International Journal of Assessment Tools in Education
2019, Vol. 6, No. 1, 92–108
https://dx.doi.org/10.21449/ijate.523586

Published at http://www.ijate.net http://dergipark.gov.tr/ijate **Research Article**

The Impact of Ignoring Multilevel Data Structure on the Estimation of Dichotomous Item Response Theory Models

Hyung Rock Lee [1,*], **Sunbok Lee** [2], **Jaeyun Sung** [3]

[1] University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR USA
[2] University of Houston, Department of Psychology, Houston, TX USA
[3] Lyon College, Department of Political Science, Batesville, AR, USA

ARTICLE HISTORY

Received: 14 November 2018

Accepted: 05 February 2019

KEYWORDS

Item Response Theory,
Rasch,
Multilevel Data,
Monte Carlo Simulation

Abstract: Applying single-level statistical models to multilevel data typically produces underestimated standard errors, which may result in misleading conclusions. This study examined the impact of ignoring multilevel data structure on the estimation of item parameters and their standard errors of the Rasch, two-, and three-parameter logistic models in item response theory (IRT) to demonstrate the degree of such underestimation in IRT. Also, the Lord's chi-square test using the underestimated standard errors was used to test differential item functioning (DIF) to show the impact of such underestimation on the practical applications of IRT. The results of simulation studies showed that, in the most severe case of multilevel data, the standard error estimate from the standard single-level IRT models was about half of the minimal asymptotic standard error, and the type I error rate of the Lord's chi-square test was inflated up to .35. The results of this study suggest that standard single-level IRT models may seriously mislead our conclusions in the presence of multilevel data, and therefore multilevel IRT models need to be considered as alternatives.

1. INTRODUCTION

In traditional statistical models, observations are typically assumed to be independent. However, the assumption of independence is quite strong and may not be tenable in practice. In educational research, for example, observations in data are often not independent because of a hierarchical data structure. It is well known that applying traditional statistical models based on the independence assumption to multilevel data may result in incorrect standard errors (Barcikowski, 1981; Tate & Wongbundhit, 1983; Satorra & Muthen, 1995; Goldstein, 1987; Julian, 2001; Finch & French, 2011). Because the use of correct standard errors is the key

CONTACT: Hyung Rock Lee ✉ rlee@uca.edu University of Central Arkansas, Department of Exercise & Sport Science, Conway, AR USA

ISSN-e: 2148-7456 /© IJATE 2019

element for valid statistical inferences such as hypothesis testing and confidence intervals, applying single-level models to multilevel data could be problematic.

Given the concerns on the use of single-level models to multilevel data, the goal of this study is to examine the extent to which multilevel data structure affects the estimation of the single-level dichotomous IRT models and their subsequent application. More specifically, two Monte Carlo simulation studies were conducted to examine 1) the impact of ignoring multilevel data structure on the estimation of item parameters and their standard errors of the standard single-level Rasch, two- (2PL), and three- (3PL) parameter logistic models in item response theory (IRT); 2) the type I error inflation of the Lord's chi-square tests based on standard errors estimated from the single-level IRT models. In the simulation study 1, item responses with multilevel data structure were generated using the Rasch, 2PL, and 3PL models formulated in the hierarchical generalized linear model (HGLM), in which items, persons, and schools were modeled in Level-1, Level-2, and Level-3, respectively (Kamata & Vaughn, 2011). In generating item responses with multilevel structure, intraclass correlation coefficients (ICCs), numbers of groups, and group sizes were manipulated. Given the item responses with multilevel structure, item parameters and their standard errors were estimated using single-level IRT models with BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). To evaluate the extent to which standard errors are underestimated, the analytical minimal standard errors (Thissen & Wainer, 1982) for item parameters in the Rasch, 2PL, and 3PL were used as reference values. In practice, the underestimated standard errors can be used for other applications such as DIF tests. In the simulation study 2, the type I error rates of two DIF tests were compared: the Lord's chi-square test (Lord, 1980) using the underestimated standard errors from the single-level IRT models and the DIF test based on the Rasch model that was formulated in the hierarchical generalized linear model (HGLM).

2. IGNORING MULTILEVEL DATA STRUCTURE IN STATISTICAL MODELS

The impact of multilevel data on the estimation of standard errors in statistical models can be illustrated by an example of cluster sampling designs (Kish, 1965), in which only a subset of primary units or clusters is randomly selected, and then secondary units are sampled within the selected primary units. Cluster sampling designs are often preferred because of cost and time effectiveness. In cluster sampling designs, respondents in the same cluster are likely to be similar to one another because they share similar contexts. From a statistical viewpoint, the similarity between respondents makes the information in data more redundant or less unique, which results in the reduction of effective sample sizes. As a result, the estimated sampling variances from cluster sampling designs are larger than the ones from simple random sampling designs. The loss of effectiveness in cluster sampling designs is measured by the design effect, which is defined as the ratio of the sampling variance in cluster sampling designs to the sampling variance in simple random sampling designs. In other words, the design effect is a correction factor to be multiplied to the sampling variance of the simple random sampling to get the actual sampling variance in cluster sampling designs (Hox, 1998). In the simplest cluster sampling design, the design effect is defined by the following equation:

$$\text{Design effect} = 1 + \left(n_g - 1\right)\rho_I, \qquad (1)$$

where n_g is the sample size in a group, and ρ_I is the intraclass correlation coefficient (ICC). The ICC provides a measure of the amount of dependency among individuals or how similar individuals are within groups. As can be seen from Equation 1, the design effect is greater than one for a non-zero ICC. Therefore, if appropriate statistical models that can accommodate cluster structure are used, the sampling variance in data from cluster sampling designs should be larger than the one from simple random sampling designs because of reduction in effective sample sizes. On the other hand, given non-zero ICCs, observed variance within clusters is typically less than observed variance between clusters since observations within cluster tend to be more similar to one another. Therefore, when observations are assumed to be independent, overall observed variance that is obtained without reflecting cluster nature tends to be underestimated, which could result in type I error inflation (Goldstein, 1987).

The effect of multilevel data structure on the estimation of statistical models has been investigated in many different settings. Barcikowski (1981) reported that the type I error rates of t-tests can be dramatically increased as the ICC increased. Also, Tate and Wongbundhit (1983) reported that the ordinary least square (OLS) regression produced unbiased parameter estimates but downwardly biased standard error estimates in the presence of multilevel data. Satorra and Muthen (1995) compared the standard maximum likelihood estimation, the robust maximum likelihood estimation, and the multilevel maximum likelihood estimation for structural equation modeling (SEM) under complex sampling designs and found that standard error estimates from the standard maximum likelihood estimation were downwardly biased. Recently, Finch and French (2011) found that applying standard approaches for differential item functioning (DIF) to multilevel data caused type I error inflation. In line with those concerns on the use of standard single-level models in the presence of multilevel data, this study was designed to explicitly show the degree to which the multilevel data structure influences the estimation of standard single-level IRT models.

3. STANDARD ERRORS IN IRT APPLICATIONS

Standard errors measure the accuracy of estimation. Using correct standard errors is an essential component for valid statistical inferences based on hypothesis tests and confidence intervals. The use of the correct standard error is also important in many IRT applications (Toland, 2008). For example, the accurate standard error estimate is important in identifying DIF items using the Lord's chi-square test in which the difference in the item parameters between the focal and reference groups is tested using the following equation:

$$\chi^2 = \frac{\left(\hat{\theta}_F - \hat{\theta}_R\right)^2}{\hat{\theta}_F^2 + \hat{\theta}_R^2}, \qquad (2)$$

Where $\hat{\theta}_F$ and $\hat{\theta}_R$ represent the parameter estimates in the focal and reference groups, and $\hat{\sigma}_F^2$ and $\hat{\sigma}_R^2$ represent the standard error estimates for $\hat{\theta}_F$ and $\hat{\theta}_R$. Because the item parameter estimates for the focal and reference groups are obtained from separate calibrations, item parameter estimates need to be transformed on a common metric using an appropriate transformation. As can be seen from Equation 2, standard error estimates affect the result of the Lord's chi-square test. Some other IRT applications also require accurate standard error estimates for item parameter estimates (Toland, 2008): the separate calibration *t*-test for DIF (Wright & Stone, 1979), the item parameter replication (IPR) method for testing non-compensatory DIF (Oshima, Raju, & Nanda, 2006), and the cumulative sum (CUSUM) procedure for the computer adaptive test (Veerkamp & Glas, 2000).

In examining the estimation for standard errors in IRT models, this study used the minimum obtainable standard errors for item parameters (Thissen & Wainer, 1982) as references values. Thissen and Wainer (1982) derived analytical asymptotic standard errors for item parameters using the inverse information matrix. Those asymptotic standard errors can be considered as the lower limits for the estimated standard errors because they are derived under the very strong assumptions which are not likely to be met in practice. Therefore, estimated standard errors are larger than the minimal asymptotic standard errors.

4. MULTILEVEL IRT MODELS

One of the assumptions in traditional IRT models is the local independence assumption in which the dependencies among item responses are assumed to be fully explained by the specified IRT model (Embretson & Reise, 2000). More specifically, two different kinds of local independence assumptions can be considered (Reckase, 2009; Jiao, Kamata, Wang, & Jin, 2012). The local item independence refers to the independence of responses for items within a specific person. Given the ability of a person, a person's response to an item does not have any influence on the probability of that person's response to another item. On the other hand, the local person independence refers to the independence of responses of persons for a specific item. Given the abilities of persons, a person's response to a specific item does not affect the probability of another person's response to that item.

Since the traditional IRT models assume a single source of the dependencies among item responses, which is the ability of a person, problems could occur when the dependencies among item responses still remain beyond what is explained by the specified IRT model. In order to fully explain the dependencies, therefore, additional sources of the dependencies need to be specified in the IRT model. For example, a common passage in a test could cause additional dependencies among item responses. In that case, the local item independence is considered to be violated. On the other hand, the local person independence could be violated in the presence of clustered data (Jiao et al., 2012). For example, the responses of students from the same school could be more similar to each other than to responses from students from other schools, even after controlling for the abilities of persons. In multilevel IRT models, the clustered data structure is considered the additional source of the dependencies among item responses (Kamata, 2001).

A simple multilevel IRT model assumes that items are nested within persons, and persons are nested within groups (Kamata & Vaughn, 2011). For example, multilevel 2PL models can be expressed as

$$P_{ijg}[Y = 1] = \frac{\exp[\alpha_i(\theta_g+\theta_{jg})+\beta_i]}{1+\exp[\alpha_i(\theta_g+\theta_{jg})+\beta_i]}, \qquad (3)$$

Where α_i and β_i are the discrimination and difficulty parameters of item i, θ_g is the mean of ability of group g, θ_{jg} is the amount of deviation from the group mean ability for a person j in a group g.

5. SIMULATION STUDY1

5.1. Simulation Designs

This simulation study was designed to examine the impact of ignoring multilevel data structure on the estimation of the Rasch, 2PL, and 3PL models. To simulate multilevel data structure, item responses were generated based on Equation 4 below (Kamata & Vaughn, 2011). The parameters and their standard errors were estimated using BILOG-MG (Zimowski et al., 1996). To make estimates comparable across replications, metric transformations were performed to put estimates on a common scale. This simulation was conducted using the R software package (R Core Team, 2013).

5.1.1. *Simulation Conditions*

The simulation conditions for multilevel data structure was manipulated in terms of the ICC, number of groups (nG), and group sizes (nW). The values of the ICC in this simulation study were set at 0, .05, .15, .25, .35, and .45 based on prior research. Hedges and Hedberg (2007) reported that the values of the ICC in educational performance research often range between .10 and .25. Snijders and Bosker (1999) reported that the values of the ICC between .05 and .20 are most common in educational research, and values greater than .20 can be considered large. Also, the numbers of groups were set at 50, 100, and 200 based on prior research (Maas & Hox, 2005; Finch & French, 2011). The group sizes or within-group sample sizes were set at 5, 15, 25, and 50, which cover the typical range of within-group sample sizes in family and educational research (Maas & Hox, 2005). In all, there were total 72 (=6 × 3 × 4) simulation conditions, and 1000 simulated data sets were replicated for each simulation condition.

5.1.2. *Data Generation*

To simulate multilevel data structure, item responses were generated based on the following equation:

$$P_{ijg}[Y=1] = r_i + (1-r_i)\frac{\exp[\alpha_i(\theta_g+\theta_{jg})+\beta_i]}{1+\exp[\alpha_i(\theta_g+\theta_{jg})+\beta_i]}, \tag{4}$$

$$\theta_{jg} \sim N(0,1), \tag{5}$$

$$\theta_g \sim N\left(0, \sigma^2_{\theta_g}\right), \tag{6}$$

which is the three-level hierarchical generalized linear model (Kamata, 2001), in which items, persons, and groups are modeled in Level-1, Level-2, and Level-3, respectively. In this simulation, the values of the difficulty parameters for seven items were set at (-3, -2, -1, 0, 1, 2, 3) so that the estimated standard errors can be compared to the minimal asymptotic standard errors tabulated in Thissen and Wainer (1982). The values of the discrimination parameters of the Rasch, 2PL, and 3PL models were set at (1, 1, 1, 1, 1, 1, 1), (1, 2, 1, 2, 1, 2, 1), and (1, 2, 1, 2, 1, 2, 1), respectively. For the 3PL model, guessing parameters were set at (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2).

The proportion of the between-group variance in the total variance, which is the ICC, was calculated based on the following equation:

$$ICC = \frac{\theta^2_{\theta_g}}{\theta^2_e+\sigma^2_{jg}+\sigma^2_{\theta_g}}, \tag{7}$$

Where σ^2_e, $\sigma^2_{\theta_{jg}}$, and $\sigma^2_{\theta_g}$ represent the variation in Level-1, Level-2, and Level-3, respectively. σ^2_e representing the Level-1 variance was set at $\pi^2/3$ following Snijders and Bosker (2011). $\sigma^2_{\theta_{jg}}$ representing the amount of deviation from the group mean ability for a person j in a group g and was set at 1. The values of $\sigma^2_{\theta_g}$, which represents the variance of the mean of abilities in a group g, can be determined given the values of the ICC.

5.1.3. *Minimal Asymptotic Standard Errors*

In examining the estimated standard errors from BILOG-MG, the minimal asymptotic standard errors (Thissen & Wainer, 1982) were used as reference values. Thissen and Wainer (1982) provided tables that contain minimal asymptotic standard errors for various values of locations, slopes, and asymptote parameters. Note that the values in the tables need to be adjusted using specific values of sample sizes.

5.1.4. *Scale Transformation*

In estimating parameters in IRT models, some parameters need to be fixed to arbitrary values to identify the models. Therefore, in IRT, independent estimates from two separate data sets can be compared only after they are expressed on a common metric (Stocking & Lord, 1983).

In this study, item parameter estimates from each replication were transformed into the metric defined by the original parameter values using the following equations (De Ayala, 2009): $\hat{\alpha}^* = \hat{\alpha}/A$, $\hat{\beta}^* = A\hat{\beta} + B$, $\hat{c}^* = \hat{c}$, where $A = S_{\hat{\beta}^*}/S_{\hat{\beta}}$, and $B = \bar{\hat{\beta}}^* - A\bar{\hat{\beta}}$; $\hat{\alpha}$, $\hat{\beta}$, and $\hat{c}$ represent the discrimination, difficulty, and guessing parameter estimates in the origianl metric; $\hat{\alpha}$, $\hat{\beta}$, and $\hat{c}$ represent corresponding estimates in the target metric; and $S_{\hat{\beta}^*}$ and $S_{\hat{\beta}}$ represent standard deviations for diffculty parameters on the target and original metric respectively. The standard error estimates were also transformed to the metric defined by the original parameter values using the following equations (Kim & Cohen, 1995):

$$SE = \sqrt{Var[\hat{\alpha}^*]} = \sqrt{Var\left[\frac{\hat{\alpha}}{A}\right]} = \frac{SE[\hat{\alpha}]}{A}, \tag{8}$$

$$SE[\hat{\beta}^*] = \sqrt{Var[\hat{\beta}^*]} = \sqrt{Var[A\hat{\beta} + B]} = A \times SE[\hat{\beta}], \tag{9}$$

Where the coefficients A and B are the ones that are defined above.

5.1.5. *Evaluation*

To evaluate the impact of multilevel data structure on the estimation of item parameters for standard IRT models, the bias was calculated using the following equation and compared across simulation conditions:

$$Bias(\hat{\theta}) = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I}(\hat{\theta}_{ri} - \theta_i)}{RI}, \tag{10}$$

Where *R* and *I* represent the number of replications and the number of items respectively. Also, the following ratio was calculated to compare the standard error estimates from BILOG-MG with the minimal asymptotic standard errors (Thissen & Wainer, 1982):

$$r = \frac{SE_B}{SE_T}, \tag{11}$$

Where SE_B and SE_T represent the standard error estimates from BILOG-MG and the minimal asymptotic standard errors, respectively.

On the other hand, the type I error inflation is also of interest when the standard errors are underestimated. To obtain a rough idea for the type I error inflation in the presence of underestimated standard errors, the theoretical type I errors of the *z*-tests for the statistical significance of time parameters were calculated in the following way. Under the assumption that item parameters following the standard normal distribution, the type I errors can be expressed as the following:

$$Type\ I\ error = 1 - \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \tag{12}$$

Now, let us express the *z*-statistic based on the standard error estimates from BILOG-MG, which is denoted by z', in terms of the *z*-statistics based on the minimal asymptotic standard error estimates, which is denoted by *z*, as the following:

$$z' = \frac{\theta}{SE_B} = \frac{\theta}{rSE_T} = \frac{z}{r}, \tag{13}$$

Under the assumption that the *z*-test based on $z = \theta/SE_T$ gives us the exact type I error based on the standard normal distribution, the theoretical type I error of the *z*-test based on $z' = \theta/SE_B$ can be calculated as the following:

$$Type\ I\ error(r) = 1 - \int_{-1.96}^{1.96} \frac{r}{\sqrt{2\pi}} e^{-\frac{(rz')^2}{2}} dz', \tag{14}$$

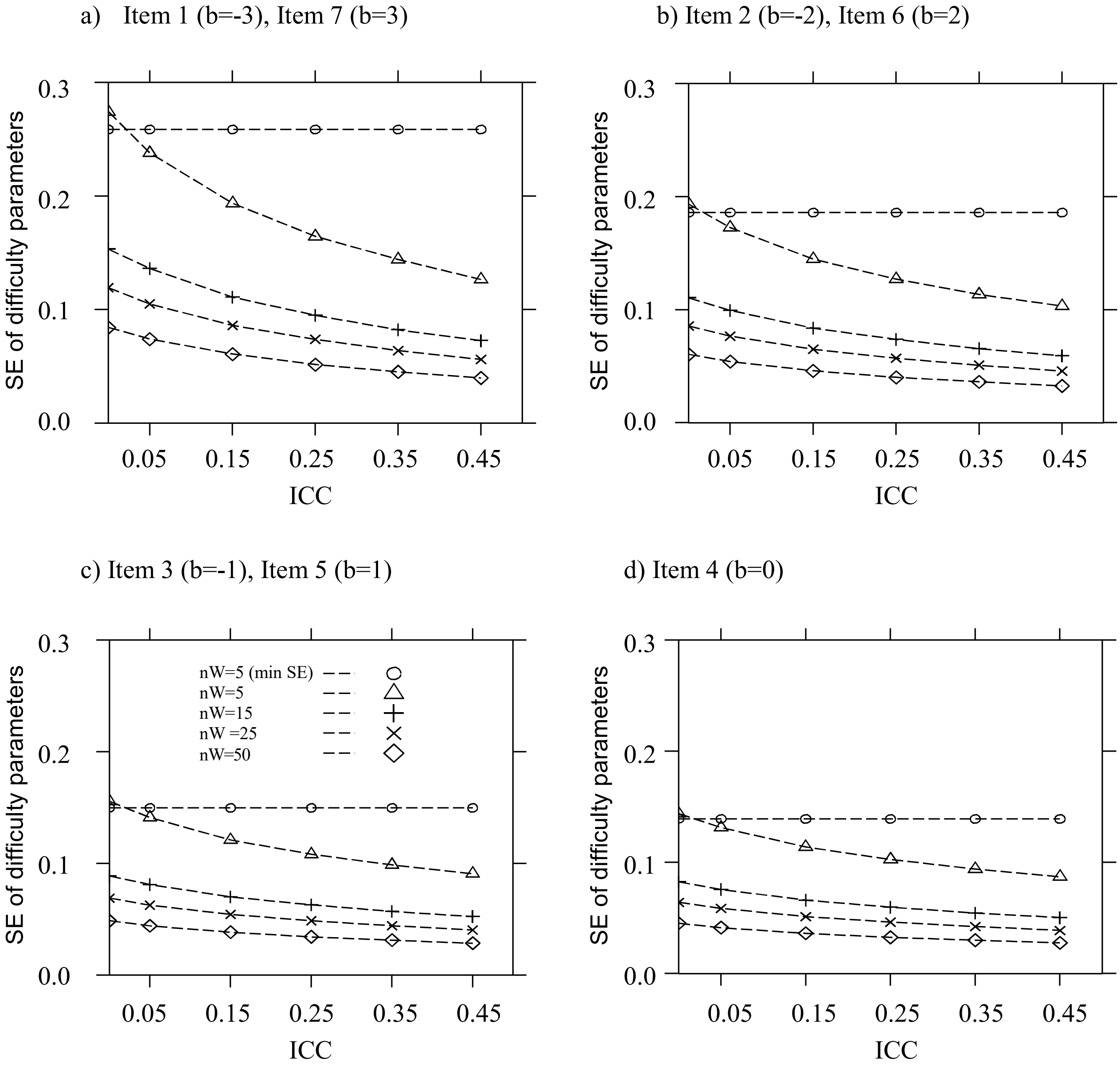
Based on Equation 14, the theoretical type I error of z-test based on the underestimated standard error from BILOG-MG can be calculated. For example, $r = 0.5$ indicates that the standard error estimate form BILOG-MG is half of the minimal asymptotic standard error. Then, the z-statistic is doubled based on Equation 13, and the theoretical type I error becomes 0.32 based on Equation 14.

5.2 Results for the Rasch Model

5.2.1. *Standard Errors of Difficulty Parameters*

The standard error estimates of the difficulty parameters in the Rasch model estimated from BILOG-MG are plotted in Figures 1 through 3 to demonstrate the influence of multilevel data structure on the estimation of standard errors. Figure 1, Figure 2, and Figure 3 show the standard error estimates for the cases where the number of groups (nG) are 50, 100, and 200, respectively. Each subplot in the figures shows the standard error estimates for a specific value of item difficulty parameters (b), and each line in the subplots shows the standard error estimates for a specific value of group sizes (nW). Because of space limitations, the ratios defined by Equation 11 are presented only for the number of groups (nG) of 50 in Table 1. In the table, the numbers in the parentheses are the type I errors for the corresponding values of r that were calculated based on Equation 14.

Figure 1. Standard error estimates of difficulty parameters in the Rasch model (BILOG, nG=50)



a) Item 1 (b=-3), Item 7 (b=3)

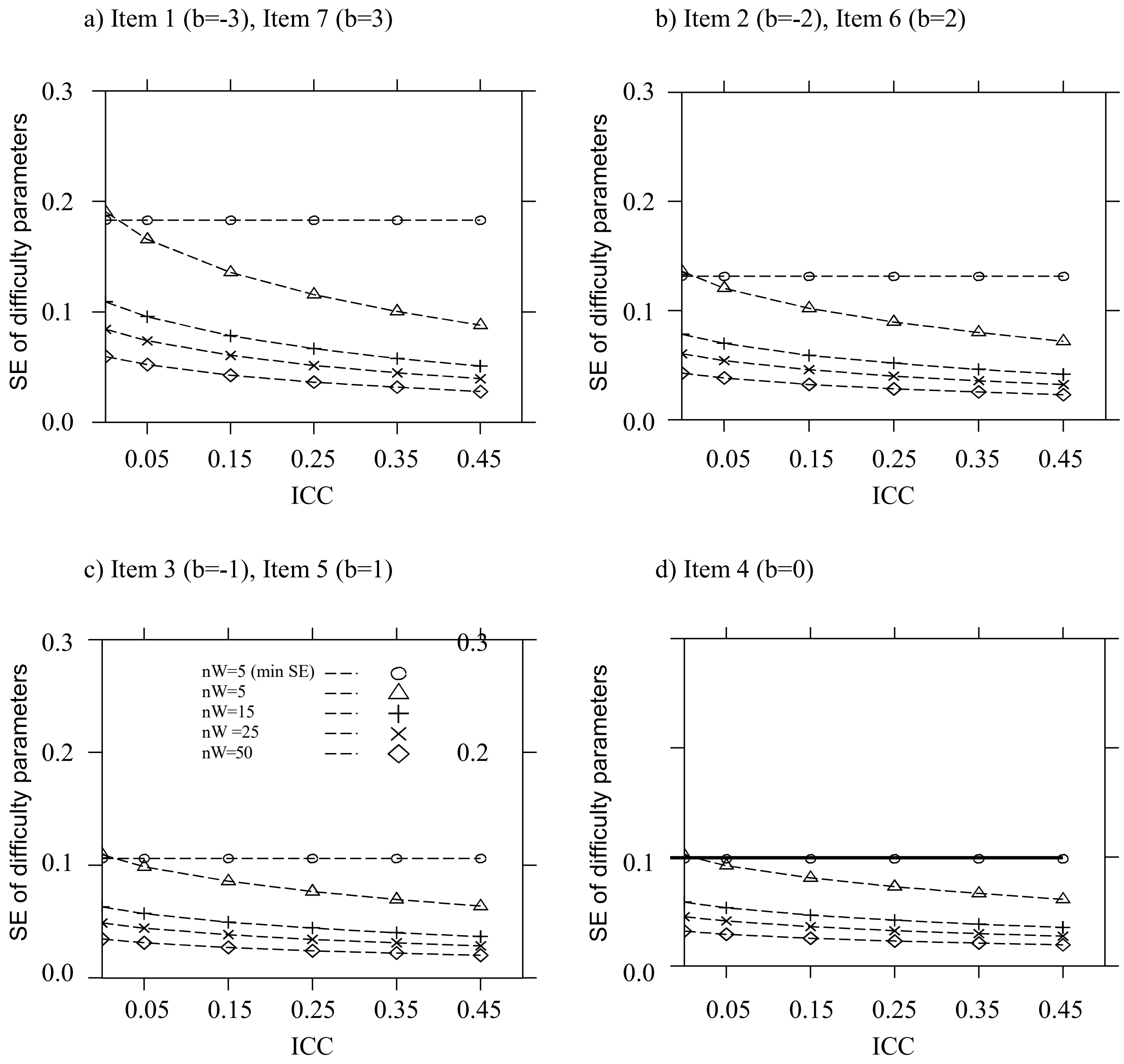
b) Item 2 (b=-2), Item 6 (b=2)

c) Item 3 (b=-1), Item 5 (b=1)

d) Item 4 (b=0)

Note. This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups (nG) is 50. For the group size (nW) of 5, the minimal asymptotic standard errors were plotted together for comparison.

Figure 2. Standard error estimates of difficulty parameters in the Rasch model (BILOG, nG=100)



a) Item 1 (b=-3), Item 7 (b=3)

b) Item 2 (b=-2), Item 6 (b=2)

c) Item 3 (b=-1), Item 5 (b=1)

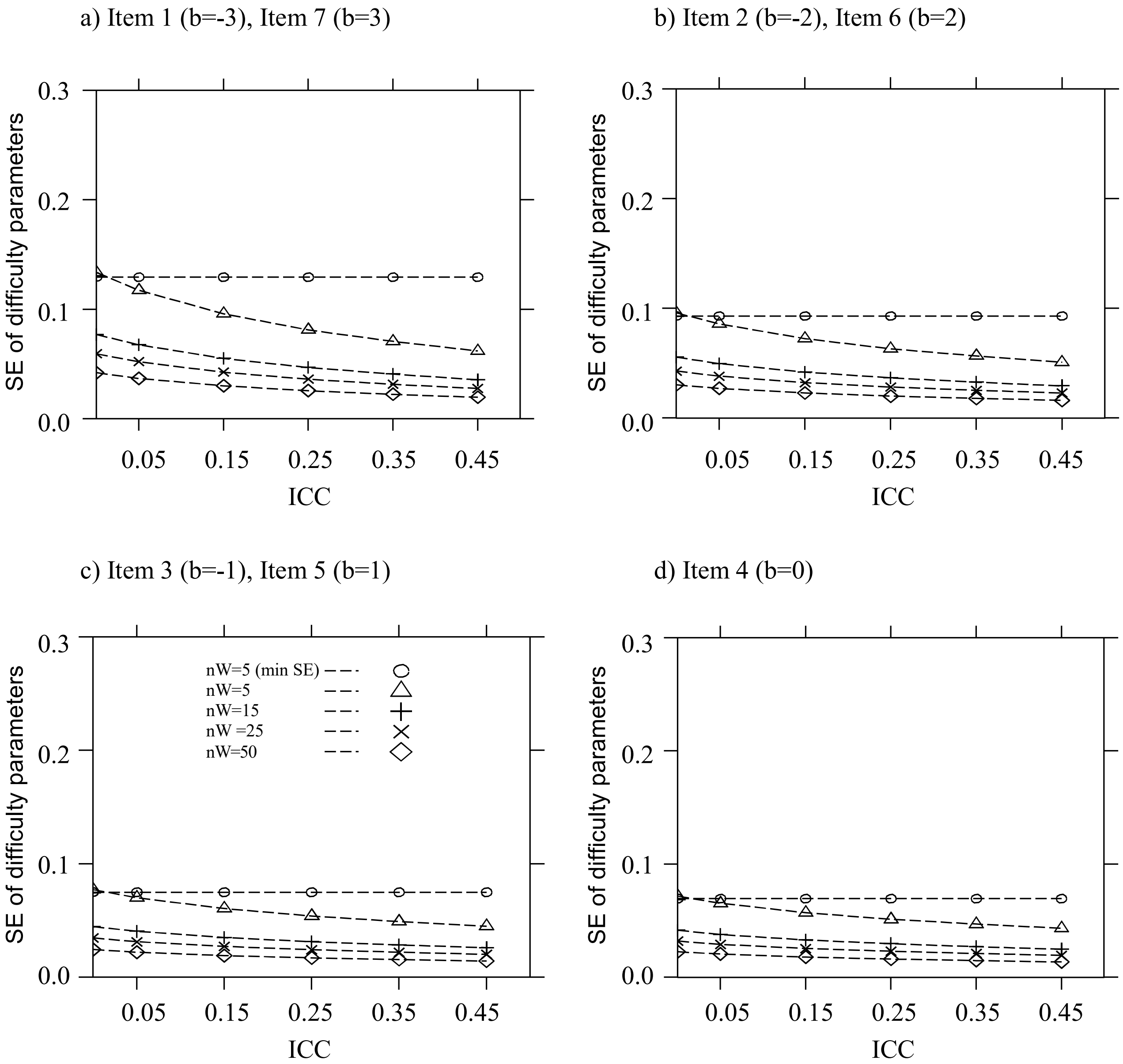
d) Item 4 (b=0)

Note. This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups (nG) is 100. For the group size (nW) of 5, the minimal asymptotic standard errors were plotted together for comparison.

Several trends can be identified from the results. Most importantly, the results show that the standard errors estimates decrease as the values of the ICC increase. The decrease is most prominent when the number of groups and the group sizes are small. For example, in Figure 1a, the standard error estimates for nG = 50, nW = 5, and b=-3 or 3 decrease from 0.2742 to 0.1265 as the values of the ICC increase from 0 to 0.45. Also, as can be seen from Table 1, the ratio r for ICC = 0.45, nG = 50, nW = 5, and b=-3 or 3 was 0.49, which indicates that the standard error estimate from the standard single-level Rasch model, which is 0.1265 in this case, is about half of the minimal asymptotic standard error. Note that the minimal asymptotic standard error for nG = 50, nW = 5, and b=-3 or 3 is 0.2586.

Secondly, the effect of the ICC on the estimation for the standard error decrease as the number of groups (nG) and the group sizes (nW) increase. For example, in Figure 1a, the standard error estimate for nW = 5 decrease more than nW = 50 as the values of the ICC increase. Also, the decrease is more prominent for nG = 50 (Figure 1) than nG = 200 (Figure 3).

Figure 3. Standard error estimates of difficulty parameters in the Rasch model (BILOG, nG=200)



Note. This figure provides a graphical illustration of changes in standard error estimates depending ICC when the number of groups (nG) IS 200. For the group size (nW) of 5, the minimal asymptotic standard errors were plotted together for comparision.

Table 1. The Ratios and Type I Errors for the Rasch Model When *n*G = 50

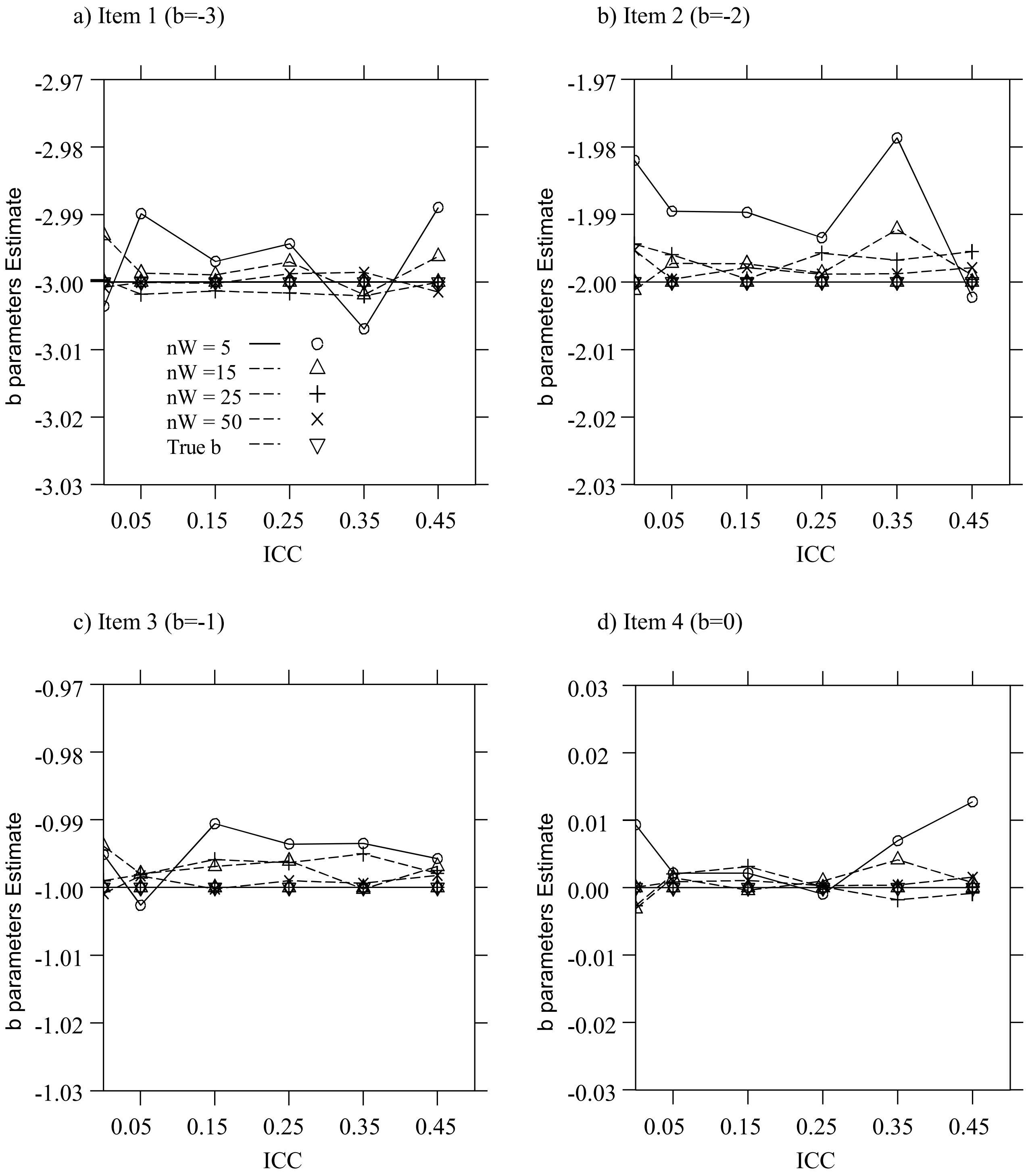
ICC	Groups	Groups Sizes	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
0.05	50	5	0.92	0.93	0.94	0.94	0.94	0.93	0.92
			(0.07)	(0.07)	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
0.05	50	15	0.91	0.92	0.94	0.94	0.94	0.92	0.91
			(0,07)	(0.07)	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)
0.05	50	25	0.91	0.92	0.94	0.94	0.94	0.92	0.91
			(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
0.05	50	50	0.90	0.92	0.93	0.94	0.93	0.92	0.90
			(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)
0.15	50	5	0.75	0.78	0.81	0.82	0.81	0.78	0.75
			(0.14)	(0.13)	(0.11)	(0.11)	(0.11)	(0.13)	(0.14)
0.15	50	15	0.74	0.78	0.81	0.82	0.81	0.78	0.74
			(0.15)	(0.13)	(0.11)	(0.11)	(0.11)	(0.13)	(0.15)
0.15	50	25	0.74	0.78	0.81	0.82	0.81	0.78	0.74
			(0.14)	(0.13)	(0.11)	(0.11)	(0.11)	(0.13)	(0.14)
0.15	50	50	0.74	0.78	0.81	0.82	0.81	0.784	0.74
			(0.15)	(0.13)	(0.11)	(0.11)	(0.11)	(0.13)	(0.14)
0.25	50	5	0.64	0.68	0.72	0.74	0.72	0.69	0.64
			(0.21)	(0.18)	(0.16)	(0.15)	(0.16)	(0.18)	(0.21)
0.25	50	15	0.64	0.69	0.73	0.74	0.73	0.69	0.64
			(0.21)	(0.18)	(0.15)	(0.15)	(0.15)	(0.18)	(0.21)
0.25	50	25	0.64	0.69	0.73	0.74	0.73	0.69	0.64
			(0.21)	(0.18)	(0.15)	(0.14)	(0.15)	(0.18)	(0.21)
0.25	50	50	0.63	0.68	0.72	0.74	0.72	0.68	0.63
			(0.22)	(0.18)	(0.16)	(0.15)	(0.16)	(0.18)	(0.22)
0.35	50	5	0.56	0.61	0.66	0.68	0.66	0.61	0.55
			(0.27)	(0.23)	(0.20)	(0.19)	(0.20)	(0.23)	(0.28)
0.35	50	15	0.55	0.61	0.66	0.68	0.66	0.61	0.55
			(0.28)	(0.23)	(0.20)	(0.18)	(0.20)	(0.23)	(0.28)
0.35	50	25	0.55	0.61	0.66	0.68	0.66	0.61	0.55
			(0.28)	(0.23)	(0.19)	(0.18)	(0.19)	(0.23)	(0.28)
0.35	50	50	0.55	0.61	0.66	0.68	0.66	0.61	0.56
			(0.28)	(0.23)	(0.19)	(0.18)	(0.19)	(0.23)	(0.28)
0.45	50	5	0.49	0.56	0.61	0.63	0.61	0.55	0.49
			(0.34)	(0.28)	(0.23)	(0.22)	(0.23)	(0.28)	(0.34)
0.45	50	15	0.49	0.55	0.61	0.63	0.61	0.55	0.49
			(0.34)	(0.28)	(0.24)	(0.22)	(0.23)	(0.28)	(0.34)
0.45	50	25	0.49	0.55	0.60	0.62	0.60	0.55	0.48
			(0.34)	(0.28)	(0.24)	(0.22)	(0.24)	(0.28)	(0.34)
0.45	50	50	0.48	0.55	0.61	0.63	0.61	0.55	0.49
			(0.34)	(0.28)	(0.24)	(0.22)	(0.23)	(0.28)	(0.34)

Notes. For each simulation condition, the numbers in the first line represent ratios *r* based on Equation 11, and the numbers in parentheses in the second line represent type I erros based on Equation 14.

5.2.2. *Biases for Difficulty Parameters*

The estimates for the item difficulty parameters were also monitored to check the influence of manipulated factors on the estimation of item difficulty parameters. Because of the space limitations, the parameter estimates are presented only for the number of groups (nG) 50 in Figure 4. In contrast to the results of standard error estimates, it seemed that the ICC did not affect the estimation of the item difficulty parameters. The figure does not show any systematic pattern, and the parameter estimates remain stable across the values of the ICC. Similarly, no systematic pattern was observed for the number of groups (nG) 100 and 200.

Figure 4. Biases of difficulty parameters in the Rasch model (BILOG, nG=50)



Note. This figure provides a graphical illustration of changes in bias of parameter estimates depending on ICC for the number of groups (nG) 50. The true values of parameters were plotted as horizontal lines.

5.3. *Results for the 2PL and 3PL*

Overall, similar patterns were observed for the 2PL and 3PL models. The standard error estimates decreased as the values of the ICC increased. Also, the estimates for item parameters were stable across different values of the ICC, and no systematic pattern was observed for the bias of parameter estimates. Because of space limitations, only parts of the results are presented in Table 2. The results for the number of groups 100 and 200 also showed similar patterns but are not presented because of the limitation of space.

Table 2. The Ratios and Type I Errors for the 2PL and 3PL Models When *n*G = 50

ICC	Groups	Groups Sizes	2PL (a=1)	2PL (b=3)	3PL (a=1)	3PL (b=3)	3PL (c=0.2)
0.05	50	5	0.61	0.48	0.54	0.75	0.57
			(0.23)	(0.34)	(0.28)	(0.13)	(0.25)
0.05	50	15	0.60	0.47	0.65	0.71	0.72
			(0.23)	(0.34)	(0.20)	(0.16)	(0.15)
0.05	50	25	0.60	0.46	0.69	0.72	0.79
			(0.23)	(0.35)	(0.17)	(0.15)	(0.11)
0.05	50	50	0.60	0.47	0.75	0.75	0.85
			(0.23)	(0.35)	(0.13)	(0.14)	(0.09)
0.15	50	5	0.58	0.44	0.43	0.57	0.59
			(0.25)	(0.38)	(0.39)	(0.26)	(0.24)
0.15	50	15	0.60	0.42	0.50	0.60	0.74
			(0.23)	(0.40)	(0.32)	(0.23)	(0.14)
0.15	50	25	0.56	0.43	0.54	0.57	0.78
			(0.26)	(0.39)	(0.28)	(0.25)	(0.12)
0.15	50	50	0.57	0.44	0.60	0.59	0.81
			(0.25)	(0.38)	(0.23)	(0.24)	(0.10)
0.25	50	5	0.53	0.40	0.34	0.53	0.60
			(0.29)	(0.43)	(0.49)	(0.29)	(0.23)
0.25	50	15	0.54	0.39	0.42	0.51	0.73
			(0.28)	(0.44)	(0.40)	(0.31)	(0.15)
0.25	50	25	0.54	0.39	0.46	0.50	0.76
			(0.28)	(0.44)	(0.36)	(0.32)	(0.13)
0.25	50	50	0.54	0.38	0.50	0.48	0.77
			(0.28)	(0.44)	(0.32)	(0.34)	(0.13)
0.35	50	5	0.51	0.37	0.29	0.46	0.61
			(0.31)	(0.45)	(0.56)	(0.36)	(0.22)
0.35	50	15	0.52	0.37	0.37	0.44	0.71
			(0.30)	(0.46)	(0.46)	(0.38)	(0.16)
0.35	50	25	0.52	0.36	0.40	0.42	0.72
			(0.30)	(0.47)	(0.42)	(0.40)	(0.15)
0.35	50	50	0.52	0.36	0.43	0.41	0.73
			(0.30)	(0.47)	(0.39)	(0.41)	(0.15)
0.45	50	5	0.50	0.37	0.26	0.42	0.61
			(0.32)	(0.46)	(0.60)	(0.40)	(0.23)
0.45	50	15	0.52	0.36	0.33	0.40	0.69
			(0.30)	(0.47)	(0.50)	(0.42)	(0.17)
0.45	50	25	0.52	0.36	0.36	0.38	0.70
			(0.30)	(0.47)	(0.47)	(0.45)	(0.16)
0.45	50	50	0.52	0.36	0.38	0.36	0.68
			(0.30)	(0.47)	(0.44)	(0.47)	(0.17)

Note. For each simulation condition, the numbers in the first line represent ratios *r* based on Equation 11, and the numbers in parentheses in the second line represent type I errors based on Equation 14.

6. SIMULATION STUDY2

6.1. *Simulation Designs*

This simulation study was to design to compare type I error rates of DIF tests using models with and without reflecting multilevel data structure. To do so, data sets were generated using the same setting of the Rasch model in the simulation study 1 with no DIF for a studied item across hypothetical binary groups. In generating data sets, item difficulty parameters were set at (-3, -2, -1, 0, 1, 2, 3), and multilevel structure was implemented with different values of the ICC, which are 0, .05, .15, .25, .35, and .45. Two different kinds of DIF tests were performed across those hypothetical groups using the Lord's chi-square test and the Rasch model formulated in hierarchical generalized linear model (HGLM). The Lord's chi-square tests were performed using parameter estimates and their standard errors estimated from BILOG-MG. Also, another DIF tests were performed based on the Rasch model that was formulated in the hierarchical generalized linear model (HGLM) in which items, persons, and groups are modeled in Level-1, Level-2, and Level-3 respectively.

6.2. *Results*

The results of DIF tests using Lord's chi-square tests and the multilevel Rasch model for the number of groups (nG) 50 are presented in Table 3. In the table, the numbers in the first line of each simulation condition represent type I errors from the Lord's chi-square tests, and the numbers in parentheses in the second line represent type I errors from the multilevel Rasch model. From the table, it can be seen that the type I error rates of the Lord's chi-square tests are inflated up to .270 as the values of the ICC increase, whereas the type I error rates of the multilevel Rasch model remain quite stable close to the nominal level of significance, which is .05.

Table 3. DIF using Lord Chi square Test vs HGLM When *n*G = 50

ICC	Groups	Groups Sizes	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
0.05	50	5	0.005	0.044	0.082	0.064	0.098	0.048	0.008
			(0.026)	(0.025)	(0.051)	(0.054)	(0.053)	(0.039)	(0.022)
0.05	50	15	0.009	0.033	0.075	0.078	0.066	0.066	0.005
			(0.027)	(0.028)	(0.050)	(0.046)	(0.034)	(0.026)	(0.021)
0.05	50	25	0.002	0.043	0.056	0.042	0.058	0.035	0.002
			(0.0274)	(0.031)	(0.053)	(0.053)	(0.048)	(0.033)	(0.022)
0.05	50	50	0.003	0.046	0.062	0.073	0.068	0.049	0.007
			(0.023)	(0.035)	(0.053)	(0.064)	(0.074)	(0.055)	(0.039)
0.15	50	5	0.021	0.116	0.136	0.130	0.114	0.105	0.026
			(0.052)	(0.081)	(0.093)	(0.080)	(0.060)	(0.053)	(0.027)
0.15	50	15	0.012	0.072	0.100	0.112	0.093	0.080	0.013
			(0.034)	(0.039)	(0.076)	(0.075)	(0.054)	(0.049)	(0.042)
0.15	50	25	0.022	0.081	0.121	0.143	0.129	0.096	0.031
			(0.051)	(0.061)	(0.066)	(0.073)	(0.051)	(0.041)	(0.039)
0.15	50	50	0.021	0.094	0.131	0.134	0.121	0.072	0.015
			(0.035)	(0.035)	(0.059)	(0.055)	(0.075)	(0.063)	(0.045)
0.25	50	5	0.028	0.094	0.112	0.130	0.129	0.105	0.020
			(0.022)	(0.025)	(0.041)	(0.055)	(0.042)	(0.045)	(0.030)
0.25	50	15	0.032	0.121	0.167	0.149	0.130	0.109	0.048
			(0.037)	(0.044)	(0.044)	(0.040)	(0.034)	(0.044)	(0.046)
0.25	50	25	0.028	0.132	0.159	0.164	0.145	0.127	0.047
			(0.030)	(0.040)	(0.047)	(0.041)	(0.029)	(0.041)	(0.029)

Table 3. Continues

0.25	50	50	0.038 (0.035)	0.143 (0.046)	0.163 (0.055)	0.164 (0.045)	0.162 (0.041)	0.122 (0.026)	0.045 (0.034)
0.35	50	5	0.069 (0.035)	0.141 (0.045)	0.193 (0.031)	0.200 (0.043)	0.213 (0.048)	0.155 (0.035)	0.068 (0.027)
0.35	50	15	0.049 (0.027)	0.138 (0.040)	0.176 (0.047)	0.176 (0.052)	0.186 (0.054)	0.133 (0.042)	0.055 (0.033)
0.35	50	25	0.065 (0.036)	0.178 (0.041)	0.195 (0.043)	0.228 (0.034)	0.227 (0.044)	0.166 (0.029)	0.088 (0.030)
0.35	50	50	0.090 (0.043)	0.163 (0.044)	0.227 (0.050)	0.230 (0.037)	0.213 (0.033)	0.167 (0.038)	0.063 (0.031)
0.45	50	5	0.182 (0.047)	0.316 (0.078)	0.353 (0.083)	0.341 (0.063)	0.317 (0.047)	0.317 (0.045)	0.197 (0.031)
0.45	50	15	0.118 (0.052)	0.253 (0.051)	0.281 (0.049)	0.258 (0.043)	0.258 (0.051)	0.236 (0.067)	0.120 (0.038)
0.45	50	25	0.137 (0.032)	0.231 (0.044)	0.268 (0.041)	0.298 (0.038)	0.282 (0.043)	0.238 (0.034)	0.116 (0.034)
0.45	50	50	0.131 (0.039)	0.204 (0.048)	0.259 (0.085)	0.270 (0.070)	0.263 (0.049)	0.213 (0.026)	0.137 (0.034)

Note. For each simulation condition, the numbers in the first line represent type I errors from Lord Chi Square tests, and the numbers in parentheses in the second line represent type I errors from HGLM.

7. DISCUSSION

It is well known that applying single-level statistical models to multilevel data may produce underestimated standard error estimates, which in turn result in invalid statistical inferences based on such underestimated standard errors. The goal of this study was to examine the impact of multilevel data structure on the estimation of standard errors in dichotomous IRT models in order to explicitly demonstrate the degree of such underestimation in IRT. Given existing and potential IRT applications in which standard error estimates for item parameters play a crucial role (Toland, 2008), it is important to understand the behavior of the standard error estimation of the IRT models in the presence of multilevel data. Our simulation study showed that the degree of underestimation could be quite huge depending on the values of the ICC. In the most severe case, where the value of the ICC was .45, the standard error estimate from

BILOG-MG was about half of the minimal asymptotic standard error; the type I error rates of the Lord's chi-square tests were inflated up to .35; and the type I error rates of hypothetical *z*-test using Equation 14 were also inflated up to .47. However, the type I error rates of DIF tests using the multilevel Rasch model were close to the nominal level of α, which is .05. Multilevel data structure did not affect item parameter estimates.

The results of this study match those of previous studies. Ignoring multilevel data structure caused underestimated standard errors in regression (Goldstein, 1987) and SEM (Satorra & Muthen, 1995). Barcikowski (1981) also found that even a small amount of the ICC can produce dramatic increases in the actual type I error of a t-test. For example, with the group size of 50, an ICC of .05, which is usually considered small, produced a type I error of .30. In IRT, Finch and French (2011) showed that the type I error of a DIF test using a standard logistic regression can be inflated in the presence of multilevel data structure. In their work, the type I error rate was inflated up to .44 when the value of the ICC was .45. Because the reason for such type I error inflation is the underestimated standard errors, in this study, we wanted to explicitly show the degree of underestimation in IRT settings.

The underestimation of standard errors is caused by the violation of the independent assumption of traditional statistical models. In the presence of multilevel data structure, individuals share common experiences due to closeness in space or time, which makes individuals within the same context more similar to one another. Therefore, observed variance within clusters is typically less than observed variance between clusters. When observations are assumed to be independent, overall variance is calculated without considering the similarity among individuals within clusters, and tends to be underestimated. In fact, as the values of ICC increase, the standard error estimates should increase if the multilevel data structure is properly handled by statistical models (Snijders & Bosker, 1999; Raudenbush, 1997).

Taken all together, the results of this study suggest that ignoring multilevel data structure in the estimation of IRT models could result in underestimated standard errors for item parameter estimates. More importantly, the extents to which standard errors are underestimated are quite huge. Many evidences from previous studies also suggest that standard error estimates in statistical models in general are quite sensitive to multilevel data structure. Therefore, ignoring multilevel data structure could result in invalid statistical inferences in IRT settings. Therefore, researchers who want to use IRT applications in which standard error estimates of item parameters play a crucial role need to check whether their data sets have multilevel data structure or not. In the presence of multilevel structure, traditional single level model could be problematic. Instead, multilevel IRT models are recommended.

ORCID

Hyung Rock Lee https://orcid.org/0000-0002-7415-9466
Sunbok Lee https://orcid.org/0000-0020-0924-7056
Jaeyun Sung https://orcid.org/0000-0001-7461-3123

8. REFERENCES

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational and Behavioral Statistics*, 6, 267–285.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guildford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Mahwah, NJ: Erlbaum.

Finch, W. H., & French, B. F. (2011). Estimation of mimic model parameters with multilevel data. *Structural Equation Modeling*, 1, 229–252.

Goldstein, H. (1987). *Multilevel statistical models*. London: Edward Arnold.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29, 60–87.*

Hox, J. (1998). Multilevel modeling: When and why. *In Classification, data analysis, and data highways* (pp. 147–154). Springer.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, *49*, 82–100.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, 8, 325–352.

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

Kamata, A., & Vaughn, B. K. (2011). Multilevel IRT modeling. *Handbook of advanced multilevel analysis* (pp. 41-57). New York, NY: Taylor and Francis Group.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of lord’s chi-square, raju’s area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8, 291–312.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

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

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 86–92.

Oshima, T., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (dfit) framework. *Journal of Educational Measurement*, 43, 1–17.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer. Satorra, A., & Muthen, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267–316.

Snijders, T. A., & Bosker, R. J. (1999). *Introduction to multilevel analysis.* London: Sage.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publishers.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.

Tate, R. L., & Wongbundhit, Y. (1983). Random versus nonrandom coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, 8, 103–120.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412.

Toland, M. D. (2008). *Determining the accuracy of item parameter standard error of estimates in bilog-mg 3*. ProQuest.

Veerkamp, W. J., & Glas, C. A. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics*, 25, 373–389.

Wright, B., & Stone, M. (1979). *Best test design: A handbook for rasch Measurement*. Chicago: MESA.

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: Scientific Software International*, 4, 10.