교육과정평가연구
The Journal of Curriculum and Evaluation
2012, Vol. 15, No. 2, pp. 241~259

Effects of Within-Group Homogeneity on Parameter Estimation of the Multilevel Rasch Model

Chanho Park
(Associate Research Fellow, KICE)*

≪ ABSTRACT ≫

If a hierarchical structure exists in educational measurement data and examinees within groups are homogeneous, a multilevel item response theory (MLIRT) model may be appropriate. Among the MLIRT models, the multilevel Rasch model is equivalent to a generalized linear mixed model (GLMM) with a logit link where person abilities are considered random effects and item difficulties fixed effects. Then the lme4 package in R can be used to fit the multilevel Rasch model. In this study, it was shown how the multilevel Rasch model can be formulated as a three-level GLMM, followed by a simulation analysis, where intraclass correlation (ICC) of latent abilities as a measure of within-group homogeneity was manipulated from low to high under the conditions of small to large numbers of examinees and items. Item parameter estimates by marginal maximum likelihood estimation (MMLE) were compared with those obtained under the GLMM framework. Biases of item parameter estimates by both methods were not evident in all conditions. However, estimation results by MMLE became proportionally less accurate as the ICC increased when the number of examinees was small. If the number of examinees was large, estimation accuracies of MMLE were acceptable even when a high level of within-group homogeneity existed. GLMM produced stable results at all levels of ICC. In all conditions, the number of examinees was more influential than the number of items.

Key Words : Within-Group Homogeneity, Multilevel Rasch Model, Intraclass Correlation, Generalized Linear Mixed Model, Marginal Maximum Likelihood Estimation

* 제1저자 및 교신저자, cpark@kice.re.kr

Ⅰ. Introduction

In education research, data sets often have a multilevel structure; that is, students are nested within classrooms, classrooms are nested within schools, schools are nested within school districts, and so on. When such a multilevel or hierarchical structure exists in the data, units at a lower level may resemble each other within a higher level group (e.g., classroom or school). For example, students attending the same school share many similarities such as living environments, parents' socioeconomic status, cultural norms, etc. while students attending different schools do not share those similarities. Then students can be homogeneous within groups and heterogeneous between groups.

The intraclass correlation (ICC) is defined as the ratio of between-group variability to the overall variability, which is again decomposed into between-group variability and within-group variability. It is an indicator of between-group heterogeneity for multilevel data. High ICCs are often of interest to educators and policy makers because it may indicate students' achievement gap due to unbalanced distribution of education resources. When a high ICC is observed for a hierarchically structured data set, a multilevel model needs to be considered (Snijders & Bosker, 1999; Raudenbush & Bryk, 2002).

As a hierarchical linear model (HLM) instead of a single-level linear regression can be suited for hierarchical data, a multilevel item response theory (MLIRT) model may replace an ordinary item response theory (IRT) model for educational measurement data when student abilities are more similar within groups (e.g., schools) than across groups. Although at least two levels are assumed for multilevel models such as HLMs, even ordinary IRT models are two-level models because item responses are nested within persons (Kamata, 2001, 2002; Raudenbush & Bryk, 2002). For MLIRT models, it is further assumed that person abilities are nested within higher-level groups. Therefore, MLIRT models have at least three levels.

Since MLIRT is a multilevel extension of IRT models (Adams, Wilson, & Wu, 1997), there can be as many MLIRT models as there are IRT models for dichotomous or polytomous items. Software programs of parameter estimation for multilevel extensions of the two- or three-parameter logistic IRT models for dichotomous items and of some models for polytomous items are available as a package (Fox, 2007) in R (R Development Core Team, 2012) and as a WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) program (Natesan, Limbers, & Varni, 2010), all of which use a Markov chain Monte Carlo (MCMC) estimation algorithm (Albert, 1992; Patz & Junker, 1999a, 1999b; Fox, 2005; Fox & Glas, 2001).

For parameter estimation of the Rasch model (Rasch, 1960), which is equivalent to the one-parameter logistic model (1PLM), and its multilevel extension, the model parameters can also be estimated using the framework of a hierarchical generalized linear model (HGLM; Raudenbush & Bryk, 2002) or a generalized linear mixed model (GLMM; McCulloch & Searle, 2001). Rijmen, Tuerlinckx, De Boeck, & Kuppens (2003) illustrated how IRT models can be understood in a nonlinear mixed model framework. Kamata (2001, 2002) successfully demonstrated that the Rasch model is an HGLM and illustrated how its parameters can be estimated using the HLM software package (Raudenbush, Bryk, Cheong, & Congdon, 2006). Doran, Bates, Bliese, and Dowling (2007) as well as De Boeck et al. (2011) also showed how to fit the multilevel Rasch model using the R package **lme4** (Bates & Maechler, 2010) for mixed-effects models. The **lme4** package can be used to estimate the parameters of the Rasch model by treating the student abilities as random effects and the item difficulties as fixed effects, which also bases the logic of marginal maximum likelihood estimation (MMLE) in IRT (Baker & Kim, 2004; Bock & Aitkin, 1981).

For the Rasch model, parameter estimates by MMLE and those from the GLMM methods are expected to be similar when there are only one group of students. If the students form several groups, however, parameter estimates by MMLE can be less accurate because only one population is generally assumed for MMLE. On the other hand, GLMM can accommodate a grouping level, and, if the grouping variable is correctly specified in the GLMM approach, it can produce more accurate estimates.

In spite of recent advances in MLIRT modeling, a multilevel structure of data is often ignored in practices of educational measurement. Also, it is not quite clear what the actual outcome will be if an ordinary IRT model is used when an MLIRT model fits. For dichotomously-scored test data, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) implementing MMLE is widely used. It is worth investigating how the performance of MMLE is compared with that of the GLMM approach as the level of within-group homogeneity varies under various conditions of sample sizes and numbers of items. In this study, more details of the multilevel Rasch model as a GLMM are discussed, followed by a simulation analysis comparing the two approaches on different levels of within-group homogeneity under the simulation conditions considered.

It is expected that model parameter recovery will somehow be less accurate if existing within-group homogeneity is ignored. However, little is known as to when we should use MLIRT instead of IRT models and how robust MMLE is when within-group homogeneity exists. It is of interest to the practitioners of educational measurement how much within-group homogeneity can be ignored to stick to MMLE since it is now the de facto standard. Although only the multilevel Rasch model is investigated, the results from this study may easily be generalized to other MLIRT models.

Ⅱ. Theoretical Background

1. Intraclass Correlation as a Measure of Within-Group Homogeneity

Even with hierarchically structured data, a multilevel analysis may not be necessary if there is no homogeneity within groups. That is, a multilevel analysis may not improve the results if students are not different across groups. Therefore, not only hierarchical structure but also the level of within-group homogeneity is important when deciding to use a multilevel model. Since the ICC is a measure of the level of within-group homogeneity for hierarchically structured data, it is useful when deciding whether to use a multilevel model (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

For an outcome variable of the hierarchically structured data, the ICC is the ratio of the higher-level variance to the total variance (Snijders & Bosker, 1999). The ICC can be defined using the following random-effects analysis of variance (ANOVA) model:

$$Y_{ij} = \mu + U_j + R_{ij}, \tag{1}$$

where Y_{ij} is the i^{th} observed value in group j, μ is the overall mean, U_j is the group-level random effects, and R_{ij} is the individual-level random effects. As random effects, both U_j and R_{ij} are assumed to be normally distributed with means of zero and variances of σ_U^2 and σ_R^2, respectively. It is further assumed that σ_U^2 and σ_R^2 are uncorrelated with each other. Thus, the total variance consists only of the two variance components. Then the ICC coefficient, ρ, is defined as the ratio of σ_U^2 to the sum of σ_U^2 and σ_R^2.

$$\rho = \frac{\sigma_U^2}{\sigma_U^2 + \sigma_R^2}. \tag{2}$$

Since the ICC is a ratio of variances, it can take only non-negative values although it is called a correlation coefficient. The ICC becomes zero when there is no group-level variance, meaning that groups are not heterogeneous at all. When there is no individual-level variance or when the individual observations are the same within groups while groups differ, the ICC becomes one. As the units in the same groups become more homogeneous, ρ also increases.

2. Multilevel Rasch Model as a GLMM

The multilevel Rasch model as a multilevel extension of the Rasch model (Adams et al., 1997) can be understood as a generalized Rasch model. Adams, Wilson, and Wang (1997) presented the multidimensional random coefficients multinomial logit model (MRCMLM) as a most generalized form of the Rasch model, which may include the multilevel Rasch model as a special case. However, the ConQuest software program (Wu, Adams, & Wilson 1997) implementing the MRCMLM can include only person level random effects (Kamata & Cheong, 2007) and may not be used for the multilevel Rasch model.

When an educational assessment data set was obtained from multiple groups of examinees, the level of achievement by examinees in the same groups is often similar while groups differ. Such between-group heterogeneity as indicated by a high ICC is an issue to be resolved in education because the high ICC may represent education gap due to different resources available to the groups. For example, lack of education opportunities in an underdeveloped region in a country often results in high ICCs.

Different group-level achievements can be modeled by allowing random effects for the group means. Since person abilities in the Rasch or IRT models assume random effects (Baker & Kim, 2004), the multilevel Rasch model can be obtained by having group-level as well as person-level components for the variance of person abilities. The formulation of the multilevel Rasch model is shown as follows. If groups do not exist or can be ignored, the Rasch model has two levels.

At Level 1:

$$p_{ij} = \frac{\exp\left(\theta_j - b_i\right)}{1 + \exp\left(\theta_j - b_i\right)}, \tag{3}$$

where p_{ij} is the probability of correct answer by person j on item i, θ_j is the latent ability of person j, and b_i is the difficulty of item i. The probability is expressed as a logistic function, which is why the Rasch model is equivalent to the 1PLM. Equation 3 can also be expressed in a logit form. Therefore, the Rasch model can be viewed as a GLMM with a logit link (Equation 4).

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \text{logit}\left(p_{ij}\right) = \theta_j - b_i. \tag{4}$$

At Level 2:

$$\theta_j = \mu + \delta_j , \tag{5}$$

where μ is the grand mean, and δ_j is the residual term representing the random effects of person abilities, which is assumed to be normally distributed with a mean of zero and variance of σ^2. Note that no distributional assumptions are necessary for fixed-effect item parameters.

Since latent abilities have no inherent metric, a constraint is required at Level 2 to identify the model. It suffices to fix μ at zero if θ is assumed to follow a normal distribution. In IRT models including the Rasch model, person abilities, θ_j, are generally considered random effects and item difficulties fixed effects (Baker & Kim, 2004; de Ayala, 2009). It is also possible to view both person ability and item difficulty parameters as fixed effects. Then another approach (e.g., fixing the mean of item parameter estimates at zero) is needed for model identification. In this study, however, all IRT models are assumed mixed-effects models.

In order to extend the Rasch model to a three-level model, we need an additional subscript k to accommodate the group level.

At Level 1:

$$\text{logit}\left(p_{ijk}\right) = \theta_{jk} - b_i , \tag{6}$$

where p_{ijk} is the probability of correct answer on item i by person j in group k, θ_{jk} is the latent ability of person j in group k. Since fixed effects are assumed on b_i, its form and interpretation are the same as in the two-level model.

At Level 2:

$$\theta_{jk} = \lambda_k + \delta_{jk} , \tag{7}$$

where λ_k is the ability mean of group k and δ_{jk} is the residual term, which is assumed to follow a normal distribution with a mean of zero and variance of $\sigma^2_{\delta k}$. That is, group variances may differ across groups.

At Level 3:

$$\lambda_k = \mu + \varepsilon_k, \tag{8}$$

where μ is the grand mean and ϵ_k represents variability among the groups with an assumption to follow N(0, σ^2_ϵ).

The three-level model requires more constraints than the two-level model for identification. First, μ is fixed at zero, and then within-group variances are made equal for all groups. That is, the k subscript is dropped for the variance of δ_{jk} (i.e., $\sigma^2_{\delta k}$ becomes σ^2_δ). Then the three-level Rasch model can be fitted as a GLMM.

Although θs are not outcome variables, the ICC of latent abilities can be defined like the following as Equation 2 was derived from Equation 1.

$$\rho = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \sigma^2_\delta}, \tag{9}$$

where σ^2_ϵ and σ^2_δ are between-group and within-group variances of θ. The ICC for θ can be estimated as such because within-group variances were constrained to be equal across the groups. The ICC of θ in Equation 9 shows how much latent abilities are clustered within groups. One of the advantages of using the GLMM framework for the multilevel Rasch model is that the ICC can also be estimated since variances are estimated at both individual and group levels.

3. Parameter Estimation of the Multilevel Rasch Model

In the original framework of the Rasch model, both person ability and item difficulty parameters were considered fixed effects (Rasch, 1960). Then the person ability and item difficulty parameters should be estimated simultaneously, or jointly, and this estimation procedure is thus named joint maximum likelihood estimation (JMLE), which is available in WINSTEPS (Linacre, 2006). Since JMLE is known to produce inconsistent parameter estimates (Baker & Kim, 2004), person ability parameters were removed from calibration by conditioning on sufficient statistics such as the sum score (conditional maximum likelihood estimation) or through integration (MMLE). Among the maximum likelihood estimation (MLE) methods available for the Rasch model, only MMLE was considered in this study because it is comparable to the estimation algorithm for the GLMM.

For the multilevel Rasch model, parameters can be estimated using an MCMC algorithm such as Gibbs sampling implemented in WinBUGS (Spiegenhalter et al., 2003). This study, however, focuses on the multilevel Rasch model as a special case of GLMMs. Thus, estimation of the multilevel Rasch model parameters is considered in the GLMM framework. Even within the GLMM framework, estimation of the multilevel Rasch model is available in other software packages such as HLM (Raudenbush et al., 2006) or SAS NLMIXED procedure (Rijmen et al., 2003), but only estimation in the free statistical computing system R (R Development Core Team, 2012) is used in this study.

Ⅲ. Method

1. Data

The simulation analysis in this study considered conditions in which the data size and the level of within-group homogeneity varied. First, the number of test items varied from small (20 items) to large (60 items). Except for such rare cases as international comparative study in which a matrix sampling design is applied, the number of test items of most testing programs fall within this range. Second, the number of examinees also varied from small (500 examinees) to medium (2,000 examinees) and large (5,000 examinees). The minimum sample size was chosen because 500 examinees were recommended for reliable estimation of the Rasch model parameters (Wright, 1977). Although there are many cases where more than 5,000 examinees take a test, 5,000 examinees were considered to be large enough for precise estimation of the parameters of the multilevel Rasch model. Also, unlike MMLE approaches where data are in a matrix format, data are arranged in a single column vector when a GLMM is applied. When 5,000 examinees take a 60-item test, there are 300,000 cases to be analyzed to estimate parameters. Thus, it may take too long to analyze a larger data set. Third, the ICC coefficient was manipulated from .1 (i.e., very low within-group homogeneity) to .9 (i.e., very high within-group homogeneity) with an increment of .1. Thus, nine conditions were considered for different levels of ICC.

The number of examinees or level 2 units within groups was fixed at 100; thus, only the number of groups varied from five to 20 or 50. Although the number of examinees within groups may influence the accuracy of parameter estimation, this study focused only on the group size and the number of items. The three simulation factors were fully crossed, producing 54 (=2×3×9) simulation conditions. Each of the 54 conditions was replicated 100 times, and MMLE (as the Rasch model) and GLMM (as the multilevel Rasch model) approaches were compared for each replication of the conditions.

The three-level model in Equations 6-8 was used to generate data sets. The *b*-parameters were generated from N(0, 1), and the λ and θ were generated from N(0, σ_{ϵ}^2) and N(0, σ_{δ}^2), respectively. Although σ_{ϵ}^2 and σ_{δ}^2 were manipulated for different levels of ICC, their sum was set at 1.0 to make the item parameter estimates comparable between GLMM and MMLE approaches because BILOG-MG fixes the mean and variance of θ at zero and one, respectively. For example, σ_{ϵ}^2 and σ_{δ}^2 were set at .8 and .2, respectively, for a condition of the ICC being .8.

2. Analysis

The items of the generated data sets were calibrated using BILOG-MG (Zimowski et al., 2003) for MMLE and also using the *lmer* function of the **lme4** R package (Bates & Maechler, 2010) under the GLMM framework. Note that the multilevel structure was not considered for MMLE while the group-level random effects were also estimated for the GLMM. For both MMLE and GLMM estimation, default settings of BILOG-MG and the *lmer* function were used.

After items were calibrated, estimated item parameters were compared with the generating parameters for bias and root mean square error (RMSE) as a measure of estimation accuracy. For each condition, bias and RMSE were calculated for each replication, and they were averaged over the 100 replications. The equations for bias and RMSE were as follows:

$$\text{bias} = \frac{\sum_{i=1}^{I}\left(\hat{b}_i - b_i\right)}{I} \tag{10}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{I}\left(\hat{b}_i - b_i\right)^2}{I}} \tag{11}$$

Unlike JMLE, MMLE is known to produce unbiased item parameter estimates; however, it was of interest if ignoring multilevel data structure introduced any biases in parameter estimates. More interestingly, it was investigated how the accuracy of item parameter estimation (i.e., RMSE of item parameters) was impacted as the level of within-group homogeneity varied.

In this study, only the recovery of item parameters was investigated. Although person ability estimates can be obtained as empirical Bayes estimates, estimation of latent abilities is often conducted in practice using an MLE algorithm after item parameter estimates are confirmed. Since the recovery of person abilities can be confounded by the accuracy of item parameter estimates, only item parameter recovery was examined.

Ⅳ. Results

<Table 1> shows biases of item parameter estimates by GLMM averaged over 100 replications, rounded to the second decimal. It shows that the parameter estimates are unbiased across all 54 conditions.

〈Table 1〉 Bias of Item Parameters Estimated by GLMM

Size*	ICC								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20/500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
60/500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20/2,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
60/2,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20/5,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
60/5,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* number of items/number of examinees

<Table 2> shows that item parameter estimates by MMLE are also almost unbiased. Item parameter estimates are unbiased when there were 5,000 examinees. On the other hand, nonzero biases were sometimes observed for MMLE when the number of simulees was 500 or 2,000. Biases were closer to zero when 60 items were simulated than when only 20 items were used. However, the differences by the number of item conditions were not as big as the differences by the number of simulees.

〈Table 2〉 Bias of Item Parameters Estimated by MMLE

Size*	ICC								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20/500	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	-0.02	0.02
60/500	0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.00
20/2,000	0.00	0.00	0.00	0.01	0.00	0.01	0.01	-0.01	0.01
60/2,000	0.00	0.00	0.00	0.00	0.00	0.01	0.00	-0.01	0.01
20/5,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
60/5,000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* number of items/number of examinees

<Table 3> and <Table 4> show RMSEs of item parameter estimates by GLMM and MMLE, respectively. <Table 3> shows that accuracy of item parameter estimation improves as the number of items and the number of examinees become larger. The effects of the number of examinees were much greater than those of the number of items. RMSEs were close across the nine ICC coefficient conditions in each row.

〈Table 3〉 RMSE of Item Parameters Estimated by GLMM

Size*	ICC								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20/500	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11
60/500	0.10	0.10	0.10	0.09	0.10	0.10	0.10	0.10	0.10
20/2,000	0.05	0.07	0.06	0.06	0.05	0.05	0.06	0.05	0.04
60/2,000	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.04	0.04
20/5,000	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.04
60/5,000	.03	0.03	0.03	0.03	0.04	0.03	0.03	0.03	0.03

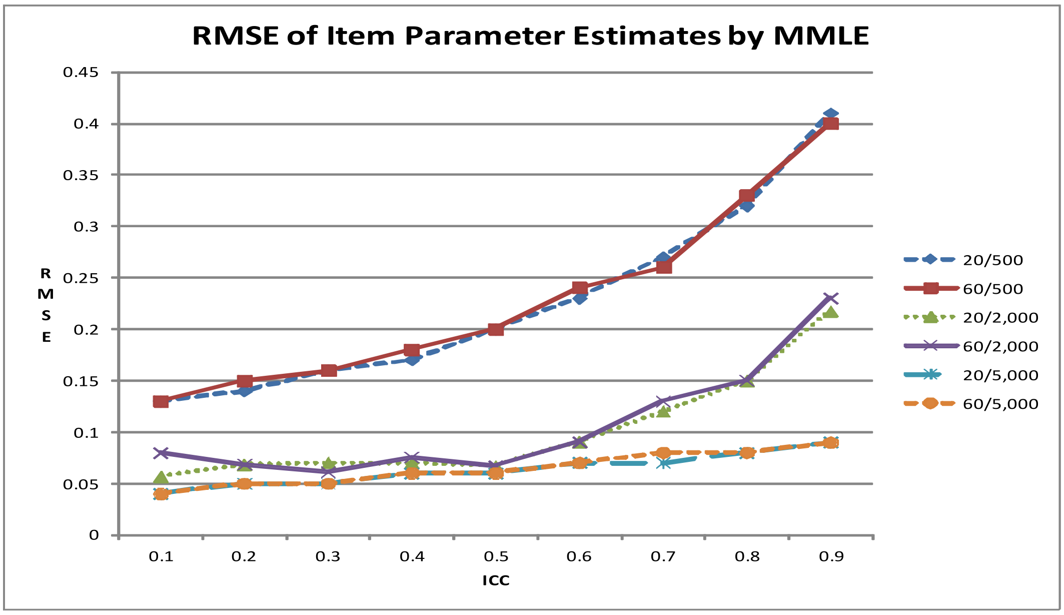
* number of items/number of examinees

〈Table 4〉 RMSE of Item Parameters Estimated by MMLE

Size*	ICC								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
20/500	0.13	0.14	0.16	0.17	0.20	0.23	0.27	0.32	0.41
60/500	0.13	0.15	0.16	0.18	0.20	0.24	0.26	0.33	0.40
20/2,000	0.06	0.07	0.07	0.07	0.07	0.09	0.12	0.15	0.22
60/2,000	0.08	0.07	0.06	0.08	0.07	0.09	0.13	0.15	0.23
20/5,000	0.04	0.05	0.05	0.06	0.06	0.07	0.07	0.08	0.09
60/5,000	0.04	0.05	0.05	0.06	0.06	0.07	0.08	0.08	0.09

* number of items/number of examinees

The most interesting results were RMSEs of item parameter estimates by MMLE shown in <Table 4>, which was also graphically displayed in 〔Figure 1〕. It shows that the number of items did not cause big differences while the number of examinees made a huge difference across the three levels (i.e., 500 or 2,000 or 5,000 examinees). When there were 500 examinees, the RMSEs showed an almost linear and steep increase. On the other hand, as ICC increased, the RMSEs were rather flat and the gap between the 500-examinee condition and the 5,000-examinee condition became larger. When there were 2,000 examinees, RMSEs were located between the two other conditions for the number of examinees (500 or 5,000). Generally, the results by the 2,000 examinee conditions were more similar to those of the 5,000 examinee conditions than those of the 500 examinee conditions. Indeed, RMSEs were quite similar between when there were 2,000 examinees and when there were 5,000 examinees except when the ICC was close to .9. Although not shown graphically, the differences beteen RMSEs by GLMM and MMLE were substantially small when a large number of examinees took the test (see <Table 3> and <Table 4>).



〔Figure 1〕 RMSE of Item Parameters Estimated by MMLE

Ⅴ. Discussion and Conclusion

As achievement gap is found among schools or units larger than schools, students within groups can be more similar within groups than across groups, which justifies the necessity of applying multilevel models when analyzing education data. With educational measurement data, an MLIRT model can be an alternative to ordinary IRT models when the level of within-group homogeneity is high. Among the MLIRT models, the multilevel Rasch model can relatively easily be applied to educational measurement data since it is equivalent to a GLMM.

In this study, first, it was shown how the Rasch model and its multilevel extension could be formulated as two- or three-level GLMMs, and how their parameters could be estimated in a GLMM framework. When item difficulty and person ability parameters were regarded as fixed and random effects, respectively, the Rasch model was equivalent to a GLMM with a logit link. The estimation procedure was also compared with an MMLE algorithm for the Rasch model. Then a simulation analysis was conducted to compare item parameter estimates by MMLE ignoring the multilevel structure of data and those of the GLMM framework accommodating a grouping variable. 54 conditions were simulated where a small/large number of test items were taken by a small/medium/large number of examinees. Also, ICCs of latent abilities were manipulated from low to high within-group homogeneity.

From the results, biases were not evident in both MMLE and GLMM. However, while RMSEs by GLMM were relatively low and rather constant with an increase of ICCs, a substantial decrease was observed as the number of examinees increased from low to high. On the other hand, RMSEs by MMLE were large and almost linearly increased as ICC coefficients became larger when only 500 simulees were used. Yet, RMSEs were small and did not increase much as ICC coefficients became higher when there were 5,000 examinees. When there were 2,000 examinees, RMSEs were similar to those by 5,000 examinees except when the ICC was close to .9.

This study showed that a multilevel analysis should be considered if measurement data were collected from multiple groups, the groups are somewhat heterogeneous, and the size of the data is small (i.e., small number of items and examinees). However, parameter estimates by MMLE, a standard procedure by most test practitioners, are also accurate and dependable if the data size is large. In particular, a large number of examinees are crucial. Practically speaking, MMLE approaches are acceptable when there are more than 2,000 examinees. If between-group heterogeneity is too evident, however, more examinees could be necessary.

The fact that the Rasch model is equivalent to a GLMM has been known to the educational measurement research community (e.g., Kamata, 2001; Raudenbush & Bryk, 2002), and the GLMM framework for the Rasch model provides a powerful estimation program freely available in R. Moreover, the Rasch model as a GLMM can easily accommodate a grouping variable in the model. In this study, the Rasch model and its multilevel extension were formulated as two- and three-level GLMMs. In addition, it was explained how the MMLE for the Rasch model is comparable to the estimation procedure for the Rasch model as a GLMM because item parameters were treated as fixed effects and person abilities as random effects in both estimation algorithms.

MLIRT models have been presented as an alternative model when the ordinary Rasch model does not fit the data. For example, Kamata and Cheong (2007) applied the multilevel Rasch model for reading and mathematics assessment data taken by students nested within schools. Very few, if any, studies investigated how much within-group heterogeneity as indicated by a high ICC can be tolerated by the Rasch model without incorporating the grouping level. Since there are numerous ways to group students in an educational assessment, the grouping structure is often ignored when analyzing data. This study provides a practical guideline for when to adopt and when not to adopt multilevel modeling approaches.

Using the GLMM framework for the multilevel Rasch model provides not only more accurate item parameter estimates but also additional advantages such as model fit indices and estimates of variances at different levels. In the output of the *lmer* function, model fit indices such as Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) are available in addition to deviance statistics. These indices are useful when nested or nonnested models are compared for goodness of fit. Also, the *lmer* function estimates variances of random effects at both individual and group levels. These variance estimates can then be used to estimate the ICC of θ as an indicator of within-group homogeneity.

Although this study guides when to use multilevel models when data are hierarchically structured, it also has limitations. First of all, more simulation conditions could be necessary. The number of items did not cause large variations in this study, but it needs to be confirmed if the number of items is indeed a nonsignificant factor. Also, the number of examinees within groups was fixed in all conditions in this study. Although it is not easy to implement different group sizes since it may exponentially increase simulation conditions, other conditions could be considered in future studies. Second, JMLE can also be compared with the other estimation methods. JMLE was not considered in this study because it is known to produce biased parameter estimates due to its inconsistency problem. However, JMLE is widely used for the Rasch family models (e.g., Linacre, 2006) and may be considered in future studies. Third, since a multi-group approach as implemented in BILOG-MG (Zimowski et al., 2003) can be used for data with multiple groups, we need to investigate if it is another viable option instead of using a GLMM.

Last, only the multilevel Rasch model was considered in this study because similar results were expected for other IRT modes; however, the expectations need to be confirmed. Considering the popularity of the three-parameter logistic model (3PLM) in educational measurement, similar analyses using an MLIRT model as an extension of the 3PLM can be informative to users of the model. Also, the effects of within-group homogeneity on the analysis of polytomously-scored items will be worth while to study.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

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

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

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

Bates, D. & Maechler, M. (2010). *lme4: Linear mixed-effects models using S4 classes. R package* (Version 0.999375-36) [Computer program]. Available from http://CRAN.R-project.org/.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

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

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the *lmer* function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1-28.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software, 20*, 1-18.

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, *58*, 145-172.

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, *20*, 1-16.

Fox, J. P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269-286.

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

Kamata, A. (2002, April). *Procedures to perform item response data analysis by HLM.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Kamata, A. & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Linacre, J. M. (2006). *A user's guide to WINSTEPS and MINISTEP: Rasch-model computer programs*. Chicago, IL.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Natesan, P., Limbers, C., & Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: Formulation and illustration. *Educational and Psychological Measurement*, *70*, 420-439.

Patz, R. J. & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Patz, R. J. & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342-366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded Edition, 1980. Chicago: Chicago University Press)

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2006). *HLM: Hierarchical Linear and Nonlinear Modeling* [Computer program]. Chicago: Scientific Software International.

R Development Core Team (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185-205.

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

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS* (Version 1.4) [Computer program]. Cambridge, UK: MRC Biostatistics Unit.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational Measurement*, *14*, 219-226.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Multi-Aspect Test Software*. Camberwell, Australia: Australian Council for Educational Research.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG* (Version 3.0) [Computer program]. Chicago: Scientific Software International.

·논문접수: 2012-05-01/ 수정본 접수: 2012-06-15/ 게재승인: 2012-06-18

요약

집단 내 동질성이 다층 라쉬 모형 모수 추정에 미치는 영향

박찬호(한국교육과정평가원, 부연구위원)

교육 검사 자료에 위계적인 구조가 존재하고 집단 내 개인이 동질적이라면, 즉 집단 간 이질성이 존재한다면 다층 문항반응이론 모형을 고려할 필요가 있다. 다층 문항반응이론의 모형 중 다층 라쉬 모형은 개인의 능력을 무선 효과로 보고 문항의 난이도를 고정 효과로 보았을 때 로짓을 연결 함수로 하는 일반화선형혼합모형으로 볼 수 있다. 따라서 R의 lme4라는 일반화선형혼합모형 패키지를 이용하여 위계적 구조를 지니는 검사 자료에 다층 라쉬 모형을 적용할 수 있다. 본 연구에서는 먼저 다층 라쉬 모형이 어떻게 3수준의 일반화선형혼합모형으로 볼 수 있는지 공식화하고, 다음으로 모의실험을 실시하였다. 모의실험에서는 피험자와 문항이 작은 수부터 큰 수로 변하는 조건에서 집단 내 동질성을 보여주는 급내상관계수를 작은 값으로부터 큰 값으로 변화시켰다. 주변최대우도추정법으로 추정한 문항 난이도와 일반화선형혼합모형을 이용한 난이도 추정치를 비교하였을 때 편파성은 모든 조건에서 어느 방법도 뚜렷하지 않았다. 그러나 주변최대우도추정법에 의한 모수 추정 결과는 피험자 수가 작을 때 급내상관계수가 커짐에 따라 비례적으로 정확도가 떨어졌다. 피험자 수가 클 때에는 집단 내 동질성이 높은 정도로 존재함에도 불구하고 주변최대우도추정법의 추정치도 좋은 결과를 보였다. 일반화선형혼합모형의 틀을 이용한 문항 난이도 추정치는 급내상관계수의 값에 상관 없이 안정적인 정확도를 보였다. 모든 조건에서 문항 수보다 피험자의 수가 더 큰 영향을 미쳤다.

주제어 : 집단 내 동질성, 다층 라쉬 모형, 급내상관계수, 일반화선형혼합모형, 주변최대우도추정법