Investigating Effect of Ignoring Hierarchical Data Structures on Accuracy of Vertical Scaling Using Mixed-Effects Rasch Model

Shudong Wang
NWEA

Hong Jiao
University of Maryland

Ying Jin
American Institute for Research

Yeow Meng Thum
NWEA

Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). April 30-May 5, 2010, Denver, Colorado.

Send correspondence to:
Shudong Wang
NWEA
5885 SW Meadows Road, Suite 200
Lake Oswego, OR 97035-3256
(503) 624-1951
Shudong.Wang@NWEA.org

Abstract

The vertical scales of large-scale achievement tests created by using item response theory (IRT) models are mostly based on cluster (or correlated) educational data in which students usually are clustered in certain groups or settings (classrooms or schools). While such application directly violated assumption of independent sample of person in IRT, the consequence of such violation is usually ignored in practice. The purpose of this study is to investigate the effect of ignoring hierarchical data structures on the accuracy of vertical scaling by using regular Rasch model and mixed-effect or multilevel Rasch Model.

Introduction

When conducting education research using educational assessment data, such as assessing the long term effectiveness of certain school or state programs for student achievement growth, we not only consider scores of each individual student, but also consider the social structure and environment in which that student is embedded. For example, while reporting individual student achievement growth is one of the most important purposes of K-12 achievement assessments, student growth at different aggregated levels (class, school, school district, and state) can also be looked at. A vertical scale connects forms constructed to assess each grade student performance across different grades and can be used to track growth over time and model changes in student achievement (Kolen & Brennan, 2006). Because of the unique properties of vertical scale, many states and standardized large-scale achievement tests (e.g., TerraNova, Iowa Test of Basic Skill, Stanford Achievement Test, and Measures of Academic Progress) report scores on a vertical scale that allows assessment of student group trends and individual growth in achievement. Using a vertical scale to indicate student progress over a period of time provides vital information about on education from the federal to individual student levels.

Although vertical scale has been widely used, many of the practical concerns in test development, scaling design, data collection, statistical method, and analysis results have to be solved to reduce both systematic and random errors in vertical scaling. In practice, when a vertical scale is constructed, IRT (Hambleton & Swaminathan, 1985) models are often used to fulfill the purpose. The use of any of IRT models is validated only under the assumptions required being satisfied for the mathematics model, for examples, the unidimensionality and local independent assumptions (Lord, 1980). One of the assumptions for equating or scaling standardized achievement tests using IRT is the independence of observations in sample. However, calibrating, equating and scaling are often conducted based on a representative sample selected using cluster sampling or stratified sampling methods. Such sampled data always involve nested data structure where individual students are nested within organizational settings, such as class or school. These dependencies between individuals and clusters cause problems for proper application of IRT.

The correlation within clusters is called intra-class correlation (ICC). The ICCs measure the dependence of the students within groups (class, school, or school district, etc.). It is the ratio of the variance (σ^2) component due to a particular group level to the total variance for individual students. There are two scenarios in discussing ICC of cluster effect of data in this study. If dependent variable can be treated as continue variables such as student test scores, then ICC can be discussed in the context of general linear model framework (Timm & Mieczkowski, 1998); if dependent variable is not continue variable such as student item score that dichotomously scored as zero or one, then ICC can be discussed in the context of generalized linear model framework (McCullagh & Neder, 1989).

Generally, an appropriate context should be given when ICC is used. The "ICC" is unambiguous when we are dealing with the random effects ANOVA – two levels, e.g., students nested within schools. All it refers to is the proportion of total variance attributable to differences in school means. However, in a 3-level setting (e.g., students, classrooms, and schools), proportion of total variance for classrooms must be clearly distinguished from the proportion of total variance for schools. For example (see Appendix A for more details) of student test score as dependent variable, if total variance is expressed in terms of a three-level hierarchical data structure (level-1: student, level-2: class, level-3: school),

$$\sigma^2 = \sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}}$$

then the ICC for students within class is

$$ICC_{\text{level-2}} = (\sigma^2_{\text{level-2}})/(\ \sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}}).$$

The ICC for student within school is

$$ICC_{\text{level-3}} = \sigma^2_{\text{level-3}}/(\ \sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}}).$$

However, if student item score as dependent variable, then the ICCs for IRT model that assume that student ability as a random variable with standardized normal distributed N(0,1) in GLM context can be expressed as in three levels, level-1: item, level-2: student, and level-3: classroom (or school),

$$ICC_{\text{level-2}} = 1/(\ \sigma^2_{\text{level-3}} + 1 + \pi^2/3), \text{ and}$$

$$ICC_{\text{level-3}} = \sigma^2_{\text{level-3}}/(\ \sigma^2_{\text{level-3}} + 1 + \pi^2/3).$$

Because probability of item responses are logistic given item parameters and ability, the individual level variance equal to $\pi^2/3$ (Goldstein, Browne, Rashash, 2002; Rashash, Steele, Browne, 2003; Snijders, Basker, 1999) or ≅ 3.29. It can be seen, in the traditional IRT calibration context, one may argue that between-examinee ICC is not zero but is fixed by default at 1.0/(1.0+ π^2/3) since the distribution of ability is assumed to be distributed normal with mean 0 and variance 1.0. The argument to be made here is that, in practice, the distribution of clusters of students is likely to be distributed with non-zero mean and variance. If students are sampled in clusters, the assumption of independence among students, a necessary condition for IRT, is violated.

Many published literatures have discussed the ICC issues in statistical field (Cochrane, 1977; Cornfield, 1978; Kish, 1965; Walsh, 1947) and medical field (Donner & Koval, 1980, 1983; Rosner, 1984; Munoz, Rosner, & Carey, 1986). Few studies have examined the dependence nature of educational data in large scale achievement context. Schochet (2005) summarized some studies on estimation of ICC for different standardized tests (Table 1). Wang (2006) conducted study on the effect of cluster data at test score level on sample size requirement for IRT calibration. Partial results of ICC for different subject areas of standardized achievement tests are listed in Table 2. Besides the effect of large-than-zero ICC large on accuracy of IRT parameter estimates, Wang pointed out that the sample sizes of equating and calibration used by states and testing vendors were, sometimes, much smaller than what IRT models would required. The degree of reduction in sample size is measured by the design effect (Deff)(Kish, 1965) or a correction factor that correct variance inflation caused by within cluster correlation for simple random samples (SRS) and cluster samples (CS):

$$\text{Deff} = \frac{\text{Variance (CS)}}{\text{Variance (SRS)}} = 1 + (\text{average cluster size -1}) * \text{ICC}.$$

It can be seen that all achievement tests show certain degree of dependence in samples. Ignoring cluster nature of educational data in applying IRT model could lead to biased parameter estimates and misleading results. Because Rasch (Rasch, 1960) model is one of the most commonly used IRT models in current achievement tests, this study used Rasch model as an example to illustrate the problems neglecting the cluster data structure in regular vertical scaling and their solutions.

The problems of mistaking CS as SRS can be coped with using multilevel models (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 1986; Goldstein, 1995; Longford, 1993; Raudenbush, 1988). Some researchers (Adams, Wilson, & Wu, 1997; Kamata, 2001;

Mellenbergh, 1994; Mislevy & Bock, 1989) have shown that IRT models can typically be treated as logistic mixed models. Mislevy and Bock (1989) applied multilevel modeling in the framework of IRT models where group-level and student-level effects were combined in a hierarchical IRT model. Adams et. al. (1997) showed that latent ability could be used as outcomes in a regression analysis. They showed that a regression model on latent ability variables could be viewed as a two-level model where the first level consisted of the item response measurement model which served as a within-student model and the second level consisted of a model on the student population distribution, which served as a between-students model. Fox and Glas (2001) introduce a general approach for binary outcomes in a strictly clustered setting (i.e., items are nested within students and students are nested within schools). This general approach entails a multilevel regression model on the latent ability variables allowing predictors on the student-level and group-level.

Many of these developments fall under the rubric of generalized linear mixed model (GLMM, McCulloch & Searle, 2001), which extend generalized linear models (GLM, includes logistic regression) by the inclusion of random effects in the predictor. Recently, Rijmen, Tuerlinckx, De Boeck, & Kuppens (2003) provide a comprehensive overview and bridge between IRT models, multilevel models, mixed models, and GLMMs. According to them, only the Rasch model (RM, Rasch, 1960) and family of Rasch models, such as *linear logistic test model* (LLTM, Scheiblechner, 1972; Fischer, 1973), the *rating scale model* (RSM; Andrich, 1978), the *linear rating scale model* (LRSM; Fischer & Parzer, 1991), the *partial credit model* (PCM; Masters, 1982), the *linear rating scale model* (LRSM; Fischer & Ponocny, 1994), and the *mixed Rasch model* (Rost, 1999), belong to the class of GLMMs. Other IRT models, such as two- and three-parameter models are not within the class of GLMMs because they include a product of parameters and no longer linear. Rasch family models have the following common properties: sufficiency of the raw scores, parallel item characteristic curves, specific objectivity, and latent additivity. Traditional RM is still widely used in education testing in which its assumptions are often violated because students are clustered with classes and schools. The mixed-effect (or multilevel) Rasch model (MERM) that explicitly recognize the clustered nature of the data and directly incorporate random effects to account for the various dependencies is used in this study. Because MERM is a special case of the mixed-effects logistic regression model (MELRM), here we present MELRM and show how it relates to MERM.

MELRM is a common choice for analysis of multilevel dichotomous data (that has value 0 or 1). The major differences between GLMM and general linear model are in two aspects. First, the distribution of dependent variable in GLMM can be non-normal, and does not have to be continuous. Secondly, dependent variable in GLMM still can be predicted from a linear combination of independent variable(s), but they are "connected" via a link function. In the GLMM context, this model utilizes the logit link, namely (De Boeck & Wilson, 2004)

$$g(\mu_{ij}) = \log it(\mu_{ij}) = \ln\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] = \eta_{ij} = \sum_{k=0}^{K}\beta_k X_{jk} + \sum_{l=0}^{L}\theta_{il} Z_{jl} \qquad (1)$$

where i for person, i=1,2,…, I; j for item, j=1,2,…, J; k for item predictors, k=0, 1,…, K; l for person predictors, l=1,2,…,L. X_{jk} is the value of predictor k for item j; Z_{jl} is the value of predictor l for item j; β_k is the fixed regression weight of predictor k and θ_{il} is the random regression weight of predictor l for person i. The equation (1) can be expressed in matrix notion

$$\eta_i = \mathbf{X\beta} + \mathbf{Z\theta_i}\,, \qquad (2)$$

here **X** is a $J \; x \; K$ design matrix for fixed effects; $\boldsymbol{\beta}$ is a $K \; x \; 1$ vector of fixed regression weights; Z is a $J \; x \; L$ design matrix for random effects and θ is $L \; x \; 1$ vector of random regression weights for person. η_{ij} is linear predictor, the conditional expectation $\mu_{ij} = E(Y_{ij} \mid \mathbf{X}, \boldsymbol{\beta}, \mathbf{Z}, \boldsymbol{\theta})$ equals $P(Y_{ij} = 1 \mid \mathbf{X}, \boldsymbol{\beta}, \mathbf{Z}, \boldsymbol{\theta})$, namely,

$$P(Y_{ij} = 1|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\theta}) = g^{-1}(\eta_{ij}) = \Psi(\eta_{ij}) \qquad (3)$$

the conditional probability of a response given the random effects (and covariate values if there is any one) and Y_{ij} is observations. where the inverse link function $g^{-1}(\eta_{ij})$ or $\Psi(\eta_{ij})$ is the logistic cumulative distribution function (cdf), namely $\Psi(\eta_{ij}) = [1 + \exp(-\eta_{ij})]^{-1}$.

RM gives the probability of a correct response to the dichotomous item j ($Y_{ij} = 1$) conditional on the random effect or ‘ability’ of subject i (θ_i):

$$p(y_{ij} = 1|\theta_i) = \Psi(\eta_{ij}) = \Psi(\theta_i - b_j) = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)} \qquad (4)$$

where b_j is the difficulty parameter for item j. Comparing (1) to (4), it can be seen that RM is special case of a random-intercepts model that includes item dummies as fixed regressors and here we call it MERM. The assumption of local independence means that, for a given test, the probabilities of given items (*j=1, 2, …,J*) for one person can be jointly determined by

$$p(Y = y|\theta) = \prod_{j=J}^{J} p(y_j|\theta) = p(y_1|\theta)p(y_2|\theta)...p(y_J|\theta) \qquad (5)$$

or the probabilities of given persons (*i= 1,2,…, I*) for one item *j* can be jointly determined by

$$p(Y_{ij} = y_{ij}|\theta) = \prod_{i=1}^{I} p(y_{ij}|\theta_i) = p(y_{1j}|\theta_1)p(y_{2j}|\theta_2)...p(y_{ij}|\theta_I) \qquad (6)$$

for one item. Cluster sample used in RM directly violates the assumption local independence assumption from person perspective, i.e., equation (6).

Though IRT models were not originally cast as GLMMs, formulating them in this way easily allows covariates to enter the model at either level (i.e., items or subjects). Kamada (2001) formulated MERM in the context of multilevel model (multilevel RM) within GLMM framework. One of multilevel RMs he proposed is three-levels Rasch model (item, student, classroom or school):

Level 1 (Item-Level) Model:

$$\begin{aligned} log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) &= \eta_{ijm} \\ &= \beta_{0jm} + \beta_{1jm} X_{1jm} + \beta_{2jm} X_{2jm} + \ldots + \beta_{(k-1)jm} X_{(k-1)jm} \\ &= \beta_{0jm} + \sum_{q=1}^{k-1} \beta_{qjm} X_{qjm} , \qquad (7) \end{aligned}$$

where i = 1,2,…,k-1for items, j = 1,2,…,n for students, m = 1,2,…,r for class. p_{ijm} is the probability that person j in class m answers item i correctly and X_{qijm} is qth dummy variable (q = 1,2,..,k-1) for the ith item for person j in class m. β_{0jm} is the effect of the reference item, and β_{qjm} is the effect of the qth item compared to the reference item.

Level 2 (student-Level) Model:

The student-level models for student j in class m are written as

$$\begin{aligned} \beta_{0jm} &= \gamma_{00m} + u_{0jm} \\ \beta_{1jm} &= \gamma_{10m} \\ \beta_{2jm} &= \gamma_{20m} \\ &: \\ \beta_{(k-1)jm} &= \gamma_{(k-1)0m}, \end{aligned}$$

where u_{0jm} ~$N(\gamma_{00m}, \tau_\gamma)$ and τ_γ, the variance of u_{0jm} within class m is assumed to be identical across classes.

Level 3 (class-Level) Model:

In this model, the intercept γ_{00m} is only term that arises across classes and item effects are constant across classes. For class m,

$$\begin{aligned} \gamma_{00m} &= \pi_{000} + r_{00m} \\ \gamma_{10m} &= \pi_{100} \\ \gamma_{20m} &= \pi_{200} \\ &: \\ \gamma_{(k-1)0m} &= \pi_{(k-1)00}, \end{aligned}$$

where r_{00m} ~$N(0, \tau_\pi)$. The combined model is

$$\begin{aligned} log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) &= \eta_{ijm} = (\gamma_{00m} + u_{0jm}) + (\gamma_{i0m}) \\ &= \pi_{000} + r_{00m} + u_{0jm} + \pi_{i00} = (r_{00m} + u_{0jm}) - (-\pi_{i00} - \pi_{000}). \qquad (8) \end{aligned}$$

The probability that person *j* in class m answers item *i* correctly is

$$p_{ijm} = \frac{1}{1+exp[-\eta_{ijm}]} = \frac{1}{1+exp\{-[(r_{00m}+u_{0jm})-(-\pi_{i00}-\pi_{000})]\}}. \qquad (9)$$

In current educational testing, little work has been done on using MERM to create vertical scales which are important in determining student achievement growth. As a matter of fact, a few attempts tried to investigate the cluster effect on vertical scale. The purpose of this study is to investigate the effect of ignoring hierarchical data structures by using RM on the accuracy of vertical scaling by using MERM in GLMM framework.

Methods

Because true cluster sampling effect is not known in practical setting, the Monte Carlo (MC) technique was used to investigate effect of ignoring cluster data structures on vertical scale by using RM and MERM. All simulation data were generated based on these models.

The simulated vertical scales were constructed across grade 4 to grade 5 using a common-person design which set up the linking in a vertical scale by using both below-grade and on-grade items to link adjacent grades (Table 3). Test length is 40 items across grades and sample size is 1000 persons per grade. To simulate the true growth patterns, two factors were manipulated with multiple levels: the mean of the ability distribution and the standard deviation (SD) of the ability distribution. The combination of these two factors determines the simulated growth patterns of student achievement.

First, for all simulations, three fixed tests across grades with 40 items per grade were used. All items were generated from *N*(*M, SD*) and once they were generated, then applied to all simulation conditions. Table 4 lists true distribution parameters that generate true item difficulties across grades.

Second, two types of samples were generated: simple random samples (SRS) and cluster samples (CS) with different ICCs (0.2, 0.3, and 0.4) for each grade. For both samples, there are 1000 observations per grade. The SRS are from *N*(*M*, *SD*) For CS, data for student in classroom/school were simulated using a multilevel model without explanatory variable at both person and classroom or school levels. 20 replications will were conducted across all conditions generated from distributions listed in Table 4.

Simulation Procedure

Two types of data set were simulated by using RM and MERM. Given parameters defined by the specifications mentioned above, the steps involved in the simulation process are as follows:

Step 1: Two samples of 1000 simulated examinees (simulees) of true abilities were generated from standard normal distribution; One is for RM, the other is for MERM. For MERM, the correlated data were generated with three different ICC values (0.2, 0.3, 0.4) under the assumption that the average cluster (class or school) size was 25 and the number of clusters was 25.

Step 2: The known item parameters (table 4) were used to calculate the probability of each simulee for both RM and MERM.

Step 3: The generated probabilities (P) were compared to a uniform (0,1) random number (RN), if the P > RN, the simulee was given a correct score (1), otherwise an incorrect score (0).

The whole process was repeated 20 times. Different random seeds were used as responses data were calibrated using the WINSTEPS and HLM, respectively. The estimated and true parameters were compared in terms of five dependent variables (see following section).

Calibration Methods

Two different software packages WINSTEPS and HLM, were used for calibration in this study. These software packages are based on different estimation methods. WINSTEPS, is the most widely-used Rasch software (Association of Test Publishers, 2001). It pays little attention to the estimation of effective standard errors for Rasch models, especially under the complex sample designs typically found in state testing programs (Cohen, Chan, Jiang, & Seburn, 2008). HLM6 software (Raudenbush, Bryk, Cheong, & Congdon, 2004) estimates model coefficients at each level of hierarchical data.

We initially also attempted to use the AM software (American Institutes for Research & Jon Cohen, 2005) to calibrate the simulated data. AM software produces IRT item parameters and uses nonparametric marginal maximum likelihood (NPMML) approach to estimate examinee proficiency on the theta scale. AM provides appropriate standard errors for multistage complex samples using a Taylor-series approximation. Cohen et al. (2008) claimed that NPMML results in more consistency of the estimators in both simple random samples and more realistic multistage samples than conditional maximum likelihoods (CML, Rasch, 1961) used in WINSTEPS. However, currently AM's IRT Model Simulation procedures do not provide person ability estimates. Running AM in the interactive mode for each replication of the simulated data separately would be too time-consuming. Therefore discussion of AM from AM was not included in this study.

The effect of ignoring correlated data is evaluated in terms of nature of data (i.e., SRS or CS) and models (or software) used (i.e., RM or MERM). RM was calibrated by using WINSTEPS and MERM was calibrated by using HLM. Table 5 depicts the design of calibration procedure.

Scaling

All vertical scaling was developed by separate calibrations for each grade score first, then adjacent grades was linked by using linear transformation such as mean/mean method (Kolen & Brennan, 2004). However, at each grade, concurrent calibration method was used for both off-grade and on-grade student responses to a total 80 items. Because each grade has 40 anchor items, here we use grade subscript indicate test form. Let θ and b on grade *K* be linearly transformed onto grade *K*+1 by

$$\theta_{(K+1,i)} = \theta_{(K,i)} + B, \qquad (10)$$

$$b_{(K+1,j)} = b_{(K,j)} + B, \qquad (11)$$

where $\theta_{(K+1,i)}$ and $\theta_{(K,i)}$ are ability θ values for individual *i* on Scale *K+1* and Scale *K*. $b_{(K+1,j)}$ and $b_{(K,j)}$ is item difficulty *b* values for item *j* on Scale *K+1* and Scale *J*. B is a constant in a linear transformation equation (intercept) for Rasch model and B is:

$$B = \mu(b_{K+1}) - \mu(b_K) = \mu(\ \theta_{K+1}) - \mu(\ \theta_K) \qquad (12)$$

Where $\mu(*_{\text{K+1}})$ and $\mu(*_{\text{K}})$ are either the mean of item or the mean of person parameters at grade level *K*+1 or *K*.

Evaluation of Recovery of Results

There are in total 3 ICCs (i.e., 0.2, 0.3, 0.4) x 2 software (WINSTEPS and HLM) x 3 grades (4, 5, and 6) = 18 simulation conditions. An another 3 independent WINSTEPS runs of grade 4, and 5, and 6 for ICC=0 data were not included in the total run.

The bias, SE, RMSE, and correlation between true and estimated parameters are used to evaluate how well true parameters are recovered for each of 6 simulation conditions. These formulas are,

$$Bias(\hat{\theta}_g) = \frac{1}{N_R}\sum_{r=1}^{N_R}\left(\frac{\sum_{i=1}^{N_p}\hat{\theta}_{gri}}{N_p} - \bar{\theta}_g\right), \qquad (11)$$

$$SE(\hat{\theta}_g) = \sqrt{\frac{1}{N_R}\sum_{r=1}^{N_R}\left(\frac{\sum_{i=1}^{N_p}\hat{\theta}_{gri}}{N_p} - \frac{1}{N_R}\frac{1}{N_p}\sum_{r=1}^{N_R}\sum_{i=1}^{N_p}\hat{\theta}_{gri}\right)^2}\ , \text{ and} \qquad (12)$$

$$RMSE(\hat{\theta}_g) = \sqrt{\frac{1}{N_R}\sum_{r=1}^{N_R}\left(\frac{\sum_{i=1}^{N_p}\hat{\theta}_{gri}}{N_p} - \bar{\theta}_g\right)^2}\ , \qquad (13)$$

where *i*, *g*, and *r* represent individual, grade, and replication, respectively. $\hat{\theta}_{gri}$ is the estimated person parameter for grade *g*, replication *r*, and person *i*. $\bar{\theta}_g$ is the mean of the generated true students' abilities in grade *g*. N_p is the number of simulated examinees and N_R is the number of replications of the simulation. Different random numbers are used as seeds for each of the 20 replications. Based on a past research suggestion (Harwell, Stone, Hsu, & Kirisci, 1996), both descriptive and inferential procedures were used to summarize the simulation results.

Results

(1) Recovery of Person and Item Parameters

Tables 6 and 7 summarize the recovery indexes (Bias, SE, RMSE, and correlations) for ability and difficulty parameters used to evaluate the calibration accuracy in different simulations. The effect of two independents variables (ICC and model) on dependent variables (parameter estimates) were analyzed using two-way analysis of variance (two-way ANOVA) for both ability and difficulty. Since the study focus on the effect of nature of data (i.e., different ICC conditions) and model (RM and MERM) on IRT estimates of item and person parameters. The results of ANOVA of means of both ability and item difficulty parameter estimations across grades are presented in Tables 8 and 9. If the overall *F*-test is statistically significant, then it means that model accounts for a significant amount of variation in the dependent variables. In general, for both ability and item parameter estimates, the means of estimates across grades are statistically significantly affected by both model and ICC factors except for grade 4 mean estimate of item difficulty. R^2 indicates the percent of the variance in the dependent explained uniquely or jointly by the independents. R^2 can also be interpreted as the proportionate reduction in error in estimating the dependent when knowing the independents. For example, for the given the simulated conditions used in this study, for grade 5 ability estimates, the model and ICC account for about 90% total variance in mean estimates of ability; while the model and ICC account for about 15% total variance in mean estimates of item difficulty. From Tables 8 and 9, it is clear that model and ICC account for more variances in ability estimates than variance in item difficulty estimates, which is not surprising for the given the fact that the focus of this study is on person side of data, not on item side of data. Results show that cluster data or correlated data and models used to fit the data have significant impact on the accuracy of both person and item parameters estimates.

Besides, the effect of independent variables on overall dependent variables and the effect of each of the independent variables on each of the dependent variables were also tested. The main effect of model and ICC factors and the interaction effect between model and ICC are also shown in Tables 8 and 9. For ability estimate, the model factor has statistically significant effect on mean of ability estimates across grades; neither the ICC factor or the interaction between model and ICC factors has statistically significant effect on ability estimates. This means that, for the given simulation conditions, whether using a model that account for correlated data or not has significant impact on the estimated ability parameter, although the degree of correlation (ICC=0.2, 0.3, or 0.4) in the correlated or cluster sample may not be matter. This result implies that the correlated or clustered data should not be treated as if they were random independent data when applying IRT model and consequences of violating assumption of IRT models should not be neglected when applying IRT model in educational setting.

Figures 2 to 5 show the Bias, SE, RMSE, and correlation for ability estimates under different ICC conditions (I1=0.2, I2=0.3, I3=0.4) and models (MR= calibrating clustered data using MERM, R-MR=calibrating clustered data using RM) across grades. The results prove that, overall, MERM model recovers the true ability estimates better in the clustered or correlated data than RM does across grades. For example, the RMSE for MR is much lower than that of R-MR across grades and ICCs. The correlations between true and estimated ability parameters for MR are higher than those for R-MR across ICCs and grades.

Although there appears to be no clear-cut result, the general trend suggests that the degree of clustering or correlating indicated by ICCs has some impact on accuracy of ability estimates across grades. For example, for grade 4, the biases of ability estimates of MR for all ICCs are smaller than those of R-MR; for grades 5 and 6, however, the biases of ability estimates of MR for ICC=0.3 and 0.4 are smaller than those of R-MR, but this is not true for ICC=0.2 condition.

The Bias, SE, RMSE, and correlation for item estimates under different ICC conditions (I1=0.2, I2=0.3, I3=0.4) and models (MR= calibrating clustered data using MERM, R-MR=calibrating clustered data using RM) across grades are depicted in Figures 6 to 9. Overall results show that accuracies of true parameter recovery for MR conditions are better than those for R-MR conditions across ICCs and grades. The differences of correlations between true and estimated *item* parameters are, however, not as significant in comparison with much the differences of correlations between true and estimated *ability* parameters.

(2) Recovery of Vertical Scale

Table 9 and Figure 10 present results of vertical scale recovery under different ICC conditions (I1=0.2, I2=0.3, I3=0.4) and models (MR= calibrating clustered data using MERM, R-MR=calibrating clustered data using RM) across grades. It can be seen that the effect of treating clusters data or correlated data as if they were random independent data when conducting vertical scaling is not ignorable in terms of vertical scaling accuracy.

Discussion and Summary

In developing methods of measuring educational growth for the purposes of accountability and AYP, a myriad of decisions must be made to deal with many aspects of educational growth. Among these decisions are the methods for demonstrating educational achievement from year to year. A vertical scale could serve such a purpose. If a vertical scale is chosen to express yearly progress, the design and process to develop the scale must be considered. How accurately a vertical scale reflects the ‘true’ AYP is a crucial issue when informing high-stake decisions. Currently, most of the vertical scales created by states or test companies are based on IRT models and the most frequently used IRT model is RM. However, the effect of clustering sampling is usually ignored in the practical applications of IRT (e.g., equating and scaling) in educational setting, where the data are usually clustered (student within classroom, classroom within school, school within school district, etc.). The consequences of treating CS data as SRS data in current vertical scaling practices result in (1) reducing effective sample size for IRT model, and (2) increase vertical scaling errors.

It is very hard, and sometimes impossible, to draw simple random samples of individual students without interfering with their normal learning process during school hours. The education law and regulation do not allow samplers to cherry pick students across classroom, school, school district, and state. One way to get around this problem is to collect a sample as large as possible and then conduct random sampling on individual students from the large sample. Unfortunately, most vendors and states don’t have the budget to collect large-enough sample to do that. Besides, to waste the rest of student data for the purposes of equating and scaling is not likely to be allowed by state and testing vendor for the same concerns. So the chance of improving sampling quality is practically slim if not impossible. However, continuing ignoring the consequences of violating IRT model assumptions in educational applications such as vertical scaling is unacceptable because actions that come out of such applications, too often, are high-stake decisions. One of the viable solutions, if there are many, is to use models such as MERM that can account for the clustering nature of the data and produce more accurate results instead of avoiding or ignoring the cluster effect of samples during the process of vertical scale development.

References

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

Association of Test Publishers. (2001). Test publisher 8.2 [Computer software]. Washington, DC: Author.

Bates, D. (2007). Linear mixed model implementation in lme4 Package. URL: ftp://ftp.auckland.ac.nz/pub/software/CRAN/doc/vignettes/lme4/Implementation.pdf.

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

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

Cochran, W. (1977). *Sampling techniques*. New York: John Wiley and Sons.

Cohen, J. Chan, Z., Jiang, T., & Seburn, M. (2008). Consistent Estimation of Rasch Item Parameters and Their Standard Errors Under Complex Sample Designs. *Applied Psychological Measurement, 32*, 289-310

Cohen, J., & the American Institutes for Research. (2002). AM statistical software (Beta version 0.06.00) [Computer software]. Washington DC: American Institutes for Research. Available from http://am.air.org

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, v108*, 2.

CTB/McGraw-Hill. (1997). *Winter norms book: TerraNova*. Monterey, CA: Author.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

de Leeuw, J., & Krefl, I.G.G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics, 11*, 57-86.

Donner, A., & Bull, S. (1983). Inferences concerning an intraclass coefficient in the one-way random effects model. *Ont. Statist. Rev. 54*, 67-82.

Donner, A., & Koval, J.J. (1980). The Estimation of Intraclass Correlation in the Analysis of Family Data. *Biometrics, 40*, 393-408.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the Multilevel Rasch

Model: With the lme4 Package. *Journal of Statistical Software, 20,* Issue 2.

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

Goldstein, H. (1995). *Multilevel statistical models,* 2nd Edtion. London: Arnold.

Goldstein, H., Browne, W., Rashash, J. (2002)

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer.

Harcourt Educational Measurement. (2002). Stanford Achievement Test, 10th Edition. San Antonio, Texas.

Hoover, H. D., Dunbar, S. D., & Frisbie, D. A. (2003). *The Iowa Tests of Basic Skills. Interpretive guide for teachers and councelors. Forms A and B. Levels 9-14. Itasca, Il*: Riverside Publishing.

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

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

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: methods and practices* (2nd ed.). New York: Springer.

Longford, N.T. (1993). *Random coefficient models*. New York, NY: Oxford University Press.

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

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300–307.

McCullagh P, Nelder J (1989). *Generalized Linear Models*. Chapman and Hall, 2nd edition.

McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models,* Wiley, New York.

Munoz, A., Rosner, B. & Carey, V. (1986). Regression analysis in the presence of heterogeneous intraclass correlations. *Biometrics, 42*, 653-58.

Mislevy, R.J., & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D. Bock (Eds.), Multilevel analysis of educational data (pp. 57-74). San Diego, CA: Academic Press.

NWEA (2003). *Technical manual for use with Measures of Academic Progress and Achievement Level Tests*. Portland, OR: Northwest Evaluation Association.

Pituch, K. A. (1999). Describing school effects with residual terms: Modeling the interaction between school practice and student background. *Evaluation Review, 23*(2), 190-211.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute of Educational Research, Copenhagen.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*, 85-116.

Rosner, B. (1984). Multivariate methods in opthalmology with application to other paired-data situations. *Biometrics, 40*, 1025-35.

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

Roberts, J. S. & Ma, Q. L. (2006). *IRT Models for the assessment of change across repeated measurements*. In R. W. Lissitz (Ed.), Longitudinal and value added models of student performance (pp.100-127). Maple Grove, MN: JAM Press.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Schochet, P. (2005). Statistical power for random assignment evaluations of education programs. Mathematic Policy Research, Inc. Princeton, NJ.

Timm, N. H. & Mieczkowski, T. A. (1997). *Univariate & multivariate general linear models theory and applications using SAS software*. SAS Institute (Cary, NC)

Walsh, J. (1947). Concerning the effects of the intra-class correlation on certain significance tests. *Annals of Mathematical Statistics,* v18.

Wang, S. (2006). *Brief study of impact of equating sample size on measurement error for catalog products. Research report*. Harcourt Assessment Inc.

Table 1. Summary of Intra-class Correlation (ICC) Estimates of Standardized Tests Based on Schochet (2005) Study.

Data Source	Standardized Test Used	Description of Data	Grade and Year	Average ICC Estimates
Longitudinal Evaluation of School Change and Performance	Stanford Achievement Test (Version 9)	71 Title I schools in 18 school districts in 7 states	3 grade in 1997 4 grade in 1998 5 grade in 1999	.18
21st Century Community Learning Centers Program	Stanford Achievement Test (Version 9)	30 schools in 12 school districts	1, 3,and 5 grades in 2002	.18
Test for America Evaluation	Iowa Test of Basic Skill (ITBS)	17 schools in six cities	2 and 4 grades in 2003	.12
Prospects Study: Figures Reported in Hefberg et al. (2004)	Comprehensive Test of Basic Skills (CTBS)	327 Title I schools in 120 school districts	3 grade in 1991	.22

Table 2. Intra-Class Correlations of Large-Scale Standardized Test across Content Areas and Grades (Class Sizes from 10 to 40)

Grade	Total Reading	Word Study Skill	Reading Comprehension	Mathematics/ Total Mathematics	Mathematics Problem Solving	Mathematics Procedures	Language	Spelling	Environment	Science	Social Science
1	0.28	0.28	0.22	0.24	0.24	0.21	0.26	0.25	0.29		
2	0.29	0.26	0.32	0.25	0.26	0.21	0.25	0.16	0.25		
3	0.17	0.14	0.28	0.20	0.17	0.23	0.16	0.09	0.29	0.18	0.20
4	0.31	0.28	0.32	0.30	0.30	0.28	0.25	0.20	0.25	0.27	0.30
5	0.29		0.31	0.27	0.25	0.26	0.25	0.16	0.29	0.30	0.30
6	0.29		0.37	0.31	0.30	0.30	0.22	0.16	0.25	0.24	0.27
7	0.33		0.21	0.34	0.28	0.39	0.28	0.18	0.29	0.27	0.26
8	0.37		0.22	0.36	0.33	0.37	0.30	0.26	0.25	0.29	0.31
9	0.20		0.33	0.18			0.20	0.14	0.29	0.25	0.22
10	0.28		0.32	0.24			0.31	0.19	0.25	0.22	0.20
11	0.43		0.22	0.27			0.38	0.34	0.29	0.31	0.25
12	0.38		0.32	0.21			0.36	0.25	0.25	0.25	0.25
Mean	0.30	0.24	0.29	0.26	0.27	0.28	0.27	0.20	0.27	0.26	0.26

Table 3. Vertical Scaling Linking Designs for On-grade and Below-grade Items.

Grade	Item		
4	G4_on		
5	G5_below	G5_on	
6		G6_below	G6_on

Table 4. True Distribution Parameters of Item Difficulties across Grades ~ N(M, SD)

Grade	On		Off	
	M	SD	M	SD
4	0	1	-.5	1
5	0.5	1	0	1
6	1.0	1	0.5	1

Table 5. Calibration Procedures

		Model	
		RM (WINSTEPS)	MERM (HLM)
Data	SRS	Yes	
	CS	Yes	Yes

Table 6. Means (over replication) of the Bias, SE, RMSE, and Correlation of Ability Parameter Estimations for Simulation Conditions

Software*	Model**	ICC	Grade	Bias	SE	RMSE	Correlation
WINSTEPS	RM	0.0	4	-0.0023	0.0030	0.0038	0.9643
		0.0	5	-0.0035	0.0020	0.0040	0.9659
		0.0	6	-0.0032	0.0023	0.0040	0.9671
HLM	RMEM	0.2	4	0.0117	0.0000	0.0117	0.9872
		0.2	5	-0.0140	0.0001	0.0140	0.9872
		0.2	6	0.0117	0.0001	0.0117	0.9869
		0.3	4	-0.0150	0.0000	0.0150	0.9878
		0.3	5	-0.0150	0.0001	0.0150	0.9876
		0.3	6	-0.0149	0.0002	0.0149	0.9876
		0.4	4	-0.0162	0.0000	0.0162	0.9882
		0.4	5	0.0960	0.0001	0.0960	0.9880
		0.4	6	0.0711	0.0002	0.0711	0.9875
WINSTEPS	RM	0.2	4	0.1078	0.1153	0.1578	0.9827
		0.2	5	0.0756	0.1188	0.1408	0.9825
		0.2	6	0.0869	0.1314	0.1575	0.9816
		0.3	4	0.0988	0.1397	0.1711	0.9827
		0.3	5	0.0909	0.1398	0.1668	0.9827
		0.3	6	0.0703	0.1458	0.1619	0.9815
		0.4	4	0.1228	0.1745	0.2134	0.9826
		0.4	5	0.1799	0.1644	0.2437	0.9826
		0.4	6	0.1417	0.1491	0.2057	0.9815

Software*: Software used to calibrate response.
Model**: Model used to generated responses.

Table 7. Means (over replication) of the Bias, SE, RMSE, and Correlation of Item Difficulty Parameter Estimations for Simulation Conditions

Software*	Model**	ICC	Grade	Bias	SE	RMSE	Correlation
WINSTEPS	RM	0.0	4	-0.0056	0.0088	0.0104	0.9978
		0.0	5	-0.0051	0.0135	0.0144	0.9967
		0.0	6	-0.0029	0.0126	0.0130	0.9965
HLM	RMEM	0.2	4	0.1176	0.1708	0.2073	0.9964
		0.2	5	-0.0569	0.1532	0.1635	0.9943
		0.2	6	-0.0269	0.1702	0.1723	0.9943
		0.3	4	0.0741	0.1668	0.1826	0.9962
		0.3	5	-0.0432	0.1460	0.1523	0.9940
		0.3	6	-0.0454	0.1553	0.1618	0.9941
		0.4	4	0.0858	0.1855	0.2044	0.9959
		0.4	5	0.0079	0.1825	0.1827	0.9937
		0.4	6	-0.0048	0.1805	0.1805	0.9937
WINSTEPS	RM	0.2	4	0.1075	0.2711	0.2917	0.9965
		0.2	5	0.0872	0.2587	0.2730	0.9945
		0.2	6	0.0978	0.2554	0.2735	0.9945
		0.3	4	0.1056	0.2688	0.2888	0.9964
		0.3	5	0.1077	0.2706	0.2912	0.9943
		0.3	6	0.0858	0.2652	0.2788	0.9943
		0.4	4	0.1355	0.2804	0.3114	0.9962
		0.4	5	0.2051	0.2648	0.3349	0.9940
		0.4	6	0.1652	0.2689	0.3156	0.9940

Software*: Software used to calibrate response.
Model**: Model used to generated responses.

Table 8. Results of ANOVA of Means of the Ability Parameter Estimations across Grades

Table	Source	Grade 4						Grade 5					Grade 6				
		df	**SS**	MS	*F*	Pr > F	R^2	SS	MS	*F*	Pr > F	R^2	SS	MS	*F*	Pr > F	R^2
ANOVA	Model_A*	5	0.424	0.085	7.650	<.0001	0.251	10.559	2.112	198.390	<.0001	0.897	34.801	6.960	652.870	<.0001	0.966
	Error	114	1.265	0.011				1.214	0.011				1.215	0.011			
	Total*	119	1.690					11.773					36.016				
Tests Effects	Model	1	0.406	0.406	36.550	<.0001		10.554	10.554	991.440	<.0001		34.799	34.799	3264.160	<.0001	
	ICC	2	0.009	0.005	0.420	0.658		0.003	0.001	0.120	0.885		0.001	0.001	0.050	0.949	
	Model*ICC	2	0.009	0.005	0.420	0.658		0.003	0.001	0.120	0.885		0.001	0.001	0.050	0.949	

Model_A*: Model_A here indicates ANOVA model, not IRT model.

Table 9. Results of ANOVA of Means of the Ability Parameter Estimations across Grades

Table	Source	Grade 4						Grade 5					Grade 6				
		df	**SS**	MS	*F*	Pr > F	R^2	SS	MS	*F*	Pr > F	R^2	SS	MS	*F*	Pr > F	R^2
ANOVA	Model_A*	5	0.048	0.010	0.170	0.972	0.008	1.012	0.202	4.000	0.002	0.149	0.695	0.139	2.700	0.024	0.106
	Error	114	6.316	0.055				5.767	0.051				5.872	0.052			
	Total*	119	6.364					6.780					6.566				
Tests Effects	Model	1	0.017	0.017	0.300	0.583		0.807	0.807	15.960	0.000		0.605	0.605	11.740	0.001	
	ICC	2	0.013	0.006	0.110	0.892		0.188	0.094	1.860	0.160		0.078	0.039	0.750	0.472	
	Model*ICC	2	0.019	0.009	0.170	0.844		0.017	0.008	0.170	0.848		0.012	0.006	0.120	0.890	

Model_A*: Model_A here indicates ANOVA model, not IRT model.

Table 10. Results of Recovery Vertical Scale (in Logit) under Different Simulation Conditions

Software*	Model**	ICC	Grade	True mean	Estimated mean
WINSTEPS	RM	0.0	4	0.0035	0.0012
		0.0	5	0.5035	0.5000
		0.0	6	1.0035	1.0003
HLM	RMEM	0.2	4	0.0706	0.0823
		0.2	5	0.5706	0.5566
		0.2	6	1.0706	1.0823
		0.3	4	0.0367	0.0217
		0.3	5	0.5367	0.5218
		0.3	6	1.0367	1.0218
		0.4	4	-0.0047	-0.0209
		0.4	5	0.4953	0.5913
		0.4	6	0.9949	1.0659
WINSTEPS	RM	0.2	4	0.0706	0.0960
		0.2	5	0.5706	0.5896
		0.2	6	1.0706	1.0751
		0.3	4	0.0367	0.1138
		0.3	5	0.5367	0.6059
		0.3	6	1.0367	1.0853
		0.4	4	-0.0047	0.1390
		0.4	5	0.4953	0.5840
		0.4	6	1.4949	1.0707

Software*: Software used to calibrate response.
Model**: Model used to generated responses.

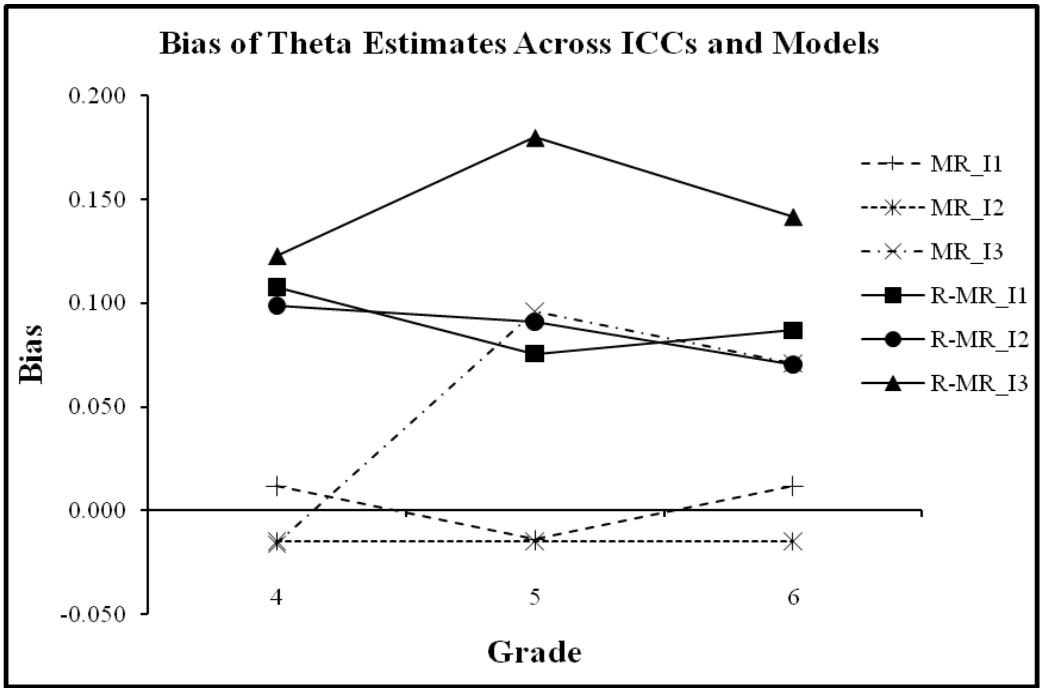

Figure 1. Bias of Ability Estimates under Different Simulation Conditions

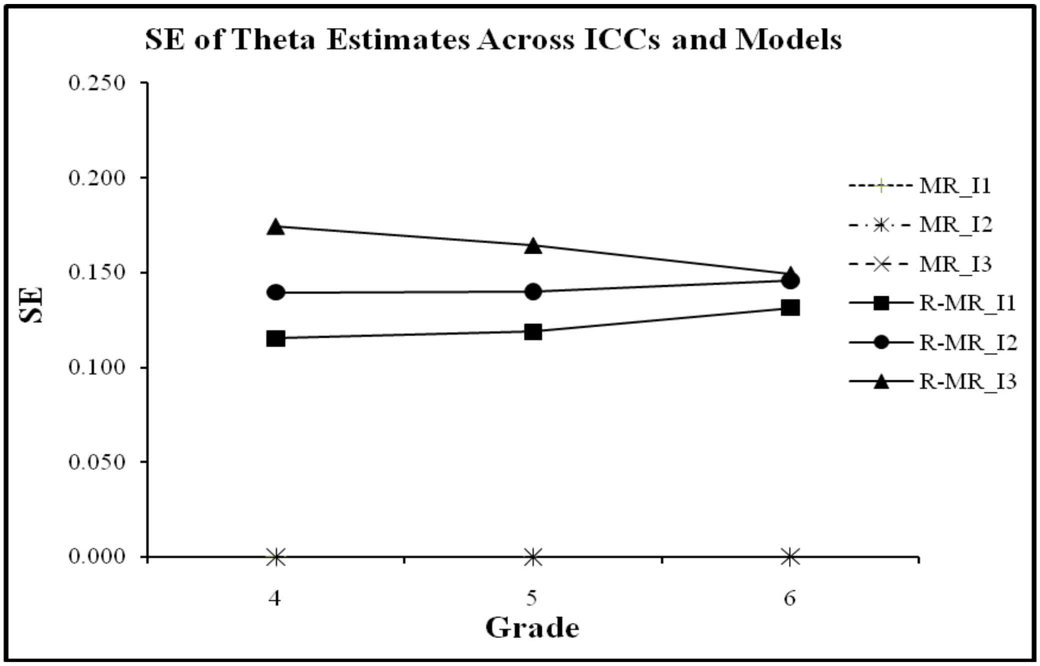

Figure 2. SE of Ability Estimates under Different Simulation Conditions

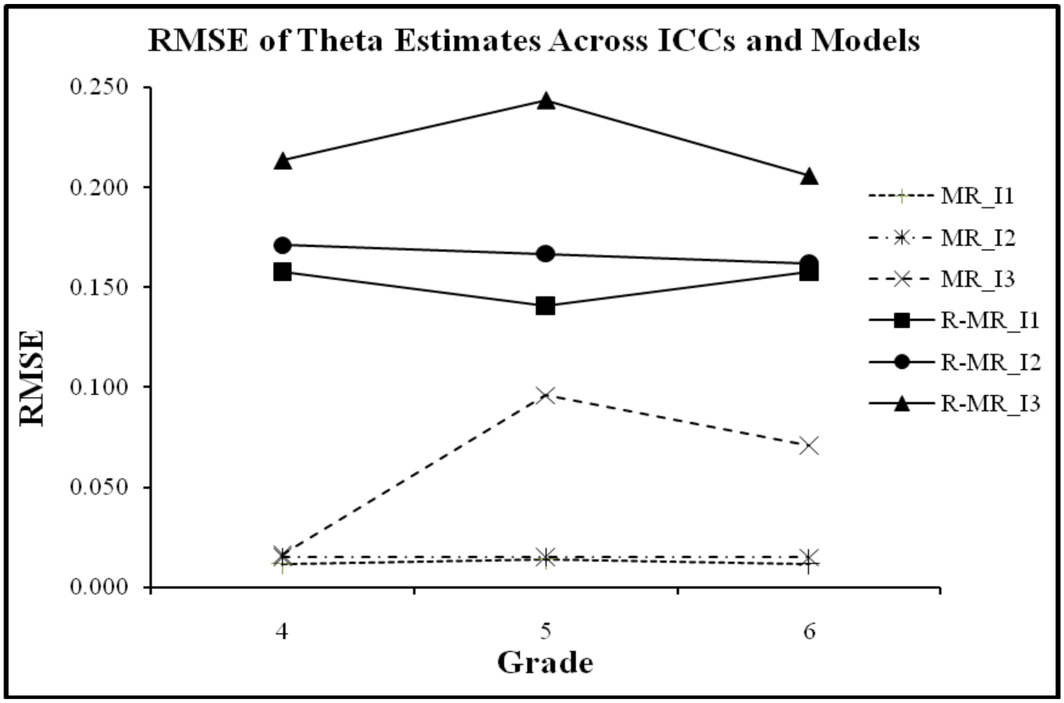

Figure 3. RMSE of Ability Estimates under Different Simulation Conditions

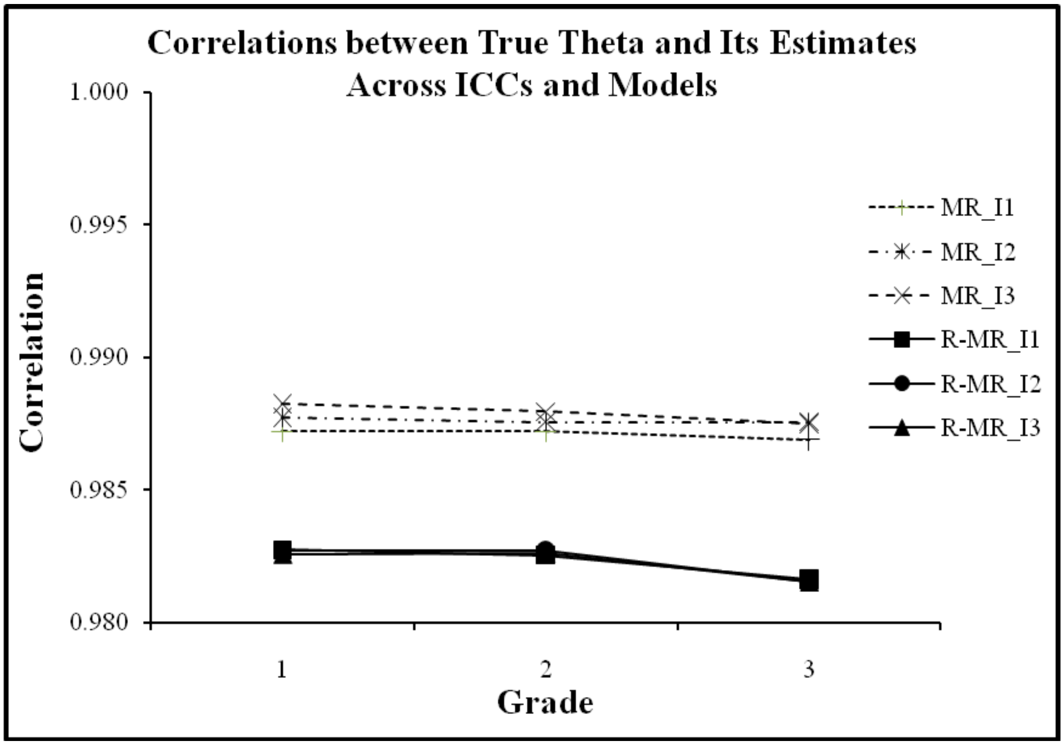

Figure 4. Correlation between True and Estimated Ability Parameters under Different Simulation Conditions

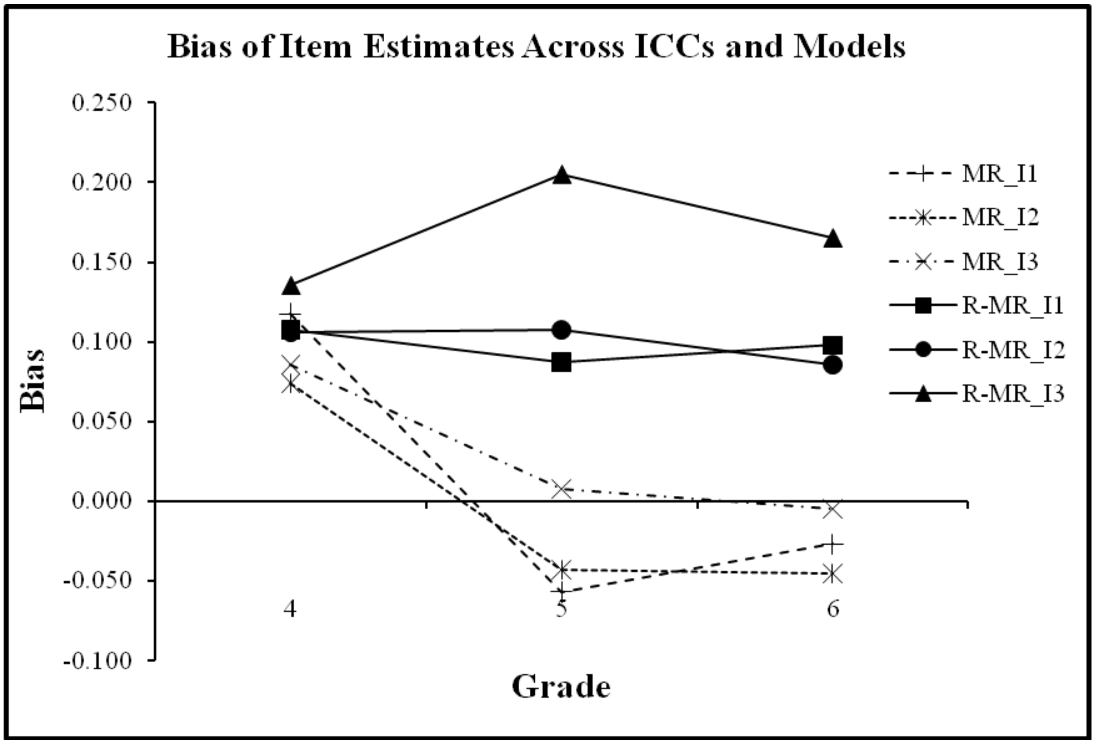

Figure 5. Bias of Item Estimates under Different Simulation Conditions

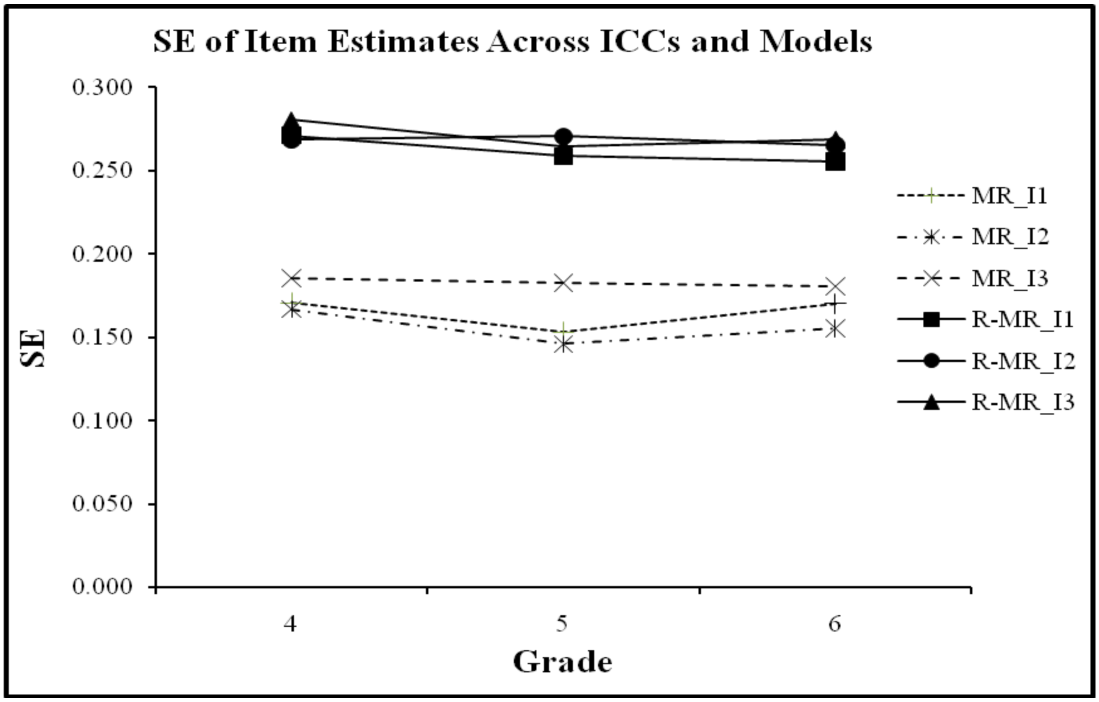

Figure 6. SE of Item Estimates under Different Simulation Conditions

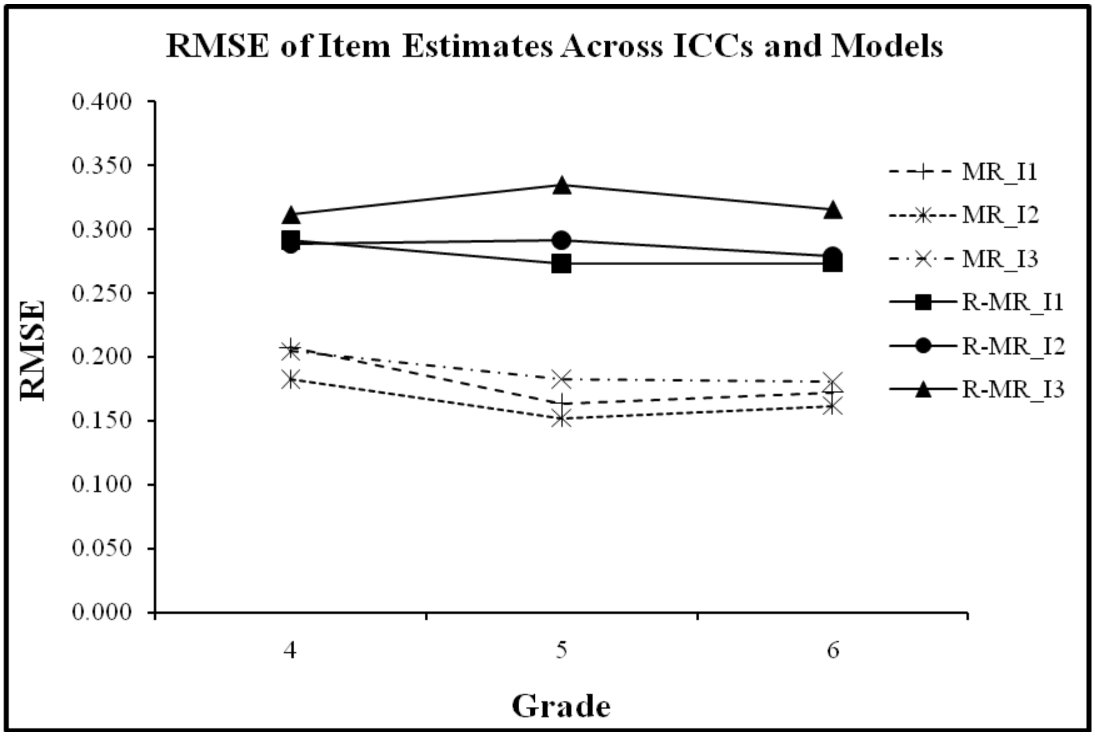

Figure 7. RMSE of Item Estimates under Different Simulation Conditions

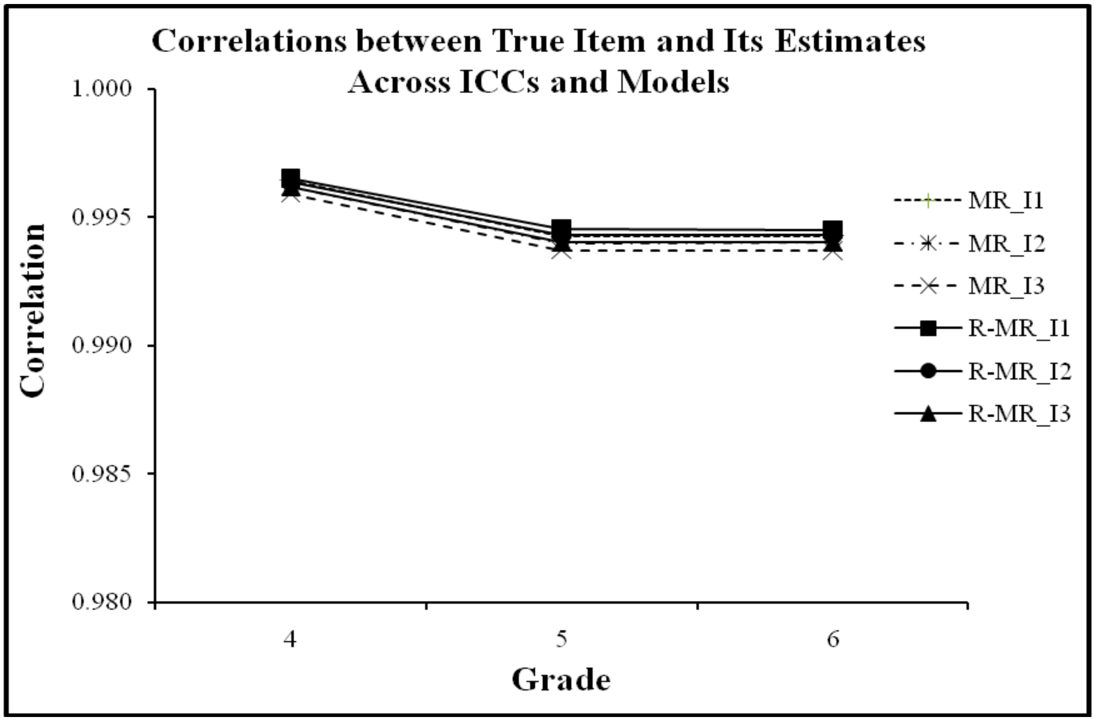

Figure 8. Correlation between True and Estimated Item Parameters under Different Simulation Conditions

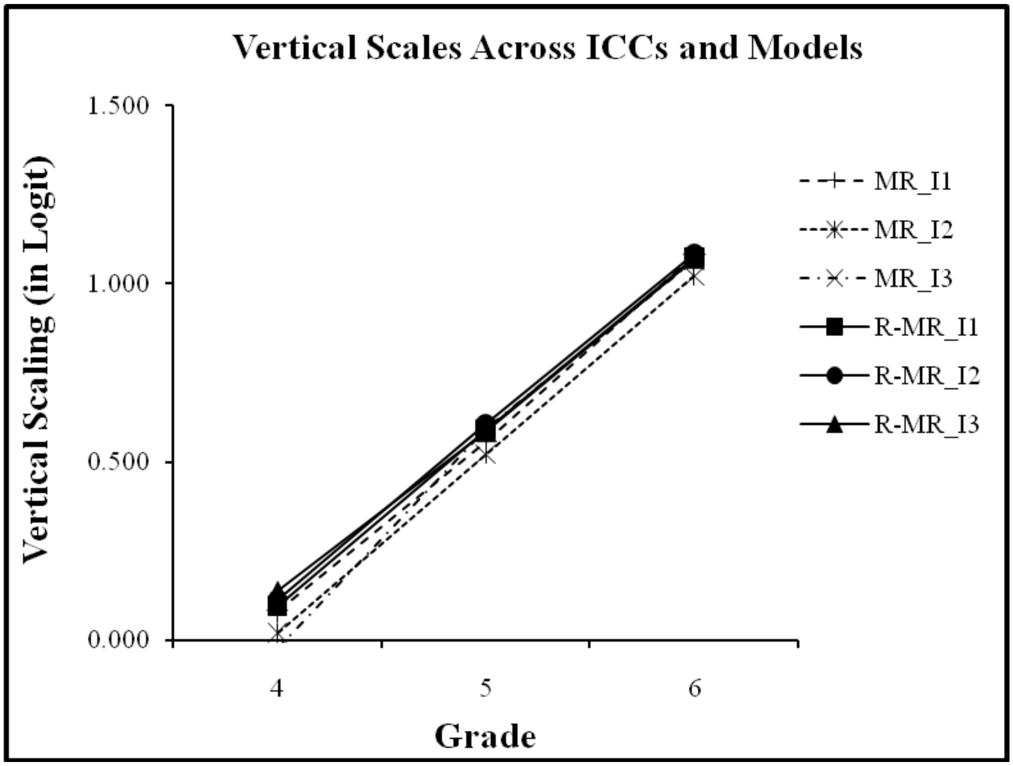

Figure 9. Recovery of Vertical Scale under Different Simulation Conditions

Appendix A

Three-Level Models (Raudenbush & Bryk, Hierarchical Linear Models: Application and Data Analysis Methods, Second Edition, page 228)

ICC in Fully Unconditional Model

<u>Student-Level Model:</u>

$Y_{ijk} = \pi_{0jk} + e_{ijk},$

where

Y_{ijk} is the score of student i in classroom j and school k;
π_{0jk} is the mean score of classroom j in school k; and
e_{ijk} is a random student effect that is the deviation of student ijk's score from classroom mean and
$e_{ijk}{\sim}N(0, \sigma^2)$.

i = 1,2,…,n_{jk} student within classroom j in school k;
j = 1,2,…,J_k classrooms within school k; and
k = 1,2,…K schools.

<u>Classroom-Level Model:</u>

Classroom mean π_{0jk} as an outcome varying randomly around some school mean:

$\pi_{0jk} = \beta_{00k} + r_{0jk},$

where
β_{00k} is the mean score in school k;
r_{0jk} is a random classroom effect, the deviation of classroom jk's mean from the school mean and
r_{0jk} ~$N(0, \tau_\pi)$. Within each school K, the variability among classroom is assumed the same.

School-Level Model:

School mean β_{00k} as varying randomly around a grand mean:

$\beta_{00k} = \gamma_{000} + u_{00k},$

where
γ_{000} is the grand mean;

u_{00k} is a random school effect, the deviation of school k's mean from the grand mean, and

$u_{00k} \sim N(0, \tau_\beta)$.

If total variance

$$\mathrm{Var}_{\mathrm{total}} = \mathrm{Var}_{\mathrm{school}} + \mathrm{Var}_{\mathrm{class}} + \mathrm{Var}_{\mathrm{student}}$$

$\mathrm{Var}_{\mathrm{school}}$ is variance between schools, which is τ_β
$\mathrm{Var}_{\mathrm{class}}$ is variance between classrooms within schools, which is τ_π
$\mathrm{Var}_{\mathrm{student}}$ is variance between students within classrooms and school, which is σ^2

So

ICC for schools is $\mathrm{ICC}_{\mathrm{school}}$

$$\mathrm{ICC}_{\mathrm{school}} = \mathrm{Var}_{\mathrm{school}} / (\mathrm{Var}_{\mathrm{school}} + \mathrm{Var}_{\mathrm{class}} + \mathrm{Var}_{\mathrm{student}}) = \tau_\beta/(\tau_\beta + \tau_\pi + \sigma^2)$$

ICC for classrooms within schools is $\mathrm{ICC}_{\mathrm{class}}$

$$\mathrm{ICC}_{\mathrm{class}} = (\mathrm{Var}_{\mathrm{class}}) / (\mathrm{Var}_{\mathrm{school}} + \mathrm{Var}_{\mathrm{class}} + \mathrm{Var}_{\mathrm{student}}) = (\tau_\pi)/(\tau_\beta + \tau_\pi + \sigma^2)$$