Detecting Cluster Bias in a Multilevel Item Response Model: A Monte Carlo Evaluation of Detection Methods and Consequences of Ignoring Cluster Bias

By

Woo-yeol Lee

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

May, 2016

Nashville, Tennessee

Approved:

Sun-Joo Cho, Ph.D.

Kristopher J. Preacher, Ph.D.

Sonya K. Sterba, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

Introduction

In many educational and psychological research settings, data often have a multilevel structure, such as students within schools or participants nested within day care centers. Further, binary outcome variables (e.g., true-false answer, present-absent symptom, endorsed-not endorsed attitude) are often assessed. When the data structure is multilevel resulting from cluster sampling or multistage sampling and the type of outcome is binary, multilevel item response models have been widely applied.

In many multilevel item response model applications, the same item discriminations or loadings are often assumed at the within-level (e.g., the student level) and the between-level (e.g., the school level). In Rasch multilevel item response models, it is assumed that item discriminations over levels are the same and constant (Kamata, 2001). In two-parameter multilevel item response models, the same item discriminations over levels are estimated (e.g., Fox & Glas, 2001; Fox, 2004, 2005, 2010; Houts & Cai, 2013; Jeon & Rabe-Hesketh, 2012). Further, the same item discriminations are assumed over levels in a multilevel extension of multiple-indicator multiple-cause (MIMIC; Jöreskog & Goldberger, 1975) approaches (Finch & French, 2011; Kim, Suh, Kim, Albanese, & Langer, 2013). The model formulation that has the same item discriminations over levels is referred to as a variance component factor model (Rabe-Hesketh, Skrondal, & Pickles, 2004).

Following the tradition of a general multilevel factor model or multilevel structural equation modeling (MSEM; McDonald, 1993; Muthén, 1991, 1994; Rabe-Hesketh, Skrondal, & Pickles, 2004), it is possible to have separate item discrimination parameters at each level of multilevel data. Jak, Oort, and Dolan (2014) used the term *cluster bias* to refer to measurement bias across clusters, and cluster bias can be interpreted as measurement bias regarding any cluster-level variable. Item discrimination differing across levels is considered evidence of *cluster bias* or a *lack of cluster invariance*. Cluster bias can exist at the test level or at the item level. Hereafter, cluster bias at the test level is called *global cluster bias*, whereas cluster bias at the item level is called *item cluster bias*.

Cluster invariance is an important assumption to test in real data applications. Within-level item discrimination indicates how strongly each item correlates with a within-level latent variable, and between-level item discrimination indicates how strongly each item correlates with a between-level latent variable. Thus, in the presence of cluster bias, the latent variables in multilevel item response models do not have the same scale or meaning over levels. In such cases, separate scores at different levels should be reported (Cronbach, 1976). When cluster invariance is assumed, it can be hypothesized that the mean of the within-level scores is approximately equal to the between-level scores (Patarapichayatham & Kamata, 2014). However, in the presence of cluster bias, it is not appropriate to report the mean individual scores instead of the between-level scores.

As noted in Muthén and Asparouhov (2013), some applications are required to have different item discriminations at each level and different numbers of latent variables at each level. For example, Härnqvist, Gustafsson, Muthén, and Nelson (1994) found “fluid” abilities highly loaded on a general factor in addition to five other residual factors at the student level, whereas “crystallized” abilities highly loaded on the general factor in addition to two other residual factors at the classroom level.

In addition, testing cluster bias is crucial for reducing the number of parameters to be estimated in multilevel item response models or MSEM with binary responses. The model with cluster invariance is a much simpler model than the model with cluster bias. In two-level data with cluster bias, for example, a model with cluster invariance can be obtained by setting equality constraints between the within-level item discriminations and the between-level item discriminations when the number of latent variables over levels is the same in the model.

Previous research on measurement invariance in multilevel data focused on testing whether the parameters of multiple-group multilevel confirmatory factor analysis (CFA) are the same across groups at the cluster level (e.g., treatment school group vs. regular school group) (Jak & Oort, 2015; Kim, Kwok, & Yoon; 2012; Mehta & Neale, 2005; Muthén, Khoo, & Gustafsson, 1997; Ryu, 2014) or at the individual level (e.g., male students vs. female students) (Jak et al., 2014; Kim, Yoon, Wen, Luo, & Kwok, 2015; Ryu, 2014). Jak and Oort (2015) reported the performance of a Wald test and a likelihood ratio test (LRT) to detect cluster bias at the cluster level with a two-level common factor model.

Compared to individual-level or cluster-level research, few studies have addressed the cluster bias over levels (e.g., the student level and the school level). Jak et al. (2014) presented a method for testing cluster bias in a two-level common factor model using a chi-square difference test and evaluated the performance of the test to detect cluster bias over levels. De Jong, Steenkamp, and Fox (2007) and Fox and Verhagen (2010) presented random item response models to test whether individual-level item parameters differ over clusters. Patarapichayatham and Kamata (2014) showed the effects of different patterns and magnitudes of item discriminations over levels on the estimates of within-level and between-level abilities in a two-parameter multilevel item response model. However, there is a lack of research on the evaluation of cluster bias detection methods and the consequences of ignoring cluster bias in terms of the accuracy of parameter estimates in the use of multilevel item response models.

Thus, the first purpose of this study is to evaluate detection methods for cluster bias. The second purpose is to show the consequences of ignoring cluster bias for the accuracy of parameter estimates and standard errors (SEs) in a two-parameter multilevel item response model because many multilevel IRT applications did not consider the cluster bias in the model. We limit our study to two-level data and a latent variable at each level that are common among educational and psychology studies. Further, in this study, the parameters of the model are estimated using marginal maximum likelihood estimation (MMLE). Accordingly, cluster bias detection methods are discussed when MMLE is used.

This paper is organized as follows. First, we specify the two-parameter multilevel item response model with and without cluster bias and present the detection methods. Second, an empirical study is shown to illustrate global and item cluster bias detection when the two-parameter multilevel item response models are used. Subsequently, a simulation study is presented to evaluate the detection methods and to show the consequences of ignoring cluster bias. We end with a summary and a discussion.

Chapter 2

Multilevel Item Response Model and Cluster Bias

To frame this data structure within the multilevel literature (e.g., Bryk & Raudenbush, 1992, Ch. 8), item responses at Level 1 are cross-classified with persons and items. Persons at Level 2 are nested within clusters at Level 3. In our specification, Level 2 is the within-level, and Level 3 is the between-level.

2.1 Multilevel Item Response Models

Multilevel Item Response Model with Cluster Bias

Figure 2.1 depicts a two-level two-parameter multilevel item response model with cluster bias, which is then specified with equations. In the figure, the squares and the ellipses represent manifest and latent variables, respectively. Item responses are specified as $[y_{jk1},\ldots,y_{jki},\ldots,y_{jkI}]'$ for person j $(j=1,\ldots,J)$, cluster k $(k=1,\ldots,K)$, and item $(i=1,\ldots,I)$. Dependency in item responses is explained by two latent variables, θ_{jk} and θ_k, for the within-level and the between-level, respectively. Each item has its own item discrimination at Level 2 and Level 3, specified as $\alpha_{i,W}$ and $\alpha_{i,B}$, respectively. An item location parameter, $\beta_{i,B}$, is specified at Level 3.

Let there be a latent response y^*_{jki} so that the observed response is 1 when $y^*_{jki} > 0$ and 0 otherwise. Assuming that

$$y^*_{jki} = \alpha_{i,W}\cdot\theta_{jk} + \alpha_{i,B}\cdot\theta_k - \beta_{i,B} + \varepsilon_{jki}, \tag{2.1}$$

where ε_{jki} is a logistic distribution with a logit link. An individual item score is the combination of the cluster mean and its deviation from the cluster mean ($E[y^*_{jki}] = E[y^*_{Bki}] + E[y^*_{W\,jki}]$) (Heck & Thomas, 2009). Accordingly, an item location is presented at Level 3.

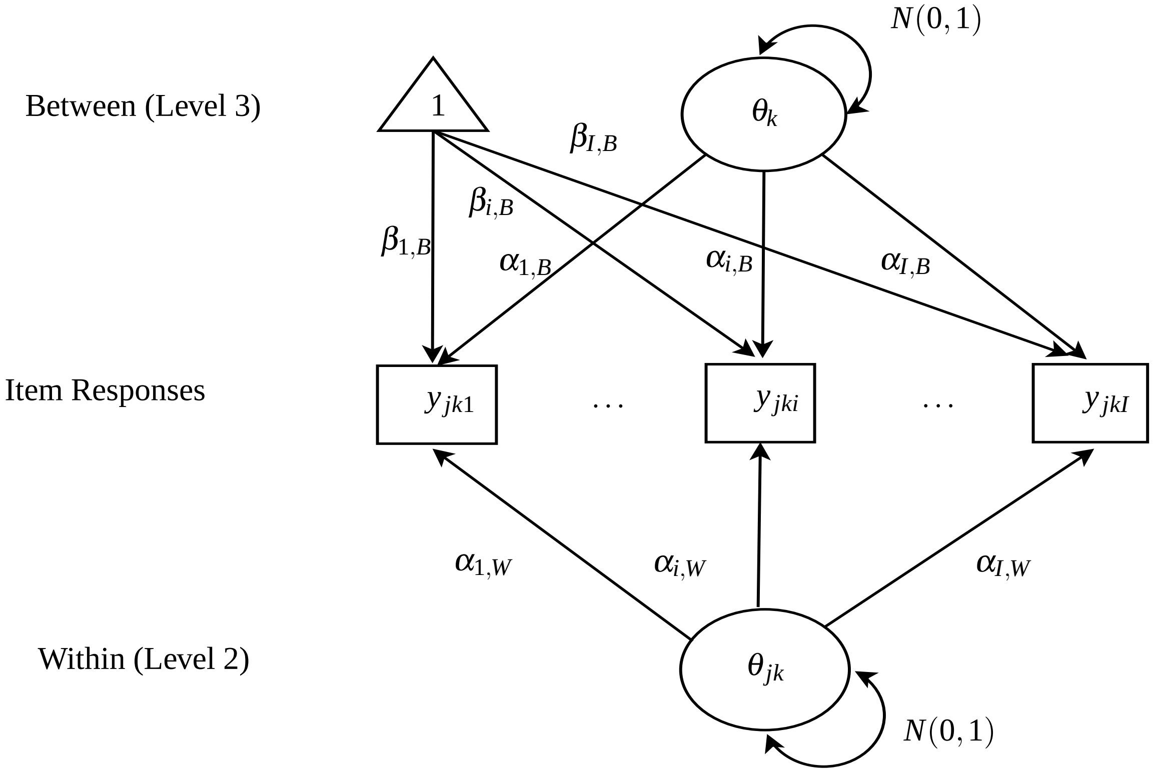

Figure 2.1: Multilevel (two-level) item response model with cluster bias

The latent response formulation in Equation 2.1 produces the model for the observed response y_{jki}. The two-parameter multilevel item response model with cluster bias is as follows:

$$\text{logit}[P(y_{jki} = 1|\theta_{jk}, \theta_k)] = \alpha_{i,W} \cdot \theta_{jk} + \alpha_{i,B} \cdot \theta_k - \beta_{i,B}. \tag{2.2}$$

To identify the model, the means and variances of the latent variables (σ^2 and τ^2) are set to 0 and 1, respectively. Alternatively, item location and discrimination for one of the items (e.g., the first item) can be set to 0 and 1, respectively, instead of setting the variances to 1: $\beta_1 = 0$, $\alpha_{1,W} = 1$, and $\alpha_{1,B} = 1$ to identify the location and scale units of parameters. This scaling of the unit variances over levels provides comparable item discrimination estimates over levels. However, this scaling does not yield equal units on a common construct. As in multigroup item response models, item parameters can be linked through anchor items over levels.

Based on Equation 2.2, an intraclass correlation (ICC) can be specified for each item to indicate the proportion of variance that is attributable to clusters. The ICC is the correlation coefficient (Corr) among the probabilities of the item responses on the logit scale for the same cluster k, but different persons j and j', and can be defined as follows:

$$\text{ICC}_i = \text{Corr}(P(y_{jki}), P(y_{j'ki})) = \frac{\text{Cov}(P(y_{jki}), P(y_{j'ki}))}{\sqrt{\text{Var}(P(y_{jki}))} \cdot \sqrt{\text{Var}(P(y_{j'ki}))}} \tag{2.3}$$

$$= \frac{\alpha_{i,B}^2 \tau^2}{\sqrt{\alpha_{i,W}^2 \sigma^2 + \alpha_{i,B}^2 \tau^2} \cdot \sqrt{\alpha_{i,W}^2 \sigma^2 + \alpha_{i,B}^2 \tau^2}}. \tag{2.4}$$

With model identification constraints, $\sigma^2 = \tau^2 = 1$, ICC_i leads to

$$\text{ICC}_i = \frac{\alpha_{i,B}^2}{\sqrt{\alpha_{i,W}^2 + \alpha_{i,B}^2} \cdot \sqrt{\alpha_{i,W}^2 + \alpha_{i,B}^2}} = \frac{\alpha_{i,B}^2}{\alpha_{i,W}^2 + \alpha_{i,B}^2}. \tag{2.5}$$

The derivation for ICC_i is shown in the Appendix.

Cluster bias. Measurement invariance is tested at the following four levels (e.g., Widaman & Reise, 1997): (a) configural invariance-the dimension and the pattern of zero and non-zero loadings (or item discriminations) are the same across groups; (b) weak invariance-the loading is invariant across groups; (c) strong invariance-the loading and the intercept (or item location) are invariant across groups; and (d) strict invariance-the loading, the intercept, and the residual variances are invariant across groups. Applying these four analyses to cluster bias, only configural invariance and weak invariance are relevant to the use of multilevel item response models. In this study, configural invariance was assumed because we set a limit of one latent variable at each level. In multilevel measurement invariance testing for clusters, cluster bias over levels involves only item discriminations for the weak invariance assumption.

With the derivation of ICC$_i$ (Equation 2.5), cluster bias (CB$_\mathrm{i}$) can be calculated as follows:

$$\mathrm{CB}_i = \alpha_{i,B} - \alpha_{i,W} = \alpha_{i,B} - \alpha_{i,B} \cdot \sqrt{\frac{1-\mathrm{ICC}_i}{\mathrm{ICC}_i}}. \tag{2.6}$$

As can be seen in Equation 2.6, cluster bias magnitude increases with decreasing ICC$_i$.

Multilevel Item Response Models with Cluster Invariance

When cluster invariance is assumed, equality constraints over levels are imposed on item discriminations in Equation 2.2: $\alpha_{i,W} = \alpha_{i,B} = \alpha_i$. Accordingly, the two-parameter multilevel item response model with cluster bias and ICC$_i$ reduces to

$$\mathrm{logit}[P(y_{jki} = 1|\theta_{jk}, \theta_k)] = \alpha_i \cdot \theta_{jk} + \alpha_i \cdot \theta_k - \beta_{i,B} = \alpha_i \cdot (\theta_{jk} + \theta_k) - \beta_{i,B}, \tag{2.7}$$

and

$$\mathrm{ICC}_i = \frac{\alpha_i^2}{\alpha_i^2 + \alpha_i^2} = 0.5. \tag{2.8}$$

In the model with cluster invariance, the variance of the latent variable at Level 2 (σ^2) can be set to 1 for model identification, and the variance of the latent variable at Level 3 (τ^2) can be estimated. Alternatively, the ICC can be presented for latent variables in the models with cluster invariance:

$$\text{ICC}_{\theta} = \frac{\tau^2}{1+\tau^2}. \tag{2.9}$$

The ICC_i is often smaller than 0.3 in cross-sectional empirical studies (e.g., Bliese, 2002; Hox, 2002; Snijders & Bosker, 1999), which indicates that $\alpha^2_{i.B}$ is likely to be smaller than $\alpha^2_{i.W}$ in many applications. In addition, for the fixed number of $\text{ICC}_i = C$, the magnitudes of item discrimination parameters with the equality constraint, $\alpha_{i,B} = \alpha_{i,W} = \alpha_i$, are expected to be larger than the magnitudes of between-level item discriminations ($\alpha_{i,B}$) as the number of items with cluster bias increases. Further, as the number of items with cluster bias increases, the magnitudes of α_i are expected to be close to $\alpha^2_{i.W}$. Thus, in the case of $\text{ICC}_i = C$, the variance of the latent variable at Level 3 (τ^2) decreases in a cluster invariance model as the number of items with cluster bias increases (with $\sigma^2 = 1$ for model identification). To put this expected result of τ^2 into an equation,

$$\text{ICC}_i = \frac{\alpha^2_{i.B}\tau^2}{\alpha^2_{i.W} + \alpha^2_{i.B}\tau^2} = C, \tag{2.10}$$

where $\alpha^2_{i.B}$ tends to be overestimated than α^2_i and $\alpha^2_{i.W}$ is close to α_i as the number of items with cluster bias increases. With this pattern, τ^2 tends to be underestimated in order to have $\text{ICC}_i = C$.

2.2 Parameter Estimation

MMLE and expected a posteriori (EAP) scoring were implemented using Mplus version 7.11 (Muthén & Muthén, 1998-2015). In Mplus, MMLE can be implemented with the MLR estimator option, which provides a test statistic and SEs using the Huber-White sandwich estimator that are robust against non-normality. Fifteen adaptive quadrature points were used for estimation and EAP scoring.

Chapter 3

Detection Methods

In this section, the LRT (also known as the chi-square difference test) and model information criteria are described as methods for detecting global cluster bias. In addition to these methods, the Wald test is described as a method for detecting item cluster bias.

3.1 Global Cluster Bias

Two models that we compared to detect global cluster bias are as follows: Model 1 (the invariance model), for all items, discrimination parameters are the same at the within-level and the between-level; and Model 2 (the global bias model), for all items, the discrimination parameters are freely estimated at the within-level and between-level.

Because the two models are nested, the LRT can be conducted. In the LRT, the approximately chi-square-distributed test statistics is -2 times the difference between the log likelihoods from the two models, with degrees of freedom equal to the difference in the number of free parameters (i.e., number of items + 1[variance at Level 3]).

The two models are also compared with model information criteria, the AIC (Akaike, 1974), the BIC (Schwarz, 1978), and the sample-size adjusted BIC (saBIC; Sclove, 1987), specified for a Model *m* as follows:

$$\text{AIC}_m = -2 \cdot LL + 2p, \tag{3.1}$$

$$\text{BIC}_m = -2 \cdot LL + p \cdot ln(J), \tag{3.2}$$

and

$$\text{saBIC}_m = -2 \cdot LL + p \cdot ln(\frac{J+2}{24}), \tag{3.3}$$

where *LL* is the log-likelihood of the estimated model, *p* is the number of estimated parameters, and J is the number of observations. In calculating the BIC, it is difficult to define the sample size (Skrondal & Rabe-Hesketh, 2004). In multilevel IRT applications, the number of persons, J, has been used for this purpose (e.g., Bartolucci, Pennoni, & Vittadini, 2011; Cho & Cohen, 2010; May, 2006). Thus, in the current study, J was chosen for the calculation of the BIC. The lowest AIC or (sa)BIC value is taken to indicate the best fitting model. See Cohen and Cho (2015) and Vrieze (2012) for reviews on using model information criteria in item response modeling and latent variable modeling.

3.2 Item Cluster Bias

The following two models can be compared to detect item cluster bias, based on the LRT, AIC, BIC, and saBIC specified in Equations 3.1-3.3: Model 1 (invariance model), for all items, the discrimination parameters are the same at the within-level and the between-level; and Model 2 (item bias model), for one item to be studied, the discrimination parameters are different at the within-level and the between-level. In the item bias model, the variance of the Level 3 latent variable can be estimated because there are anchor items over the levels. Thus, there is one degree of freedom for the LRT.

To detect item cluster bias, a Wald test for each item can be implemented. For an item i, $z = \frac{(\widehat{\alpha}_{iW} - \widehat{\alpha}_{iB}) - 0}{SE_{(\widehat{\alpha}_{iW} - \widehat{\alpha}_{iB})}}$ can be used to test whether $H_0 : a_i = \alpha_{iW} - \alpha_{iB} = 0$ can be rejected at the 0.05 level. A two-tailed test was implemented because item discrimination parameters can range from positive infinity to negative infinity (Baker & Kim, 2004), and thus, α_i ranges from positive infinity to negative infinity. When an item is tested using the Wald test, other items are assumed to be anchor items for scale comparability over levels. Thus, the variance of the Level 3 latent variable can be estimated in the detection of item cluster bias. It is expected that the performances of the Wald test and the LRT are similar because the Wald test is asymptotically equivalent to the LRT (Engle, 1984).

Chapter 4

Empirical Study

4.1 Data

To illustrate global and item cluster bias detection in the use of a two-parameter multilevel item response model, we chose a data set collected by Doolaard (1999) and previously analyzed by Fox and Glas (2001) and Vermunt (2007) using the two-parameter multilevel item response models. The data are from an 18-item math test taken by 2,156 students in 97 schools in the Netherlands. The average cluster size (i.e., the number of students for each school) was 22.22 (standard deviation = 10.31, range=[10, 66]).

4.2 Analysis

The data set was analyzed twice, once with the cluster invariance model and again with the cluster bias model. Vermunt (2007) showed that a two-parameter item response model with the same item discriminations over levels fits the same data better than Rasch model. Fox and Glas (2001) analyzed the same data using a two-parameter two-level item response model with unidimensionality at each level (assuming cluster invariance). Thus, we assume that configural invariance holds across two levels (i.e., students at Level 2 and schools at Level 3) to illustrate cluster bias detection methods and to compare results between the models with cluster invariance and with cluster bias.

4.3 Results

Table 4.1 presents the results of global and item cluster bias detection. There was evidence of global cluster bias based on the LRT (chi-square value=95.29, df=17, p-value=0.000), AIC, and saBIC. However, the BIC suggested evidence of global cluster

Table 4.1: An Empirical Study: Results of Cluster Bias Detection

	Num.	*LL*	Wald(SE)	LRT(df)	AIC	BIC	saBIC
Cluster Invariance	37	-20071.795	-	-	40217.589	40427.602*	40310.048
Cluster Bias	54	-20024.150	-	95.290(17)	40156.300*	40462.805	40291.240*
Item 1	38	-20065.873	0.541*(0.146)	11.844*(1)	40207.745*	40423.434*	40302.703*
Item 2	38	-20067.790	0.534*(0.132)	8.010*(1)	40211.580*	40427.269*	40306.538*
Item 3	38	-20071.701	-0.058(0.148)	0.188(1)	40219.403	40435.091	40314.360
Item 4	38	-20070.832	0.211(0.162)	1.926(1)	40217.664	40433.352	40312.621
Item 5	38	-20071.156	-0.141(0.146)	1.278(1)	40218.311	40434.000	40313.269
Item 6	38	-20067.765	-0.345*(0.140)	8.060*(1)	40211.531*	40427.219*	40306.488*
Item 7	38	-20071.685	-0.060(0.132)	0.220(1)	40219.371	40435.059	40314.328
Item 8	38	-20070.772	-0.252(0.190)	2.046(1)	40217.543*	40433.232	40312.501
Item 9	38	-20071.427	-0.108(0.139)	0.736(1)	40218.854	40434.542	40313.811
Item 10	38	-20070.918	0.194(0.152)	1.754(1)	40217.836	40433.525	40312.794
Item 11	38	-20071.669	0.063(0.213)	0.252(1)	40219.338	40435.026	40314.295
Item 12	38	-20039.835	-1.006*(0.136)	63.920*(1)	40155.669*	40371.358*	40250.627*
Item 13	38	-20071.642	-0.115(0.306)	0.306(1)	40219.284	40434.973	40314.242
Item 14	38	-20071.781	0.022(0.165)	0.028(1)	40219.563	40435.251	40314.520
Item 15	38	-20069.259	0.324(0.172)	5.072*(1)	40214.517*	40430.206	40309.475*
Item 16	38	-20070.947	0.172(0.145)	1.696(1)	40217.894	40433.583	40312.852
Item 17	38	-20071.282	-0.219(0.222)	1.026(1)	40218.564	40434.252	40313.521
Item 18	38	-20071.780	-0.019(0.175)	0.030(1)	40219.560	40435.248	40314.517

Note. * indicates cluster bias.

invariance. For cluster bias at the item level, Items 1, 2, 6, and 12 were detected for item cluster bias based on all criteria. Item 15 was detected as an item having cluster bias based on the LRT, AIC, and saBIC, and Item 8 was detected as an item having cluster bias based only on the AIC.

Table 4.2 shows the item parameter estimates and the ICC for the models with and without cluster bias. The item locations from the two models were similar (correlation=0.999). With equality constraints over levels on item discriminations in the model with cluster invariance, the item discrimination estimates are considered between-level item discriminations. Compared to between-level item discriminations in the cluster bias model, they were overestimated in the invariance model. The SEs for the item discrimination estimates in the cluster invariance model were smaller than those of the cluster bias model. In the invariance model, the variance estimate of Level 3 was 0.351, which was an expected result because ICC_i in the invariance model was larger than that in the bias model.

Figure 4.1 presents the IRT scale score comparisons between the model with cluster bias and the model with cluster invariance. At Level 2 (the student level), it appears that

Table 4.2: An Empirical Study: Results of Item Parameter Estimates from Models with Cluster Bias and Cluster Invariance

Item	Cluster Bias				Cluster Invariance		
	$\alpha_{i,W}$	$\alpha_{i,B}$	$\beta_{i,B}$	ICC	α_i	$\beta_{i,B}$	ICC
Item 1	1.161(0.121)	0.945(0.138)	-0.509(0.176)	0.399	1.283(0.092)	-0.539(0.134)	0.5
Item 2	1.220(0.145)	0.920(0.176)	-1.408(0.172)	0.363	1.318(0.099)	-1.435(0.143)	0.5
Item 3	1.133(0.125)	0.655(0.124)	-0.186(0.122)	0.250	1.120(0.077)	-0.183(0.119)	0.5
Item 4	1.121(0.146)	0.812(0.191)	-1.609(0.168)	0.344	1.195(0.115)	-1.630(0.142)	0.5
Item 5	0.731(0.111)	0.388(0.124)	-0.891(0.101)	0.220	0.701(0.084)	-0.883(0.094)	0.5
Item 6	0.874(0.105)	0.368(0.076)	0.051(0.090)	0.151	0.791(0.067)	0.066(0.099)	0.5
Item 7	1.119(0.116)	0.568(0.118)	-0.734(0.132)	0.205	1.069(0.091)	-0.721(0.121)	0.5
Item 8	1.162(0.149)	0.528(0.145)	-2.333(0.164)	0.171	1.066(0.095)	-2.292(0.144)	0.5
Item 9	0.979(0.102)	0.476(0.098)	-0.459(0.096)	0.191	0.927(0.071)	-0.446(0.097)	0.5
Item 10	0.814(0.126)	0.596(0.111)	-1.274(0.137)	0.349	0.874(0.082)	-1.290(0.111)	0.5
Item 11	0.981(0.140)	0.585(0.152)	-1.389(0.156)	0.262	0.978(0.086)	-1.388(0.119)	0.5
Item 12	0.968(0.097)	0.013(0.078)	-0.045(0.067)	0.000	0.650(0.054)	0.009(0.094)	0.5
Item 13	1.136(0.242)	0.607(0.189)	-2.934(0.224)	0.222	1.104(0.160)	-2.926(0.180)	0.5
Item 14	1.595(0.142)	0.935(0.219)	-1.835(0.194)	0.256	1.582(0.118)	-1.830(0.185)	0.5
Item 15	1.075(0.152)	0.732(0.144)	-1.151(0.139)	0.317	1.126(0.113)	-1.166(0.115)	0.5
Item 16	0.756(0.089)	0.472(0.102)	-0.869(0.096)	0.280	0.767(0.073)	-0.872(0.084)	0.5
Item 17	1.334(0.185)	0.725(0.134)	-2.563(0.169)	0.228	1.295(0.113)	-2.549(0.165)	0.5
Item 18	1.057(0.142)	0.604(0.122)	-0.463(0.114)	0.246	1.039(0.077)	-0.459(0.109)	0.5

Note. Estimates are not on the same scale.

the scores from the two models were similar (correlation=0.999). At Level 3 (the school level), the scores for the model with invariance were higher than the scores for the model with cluster bias at the lower level of the IRT scale scores. However, the reverse pattern was found at the higher level of the IRT scale scores. Figure 4.2 reports the SE comparisons between the model with cluster bias and the model with cluster invariance. The SEs for the model with cluster invariance were larger than those for the model with cluster bias at Level 2, whereas the SEs for the model with cluster invariance were smaller than for the model with cluster bias at Level 3. The smaller SEs in the model with cluster invariance may be due to overestimated item discriminations at Level 3 with equality constraints.

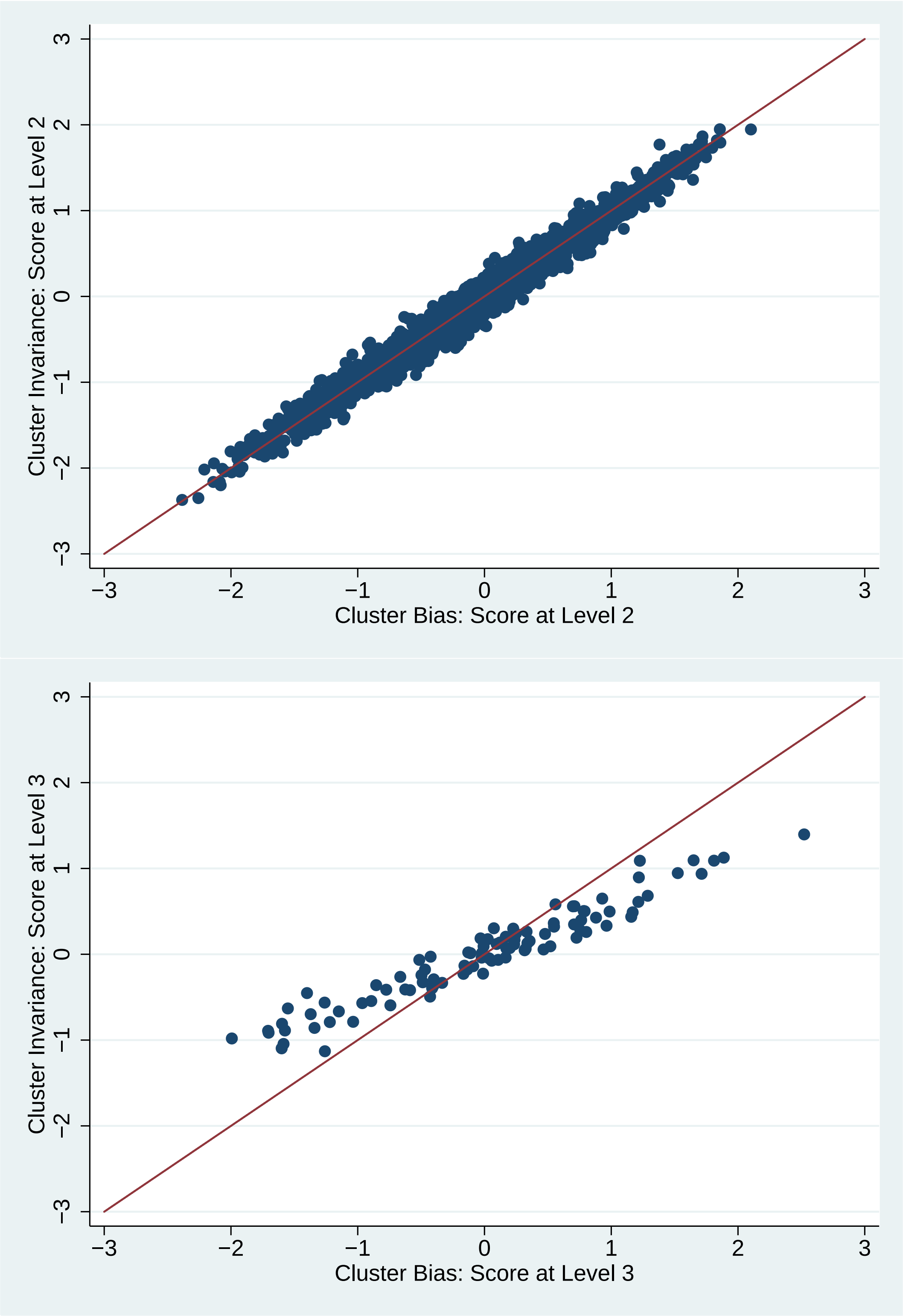

Figure 4.1: Score comparisons between cluster bias (x-axis) and cluster invariance (y-axis) at Level 2 (top) and at Level 3 (bottom)

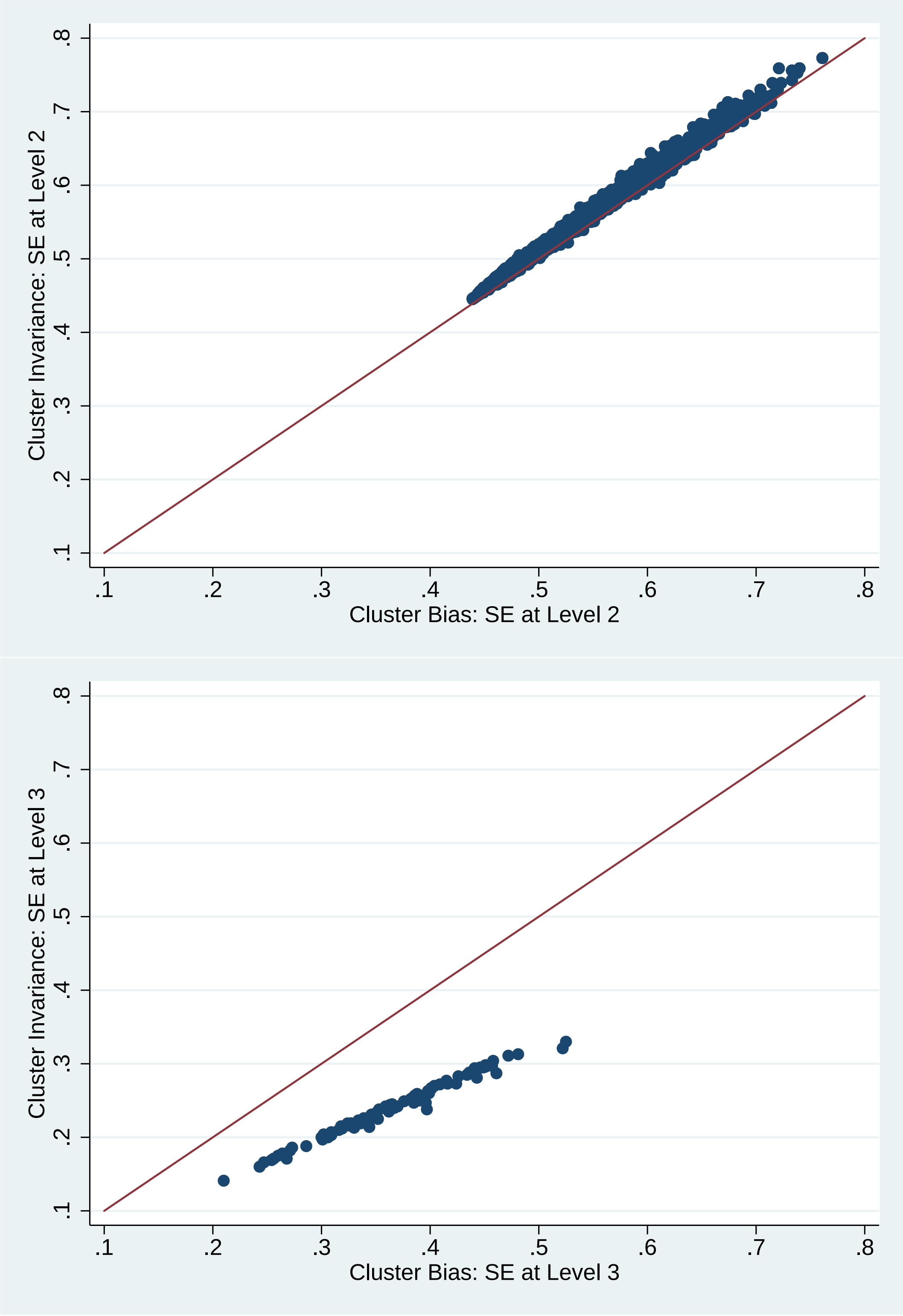

Figure 4.2: Standard error (SE) comparisons between cluster bias (x-axis) and cluster invariance (y-axis) at Level 2 (top) and at Level 3 (bottom)

Chapter 5

Simulation Study

Cluster bias and cluster invariance conditions were generated to evaluate detection methods for Type I error rates and power, and to present the consequences of ignoring cluster bias. For cluster bias conditions, the population data-generating model is a two-parameter multilevel item response model with cluster bias (Equation 2.2). An R program (R Core Team, 2015) was used to generate data sets.

We selected simulation conditions that may affect the parameters of multilevel modeling from previous research (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Preacher, Zhang, & Zyphur, 2011). The simulation conditions include the number of clusters, cluster size (i.e., the number of individuals per cluster), and the ICC. Further, simulation conditions related to cluster bias such as the magnitudes of cluster bias and the number of items that have cluster bias were also considered (e.g., Patarapichayatham & Kamata, 2014). As shown in Equation 2.5, the degree of the ICC reflects the cluster bias magnitudes. In this study, the ICC was chosen as a simulation condition instead of cluster bias magnitudes. In the following, we explain these selected conditions in more detail.

5.1 Simulation Condition

Number of clusters.

The number of clusters was set to K = 50 or 100. The smaller size of 50 was chosen in accordance with what is found in many intervention studies (e.g., Bottge et al., 2015). The larger size of 100 represents the magnitude found in national or international assessments (e.g., National Assessment of Educational Progress and the Trends in International Mathematics and Science Study).

Cluster size.

Balanced cluster sizes were selected as n_k = 20 or 50; both are commonly found in multilevel studies (e.g., Preacher et al., 2011).

Given a selected number of clusters and cluster size, the total number of individuals results in four different sample sizes, J = 1,000, 2,000, 2,500, or 5,000.

ICC (cluster bias magnitude).

The ICC was set at ICC = .05, .10, or .30. ICC values are rarely greater than .30 in educational and organizational studies (e.g., Bliese, 2002). As considered in Preacher et al. (2011), values of .05, .10, and .30 represent small, medium, and large ICCs, respectively.

Number of items that have cluster bias.

Twenty percent, 50%, and 100% of the items (4 items, 10 items, and 20 items, respectively) were considered as the number of items that have cluster bias. The first 16 items and 10 items for the 20% and 50% conditions, respectively, were set to have cluster invariance.

As a fixed condition, the number of item parameters was set at 20. Item parameters for the model with cluster invariance (α_i and β_i) were generated. Cluster bias was introduced to item discriminations at Level 3 ($\alpha_{i,B}$) for the model with cluster bias and was manipulated by the ICC (using Equation 2.5). Item location (β_i) was generated from a standard normal distribution, and item discrimination ($\alpha_{i,W}$) was generated from a log-normal distribution with a mean of 0 and a variance of .25 used as a prior distribution in the BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996). Latent variables (θ_{jk} and θ_k) were generated with a standard normal distribution to adhere to the model identification constraints.

For cluster bias, 500 replications were simulated for each of the 36 conditions (= 2 levels for the number of clusters × 2 levels for cluster sizes × 3 degrees of ICC × 3 levels for the number of items that have cluster bias) to show the performance of the detection methods. To show the consequences of ignoring cluster bias and the model comparison approaches, each generated data set per condition was analyzed twice, once with the model with cluster invariance and with the model with cluster bias. Thus, the total number of runs was 36,000 (= 36 conditions × 500 replications × 2 models). For detecting item cluster bias, the models with item cluster bias were compared with the invariance model. The number of item cluster bias models was the same as the number of items that have cluster bias: 4, 10, and 20 models for the 20%, 50%, and 100% bias conditions, respectively. The total number of runs was 24,000 for the 20% bias condition, 60,000 for the 50% bias condition, and 120,000 for the 100% bias condition (= 12 conditions × 500 replications × each number of models). Four additional conditions were conducted to show Type I error rates of detection methods at different levels of sample sizes (2 levels of cluster size and 2 levels of the number of clusters). The invariance model and the global bias model were compared in these conditions. Therefore, an additional 4,000 (= 4 conditions × 500 replications × 2 models) were considered for Type 1 error rates in the case of global bias.

5.2 Evaluation Measure

The performance of the detection methods was evaluated for Type I error rates and power. Type I error rates were defined as the ratio of the number of times that global cluster bias was incorrectly identified by the detection methods across replications when no bias was simulated. Power was defined as the proportion of the number of times that global and item cluster bias was correctly identified by the detection methods across replications when cluster bias was simulated.

To show the consequences of ignoring cluster bias, bias and the root mean square error (RMSE) for the item parameter estimates, and the IRT scale scores were calculated for the results of the model with cluster invariance. Bias for the item location estimates and IRT scale scores were considered instead of relative bias because the relative bias can be misleading when the true parameters to be used in the denominator are close to 0. However, for within- and between-item discriminations that are not near 0 values in the denominator, percentage relative bias was considered instead of bias to present the acceptable degree with an empirical cutoff (i.e., 10%).

To evaluate the SEs for each kind of parameter estimate, the relative difference of the estimated SE (Hoogland & Boomsma, 1998) for each model was calculated and compared between the cluster invariance model and the global bias model:

$$\text{SE} = \frac{\widehat{\text{SE}} - \sigma}{\sigma}, \tag{5.1}$$

where $\widehat{\text{SE}}$ is the average SE across replications and σ is the standard deviation of the parameter estimates across the replications.

5.3 Result Hypotheses

In this section, the expected results are described for the detection methods and the accuracy of the parameter estimates and SEs.

Model selection.

The LRT and the information criteria are used to detect global cluster bias. Deviance becomes smaller as a model has more free parameters to estimate. Therefore, the deviance of a global cluster bias model is always smaller than that of an invariance model. The difference in deviance between the two models approximately follows the chi-square distribution with the difference in the number of free parameters as the degrees of freedom. If the global cluster bias model has a better fit than the invariance model, the LRT result will be significant. The result of the LRT becomes significant when the difference in deviances is large and the degrees of freedom are small. The difference in the number of parameters is smaller in conditions with a small number of items that have cluster bias. Therefore, the LRT is more likely to show significant results in conditions with a small number of items that have cluster bias, controlling for deviance. The difference in the cluster bias magnitude between the global cluster bias model and the invariance model is greater in low ICC conditions, which results in a large difference between the deviances. Accordingly, the LRT is more likely to show significant results in low ICC conditions.

The AIC tends to select the more complicated model (the global cluster bias model) when the difference in the number of parameters is smaller between the two models (e.g., Burnham & Anderson, 2002). Therefore, the AIC is expected to select the global cluster bias model compared to the invariance model in conditions with a small number of items that have cluster bias when the deviance of the two models is the same. In addition, when only the ICC differs between the two models, the deviance can differ between the two models even though the number of parameters is the same. Therefore, the AIC is expected to select the global cluster bias model in low ICC conditions, controlling for other factors. Further, as the sample size becomes larger, the deviance also becomes larger, controlling for the number of parameters. Because the AIC does not take into account the sample size, the AIC is expected to select the global cluster bias model in larger sample size conditions, controlling for other factors.

The BIC penalizes more than the AIC when the sample size becomes larger than 8 ($ln(8) = 2.08$). The sample size is always larger than 8 in the simulation conditions. In addition, a more drastic increase in log-likelihood is required before a complex model (the global cluster bias model) is chosen over a simple model (the invariance model). Therefore, the BIC is expected to select the global cluster bias model less often than the AIC. The BIC also takes into account the number of parameters in the penalty term; thus the expected pattern is similar to AIC: the global cluster bias model will be selected when there is a small number of items having cluster bias and low ICC conditions, controlling for other factors. In spite of the penalty term for the sample size, previous research has shown that the accuracy of the information criteria improves when the sample size is large (e.g., Lin & Dayton, 1997).

The Wald test is used to detect item cluster bias. The result of the Wald test is significant when the cluster bias magnitude is large and the SE is small. Therefore, it is likely that the power will be higher in low ICC conditions (the cluster bias magnitude is large) and large sample sizes conditions (the SE is small). Information criteria are also used to detect item cluster bias. The general pattern of results by condition is expected to mimic that presented for global cluster bias. That is, higher power is expected in the low ICC and small number of items having cluster item bias conditions.

Consequences of ignoring cluster bias.

The invariance model has the same item discrimination parameter estimates over levels even when the true parameters are different. This equality constraint makes item discrimination parameter estimates in the invariance model more biased than those in the global cluster bias model. Because between-level item discriminations are smaller than within-level item discriminations (in our simulation design and expected in the empirical study), it is expected that the between-level item discriminations are overestimated. Under the invariance model, bias is expected to be larger in low ICC conditions because the cluster bias magnitude is larger. In addition, in the invariance model, bias is expected to be larger in large cluster bias item conditions because more items exhibit cluster bias.

As shown in Equation 1, the probability of a correct response is expressed as the difference between the item location parameter and the IRT scale score weighted by the item discrimination parameter. The item location parameter estimates will not be biased if the mean of the weighted IRT scale scores is assumed to be the same as in the population. The distribution of the true IRT scale scores at each level follows the standard normal distribution. The mean of the weighted IRT scale scores is not affected by the item discrimination parameter because the mean of the true IRT scale scores is 0. The mean of the IRT scale scores for each model was assumed to be 0 for model identification in the invariance model and global cluster bias model. Because the mean of the IRT scale scores is assumed to be 0 and the mean of the true population is also 0, the item location parameter estimates are not biased in either model.

As the bias becomes larger, the RMSE also becomes larger when variance is controlled. As the bias for the item discrimination parameter is larger for the invariance model than for the global cluster bias model, the RMSE is also expected to be larger for the invariance model. For the invariance model, the RMSE is expected to be larger in conditions with a low ICC and a large number of cluster bias items because the estimates are more biased in those conditions. For the global cluster bias model, the RMSE is not influenced by those conditions because the estimates are not biased. As the sample size increases, the SE of the estimates decreases. Under both models, the RMSE is expected to be smaller in large sample size conditions, controlling for bias. Because the item location parameter estimates are not biased, the RMSE is influenced only by the sample size. The item location parameter estimates are expected to be smaller in large sample size conditions in both models.

When MMLE is used, IRT scale scores are predicted based on the item parameter estimates. Item parameter estimates are more biased and have larger RMSE in the invariance model than in the true model. As a result, the IRT scale scores are expected to be more biased and have larger RMSE in the invariance model than in the the global cluster bias model. Within each model, the bias is expected to be larger in conditions with a low ICC and a large number of cluster bias items because the bias of the item parameters is larger in those conditions, when the variability of the estimates is the same between the two models. The RMSE is expected to be larger in low ICC, large number of cluster bias items, and smaller sample size conditions because the RMSE of the item parameters is larger in those conditions.

The differences in the SE between the invariance and cluster bias models are expected to be stronger as the ICC and the number of items with cluster bias are larger.

5.4 Results

No convergence problem occurred during the estimation process. Below, we first show the results for the detection methods. Subsequently, the results for the consequences of ignoring cluster bias are presented in terms of bias and the RMSE for the item parameter estimates and the IRT scale score.

Results of Detection Methods

Type I error rates. Table 5.1 shows the Type I error rates for the detection methods. To investigate the Type I error rates of the detection methods, the invariance model was compared with the global bias model when there was no cluster bias.

The Type I error rate for the LRT and the AIC was mainly affected by cluster size, n_k. For the LRT, the Type I error rate was 0.040 in the level of $n_k = 50$, whereas it marginally exceeds in the level of $n_k = 20$. The AIC fell below the nominal significance level: 0.008 with $n_k = 20$ and 0.006 with $n_k = 50$. The Type I error rate was 0 across all conditions for the BIC and the saBIC.

Power. In this subsection, the power of the detection methods is presented for global cluster bias and item cluster bias, respectively.

Global cluster bias. Table 5.1 and Figure 5.1 (top) present the result of power based on the LRT and the information criteria. In the 20% and 50% bias conditions, all criteria yielded acceptable power for all conditions (>0.994) except two conditions for the BIC. The power for the two conditions based on the BIC was 0.832 and 0.606 in ICC$=0.3$, $K = 50$, $n_k = 20$ in the presence of 20% and 50% bias. In the 100% bias condition, all methods did not successfully detect the true model, the invariance model. Power ranged from 0.036 to 0.190 for the LRT, less than or equal to 0.028 for the AIC. Power for the saBIC was 0 for all conditions. The power of the LRT and the AIC decreased with increasing sample size and increasing ICC.

Item cluster bias. Table 5.2 and Figure 5.1 (bottom) present the power results for cluster bias at the item level. The power of the AIC was the highest across all conditions among the five detection methods. The Wald test, LRT, and saBIC showed similar power across all conditions. The BIC showed the lowest power. In the 20% and 50% bias conditions, the following patterns were evident. All methods but the BIC showed acceptable (>0.800) power in the ICC=0.05 and ICC=0.1 conditions. When the ICC=0.3, all methods failed to have acceptable power, except for the largest sample size (i.e., $K = 100$, $n_k = 50$). For

Table 5.1: Simulation Study: Type I error and Power for Global Cluster Bias

DIF%	ICC	K	n_k	LRT	AIC	BIC	saBIC
0	.50	50	20	**0.060**	**0.008**	**0.000**	**0.000**
0	.50	50	50	**0.040**	**0.006**	**0.000**	**0.000**
0	.50	100	20	**0.056**	**0.008**	**0.000**	**0.000**
0	.50	100	50	**0.040**	**0.006**	**0.000**	**0.000**
20	.05	50	20	1.000	1.000	1.000	1.000
20	.05	50	50	1.000	1.000	1.000	1.000
20	.05	100	20	1.000	1.000	1.000	1.000
20	.05	100	50	1.000	1.000	1.000	1.000
20	.1	50	20	1.000	1.000	1.000	1.000
20	.1	50	50	1.000	1.000	1.000	1.000
20	.1	100	20	1.000	1.000	1.000	1.000
20	.1	100	50	1.000	1.000	1.000	1.000
20	.3	50	20	1.000	1.000	0.832	0.994
20	.3	50	50	1.000	1.000	1.000	1.000
20	.3	100	20	1.000	1.000	0.998	1.000
20	.3	100	50	1.000	1.000	1.000	1.000
50	.05	50	20	1.000	1.000	1.000	1.000
50	.05	50	50	1.000	1.000	1.000	1.000
50	.05	100	20	1.000	1.000	1.000	1.000
50	.05	100	50	0.998	1.000	1.000	1.000
50	.1	50	20	1.000	1.000	1.000	1.000
50	.1	50	50	1.000	1.000	1.000	1.000
50	.1	100	20	1.000	1.000	1.000	1.000
50	.1	100	50	1.000	1.000	1.000	1.000
50	.3	50	20	1.000	1.000	0.606	0.996
50	.3	50	50	1.000	1.000	1.000	1.000
50	.3	100	20	1.000	1.000	0.996	1.000
50	.3	100	50	1.000	1.000	1.000	1.000
100	.05	50	20	0.190	0.028	0.000	0.000
100	.05	50	50	0.056	0.006	0.000	0.000
100	.05	100	20	0.098	0.014	0.000	0.000
100	.05	100	50	0.060	0.004	0.000	0.000
100	.1	50	20	0.086	0.008	0.000	0.000
100	.1	50	50	0.068	0.002	0.000	0.000
100	.1	100	20	0.062	0.004	0.000	0.000
100	.1	100	50	0.036	0.006	0.000	0.000
100	.3	50	20	0.068	0.006	0.000	0.000
100	.3	50	50	0.038	0.000	0.000	0.000
100	.3	100	20	0.052	0.004	0.000	0.000
100	.3	100	50	0.050	0.006	0.000	0.000

Note. Type I error rates in bold

Table 5.2: Simulation Study: Power for Item Cluster Bias

	bias%	ICC	K	n_k	Wald	LRT	AIC	BIC	saBIC
	20	.05	50	20	0.971	0.971	0.992	0.915	0.971
	20	.05	50	50	1.000	1.000	1.000	0.999	1.000
	20	.05	100	20	0.997	0.996	1.000	0.983	0.995
	20	.05	100	50	1.000	1.000	1.000	1.000	1.000
	20	.1	50	20	0.937	0.935	0.976	0.830	0.938
	20	.1	50	50	0.998	0.999	1.000	0.983	0.997
	20	.1	100	20	0.991	0.991	0.998	0.956	0.989
	20	.1	100	50	1.000	1.000	1.000	1.000	1.000
	20	.3	50	20	0.597	0.587	0.738	0.433	0.595
	20	.3	50	50	0.834	0.825	0.920	0.628	0.783
	20	.3	100	20	0.747	0.752	0.860	0.541	0.710
	20	.3	100	50	0.950	0.951	0.982	0.828	0.918
	50	.05	50	20	0.878	0.874	0.941	0.767	0.878
	50	.05	50	50	0.978	0.975	0.991	0.923	0.967
	50	.05	100	20	0.955	0.953	0.981	0.885	0.944
	50	.05	100	50	0.998	0.997	1.000	0.986	0.997
	50	.1	50	20	0.804	0.803	0.886	0.680	0.807
	50	.1	50	50	0.943	0.942	0.974	0.856	0.927
	50	.1	100	20	0.915	0.915	0.958	0.811	0.899
	50	.1	100	50	0.994	0.994	0.999	0.965	0.989
	50	.3	50	20	0.473	0.467	0.627	0.305	0.475
	50	.3	50	50	0.683	0.667	0.786	0.485	0.622
	50	.3	100	20	0.605	0.607	0.741	0.428	0.567
	50	.3	100	50	0.901	0.824	0.899	0.649	0.764
	100	.05	50	20	0.068	0.051	0.152	0.009	0.054
	100	.05	50	50	0.072	0.047	0.152	0.004	0.029
	100	.05	100	20	0.062	0.050	0.156	0.005	0.036
	100	.05	100	50	0.060	0.048	0.149	0.003	0.020
	100	.1	50	20	0.064	0.050	0.157	0.008	0.053
	100	.1	50	50	0.074	0.048	0.160	0.005	0.030
	100	.1	100	20	0.063	0.054	0.159	0.005	0.038
	100	.1	100	50	0.060	0.051	0.159	0.004	0.021
	100	.3	50	20	0.062	0.050	0.150	0.009	0.053
	100	.3	50	50	0.064	0.046	0.152	0.004	0.028
	100	.3	100	20	0.052	0.046	0.152	0.005	0.032
	100	.3	100	50	0.059	0.052	0.159	0.004	0.022
Average									
bias%									
20%					0.918	0.917	0.955	0.841	0.908
50%					0.844	0.835	0.899	0.728	0.820
100%					0.063	0.049	0.155	0.005	0.035
ICC									
0.05					0.618	0.612	0.655	0.575	0.607
0.1					0.603	0.598	0.648	0.546	0.591
0.3					0.502	0.489	0.597	0.360	0.464
K									
50					0.552	0.544	0.619	0.465	0.537
100					0.600	0.594	0.650	0.592	0.576
n_k									
20					0.569	0.564	0.646	0.476	0.557
50					0.648	0.637	0.693	0.574	0.617

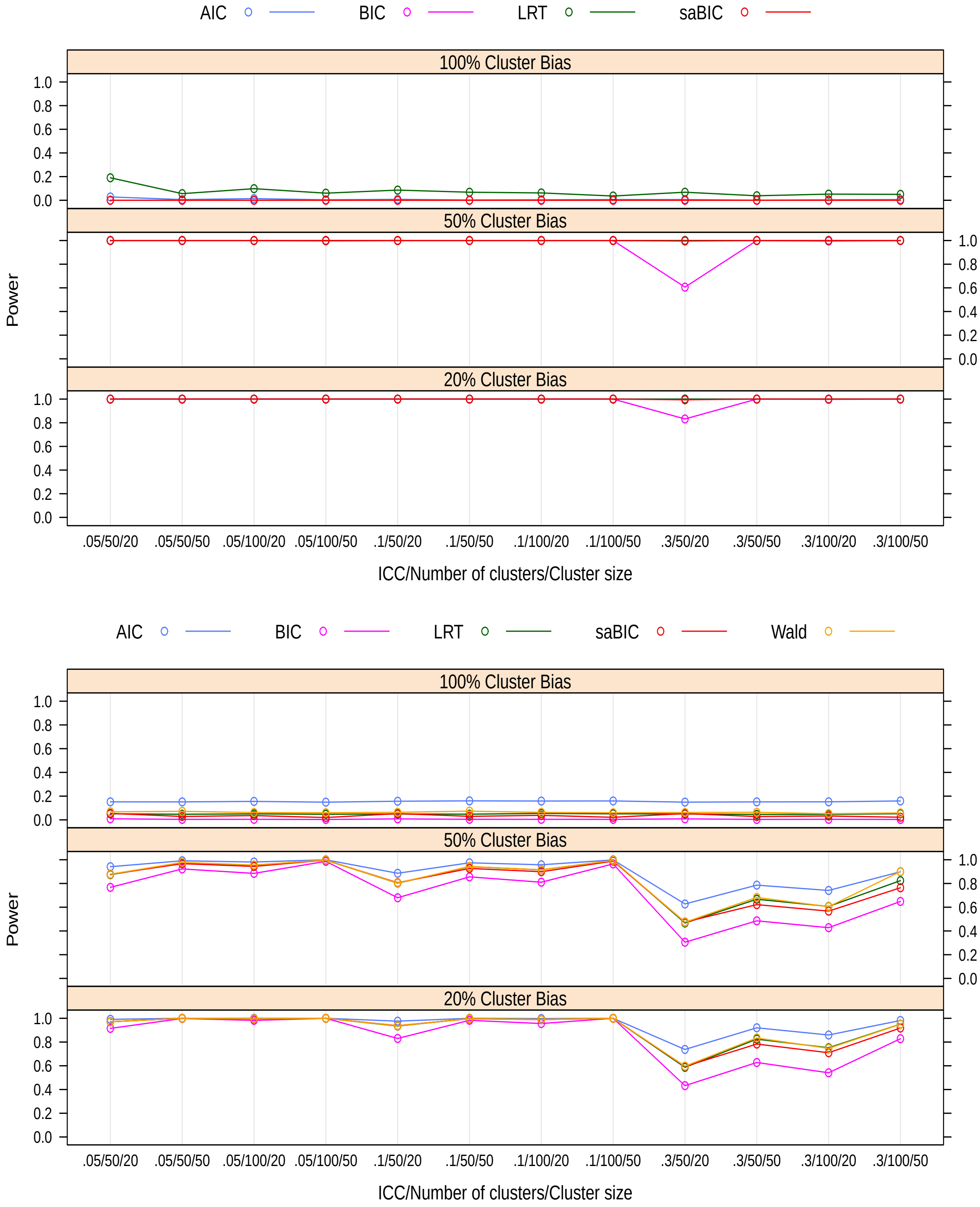

Figure 5.1: power: Global cluster bias (top) and item cluster bias (bottom)

the largest sample size, the AIC, LRT, and Wald test showed power over 0.800. For all criteria, power decreased with the increasing number of biased items and ICC, whereas power increased with the increasing number of clusters and cluster sizes. In the 100% bias condition, all methods showed low power, and power was not largely affected by any conditions of the number of clusters, cluster size, or ICC. The power for the saBIC, Wald test, LRT, and BIC was less than or equal to 0.074. The power for the AIC was less than or equal to 0.160.

Results for Consequences of Ignoring Cluster Bias

In this section, the accuracy of the item parameter estimates and the IRT scale scores and that of the SEs are presented for the invariance model and the global bias model in the presence of cluster bias. Accordingly, the results of the invariance model are for the consequence of ignoring cluster bias. Results of the global bias model are reported for comparison purposes.

Item parameter estimates.

Bias. The bias for the item discrimination parameters is shown in Table 5.3 and Figure 5.2. Relative bias was reported for item discrimination estimates, and absolute bias was reported for item location estimates. Regarding within-level item discrimination parameters, the relative bias in all conditions was less than 10% for the invariance model and the global bias model. The performance of the global bias model was superior to that of the invariance model in the 20% and 50% conditions except the 20% bias and ICC=0.3 conditions. For those conditions, the invariance model showed smaller bias than the global bias model because the degree of bias was ignorable. These results indicate that ignoring cluster bias led to inaccurate results in within-level item discrimination estimates unless the number of cluster bias items was small and the ICC was high (the degree of bias was inversely related to ICC). When cluster bias was ignored as in the invariance model, the bias was negative (i.e., the within-level item discrimination parameters were underestimated) and increased with number of bias items in 20% and 50% conditions.

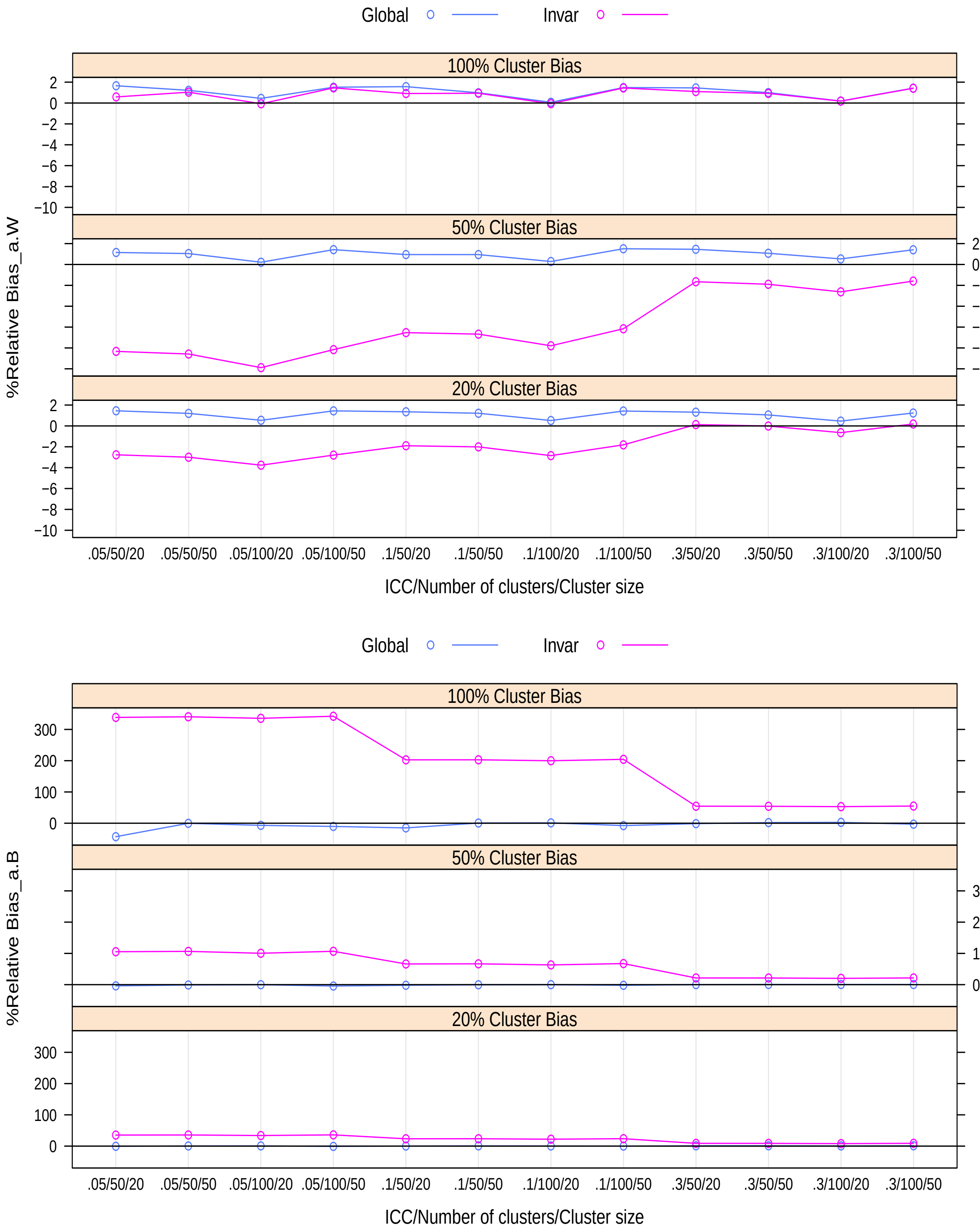

Figure 5.2: Accuracy: Percentage relative bias for within-level item discrimination parameters ($\alpha_{i,W}$) (top) and between-level item discrimination parameters ($\alpha_{i,B}$) (bottom)

However, in the 100% condition, the bias in the invariance model was unexpectedly smaller than the bias in the global bias model. We further investigated the item discrimination parameter estimates in the invariance model to investigate these unexpected results. When cluster invariance items exist over levels as in the 20% and 50% conditions, the within-level item discrimination estimates were close to an average of the within- and between-item discrimination parameters in the invariance model. However, when all item discriminations differed between levels as in the 100% condition, the within-level item discrimination estimates were close to the within-level item discrimination parameters instead of the average of the within- and between-level item discrimination estimates.

Regarding the between-level item discrimination parameters, the bias was dramatically higher for the invariance models than for the global bias model across all conditions, which suggested that ignoring cluster bias would be problematic in terms of the accuracy of the between-level item discrimination estimates. As expected, the between-level item discrimination parameters were overestimated in ignoring cluster bias. For the global bias models, the relative bias was acceptable ($<$10%) in most conditions, except the condition when the number of cluster bias items was 100%, ICC was low, and the sample size was small. The invariance model showed acceptably small bias only with 20% cluster bias and 0.3 ICC conditions. The bias was positive (i.e., the between-level item discrimination parameters were overestimated) and decreased with a higher ICC and a decreasing number of cluster bias items.

Regarding the item location parameters, in most conditions the bias was small regardless of the model. The degree of bias was less than or equal to .033 across all conditions.

RMSE. Concerning the within-level item discrimination RMSEs (see Figure 5.3 top), the global bias model outperformed the invariance model in the 20% and 50% bias conditions. The global bias model showed small RMSEs in most conditions (from 0.047 to 0.115), whereas the invariance model showed larger RMSEs for low ICC (i.e., a higher bias magnitude in item discriminations) and a larger number of cluster bias items. However, in the 100% bias condition, the RMSE in the invariance model was a bit higher than the RMSE in the global bias model. This unexpected pattern is from the fact that there was smaller bias in the invariance model than in the global bias model (as discussed above) and there was a smaller number of item discrimination parameters to be estimated in the invariance model than in the cluster bias model (i.e., smaller variability). The RMSE decreased with the increasing number of clusters and cluster size.

Regarding the between-level item discrimination parameters (see Figure 5.3 bottom), the global bias model showed better performance in all conditions, which implies that the result interpretation for the between-level item discrimination estimates can be misleading when ignoring cluster bias. Similar to the within-level item discrimination parameters, a higher ICC was positively associated with a smaller RMSE for the invariance model in the 20% and 50% conditions. In the 100% condition, the invariance model had a noticeably high RMSE when global bias was used because of the large degree of bias. In this condition, the RMSE was not affected by sample size.

Regarding RMSEs for item location estimates, the global bias model and the invariance model yielded comparable values in every condition.

Relative bias of SE. Regarding the SE for the within-level item discrimination parameter estimates, acceptable levels of bias for the SE were obtained in almost every condition for the global bias model, whereas unacceptable bias was found for the invariance model in low ICC (0.05 and 0.1) and large cluster size ($n_k = 50$) conditions. [1] Regarding the SE for the between-level item discrimination parameter estimates, the relative bias of the SE for the invariance model was identical to that for the within-level discrimination parameter because of the equality constraint. The global bias model showed acceptable relative bias in the 20% and 50% bias conditions, but unacceptable SEs were obtained in the 100%

[1] Increasing relative bias of SE in a large cluster size might be counter-intuitive. This result resulted in the characteristics of relative bias. The standard deviation of estimates decreased as the total number of individuals increased, whereas the average of estimate SE decreased both the number of individuals and the number of clusters increased. Thus, controlling for the number of individuals, the relative bias of SE was larger in large cluster size conditions.

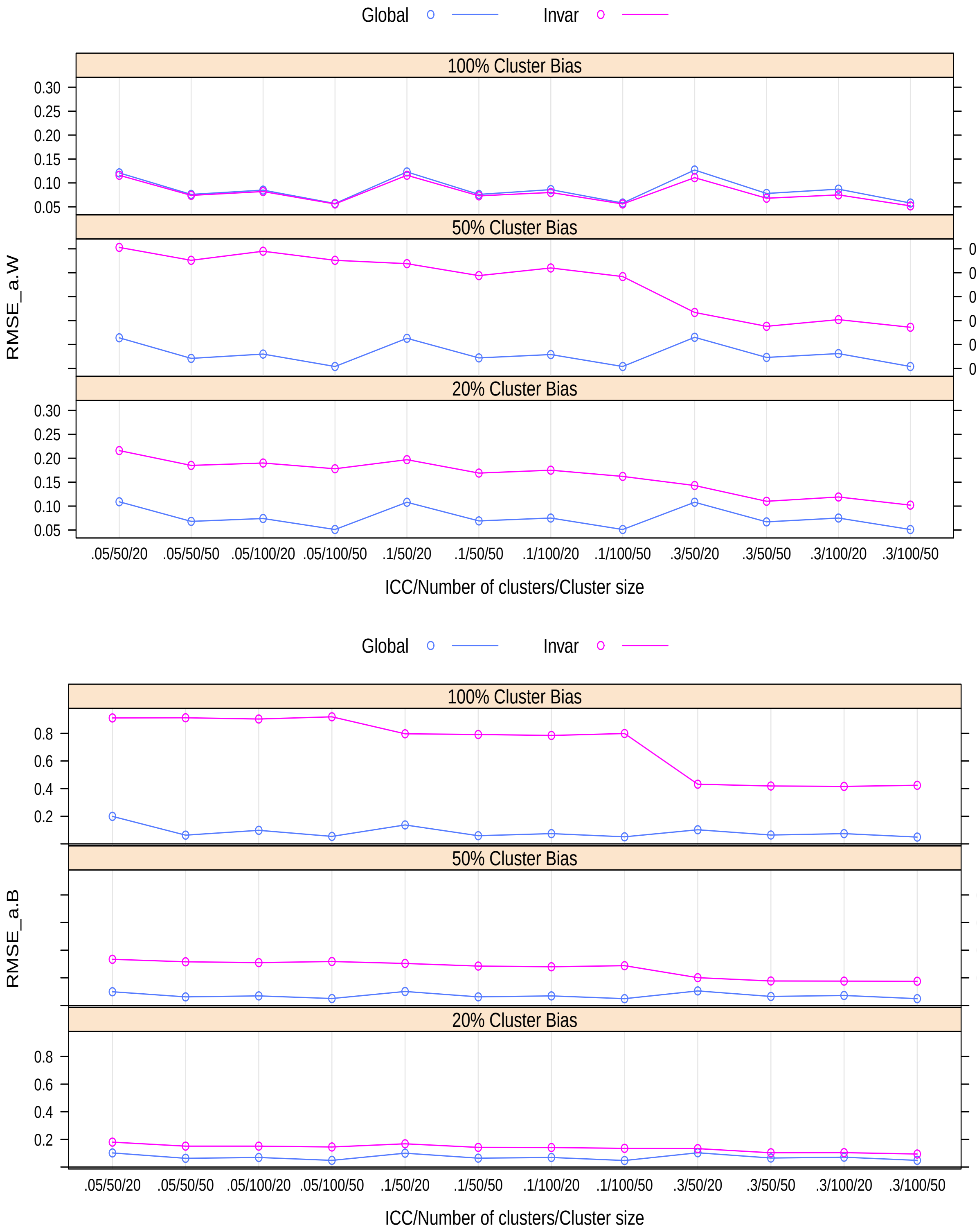

Figure 5.3: Accuracy: RMSE for within-level item discrimination parameters ($\alpha_{i,W}$) (top) and between-level item discrimination parameters ($\alpha_{i,B}$) (bottom)

bias condition. The SE was severely overestimated the most in the low ICC (0.05) and the smallest sample size ($n_k = 20$ and $K = 50$) condition. Concerning the item location parameter, both models yielded overestimated SEs in every condition. The relative bias increased with increasing sample sizes.

IRT scale score precision. When there was no cluster bias, the IRT scale score did not show any difference between the invariance and global bias models in terms of bias and RMSE. Below, the results are interpreted for cluster bias conditions (i.e., 20%, 50%, and 100%).

Bias. Table 5.4 presents the accuracy of IRT scale scores (also see Figure 5.4). The invariance model and the global bias model yielded comparable values in almost every condition: from -0.033 to 0.018 for the invariance model and from -0.033 to 0.043 for the global bias model. Bias was mainly affected by the total sample sizes (i.e., J=number of clusters × cluster size). For example, the average bias of the IRT scale scores at the within-level across conditions within the same total sample size was 0.022, -0.033, -0.021, and 0.001 in J=1000, 2500, 2000, and 5000, respectively, and the bias did not differ between the invariance and global bias models.

RMSE. Regarding the IRT scale scores at the within- and between-levels (see Figure 5.5), the RMSE for the invariance model was higher than that of the global bias model except the 100% bias condition. The RMSE decreased with increasing sample size (J) and higher ICC in both models. In the 100% bias condition, the RMSE ranged from .366 to .760 for the invariance model and from .237 to .967 for the global bias model. This extreme RMSE at the between-level resulted from different reasons in the invariance and global bias models. In the invariance model, the estimated variance of the scores at the between-level was largely underestimated. Thus, all IRT scores at the between-level shrank to the mean. In the global bias model, there was no problem with the variance of the latent variable at the between-level because it was fixed to 1 for model identification. Instead, the item discrimination parameter estimates were close to 0 as the ICC decreased (range=[0.089,0.528]

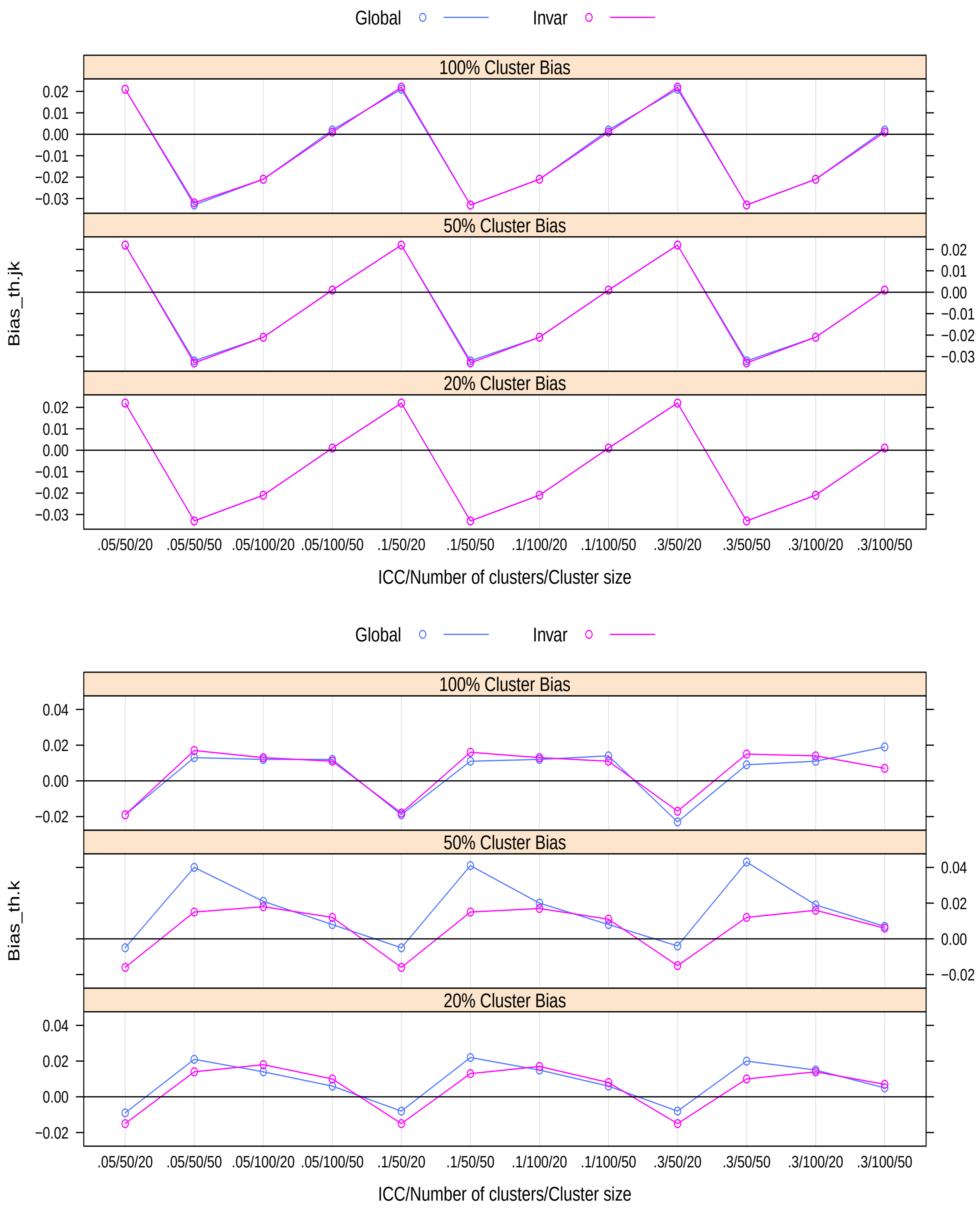

Figure 5.4: Accuracy: Bias for IRT scale scores at Level 2 (θ_{jk}) (top) and at Level 3 (θ_k) (bottom)

for ICC=0.05; range=[0.129,0.767] for ICC=0.1; range=[0.254,1.506] for ICC=0.3). It was because the true parameters at the between-level for items with a low ICC had values close 0 according to the formula we used for the ICC calculation as in Equation (5).

Relative bias of SE. Across all conditions, the estimated SE was less accurate for both the within- and between-levels in the invariance model than the global bias model. The accuracy of the SE estimation for the global bias model worsened as ICC increased, whereas that for the invariance model was not affected by the ICC. In the 100% bias condition, the invariance model revealed that the relative bias of the SE at the between-level decreased as ICC increased.

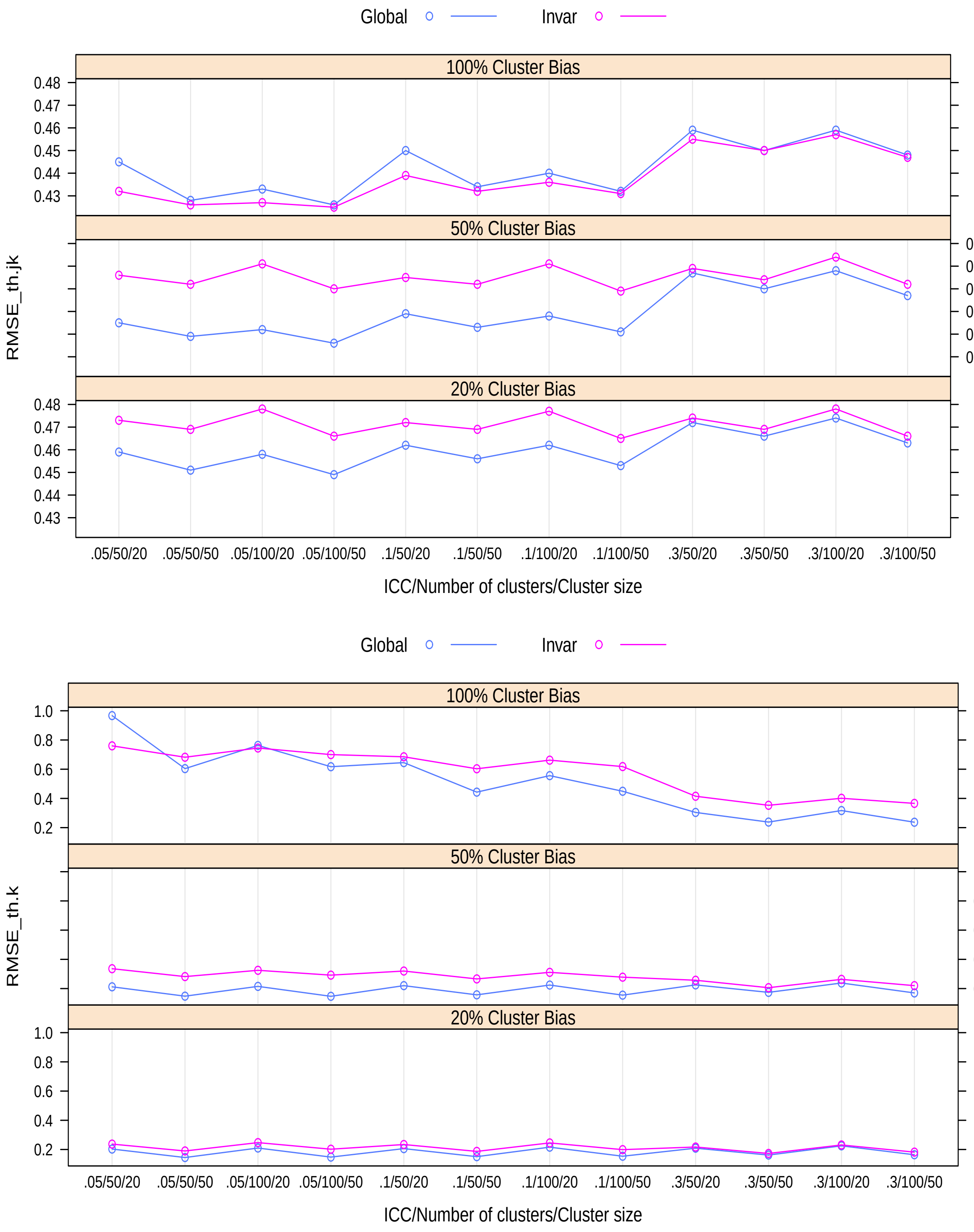

Figure 5.5: Accuracy: RMSE for IRT scale scores at Level 2 (θ_{jk}) (top) and at Level 3 (θ_k) (bottom)

Table 5.3: Simulation Study: Accuracy of Item Parameter Estimates

				Bias*						RMSE						SE					
				Invar.			Bias			Invar.			Bias			Invar.			Bias		
DIF%	ICC	K	n_k	α_W	α_B	β	α_W	α_B	β	α_W	α_B	β	α_W	α_B	β	α_W	α_B	β	α_W	α_B	β
0	.50	50	20	1.27	1.27	0.020	1.53	-1.93	0.012	0.103	0.103	0.093	0.131	0.114	0.090	0.040	0.040	1.143	0.023	0.457	1.187
0	.50	50	50	0.91	0.91	-0.030	1.12	-2.59	-0.025	0.065	0.065	0.066	0.081	0.076	0.062	0.028	0.028	2.128	0.014	0.991	2.224
0	.50	100	20	0.68	0.68	-0.009	0.79	0.54	-0.006	0.071	0.071	0.066	0.092	0.079	0.064	0.036	0.036	1.149	0.022	0.361	1.199
0	.50	100	50	1.34	1.34	0.014	1.36	-3.65	0.015	0.049	0.049	0.049	0.060	0.064	0.044	0.028	0.028	1.588	0.021	0.965	2.034
20	.05	50	20	-2.77	35.34	0.017	1.45	-0.68	0.025	0.216	0.180	0.092	0.109	0.102	0.094	0.073	0.073	0.966	0.023	0.019	0.966
20	.05	50	50	-3.00	35.66	-0.030	1.20	0.42	-0.020	0.185	0.151	0.068	0.068	0.063	0.060	0.138	0.138	1.786	0.026	0.035	1.814
20	.05	100	20	-3.76	33.84	-0.006	0.54	0.65	-0.008	0.190	0.151	0.067	0.074	0.069	0.064	0.074	0.074	0.997	0.032	0.036	0.993
20	.05	100	50	-2.80	35.88	0.017	1.44	-0.96	0.017	0.178	0.145	0.048	0.051	0.048	0.045	0.170	0.170	1.621	0.051	0.067	1.676
20	.1	50	20	-1.90	23.40	0.016	1.36	0.15	0.026	0.197	0.168	0.091	0.108	0.100	0.094	0.062	0.062	0.993	0.032	0.030	0.990
20	.1	50	50	-2.01	23.50	-0.032	1.21	0.67	-0.020	0.169	0.142	0.069	0.069	0.064	0.061	0.102	0.102	1.763	0.013	0.017	1.794
20	.1	100	20	-2.85	22.06	-0.005	0.52	0.44	-0.006	0.175	0.141	0.066	0.075	0.069	0.065	0.056	0.056	0.998	0.028	0.033	0.995
20	.1	100	50	-1.81	23.69	0.015	1.43	0.03	0.017	0.162	0.135	0.046	0.051	0.047	0.045	0.140	0.140	1.672	0.053	0.062	1.691
20	.3	50	20	0.13	8.79	0.017	1.32	0.85	0.026	0.143	0.133	0.092	0.108	0.103	0.095	0.040	0.040	1.045	0.034	0.026	1.043
20	.3	50	50	-0.01	8.68	-0.033	1.05	0.97	-0.021	0.110	0.103	0.068	0.067	0.065	0.062	0.044	0.044	1.923	0.032	0.021	1.936
20	.3	100	20	-0.64	7.90	-0.006	0.46	0.55	-0.005	0.119	0.104	0.065	0.075	0.071	0.064	0.047	0.047	1.072	0.041	0.039	1.075
20	.3	100	50	0.19	8.87	0.018	1.24	0.97	0.017	0.102	0.094	0.045	0.051	0.048	0.046	0.051	0.051	1.833	0.026	0.030	1.780
50	.05	50	20	-8.33	105.39	0.018	1.15	-3.89	0.028	0.303	0.334	0.090	0.114	0.099	0.094	0.126	0.126	0.743	0.028	0.026	0.731
50	.05	50	50	-8.58	106.46	-0.029	1.05	-1.09	-0.009	0.276	0.316	0.069	0.071	0.062	0.060	0.273	0.273	1.290	0.034	0.032	1.306
50	.05	100	20	-9.89	100.43	-0.003	0.21	-0.14	-0.004	0.295	0.310	0.066	0.080	0.069	0.062	0.102	0.102	0.806	0.017	0.020	0.787
50	.05	100	50	-8.16	106.75	0.016	1.42	-4.41	0.016	0.276	0.318	0.048	0.054	0.050	0.043	0.267	0.267	1.236	0.041	0.058	1.321
50	.1	50	20	-6.53	66.21	0.017	0.95	-2.02	0.028	0.269	0.304	0.090	0.113	0.101	0.094	0.093	0.093	0.771	0.039	0.018	0.766
50	.1	50	50	-6.67	66.66	-0.028	0.95	-0.39	-0.004	0.244	0.285	0.066	0.072	0.062	0.058	0.205	0.205	1.443	0.017	0.030	1.458
50	.1	100	20	-7.79	63.18	-0.005	0.28	0.05	-0.004	0.260	0.280	0.065	0.079	0.069	0.063	0.101	0.101	0.835	0.042	0.037	0.826
50	.1	100	50	-6.16	67.38	0.016	1.51	-2.15	0.017	0.242	0.288	0.046	0.054	0.049	0.044	0.203	0.203	1.276	0.031	0.048	1.376
50	.3	50	20	-1.65	21.41	0.019	1.46	0.12	0.031	0.167	0.201	0.091	0.115	0.105	0.096	0.057	0.057	0.942	0.037	0.031	0.939
50	.3	50	50	-1.90	21.29	-0.033	1.07	0.67	0.001	0.138	0.177	0.067	0.073	0.065	0.058	0.056	0.056	1.707	0.017	0.010	1.688
50	.3	100	20	-2.62	20.13	-0.005	0.53	0.54	-0.003	0.152	0.176	0.063	0.081	0.072	0.063	0.042	0.042	0.987	0.026	0.043	0.993
50	.3	100	50	-1.59	21.54	0.014	1.41	0.50	0.017	0.136	0.175	0.042	0.054	0.049	0.044	0.067	0.067	1.663	0.032	0.025	1.683
100	.05	50	20	0.58	338.44	0.020	1.66	-43.07	0.021	0.116	0.912	0.087	0.121	0.199	0.088	0.026	0.026	0.158	0.105	0.928	0.172
100	.05	50	50	1.04	340.42	-0.030	1.22	-0.34	-0.034	0.074	0.913	0.062	0.076	0.063	0.065	0.017	0.017	0.333	0.019	0.399	0.328
100	.05	100	20	-0.07	335.58	-0.016	0.45	-6.85	-0.018	0.082	0.904	0.062	0.085	0.098	0.063	0.024	0.024	0.200	0.067	0.535	0.201
100	.05	100	50	1.45	342.22	0.010	1.52	-10.28	0.010	0.056	0.920	0.039	0.057	0.054	0.039	0.026	0.026	0.306	0.029	0.307	0.304
100	.1	50	20	0.91	202.74	0.020	1.57	-15.01	0.019	0.116	0.797	0.088	0.123	0.137	0.088	0.023	0.023	0.249	0.061	0.496	0.250
100	.1	50	50	0.94	202.81	-0.029	0.98	0.59	-0.033	0.073	0.792	0.061	0.076	0.059	0.063	0.013	0.013	0.561	0.012	0.377	0.559
100	.1	100	20	-0.06	199.81	-0.015	0.07	1.15	-0.017	0.080	0.785	0.061	0.086	0.074	0.062	0.030	0.030	0.316	0.029	0.231	0.316
100	.1	100	50	1.44	204.33	0.012	1.48	-7.63	0.013	0.056	0.799	0.040	0.058	0.051	0.041	0.028	0.028	0.479	0.027	0.308	0.477
100	.3	50	20	1.10	54.43	0.019	1.45	-1.36	0.013	0.111	0.432	0.089	0.127	0.102	0.088	0.020	0.020	0.671	0.019	0.361	0.682
100	.3	50	50	0.92	54.16	-0.028	1.00	1.97	-0.033	0.068	0.419	0.063	0.078	0.064	0.066	0.028	0.028	1.274	0.023	0.761	1.289
100	.3	100	20	0.18	53.03	-0.010	0.19	2.76	-0.014	0.075	0.416	0.062	0.087	0.074	0.063	0.036	0.036	0.702	0.036	0.283	0.716
100	.3	100	50	1.41	54.91	0.012	1.43	-3.06	0.025	0.052	0.424	0.041	0.058	0.049	0.047	0.049	0.049	1.232	0.036	0.670	1.243
Avg.																					
bias%																					
20%				-1.77	22.30	-0.001	1.10	0.34	0.004	0.162	0.137	0.068	0.076	0.071	0.066	0.083	0.083	1.389	0.032	0.035	1.396
50%				-5.82	63.90	0.000	1.00	-1.02	0.009	0.230	0.264	0.067	0.080	0.071	0.065	0.133	0.133	1.142	0.030	0.031	1.156
100%				0.82	198.57	-0.003	1.08	-6.76	-0.004	0.080	0.709	0.063	0.086	0.085	0.064	0.027	0.027	0.540	0.039	0.471	0.545
ICC																					
0.05				-3.41	147.42	-0.001	1.02	-5.43	0.002	0.173	0.427	0.061	0.074	0.075	0.060	0.101	0.101	0.803	0.036	0.189	0.815
0.1				-2.50	89.67	-0.001	0.95	-1.86	0.003	0.157	0.378	0.061	0.074	0.068	0.060	0.081	0.081	0.874	0.029	0.130	0.885
0.3				-0.37	27.93	-0.001	1.05	0.46	0.004	0.114	0.238	0.066	0.081	0.072	0.066	0.045	0.045	1.254	0.030	0.192	1.256
K																					
50				-1.99	90.31	-0.006	1.16	-3.23	0.002	0.157	0.356	0.074	0.089	0.085	0.073	0.073	0.073	0.980	0.030	0.190	0.985
100				-2.29	89.55	0.003	0.85	-1.47	0.004	0.141	0.346	0.051	0.064	0.058	0.051	0.080	0.080	0.960	0.034	0.149	0.971
n_k																					
20				-2.55	94.01	0.005	0.87	-3.65	0.008	0.170	0.374	0.077	0.098	0.095	0.078	0.057	0.057	0.747	0.039	0.177	0.747
50				-1.96	95.85	-0.008	1.26	-1.31	-0.001	0.144	0.366	0.055	0.063	0.056	0.053	0.104	0.104	1.300	0.029	0.181	1.318

Note. *: Relative percentage bias was reported for item discrimination estimates.

Table 5.4: Simulation Study: Accuracy of IRT Scale Score

				Bias				RMSE				SE			
				Invar.		Bias		Invar.		Bias		Invar.		Bias	
DIF%	ICC	K	n_k	θ_{jk}	θ_k	θ_{jk}	θ_k	θ_{jk}	θ_k	θ_{jk}	θ_k	θ_{jk}	θ_k	θ_{jk}	θ_k
0	.50	50	20	0.022	-0.013	0.021	-0.020	0.477	0.199	0.481	0.206	0.350	1.383	0.327	1.088
0	.50	50	50	-0.033	0.013	-0.033	0.017	0.473	0.159	0.474	0.162	0.272	1.357	0.266	1.253
0	.50	100	20	-0.021	0.012	-0.021	0.015	0.482	0.216	0.483	0.219	0.354	1.357	0.343	1.212
0	.50	100	50	0.001	0.002	0.001	0.003	0.470	0.160	0.470	0.157	0.269	1.295	0.267	1.300
20	.05	50	20	0.022	-0.015	0.022	-0.009	0.473	0.237	0.459	0.203	0.350	1.370	0.232	0.370
20	.05	50	50	-0.033	0.014	-0.033	0.021	0.469	0.190	0.451	0.145	0.275	1.353	0.210	0.366
20	.05	100	20	-0.021	0.018	-0.021	0.014	0.478	0.247	0.458	0.210	0.355	1.365	0.242	0.392
20	.05	100	50	0.001	0.010	0.001	0.006	0.466	0.202	0.449	0.148	0.272	1.354	0.206	0.349
20	.1	50	20	0.022	-0.015	0.022	-0.008	0.472	0.234	0.462	0.206	0.352	1.400	0.253	0.477
20	.1	50	50	-0.033	0.013	-0.033	0.022	0.469	0.187	0.456	0.151	0.271	1.351	0.218	0.452
20	.1	100	20	-0.021	0.017	-0.021	0.015	0.477	0.245	0.462	0.216	0.355	1.369	0.259	0.488
20	.1	100	50	0.001	0.008	0.001	0.006	0.465	0.199	0.453	0.154	0.270	1.365	0.217	0.439
20	.3	50	20	0.022	-0.015	0.022	-0.008	0.474	0.217	0.472	0.209	0.351	1.413	0.311	0.898
20	.3	50	50	-0.033	0.010	-0.033	0.020	0.469	0.173	0.466	0.162	0.271	1.379	0.252	0.925
20	.3	100	20	-0.021	0.014	-0.021	0.015	0.478	0.231	0.474	0.225	0.354	1.380	0.320	0.939
20	.3	100	50	0.001	0.007	0.001	0.005	0.466	0.182	0.463	0.164	0.269	1.370	0.250	0.894
50	.05	50	20	0.022	-0.016	0.022	-0.005	0.466	0.336	0.445	0.212	0.351	1.450	0.185	0.193
50	.05	50	50	-0.033	0.015	-0.032	0.040	0.462	0.282	0.439	0.148	0.273	1.401	0.174	0.203
50	.05	100	20	-0.021	0.018	-0.021	0.021	0.471	0.325	0.442	0.215	0.357	1.409	0.191	0.223
50	.05	100	50	0.001	0.012	0.001	0.008	0.460	0.292	0.436	0.147	0.272	1.398	0.173	0.197
50	.1	50	20	0.022	-0.016	0.022	-0.005	0.465	0.320	0.449	0.220	0.351	1.440	0.203	0.261
50	.1	50	50	-0.033	0.015	-0.032	0.041	0.462	0.266	0.443	0.157	0.270	1.392	0.186	0.271
50	.1	100	20	-0.021	0.017	-0.021	0.020	0.471	0.311	0.448	0.224	0.353	1.414	0.208	0.291
50	.1	100	50	0.001	0.011	0.001	0.008	0.459	0.278	0.441	0.155	0.268	1.403	0.183	0.272
50	.3	50	20	0.022	-0.015	0.022	-0.004	0.469	0.257	0.467	0.225	0.347	1.427	0.278	0.688
50	.3	50	50	-0.033	0.012	-0.032	0.043	0.464	0.206	0.460	0.174	0.266	1.390	0.231	0.717
50	.3	100	20	-0.021	0.016	-0.021	0.019	0.474	0.263	0.468	0.238	0.348	1.388	0.287	0.748
50	.3	100	50	0.001	0.006	0.001	0.007	0.462	0.220	0.457	0.170	0.264	1.392	0.229	0.735
100	.05	50	20	0.021	-0.019	0.021	-0.019	0.432	0.760	0.445	0.967	0.211	2.646	0.146	-0.068
100	.05	50	50	-0.032	0.017	-0.033	0.013	0.426	0.682	0.428	0.604	0.196	2.007	0.179	0.696
100	.05	100	20	-0.021	0.013	-0.021	0.012	0.427	0.745	0.433	0.763	0.229	2.542	0.184	0.493
100	.05	100	50	0.001	0.011	0.002	0.012	0.425	0.700	0.426	0.617	0.190	2.114	0.180	1.044
100	.1	50	20	0.022	-0.018	0.021	-0.019	0.439	0.685	0.450	0.645	0.250	2.135	0.187	0.319
100	.1	50	50	-0.033	0.016	-0.033	0.011	0.432	0.603	0.434	0.443	0.208	1.749	0.197	1.018
100	.1	100	20	-0.021	0.013	-0.021	0.012	0.436	0.662	0.440	0.556	0.263	2.028	0.238	1.022
100	.1	100	50	0.001	0.011	0.002	0.014	0.431	0.618	0.432	0.449	0.203	1.804	0.197	1.304
100	.3	50	20	0.022	-0.017	0.021	-0.023	0.455	0.415	0.459	0.304	0.308	1.588	0.280	1.029
100	.3	50	50	-0.033	0.015	-0.033	0.009	0.450	0.353	0.450	0.238	0.236	1.509	0.231	1.291
100	.3	100	20	-0.021	0.014	-0.021	0.011	0.457	0.401	0.459	0.317	0.315	1.542	0.302	1.256
100	.3	100	50	0.001	0.007	0.002	0.019	0.447	0.366	0.448	0.237	0.232	1.498	0.230	1.374
Avg.															
bias%															
20%				-0.008	0.005	-0.008	0.008	0.471	0.212	0.460	0.183	0.312	1.372	0.248	0.582
50%				-0.008	0.006	-0.007	0.016	0.466	0.280	0.450	0.190	0.310	1.409	0.211	0.400
100%				-0.008	0.005	-0.008	0.005	0.438	0.583	0.442	0.512	0.237	1.930	0.213	0.898
ICC															
0.05				-0.007	0.006	-0.007	0.009	0.420	0.385	0.409	0.337	0.256	1.570	0.177	0.343
0.1				-0.007	0.006	-0.007	0.009	0.421	0.354	0.413	0.275	0.263	1.450	0.196	0.509
0.3				-0.008	0.005	-0.008	0.009	0.464	0.274	0.462	0.222	0.297	1.440	0.267	0.958
K															
50				-0.005	-0.001	-0.005	0.006	0.434	0.337	0.428	0.285	0.270	1.495	0.208	0.532
100				-0.009	0.012	-0.009	0.012	0.434	0.342	0.426	0.274	0.272	1.481	0.216	0.656
n_k															
20				0.000	0.000	0.000	0.002	0.462	0.383	0.455	0.342	0.322	1.628	0.239	0.557
50				-0.016	0.012	-0.016	0.017	0.455	0.333	0.446	0.248	0.250	1.513	0.208	0.697

Chapter 6

Summary and Discussion

The first purpose of this study was to evaluate the performance of the model selection criteria to detect global bias and item cluster bias. Overall, the expected simulation results were found except the case in which all items exhibited cluster bias. LRT generally revealed acceptable Type I error rates, whereas all model selection criteria revealed acceptable power to detect global cluster bias when some portion of items exhibits cluster bias (e.g., 20% and 50%). One exception for power was the BIC with a small sample size and high ICC (i.e., small cluster bias magnitude in discriminations). In addition, different detection methods showed different power regarding item cluster bias. As expected, the AIC showed the highest power among the information criteria we compared. When there is cluster bias for all items, unexpectedly, the detection methods of global bias and item cluster bias were problematic using all detection methods we considered.

The second purpose of this study was to show the consequences of ignoring cluster bias in terms of the accuracy of the parameter estimates and SEs. As expected, the bias and the RMSE of the within-level and between-level item discrimination parameter estimates were mainly problematic when a portion of items have cluster bias (e.g., 20% and 50%). Ignoring cluster bias would be acceptable only when a small portion of the items have cluster bias (e.g., 20%) and a high ICC (small bias magnitude in discrimination). Because of the equality constraints used when ignoring cluster bias, between-level item discriminations tend to be overestimated when they are not as high as within-level item discriminations (which is commonly true because the ICC is smaller than .5 in most applications). Further, unacceptable SEs of the item discrimination estimates were found when ignoring cluster bias unless there is small ICC and large cluster size. Regarding IRT scale scores, the overall accuracy (quantified RMSE) was low in ignoring cluster bias. However, bias was unexpectedly comparable between models with and without taking into account cluster bias. The SEs of the IRT scale scores were not precise in ignoring cluster bias in all conditions we considered.

The results from the present study provide implications for evaluating the detection methods and the consequences of ignoring cluster bias in multilevel item response models. First, two forms of the BIC performed differently depending on the number of cases used in the equation. The BIC with the total sample size (J=the number of clusters × cluster size) as the number of cases showed the lowest power among the criteria. The power for the BIC increased as the sample size became larger, but the power was still not adequate ($<$.80) in the condition with a large number of bias items (50%) and a high ICC (0.3). In contrast, the performance of the saBIC was comparable to that of the Wald test or the LRT. The saBIC includes a smaller penalty terms than the BIC does in the formula. Otherwise, the BIC with the number of clusters as the number of cases might be a better indicator of a multilevel item response model. Yu and Park (2014) reported that using the number of clusters leads to a better performance than total sample size in the BIC for multilevel latent class models. Taken together, based on our study, the total number of individuals is not recommended for BIC calculation in multilevel item response models. Instead, the saBIC or the BIC with the number of clusters is recommended.

Second, the power for item cluster bias was unexpectedly low when all items have cluster bias. As shown in this study, largely overestimated between-level item discriminations are expected in ignoring a large amount of cluster bias in the invariance model, which is the baseline model in detecting global bias and item bias. We chose the invariance model as a baseline model to follow IRT differential item functioning (DIF) detection method convention. When a large number of bias items is suspected, the baseline model (the invariance model) is a misspecified model. In such a case, a global bias model is more appropriate than the invariance model as a baseline model. Thus, when a large number of cluster item bias is found (e.g., larger than 50%), comparing item cluster bias detection results between two different baseline models, invariance and global bias models is recommended.

There are methodological limitations to the present study. First, the simulation conditions employed in the study are limited to two-level data and one latent variable at each level. More extensive simulations that vary these limited conditions should be conducted to make solid generalizations. Second, as mentioned earlier, there are alternative approaches for testing cluster bias for item discriminations, such as a random item response modeling approach using Bayesian analysis (De Jong, Steenkamp, & Fox, 2007; Fox & Verhagen, 2010). The random item response modeling approach was not considered in this study because our focus was on the model selection methods with MMLE. Comparing the model selection methods considered in this study to the random item response model approach is also left as a future study.

In spite of the methodological limitations, this study illustrates and evaluates model selection methods for global and item cluster bias in a common multilevel data structure - one found in empirical studies - and in the use of MMLE (which is a more common estimation method than Bayesian analysis in current IRT applications). As summarized earlier, LRT provides adequate Type I error rates and power for detecting global cluster bias, whereas the AIC is generally recommended for detecting item cluster bias. We showed that ignoring cluster bias is of concern for between-level item discriminations used for understanding constructs in multilevel data. Given our simulation results, we recommend testing global and item cluster bias as part of the analysis steps applied to multilevel item response models.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Baker, F. B. & Kim, S.-H, (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics, 36,* 491-522.

Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 401-445). San Francisco, CA: Jossey-Bass.

Bottge, B. A., Toland, M. D., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., & Ma, X. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children, 81,* 158-175.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach.* New York, NY: Springer.

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336-370.

Cohen, A. S., & Cho, S.-J. (2015). Information criteria. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory, models, statistical tools, and applications.* Boca Raton, FL: Chapman & Hall/CRC Press.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis.* Stanford, CA: Stanford University, Evaluation Consortium.

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in crossnational consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260-278.

Doolaard, S. (1999). *Schools in change or schools in chains?: The development of educational effectiveness in a longitudinal perspective.* Enschede, Netherlands: Twente University Press.

Engle, R. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics* (vol.2, pp. 775-826). New York, NY: Elsevier.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 18,* 229-252.

Fox, J.-P. (2004). Modelling response error in school effectiveness research. *Statistica Neerlandica, 58,* 138-160.

Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics, 30,* 189-212.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika, 66,* 271-288.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt & J. Billiet (Eds.), *Cross-Cultural Analysis: Methods and Applications* (pp.467-488). New York, NY: Routledge.

Härnqvist, K., Gustafsson, J. E., Muthén, B. O., & Nelson, G. (1994). Hierarchical models of ability at class and individual levels. *Intelligence, 18,* 165-187.

Heck, R. H., & Thomas, S. L. (2009). *Structural equation modeling* (2nd ed.). London: Routledge.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research, 26,* 329-367.

Houts, C. R., & Cai, L. (2013). *flexMIRT R users manual version 2: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ:Lawrence Erlbaum Associates.

Jac, S., & Oort, F. J. (2015). On the power of the test for cluster bias. *British Journal of Mathematical and Statistical Psychology, 68*, 434-455.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling, 21*, 31-39.

Jeon, M., & Rabe-Hesketh, S. (2012). Profile-likelihood approach for estimating generalized linear mixed models with factor structures. *Journal of Educational and Behvior Statistics, 37,* 518-542.

Jöreskog, K. G., & Goldberger, A. S.(1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70,* 631-639.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79-93.

Kim, E. S., Kwok, O.-M., & Yoon, M, (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 19,* 250-267.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O.-M. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC Models. *Structural Equation Modeling: A Multidisciplinary Journal, 22,* 603-616.

Kim, S-.Y., Suh, Y., Kim, J.-S., Albanese, M. A., & Langer, M. M. (2013). Single and multiple ability estimation in the SEM framework: A noninformative Bayesian estimation approach. *Multivariate Behavioral Research, 48,* 563-591.

Lin, T. H., & Dayton, C. M. (1997). Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics, 22,* 249-264.

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

Ludtke, O., Marsh, H. W., Robitzsch, A, & Trautwein U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16,* 444-467.

May, H. (2006). A multilevel Bayesian IRT method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics,* *31,* 63-79.

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika, 58,* 575-585.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259-284.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376-398.

Muthén, B. O., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-time point example. Technical report. Retrieved from http://www.statmodel.com/download/IRT1Version2.pdf

Muthén, B. O., Khoo, S. T., & Gustafsson, J. E. (1997). Multilevel latent variable modeling in multiple populations. Technical report. Retrieved from

http://www.statmodel.com/bmuthen/articles/Article_074.pdf

Muthén, L. K. & Muthén, B. O. (1998-2015). *Mplus Users Guide. Seventh Edition.* Los Angeles, CA: Muthén & Muthén.

Patarapichayatham, C. & Kamata, A. (2014). Effects of differential item discriminations between individual-level and cluster-level under the multilevel item response theory model. *Open Journal of Applied Sciences, 4,* 425-432.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18,* 161-182.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167-190.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathmatical and Statistical Psychology, 67,* 172-194.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333-343.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Vermunt, J. K. (2007). *Multilevel mixture item response theory models: An application in education testing.* Bulletin of the International Statistical Institute, 56th Session, paper #1253, 1-4. ISI 2007: Lisboa, Portugal.

Vriese, S. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228-243.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances* *from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Yu, H.-T., & Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class Models. *Multivariate Behavioral Research, 49*, 232-244.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago, IL: Scientific Software.