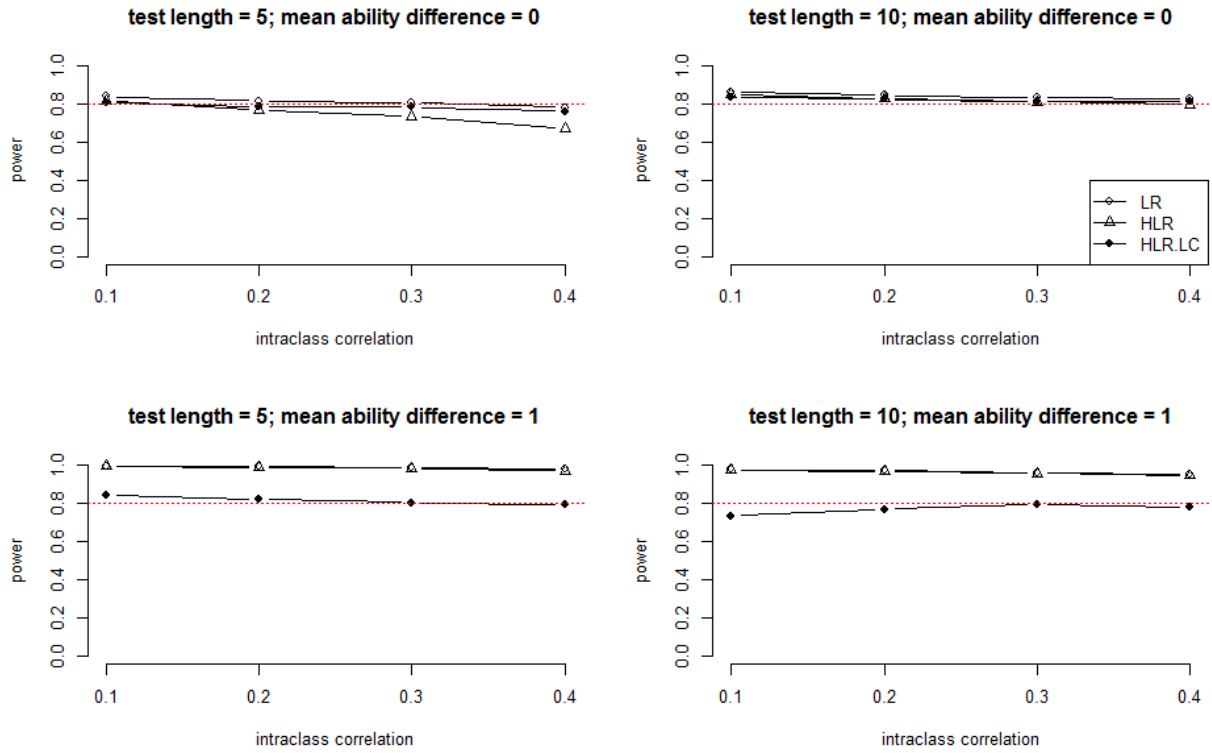

Figure 4. Effects of test length and MAD at each level of intraclass correlation

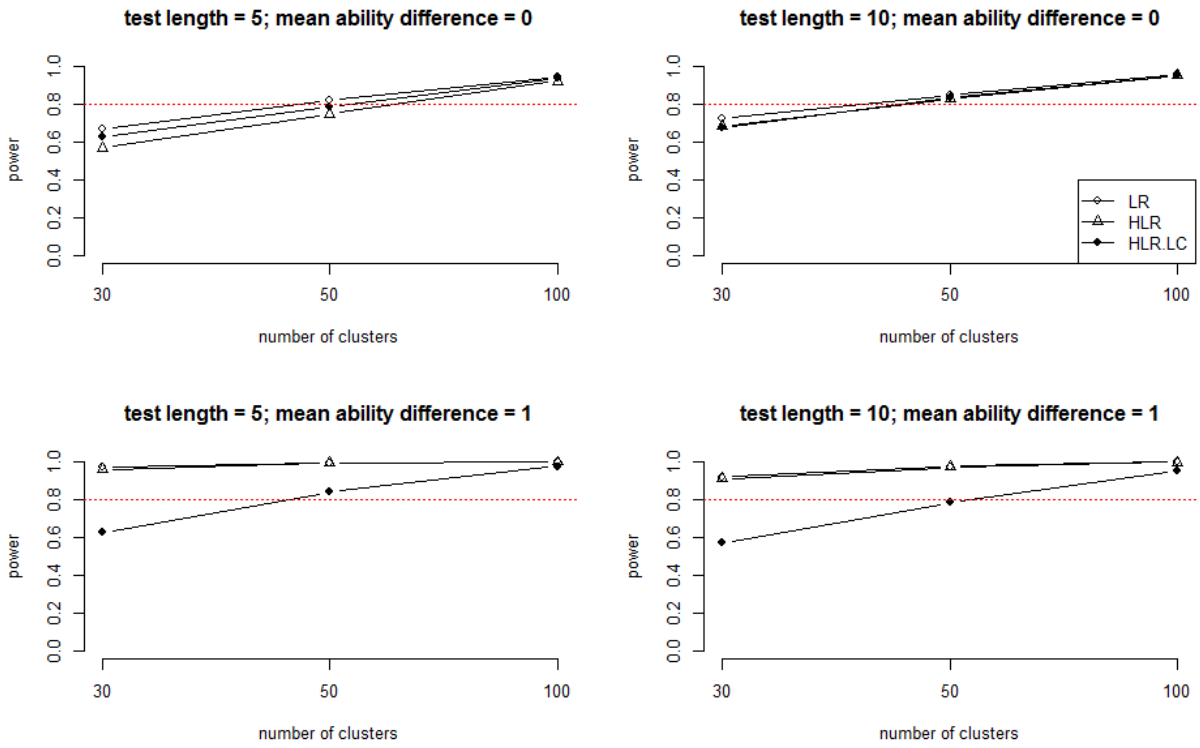

Figure 5. Effects of test length and MAD at each level of number of clusters

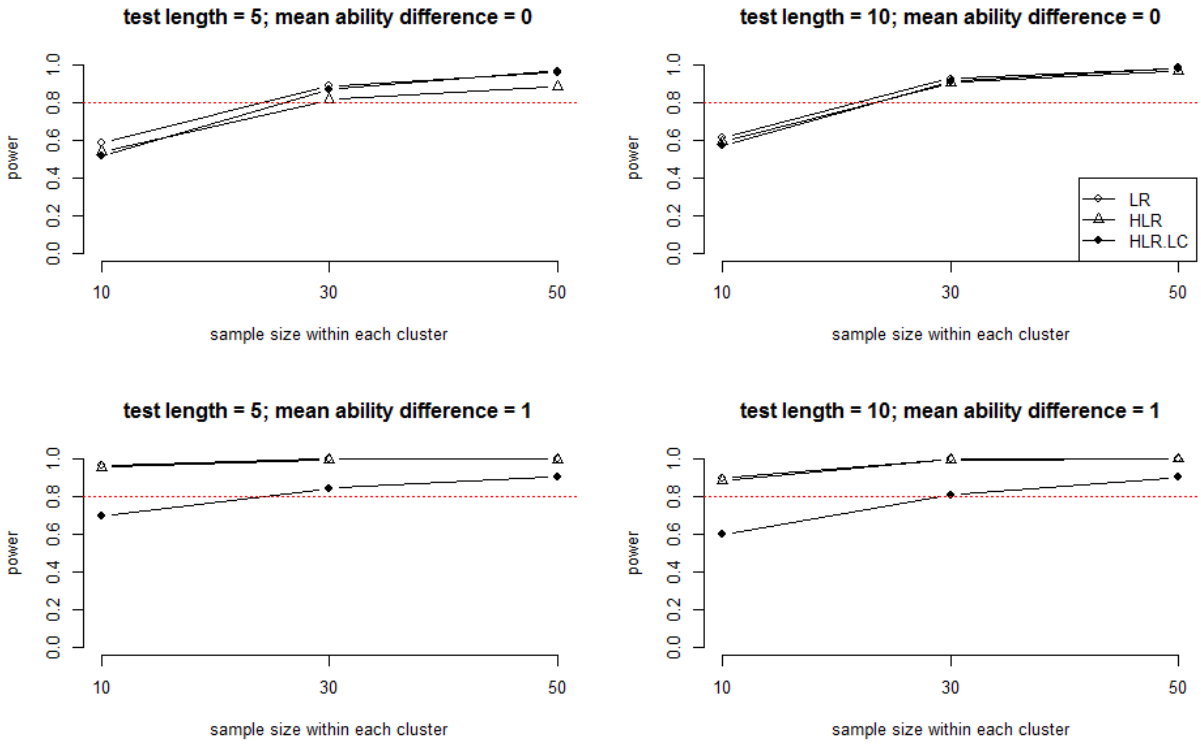

Figure 6. Effects of test length and MAD at each level of sample size within each cluster

7. Discussion

The current study evaluated two hypotheses regarding the comparative performance of LR, HLR, and HLR-LC with multilevel data. The first hypothesis stated that LR may not perform equivalently as HLR even when ρ was small to medium, which was confirmed by the results of the current study. Jin et al. (2014) have shown that LR performed equivalently as HLR with multilevel data when ρ was small to medium with no MAD between groups for relatively longer tests. The results of the current study showed that HLR outperformed LR regardless of the magnitude of ρ when MAD was present with shorter tests. As previously discussed, short tests might suffer from low reliability, therefore affecting the matching ability of the covariate in DIF analysis. The large effect of MAD on type I error inflation of LR indicated that information provided by short tests was not sufficient to match θ across groups, so that the true difference in θ between groups was contaminated by DIF, leading to inflated type I error rate.

The findings of the current study have important practical implications for both educational and psychological assessments. Researchers have proposed methods to optimize reliability when shorter tests were desirable (*e.g.*, saving administration time, Raykov, 2014) or inevitable (*e.g.*, long tests consisting of several sub-domains of the latent trait). Prilleltensky et al. (2015) recently developed a 21-item scale measuring the overall well-being. Seven sub-domains (*e.g.*, psychological well-being) were included in the scale and each sub-domain was measured by only 3 items. For scales like these, DIF analysis employing standard DIF methods (*e.g.*, LR) on each sub-domain would be problematic, especially when such a scale is to be used to collect data for cross-country comparisons, where MAD is likely to occur due to different cultural characteristics. Complex modeling of DIF can be conducted to account for the unidimensionality and independence assumption violations, for example, multilevel-multidimensional IRT DIF analysis (Walker, Zhang, & Surber, 2008), or multilevel testlet analysis (Jiao et al., 2012). These types of analyses, however, require more technical work from applied researchers.

The current study provided a relatively simple yet effective DIF method when MAD was present for short tests with multilevel data. The evaluation of the second hypothesis confirmed that HLR-LC was a more effective DIF method than both LR and HLR under most of the simulated conditions. From a practical perspective, implementation of HLR-LC is especially beneficial to situations where MAD is present for large scale analysis. For example, TIMSS was conducted in more than sixty countries, and the data collected were multilevel in nature. Liu et al. (2006) showed that the intraclass correlation of TIMSS mathematics scores could be as high as 0.62 (*e.g.*, Hong Kong). Previous DIF analysis on TIMSS scores used standard DIF methods (*e.g.*, Mantel-Haenszel test or LR) without taking the multilevel data structure into account (Innabi & Dodeen, 2006; Klieme & Baumert, 2001; Wu & Ercikan, 2006). Results of these studies might be questionable unless the magnitude of intraclass correlation was examined and MAD was negligible. The current study provided supporting evidence for implementing HLR-LC in large scale cross-country comparison studies, where HLR-LC can help maintain type I error rate at the nominal level, and distinguish DIF from true MAD between groups as well.

Based on the results of the current study, another practical implication for researchers is that short tests for cross-country comparison studies are generally not recommended, especially when MAD is present. In the current study, DIF contamination was not manipulated, a single studied item was included with the rest of the items being DIF-free. In practice, it is possible that multiple items can be DIF-present within a short test. Although item purification procedures can be conducted in advance (Wang, Shih, & Yang, 2009) and the purified covariate (*i.e.*, total score computed after excluding DIF-present items) can be used in HLR-LC to match θ between groups, with tests being short, the purified covariate might be difficult to cover a wide range of θ. One possible solution is to use multiple indicators multiple causes (MIMIC) model, which is robust against DIF contamination (Finch, 2005). In addition, the MIMIC model, like the HLR-LC method, is also a latent variable approach where group members are matched based on θ instead of on observed total scores as in LR and HLR. To ensure the accuracy of DIF analysis under frequently observed situations in practice as much as possible, future research can extend the current study by comparing these two latent variable approaches after including DIF contamination as an additional factor. Finally, future research can be further beneficial to the literature on DIF when effect size estimates of the three DIF methods are examined and compared to control for the meaningless high power due to inflated type I error rate.

References

Asparouhov, T., & Muthén, B. (2006). Constructing covariates in multilevel regression. *Mplus Web Notes: No. 11*. Retrieved from http://www.statmodel.com

Bock, R. D. (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118619179

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. TX: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. https://doi.org/10.1007/BF02310555

DeMars, C. E. (2009). Modification of the Mantel–Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*, 149-170. https://doi.org/10.3102/1076998607313923

Fidalgo, A. M., Mellenbergh, G. J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*(3), 43-53.

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. https://doi.org/10.1177/0146621605275728

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for

multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299-317. https://doi.org/10.1111/j.1745-3984.2010.00115.x

French, B. F., & Finch, W. H. (2013). Extensions of Mantel-Haenszel for multilevel DIF Detection. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0013164412472341

Hagiwara, Y., & Matsubara, K. (2012). *A DIF analysis of TIMSS = 2007 assessment in Physics and Chemistry Focusing on the matching of the test items and the curricula: The comparison of Japanese and Korean English Grader*s. Paper presented at the 10th annual conference of the Japan Association for Research on Testing, Tokyo.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*, 341-370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning grouprandomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.org/10.3102/0162373707299706

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. https://doi.org/10.1111/j.1949-8594.2006.tb17753.x

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100. https://doi.org/10.1111/j.1745-3984.2011.00161.x

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. https://doi.org/10.1177/0013164413497572

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349. https://doi.org/10.1207/S15324818AME1404_2

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. https://doi.org/10.1007/BF03173189

Liu, Y., Wu, A. D., & Zumbo, B. D. (2006). The relation between outside of school factors and mathematics achievement: A cross-country study among the U.S. and five top-performing Asian countries. *Journal of Educational Research & Policy Studies, 6*(1), 1-35.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*(3), 203-229. https://doi.org/10.1037/a0012869

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Murphy, K., & Davidshofer, C. (1988). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice Hall.

Muthén, L. K., & Muthén, B. O. (2013). *Mplus: Statistical Analysis with Latent Variables* (version 7.1) [Computer software]. Los Angeles, CA: Author.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107-124. https://doi.org/10.1080/10705519809540095

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Los Angeles, CA: Sage. https://doi.org/10.4135/9781412993913

Prilleltensky, I., Dietz, S., Prilleltensky, O., Myers, N. D., Rubenstein, C., Jin, Y., & McMahon, A. (2015). Assessing Multidimensional Well-Being: Development and Validation of the ICOPPE Scale. *Journal of Community Psychology, 43*(2), 199-226. https://doi.org/10.1002/jcop.21674

R Core Team. (2013). *R: A Language and Environment for Statistical Computing* (version 3.0.2) [Computer software]. Vienna, Austria: Author.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raykov, T. (2014). *On Optimal Shortening of Psychometric Scales*. Paper presented at the 79th Annual Meeting of the Psychometric Society, Madison, WI.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item format. *Applied Measurement in Education, 25*(3), 246-280. https://doi.org/10.1080/08957347.2012.687650

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*(2), 162-181. https://doi.org/10.1080/08957340801926201

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC Method With Scale Purification

for Detecting Differential Item Functioning. *Educational and Psychological Measurement, 69*(5), 713-731. https://doi.org/10.1177/0013164409332228

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27. https://doi.org/10.1080/00273170802620121

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. https://doi.org/10.1207/s15327574ijt0603_5

Zimmerman, D. W., & Williams, R. H. (1977). The theory of test validity and correlated errors of measurement. *Journal of Educational Measurement, 19*, 125-134. https://doi.org/10.1016/0022-2496(77)90063-3

Copyright Disclaimer