Effect of Person Cluster on Accuracy of Ability Estimation of Computerized Adaptive Testing in K-12 Education Assessment

Shudong Wang
NWEA

Hong Jiao
University of Maryland

Wei He
NWEA

Paper presented at the annual meeting of the American Educational Research Association (AERA). April 8 –April 12, 2011, New Orleans, Louisiana.

Send correspondence to:
Shudong Wang
Northwest Evaluation Association (NWEA)
121 NW Everett St.
Portland, OR 97206
Shudong.Wang@NWEA.org

Effect of Person Cluster on Accuracy of Ability Estimation of Computerized Adaptive Testing in K-12 Education Assessment

Abstract

The ability estimation procedure is one of the most important components in a computerized adaptive testing (CAT) system. Currently, all CATs that provide K-12 student scores are based on the item response theory (IRT) model(s); while such application directly violates the assumption of independent sample of a person in IRT models because ability estimation is mostly based on cluster (or correlated) educational data in which students usually are clustered in certain groups or settings (classrooms or schools). The consequences of such violations are commonly ignored. The purpose of this study is to investigate the effect of ignoring hierarchical data structures of students sample on the accuracy of ability estimation by using a regular Rasch model. Results show that ICCs have not only statistically significant effect on the accuracy of a person's ability estimation, but also large effect sizes.

I. Introduction

The Race to the Top Fund has put tremendous pressures on states to develop high-quality and high-utility assessment systems that can measure comparable academic achievement across states. Recently, the CAT has been seen as a particularly effective method in measuring an individual student's growth over time in K-12 assessment (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). Right now, besides Oregon, Delaware, and Idaho which are using CAT based on Rasch model (RM) in their state assessments, many other states (Georgia, Hawaii, Maryland, North Carolina, South Dakota, Utah, and Virginia) are also in various stages of CAT development. As a matter of fact, among the two consortia, SMARTER Balanced Assessment Consortium (SBAC) consists of 31 states, and Partnership for the Assessment of Readiness for College and Careers (PARCC) consists of 26 states, one of them (SBAC) is committed to a computer adaptive model because it represents a unique opportunity to create a large-scale assessment system that provides maximally accurate achievement results for each student (Race to the Top Assessment Program, 2010).

The CAT is of considerable interest to states right now because of its advantages, such as a short test, immediate feedback on student scores, better reliability, and accuracy (Lord, 1977; Kingsbury & Weiss, 1983; Steinberg, & Thissen, 1990) over traditional paper-pencil tests. Its unique advantages in K-12 assessment also include cost saving, multiple testing opportunities for formative and interim assessments, and better validity (Way, 2006).

In practice, when a CAT is used, IRT models (Hambleton & Swaminathan, 1985) are used to fulfill the purpose of estimating student's provisional ability. However, the use of any of IRT models is valid only when the IRT assumptions have been met. One of the IRT assumptions is the independence of observations in the sample or population, i.e., the persons should be sampled from simple random sample (SRS). This independence of observations assumption can be met in many situations such as in licensure, certification, and admission CATs, where examine can be regarded as independent each other. However, sample or population in an educational setting always involve a nested data structure where individual students are nested within organizational settings, such as a class or school. The dependencies between individuals in cluster sample (CS) pose unique challenge for proper application of CAT in educational setting, especially for the accuracy of ability estimation. Figure 1 presents both SRS and CS sampling designs.

The unique feature of educational sample is the cluster effect in which group students learn or study together so they share certain characteristics as a group. The correlation within cluster is called the intra-class correlation (ICC). If the ICC is nonzero, the assumption of independence, one of the necessary conditions 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 (Rosner, 1984; Munoz, Rosner, & Carey, 1986). Few studies have examined the dependence nature of educational data in large scale achievement context. A Schochet (2005) study shows that all achievement tests have certain degrees of dependence in samples. Wang (2006) conducted study on the effect of cluster data at test score level on sample size requirement for IRT calibration. A study by Wang, Jiao, Jin & Thum (2010) shows that degrees of dependence in educational data could lead to a biased person parameter estimation and misleading results in vertical scaling.

The problems of mistaking a cluster sample (CS) as a simple random sample (SRS) can be coped with by using multilevel models (Bryk & Raudenbush, 1992; Goldstein, 1995; Longford, 1993). Some researchers (Adams, Wilson, & Wu, 1997; Kamata, 2001; 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. Fox and Glas (2001) introduced a general approach for binary outcomes in a strictly clustered setting (i.e., items nested within students and students are nested within schools). 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) provided 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 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. The mixed-effect (or multilevel) Rasch model (MRM) 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. The MRM 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, the 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),

$$g(\mu_{ij}) = logit(\mu_{ij}) = ln\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] = \eta_{ij} = \sum_{k=0}^{K}\beta_k X_{ik} + \sum_{l=0}^{L}\theta_{jl}Z_{il} = \mathbf{X\beta} + \mathbf{Z\theta_j}, \tag{1}$$

where i for item, i=1,2,…, I; j for person, 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_{il} is the value of predictor l for item i; β_k is the fixed regression weight of predictor k and θ_{jl} is the random regression weight of predictor l for person j. 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,

$$P(Y_{ij} = 1|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{\theta}) = g^{-1}(\eta_{ij}) = \Psi(\eta_{ij}) \tag{2}$$

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 i ($Y_{ij} = 1$) conditional on the random effect or 'ability' of subject j (θ_j):

$$p(y_{ij} = 1|\theta_i) = \Psi(\eta_{ij}) = \Psi(\theta_j\text{-b}_i) = \frac{\exp(\theta_j\text{-b}_i)}{1+\exp(\theta_j\text{-b}_i)} \tag{3}$$

where b_i is the difficulty parameter for item i. Comparing (2) to (3), it can be seen that RM is special case of a random-intercepts model that includes item dummies as fixed regressors. Although 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) also formulated MERM in the context of multilevel model (multilevel RM) within GLMM framework.

Given that recently the CAT in K-12 assessment has rapidly grown and little work has been done on evaluating the accuracy of ability estimation of CAT with clustered data, the urgent need for better understanding the impact of educational cluster data on CAT quality can't be overstated. The purpose of this study is to investigate the effect of ignoring hierarchical data of structures of student sample by using regular RM on the accuracy of ability estimation in CAT environment.

II. Methods and Data Source

A Monte Carlo (MC) technique used to investigate effect of ignoring cluster data structures on the accuracy of ability estimation in CAT in this study. Two types of data sets will be simulated by using RM and MRM in fixed test length (30 and 50 dichotomous items) CAT.

2.1. Simulation of Cluster Data

The cluster data structures are simulated and different clusters effect in data are measured by ICCs. For this study, the student response score is our interest and used as dependent variable, proportion of total variance for given three levels, level-1: item, level-2: student, and level-3. According to model (1) and Kamata (2001), for MRM model,

$$\begin{aligned} log\left(\frac{p_{ijm}}{1-p_{ijm}}\right) &= \eta_{ijm} = \beta_{\mathrm{1jm}} X_{\mathrm{1jm}} + \beta_{\mathrm{2jm}} X_{\mathrm{2jm}} + \ldots + \beta_{\mathrm{(k\text{-}1)jm}} X_{\mathrm{(k\text{-}1)jm}} + \beta_{\mathrm{0jm}} = \sum_{q=1}^{k-1} \beta_{\mathrm{qjm}} X_{\mathrm{qjm}} + \beta_{\mathrm{0jm}} \\ &= (\gamma_{00m} + u_{0jm}) + (\gamma_{i0m}) = \pi_{000} + r_{00m} + u_{0jm} + \pi_{i00} = (r_{00m} + u_{0jm}) - (-\pi_{i00} - \pi_{000}) \\ &= \theta_{jm} - b_i, \end{aligned} \tag{5}$$

where η_{ijm} is linear predictor for item *i* of student *j* in class *m*; p_{ijm} is the probability that person *j* in class m answers item *i* correctly and X_{qijm} is *q*th dummy variable ($q = 1,2,..,k$-1) for the *i*th item for person *j* in class m. β_{0jm} is the effect of the reference item, and β_{qjm} is the effect of the *q*th item compared to the reference item. Because the level-2 (student-level) models for student *j* in class *m* are written as $\beta_{0jm} = \gamma_{00m} + u_{0jm}$, $\beta_{1jm} = \gamma_{10m}$, $\beta_{2jm} = \gamma_{20m}$, …, $\beta_{(k\text{-}1)jm} = \gamma_{(k\text{-}1)0m}$. Where $u_{0jm} \sim N(\gamma_{00m}, \tau_\gamma)$ and τ_γ ($\sigma^2_{\text{level-2}}$), the variance of u_{0jm} within class m is assumed to be identical across classes. In level 3 (class-Level) model, the intercept γ_{00m} is only term that arises across classes and item effects are constant across classes. For class m, $\gamma_{00m} = \pi_{000} + r_{00m}$, $\gamma_{10m} = \pi_{100}$, $\gamma_{20m} = \pi_{200}$, $\gamma_{(k\text{-}1)0m} = \pi_{(k\text{-}1)00}$, where $r_{00m} \sim N(0, \tau_\pi)$. τ_π is $\sigma^2_{\text{level-3}}$. So if let $\theta_{jm} = r_{00m} + u_{0jm}$ which means $\theta_{jm} \sim \mathrm{N}(\gamma_{00m}, \tau_\gamma)$ and $\gamma_{00m} \sim N(0, \tau_\pi)$, and also let $b_i = -\pi_{i00} - \pi_{000}$, then we have (5). Item effect is treated as fixed effect in GLMM framework, there is not supposed to have variance at item level, however, 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, and we label it as $\sigma^2_{\text{level-1}}$. The total variance and ICCs can be expressed as

$$\sigma^2 = \sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}} = \tau_\pi + \tau_\gamma + \pi^2/3$$

$$\text{ICC}_{\text{level-2}} \text{ (between students)} = \sigma^2_{\text{level-2}}/(\ \sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}}) = \tau_\gamma / (\tau_\pi + \tau_\gamma + \pi^2/3) \quad (6)$$

$$\text{ICC}_{\text{level-3}} \text{ (between class)} = \sigma^2_{\text{level-3}}/(\sigma^2_{\text{level-3}} + \sigma^2_{\text{level-2}} + \sigma^2_{\text{level-1}}) = \tau_\pi / (\tau_\pi + \tau_\gamma + \pi^2/3). \quad (7)$$

For traditional Rasch model, the assumption is that student ability as a random variable with standardized normal distributed N(0,1), in GLMM framework from (6), which means τ_γ has to be 1 and τ_π has to be 0. Even though, $\text{ICC}_{\text{level-2}} \neq 0$ but remains a constant value $1/(1+\pi^2/3)=$ 0.2331.

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 $1.0/(1.0+ \pi^2/3)$ since the distribution of ability is assumed to be distributed normal with mean 0 and variance 1.0 (Thum & Wang, 2011). The argument 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. The cluster effects of simulated data with 1000 examinees are generated by using different values of τ_π ($\tau_\pi \neq 0$) and ICC values are presented in Table 1.

Besides, ICCs, one sizes (1000 items) of item bank are used in simulation and all items difficulty parameters are generated from standard normal distribution for both RM and MRM models. Two ability estimation methods (Wang & Wang, 2001) are used to estimate examinee ability; maximum likelihood estimate (MLE) and expected a posteriori estimate (EAP). Because this study focuses only on the accuracy of ability parameters recovery under different type of data structures, content balance and item exposure control is not a concern in this study. However, at each step of CAT during the test, each item is randomly selected from a selected group items based on item information that have item difficulty range from 0.1 logit below provisional ability estimate to 0.1 logit above provisional ability estimate. Simulation procedure takes the following steps:

Step 1: To start the test, an initial ability estimate of 0.0 is assumed. The maximum information item selection algorithm is used to select next item among a group items that have item difficulty range from 0.1 logit below provisional ability estimate to 0.1 logit above provisional ability estimate.

Step 2. After an item is selected, a response based on the simulee's true ability is generated from two models. One is for RM, the other is for MRM. For RM, ICC=0 and For MRM, the correlated data were generated with two different ICC values (0.2, 0.4) under the assumption that the average cluster (class or school) size was 25 and the number of clusters was 40 so there are 1000 examinees.

Step 3. After a response is generated, the provisional ability level is estimated using one of two ability estimation methods (MLE and EAP). The provisional estimate using EAP after the first item was based on a normal prior. Based on this provisional estimate, the next item is selected the procedure described in step 1.

Step 4. Step 2 and step 3 are repeated until a termination criterion is researched. Fixed test length stopping rule is used to terminate the test.

Both descriptive methods and inferential procedures are used to analyze the results from the simulation. Total 10 replications are conducted in this study.

2.2 Design of Study

2.2.1 Independent variables

Independent variables in this study include ICC (three levels: 0, 0.2, and 0.4), estimation method (two levels: MLE and EAP) and test length (two levels: 30 and 50). For the purpose of verification, test length 1000 (total item bank) is also used for each of conditions but not included as the level of test length variable.

2.2.2 Dependent variables

There are varieties of statistics that can be used to evaluate how well true parameters are recovered for each of the simulation conditions. Five dependent (criterion) variables used in this study are: correlations between true and estimated parameters, biases, absolute biases (Abias), standard errors (SEs), root mean square errors (RMSEs). These criteria are used to examine the effects of the manipulated independent variables described in the last subsection to provide complementary evidence. For each person j, the conditional bias (Abias), SE, and RMSE of an estimator $\widehat{\theta}_j$ across size R (r=1, 2, … , R) of replications (here R=10) can be expressed as following:

$$\mathrm{Bias}(\hat{\theta}_{\mathrm{j}}) = \mathrm{E}(\hat{\theta}_{\mathrm{j}}) - \theta_{\mathrm{j}} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \hat{\theta}_{\mathrm{rj}} - \theta_{\mathrm{j}} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \hat{\theta}_{\mathrm{rj}} - \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \theta_{\mathrm{j}} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} (\hat{\theta}_{\mathrm{rj}} - \theta_{\mathrm{j}}) \qquad (8)$$

$$\mathrm{Abias}(\hat{\theta}_{\mathrm{j}}) = |E(\hat{\theta}_j) - \theta_j| = \tfrac{1}{R}\textstyle\sum_{r=1}^{R} |\hat{\theta}_{rj} - \theta_j| \qquad (9)$$

$$SE(\hat{\theta}_j) = \sqrt{Var(\hat{\theta}_j)} = \sqrt{E[(\hat{\theta}_j - E(\hat{\theta}_j))^2]} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{rj} - \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{rj}\right)^2} \qquad (10)$$

Where $\hat{\theta}_j$ is the estimated person ability and θ_j is true person ability of person j.

$$RMSE(\hat{\theta}_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{rj} - \theta_j)^2} \qquad (11)$$

There is the relationship between MSE (=RMSE^2), SE, and bias:

$$MSE(\hat{\theta}_j) = E\left[(\hat{\theta}_j - \theta_j)^2\right] = E\left[\left(\hat{\theta}_{\mathrm{j}} - E(\hat{\theta}_{\mathrm{j}})\right)^2 + \left(E(\hat{\theta}_{\mathrm{j}}) - \theta_j\right)^2\right] = Var(\hat{\theta}_{\mathrm{j}}) + Bias^2(\hat{\theta}_{\mathrm{j}}) \quad (12)$$

This relationship can be used to verify the correctness of calculation of each criterion index. Besides bias, absolute bias (absolute value of bias) is used because the direction of bias (positive or negative) is a function of either person ability or item difficulty. The average of bias, Abias, SE, and RMSE across N persons (j=1, 2, …, N) can be described as following:

$$\mathrm{Bias}(\hat{\theta}) = \frac{1}{N}\frac{1}{R}\sum_{j=1}^{N}\sum_{r=1}^{R}(\hat{\theta}_{rj} - \theta_j) \qquad (13)$$

$$\mathrm{Abias}(\hat{\theta}) = \frac{1}{N}\frac{1}{R}\sum_{j=1}^{N}\sum_{r=1}^{R}|\hat{\theta}_{rj} - \theta_j| \qquad (14)$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{N}\frac{1}{R}\sum_{j=1}^{N}\sum_{r=1}^{R}\left(\hat{\theta}_{rj} - \frac{1}{N}\sum_{r=1}^{R}\hat{\theta}_{rj}\right)^2} \qquad (15)$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{N}\frac{1}{R}\sum_{j=1}^{N}\sum_{r=1}^{R}\left(\hat{\theta}_{rj} - \theta_j\right)^2}. \qquad (16)$$

Where θ is the true ability of simulees, which was used to generate responses in the simulation, $\hat{\theta}_r$ is the estimated ability for the *r*th replication, and R is the number of replications. The relationship among average bias, SE, and RMSE in (12) is no longer true for average bias, SE, and RMSE.

These dependent (criterion) variables (equation 13-16) are used for two set simulation design. First design include three design factors of ICC (3 levels), methods (3 levels), and test length (2 levels); Second design still keep first two factors (ICC and methods) but test length is equal to total item bank. There are a total [2(method) x 2(test length) x 3(ICC) + 2(method) x 1(full bank) x 3(ICC)] x10 (replication) = 180 calibrations conducted in this study. The Table 1 shows detail information of simulation design.

III. Results

This study investigates the effect of ICC, method, and test length on the accuracy of personal parameter estimation in the CAT. Parameter recovery is evaluated by comparing the estimates to the true (generated) parameters in terms of five dependent variables: correlations, bias, Abias, SE, and RMSE. The descriptive statistics, such as tabular summaries and graphical presentations, are used to present these dependent variables.

2.1 Descriptive Statistics of Conditional Dependent Variables

2.1.1 Correlations among true and estimated person parameters.

Tables 2 show average Pearson's correlation coefficients true and estimated person parameters under different test lengths and ICC across different calibration methods and replication. The results show that as test length increases, correlation coefficient increase across different factors and test with 1000 items (total item in item bank) recovers the best among different test length. Apparently, ICC has a little effect on the recovery.

2.1.2 Conditional Bias

Figures 3 and 4 depict the conditional biases along the theta scale under different test lengths and ICCs with one replication for MLE and EAP estimation methods. It is clear that Bayesian method (EAP) show “inward” biases, which means that for low ability, EAP over-estimate true parameters and for high ability, EAP under-estimate true parameter. The MLE methods supposes to show a slight “outward” biases, which is opposite to EAP, but result do not show this trend very clear. The effects of ICC and test length on biases are not clear from the Figures.

2.1.3 Conditional SE

The conditional SEs under different test lengths and ICCs with one replication for MLE and EAP estimation methods are presented in Figures 4 and 5. Results show that the SE values with $\text{ICC}\neq 0$ are lower than that SE values when $\text{ICC}=0$ and EAP has less SEs than MLE.

2.1.4 Conditional RMSE

SEs under different test lengths and ICCs with one replication for MLE and EAP estimation methods are presented in Figures 6 and 7. The RMSE of MLE has flat distributions that that of EAP.

2.1.5 Average Dependent Variables of Bias, SE, and RMSE

The average dependent variables (bias, abias, SE, and RMSE) are computed by using equations 13 to 16 and correlation remain the same. Five average dependent variables (correlation, bias, abias, SE, and RMSE) under different simulation conditions over 10 replications are presented in Table 3. In general, the correlations decrease as test length decrease. The estimation accuracy measured by bias, SE, and RMSE decreases as ICC increase, however, ICC has more impact on the accuracy of estimation than test length and method.

2.2 Inferential Statistics of Average Dependent Variables

Another way to statistically check the effect of independent variables (ICC, method, and length) on the accuracy is to conduct ANOVA analysis on the simulation results.

2.2.1 Statistical Hypotheses

For this study, the five dependent variables (correlation, bias, abias, SE, RMSE) are used to evaluate the accuracy of item calibration, based on research questions proposed in introduction section, the statistical null hypotheses are presented as follows.

1) There are no effects of ICC on the accuracy of person parameter estimations when some or all of the dependent variables: correlation, bias, abias, SE, and RMSE are used in different simulation conditions.

2) There are no effects of test length on the accuracy of item parameter estimations when some or all of the dependent variables: correlation, bias, abias, SE, and RMSE are used in different simulation conditions.

3) There are no effects of calibration method on the accuracy of person parameter estimations when some or all of the dependent variables: correlation, bias, abias, SE, and RMSE are used in different simulation conditions.

4) There are no three-way interaction effects between any of two factors mentioned above when some or all of the dependent variables: correlation, bias, abias, SE, and RMSE are used in different simulation conditions.

5) There are no two-way interaction effects between any of two factors mentioned above when some or all of the dependent variables: correlation, bias, abias, SE, and RMSE are used in different simulation conditions.

Because the Monte Carlo study is really a statistical sampling experiment with an underlying model, the number of replications in this Monte Carlo study is the analogue of sample size. In this study, in order to have adequate power for the statistical tests in the Monte Carlo study to detect effects of interest, each simulated condition has been replicated 10 times. The simulation results of dependent variables from three-way (ICC, calibration method, test length) crossed ANOVA are presented. Both test statistics and effect sizes are used to determine levels of significant effects. The magnitude of significant effects is estimated using eta-squared η^2 (empirical η^2 as an effect size estimate). Following the advice of Cohen (Cohen, 1988), the effect size in terms of η^2 had been classified as: (a) no effect ($\eta^2 < 0.0099 \approx 0.01$), (b) small effect ($0.01 < \eta^2 < 0.0588 \approx 0.06$), (c) medium effect ($0.06 < \eta^2 < 0.1379 \approx 0.14$), and (d) large effect ($\eta^2 > 0.14$).

2.2.2 ANOVA Results

Tables 4 to 13 show both results of the three-way ANOVA for average of correlation1, bias1, abias1, SE1, and RMSE1 that account the test length difference and two-way ANOVA for average of correlation2, bias2, abias2, SE2, and RMSE2 that do not account for the test length difference. Using $\alpha = 0.05$ for each hypothesis tested (means of each dependent variables are equal across ICC, method, and length). Following the advice of Cohen (Cohen, 1988), the effect size in terms of η^2 had been classified as: (a) no effect ($\eta^2 < 0.0099 \approx 0.01$), (b) small effect ($0.01 < \eta^2 < 0.0588 \approx 0.06$), (c) medium effect ($0.06 < \eta^2 < 0.1379 \approx 0.14$), and (d) large effect ($\eta^2 > 0.14$).

2.2.2.1 Results of Three-way ANOVA

For the three-way ANOVA, three main effects are ICC(I), Length (L), and method (M). effect (I x L x M). For all dependent variables (correlation1, bias1, abias1, SE1, and RMSE1), all interaction effects (one three-way and three two-way) are not statistically significant; except for correlation1, all main effects of L and M of dependent variables are not statistically significant except for SE1 but all main effects of ICC of dependent variables are statistically significant. For correlation1, one main effect of M is statistically significant and for SE1, all three main effects are statistically significant. In terms of η^2 (total variance in the dependent variable that is explained by independent variables), the main effect I accounts for most of the variance. More specifically, 48.4% of the total sum of squares of the bias1, 53.0% of the total sum of squares of the abias1, 74.9% of the total sum of squares of the SE1, and 62.8% of the total sum of squares of the RMSE1 for the person estimation are due to ICC effect. All effect sizes for ICC are in the large ranges ($\eta^2 > 0.14$). Although main effect of L and M for SE1 are statistically significant, but two main effect account for much less total variations (2.7% for L and 7.9% for M) comparing to ICC. For correlations1, main effect of L is statistically significant and L account for 90.7% of total variation.

2.2.2.2 Results of Two-way ANOVA

Among two main effects (I and M) and one interaction effect (I x M) for bias2, abias2, SE2, and RMSE2, only main effect I is statistically significant at $\alpha = 0.05$ level. Again, 6.0% of the total sum of squares of bias2, 49.2% of the total sum of squares of abias2, 81.1.0% of the total sum of squares of SE2, 59.6% of the total sum of squares of RMSE2 for the person estimation are due to ICC effect. For correlations2, main effect of L is statistically significant and L account for 94.5% of total variation.

In general, the results for average dependent variables are consistent with the results of conditional dependent variables. The factor of ICC has the most influence on bias1, abias1, SE1, RMSE1, bias2, abias2, SE2, and RMSE2 because it accounts for more than half percent of the total variations. On the other hand, the factor of method has the most influence on correlation.

2.3 Summary of Results

In general, for both conditional and average indices of dependent variables (bias, abias, SE, RMSE) increase as the values of ICC increase and as the value of test length decrease. Among all manipulated independent variables (ICC, method, and length), the factor ICC has most impact on the accuracy of person ability estimation.

Educational Importance of the Study

The ability estimation procedure is one of the most important components in a computerized adaptive testing (CAT) system. The accuracy of ability estimation methods used in CAT has a significant impact on the quality of CAT testing because it affects not only the final score reported, but also the item selection and test termination. For decades, the computerized adaptive tests have been widely used in licensure, certification and admission purposes; and recently, the demanding on CAT in K-12 education has increased tremendously. However, the impact of the significant part of examinee characteristics difference between K-12 test and other types of tests on ability estimation in CAT has been totally neglected by both CAT researchers and developers. Unlike in licensure, certification, and admission tests where students can be regarded as random independent samples, students in K-12 education are clustered into larger units, such as class, school, school district, and so on. So far in practice, the consequence of such violation in regular IRT application such as CAT is usually ignored. This study is the first study that attempts to examine the consequence of such common practice.

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.

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: Wiley .

Cohen, J. (1988). Statistical Power Analysis for the Behavior Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology, 108*(2), 100-102.

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

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57*, 266-279.

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

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed*.), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-238). New York: Academic Press.

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

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

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95-100.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed*.), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224-236). New York: Academic Press.

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.

Munoz, A., Rosner, B., and Carey, V. (1986). Regression analysis in the presence of heterogeneous intraclass correlations. *Biometrics, 42*, 653-658.

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.

Rosner B. (1984). Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics, 40*, 1025-1035.

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

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement, 7*, 405-425.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., and Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Walsh, J. (1947). Concerning the effects of the intra-class correlation on certain significance tests. *Annals of Mathematical Statistics, 18*, xxx-yyy.

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

Wang, S., Jiao, H., Jin, Y., & Thum Y. M. (2010). *Effect of Ignoring Hierarchical Data Structures on Accuracy of Vertical Scaling: A Mixed-Effects Rasch Model Approach*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimation of ability for a polytomous model in CAT. *Applied Psychological Measurement, 25*, 1-15.

Washington State, on behalf of the SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program Application for New Grants*. Retrieved from http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf.

Way, W. (2006). *Online Testing Research: Information and Guiding Transitions to Computerized Assessments*. A white paper from Pearson Educational Measurement.

Way, W., Twing, J., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the Common Core Assessments*. Retrieved June 11, 2010, from www.ets.org/s/commonassessments/pdf/ AdaptiveTesting.pdf.

Thum, Y. M., & Wang, S. (2011). *A Rasch model for item calibration with clustered samples of examinees*. Paper presented at the Annual Meetings of the American Educational Research Association, New Orleans.

Table 1. Designs of Simulation Condition

Simulation Design	Simulation Conditions	Estimation Method	Test Length	$ICC_{level-2}$	$ICC_{level-3}$	$\sigma^{2}_{level-2}$	$\sigma^{2}_{level-3}$
First	1	MLE	30	.2331	0	1	0
	2	MLE	30	0.8	0.2	0.8	0.2
	3	MLE	30	0.8	0.4	0.666	0.333
	4	MLE	50	.2331	0	1	0
	5	MLE	50	0.8	0.2	0.8	0.2
	6	MLE	50	0.8	0.4	0.666	0.333
	7	EAP	30	.2331	0	1	0
	8	EAP	30	0.8	0.2	0.8	0.2
	9	EAP	30	0.8	0.4	0.666	0.333
	10	EAP	50	.2331	0	1	0
	11	EAP	50	0.8	0.2	0.8	0.2
	12	EAP	50	0.8	0.4	0.666	0.333
Second	1	MLE	1000	.2331	0	1	0
	2	MLE	1000	0.8	0.2	0.8	0.2
	3	MLE	1000	0.8	0.4	0.666	0.333
	4	EAP	1000	.2331	0	1	0
	5	EAP	1000	0.8	0.2	0.8	0.2
	6	EAP	1000	0.8	0.4	0.666	0.333

Table 2. Average Pearson's Correlation Coefficients Between True and Estimated Person Ability across Different Calibration Methods and Replications

Design	N	Level of Length	$ICC_{level-3}$	Mean	SD
1	20	30	0	0.933	0.0032
	20	30	0.2	0.9329	0.005
	20	30	0.4	0.9321	0.0064
	20	50	0	0.9599	0.0019
	20	50	0.2	0.9596	0.0033
	20	50	0.4	0.9578	0.0043
2	20	1000	0	0.9947	0.0025
	20	1000	0.2	0.9946	0.0026
	20	1000	0.4	0.9945	0.0027

Table 3. Average Dependent Variable across ICCs, Estimation Methods, and Test Lengths over 10 Replications

ICC	Method	Length	N	Cor1	Cor2	Bias1	Bias2	aBias1	aBias2	SE1	SE2	RMSE1	RMSE2
0	MLE	30	10	0.931854	0.997068	0.031749	-0.04157	0.431598	0.294916	0.213736	0.08435	0.303168	0.159266
0	MLE	50	10	0.960008	0.997068	0.03063	-0.04157	0.376586	0.294916	0.151553	0.08435	0.232518	0.159266
0	EAP	30	10	0.934191	0.992308	0.025262	-0.04104	0.401423	0.307122	0.156972	0.091248	0.262128	0.16754
0	EAP	50	10	0.95982	0.992308	0.023333	-0.04104	0.367485	0.307122	0.130756	0.091248	0.22111	0.16754
0.2	MLE	30	10	0.931505	0.99707	-0.12314	-0.0846	0.638257	0.558652	0.363125	0.23821	0.644933	0.513556
0.2	MLE	50	10	0.959197	0.99707	-0.12333	-0.0846	0.609826	0.558652	0.312996	0.23821	0.590002	0.513556
0.2	EAP	30	10	0.934228	0.992189	-0.12396	-0.0842	0.596237	0.564643	0.273242	0.242161	0.56581	0.5189
0.2	EAP	50	10	0.95993	0.992189	-0.12511	-0.0842	0.579344	0.564643	0.25914	0.242161	0.538709	0.5189
0.4	MLE	30	10	0.932491	0.996994	-0.1713	-0.09505	0.745248	0.68064	0.473659	0.340206	0.885459	0.752522
0.4	MLE	50	10	0.958335	0.996994	-0.1714	-0.09505	0.715688	0.68064	0.41601	0.340206	0.815099	0.752522
0.4	EAP	30	10	0.931799	0.991928	-0.16984	-0.09439	0.696526	0.686843	0.358068	0.34559	0.770991	0.759412
0.4	EAP	50	10	0.957239	0.991928	-0.16984	-0.09439	0.685768	0.686843	0.351932	0.34559	0.754326	0.759412

Note: Cor1 is for test length 30 and 50, Cor2 is for test length 1000 (Whole item bank), the rest labels for other dependent variables are the same.

Table 4. Results of Three-way ANOVA of Correlation1

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
Main Effects						
ICC (I)	2	5E-05	3E-05	1.38	0.255	0.0022
Length (L)	1	0.0209	0.0209	1119.6	<.0001	0.9066
Method (M)	1	1E-05	1E-05	0.65	0.422	0.0005
Interaction Effects						
I x L	2	9E-06	5E-06	0.24	0.7856	0.0004
I x M	2	4E-05	2E-05	1	0.3725	0.0016
L x M	1	2E-05	2E-05	1.08	0.3011	0.0009
I x L x M	2	6E-06	3E-06	0.16	0.8501	0.0003

Table 5. Results of Two-way ANOVA of Correlation2 (Full Bank))

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
ICC (I)	2	1E-06	6E-07	1.56	0.214	0.0015
Method (M)	1	0.0007	0.0007	2028.1	<.0001	0.9448
I x M	2	5E-07	2E-07	0.66	0.5164	0.0006

Table 6. Results of Three-way ANOVA of Bias1

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
Main Effects						
ICC (I)	2	0.8601	0.4301	50.75	<.0001	0.4844
Length (L)	1	2E-05	2E-05	0	0.9646	0
Method (M)	1	0.0001	0.0001	0.02	0.8948	0.0001
Interaction Effects						
I x L	2	1E-05	5E-06	0	0.9994	0
I x M	2	0.0004	0.0002	0.02	0.9787	0
L x M	1	2E-06	2E-06	0	0.9868	0
I x L x M	2	2E-06	8E-07	0	0.9999	0

Table 7. Results of Two-way ANOVA of Bias2 (Full Bank))

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
ICC (I)	2	0.0642	0.0321	3.65	0.029	0.0602
Method (M)	1	9E-06	9E-06	0	0.9752	0
I x M	2	3E-07	2E-07	0	1	0

Table 8. Results of Three-way ANOVA of aBias1

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
Main Effects						
ICC (I)	2	2.0799	1.0399	63.06	<.0001	0.53
Length (L)	1	0.0254	0.0254	1.54	0.2172	0.0065
Method (M)	1	0.0302	0.0302	1.83	0.1787	0.0077
Interaction Effects						
I x L	2	0.0036	0.0018	0.11	0.8973	0.0009
I x M	2	0.0022	0.0011	0.07	0.9343	0.0006
L x M	1	0.0022	0.0022	0.13	0.7155	0.0006
I x L x M	2	0.0001	6E-05	0	0.9962	0

Table 9. Results of Two-way ANOVA of aBias2 (Full Bank))

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
ICC (I)	2	3.0575	1.5287	55.23	<.0001	0.4919
Method (M)	1	0.002	0.002	0.07	0.7894	0.0003
I x M	2	0.0002	0.0001	0	0.9955	0

Table 10. Results of Three-way ANOVA of SE1

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
Main Effects						
ICC (I)	2	1.054	0.527	306.42	<.0001	0.7492
Length (L)	1	0.0385	0.0385	22.37	<.0001	0.0273
Method (M)	1	0.1112	0.1112	64.64	<.0001	0.079
Interaction Effects						
I x L	2	0.005	0.0025	1.45	0.2388	0.0035
I x M	2	0.0031	0.0016	0.9	0.4088	0.0022
L x M	1	0.0093	0.0093	5.38	0.0223	0.0066
I x L x M	2	0.0001	6E-05	0.03	0.9671	0.0001

Table 11. Results of Two-way ANOVA of SE2 (Full Bank))

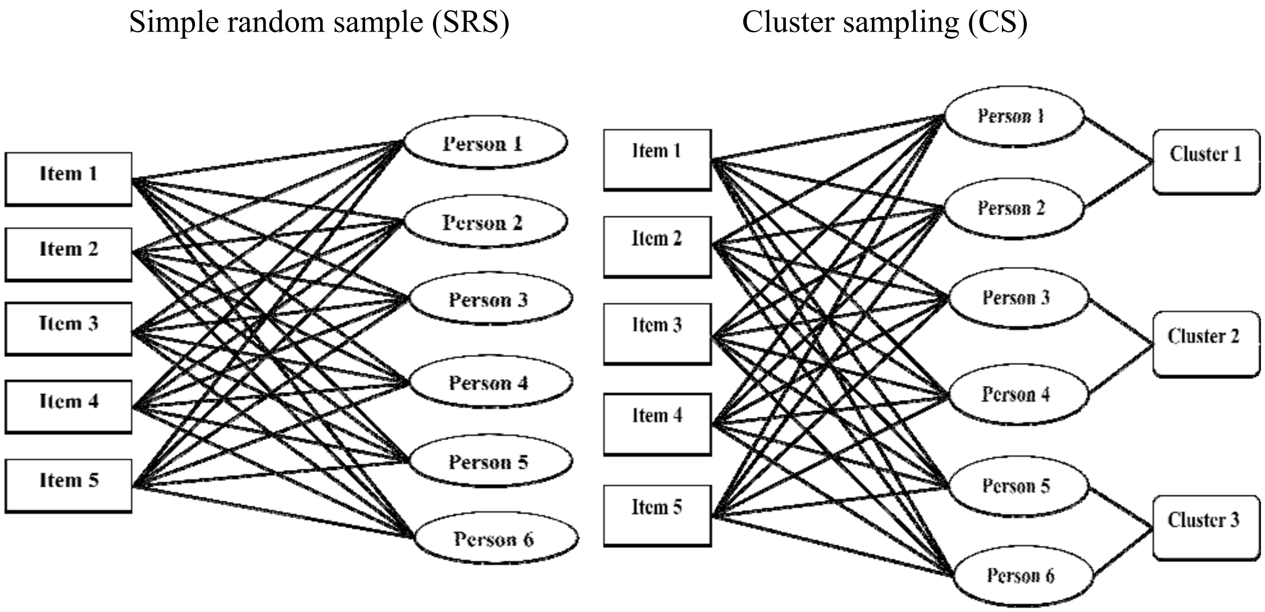
Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
ICC (I)	2	1.318	0.659	244.79	<.0001	0.8107
Method (M)	1	0.0009	0.0009	0.33	0.569	0.0005
I x M	2	4E-05	2E-05	0.01	0.992	0

Table 12. Results of Three-way ANOVA of RMSE1

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
Main Effects						
ICC (I)	2	2.9275	1.4637	62.79	<.0001	0.5277
Length (L)	1	0.0445	0.0445	1.91	0.1699	0.008
Method (M)	1	0.0454	0.0454	1.95	0.1657	0.0082
Interaction Effects						
I x L	2	0.0048	0.0024	0.1	0.9014	0.0009
I x M	2	0.0029	0.0015	0.06	0.9389	0.0005
L x M	1	0.0045	0.0045	0.19	0.6616	0.0008
I x L x M	2	0.0001	7E-05	0	0.9972	0

Table 13. Results of Two-way ANOVA of RMSE2 (Full Bank))

Source	DF	Type I SS	Mean Square	F-Value	Pr > F	η2
ICC (I)	2	7.1079	3.5539	84.16	<.0001	0.5961
Method (M)	1	0.0014	0.0014	0.03	0.8557	0.0001
I x M	2	4E-05	2E-05	0	0.9995	0



Simple random sample (SRS)

Cluster sampling (CS)

Figure 1. Sampling Design with Different Units of Sampling Frame (SRS and CS)

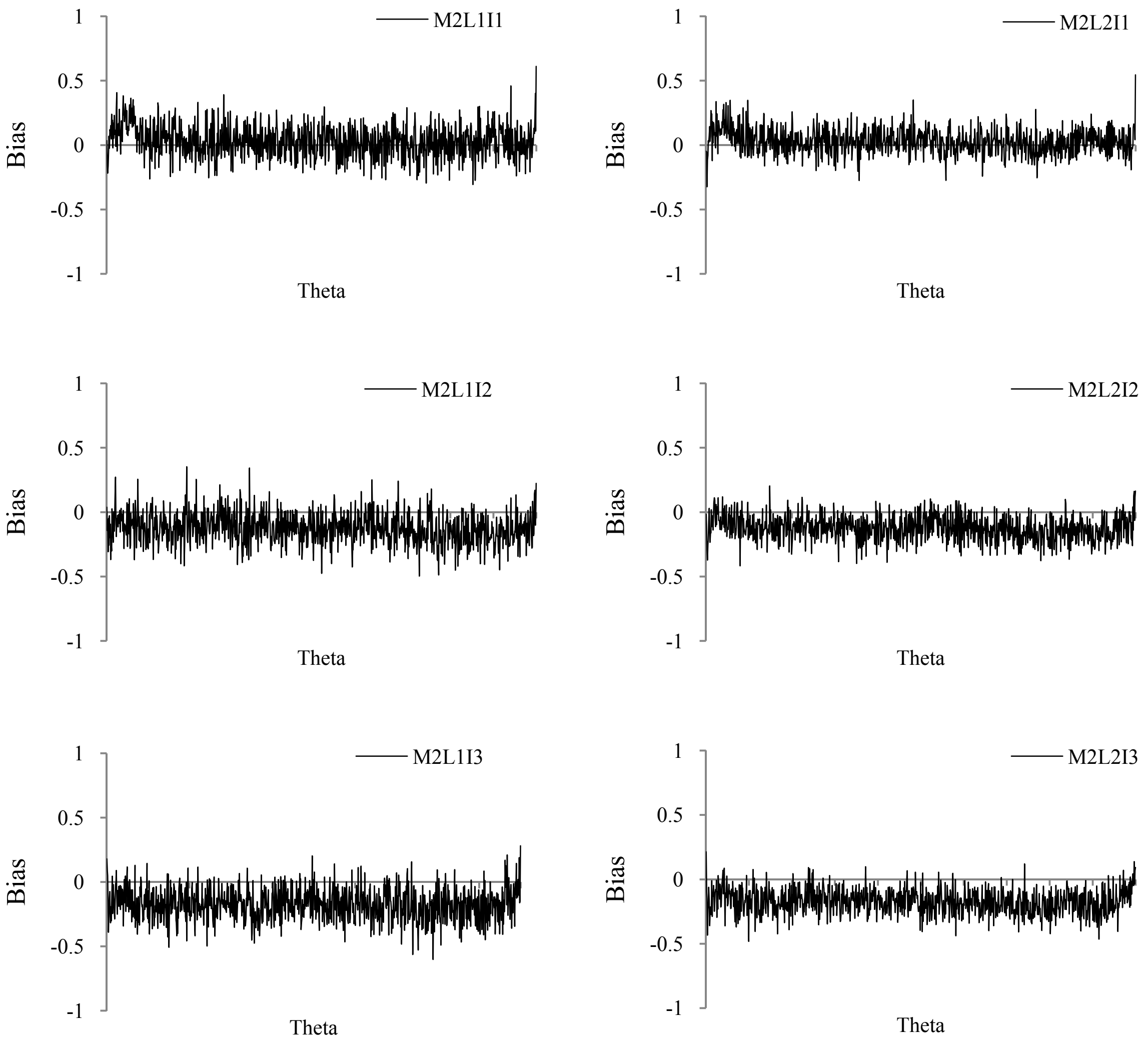


Figure 2. Distribution of Conditional Bias Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using MLE (M=2) with One Replication

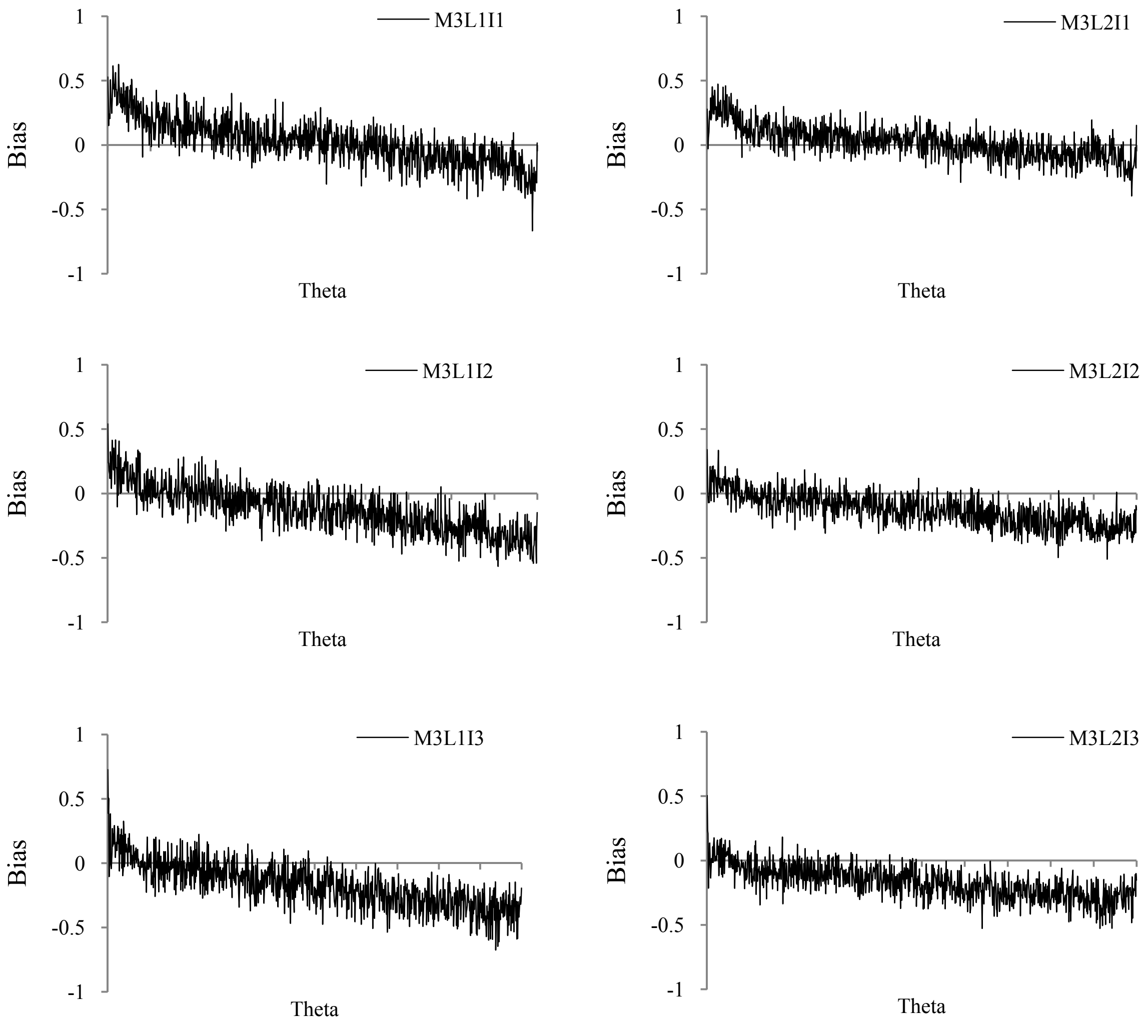


Figure 3. Distribution of Conditional Bias Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using EAP (M=3) with One Replication

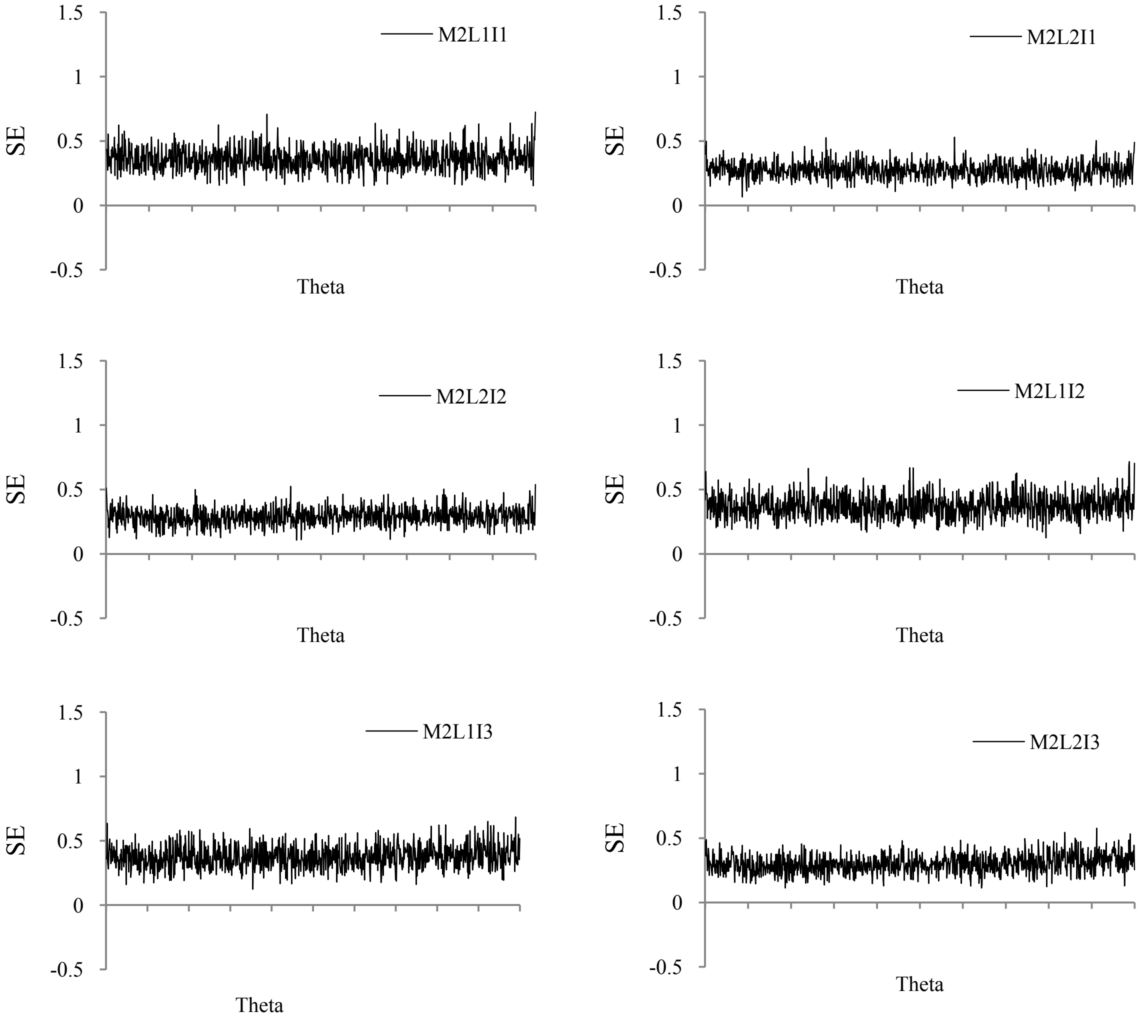


Figure 4. Distribution of Conditional SE Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using MLE (M=2) with One Replication

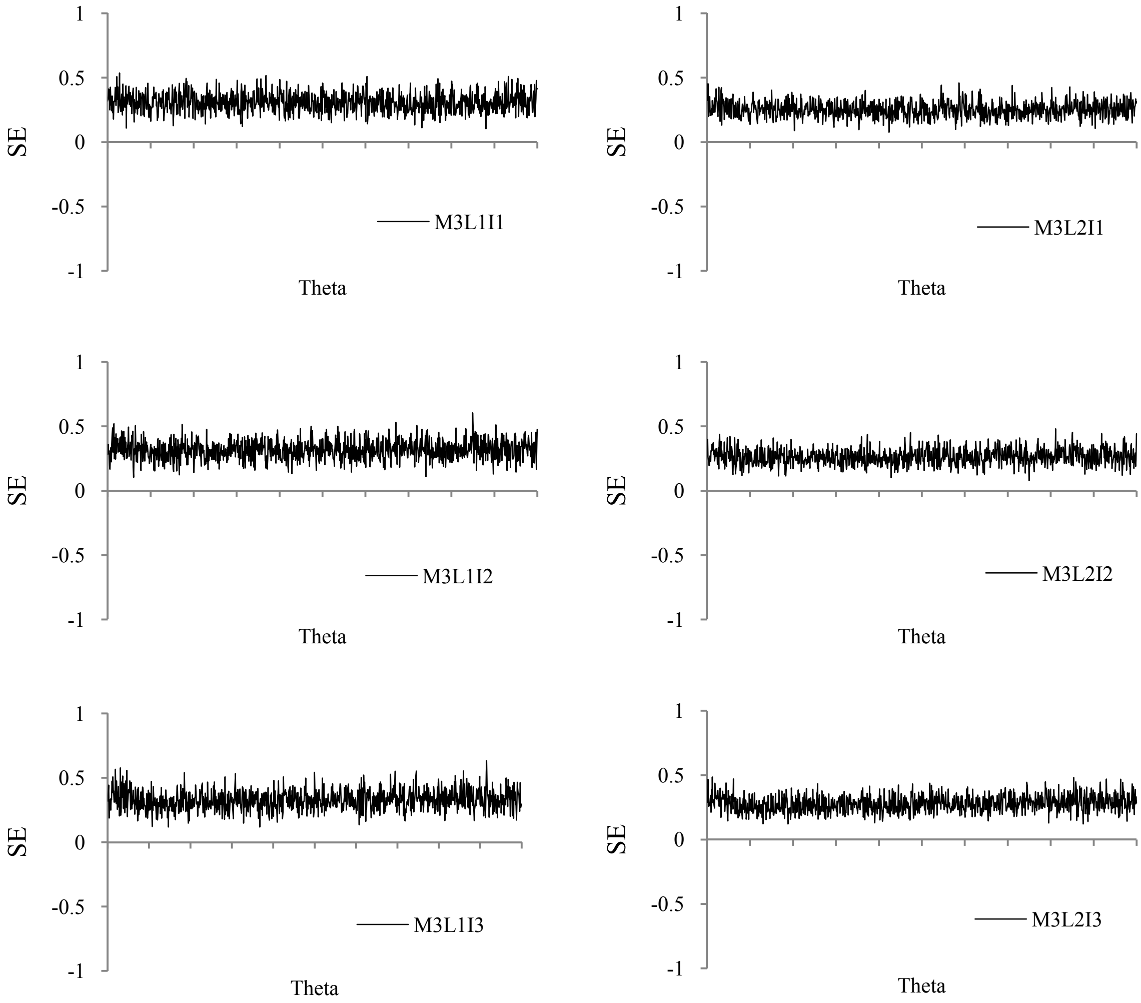


Figure 5. Distribution of Conditional SE Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using EAP (M=3) with One Replication

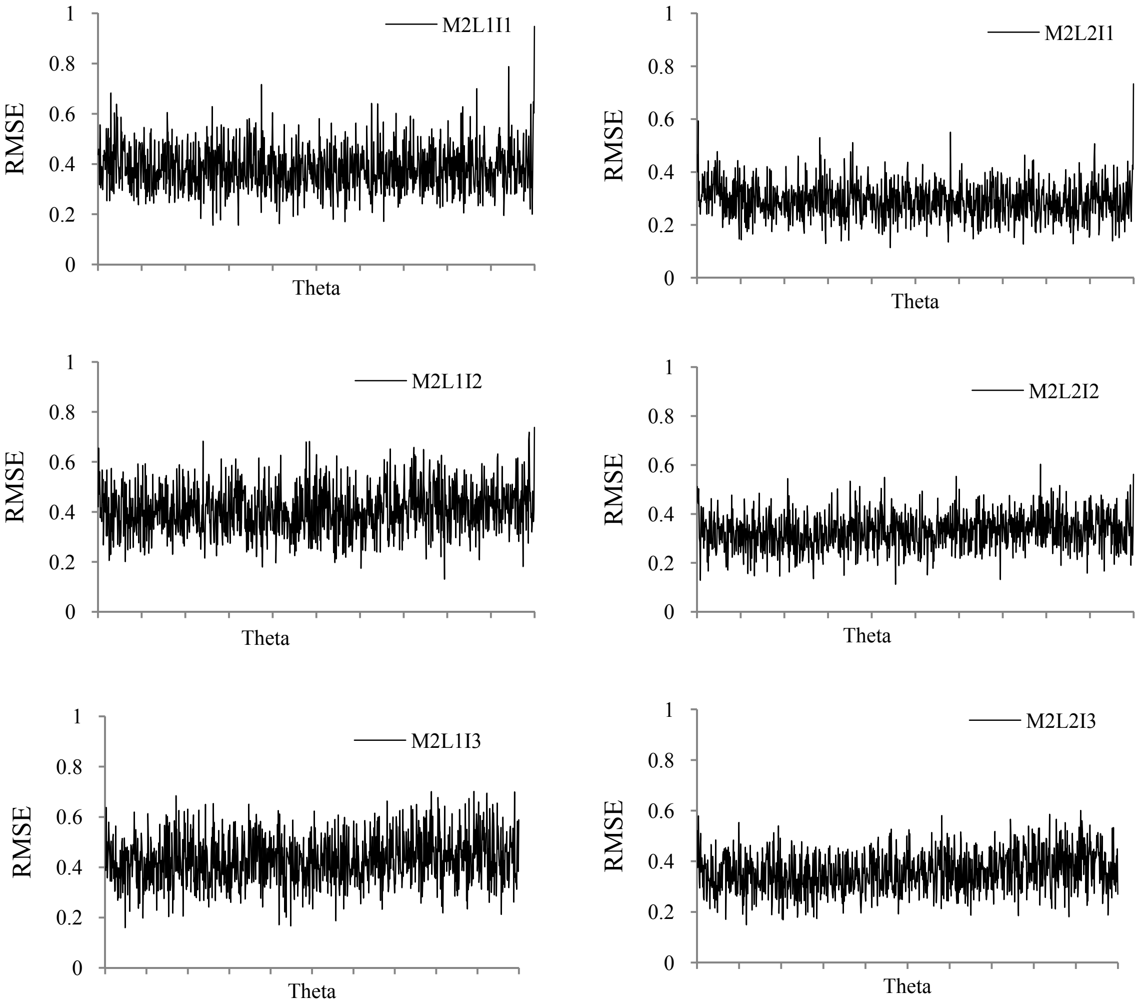


Figure 6. Distribution of Conditional RMSE Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using MLE (M=2) with One Replication

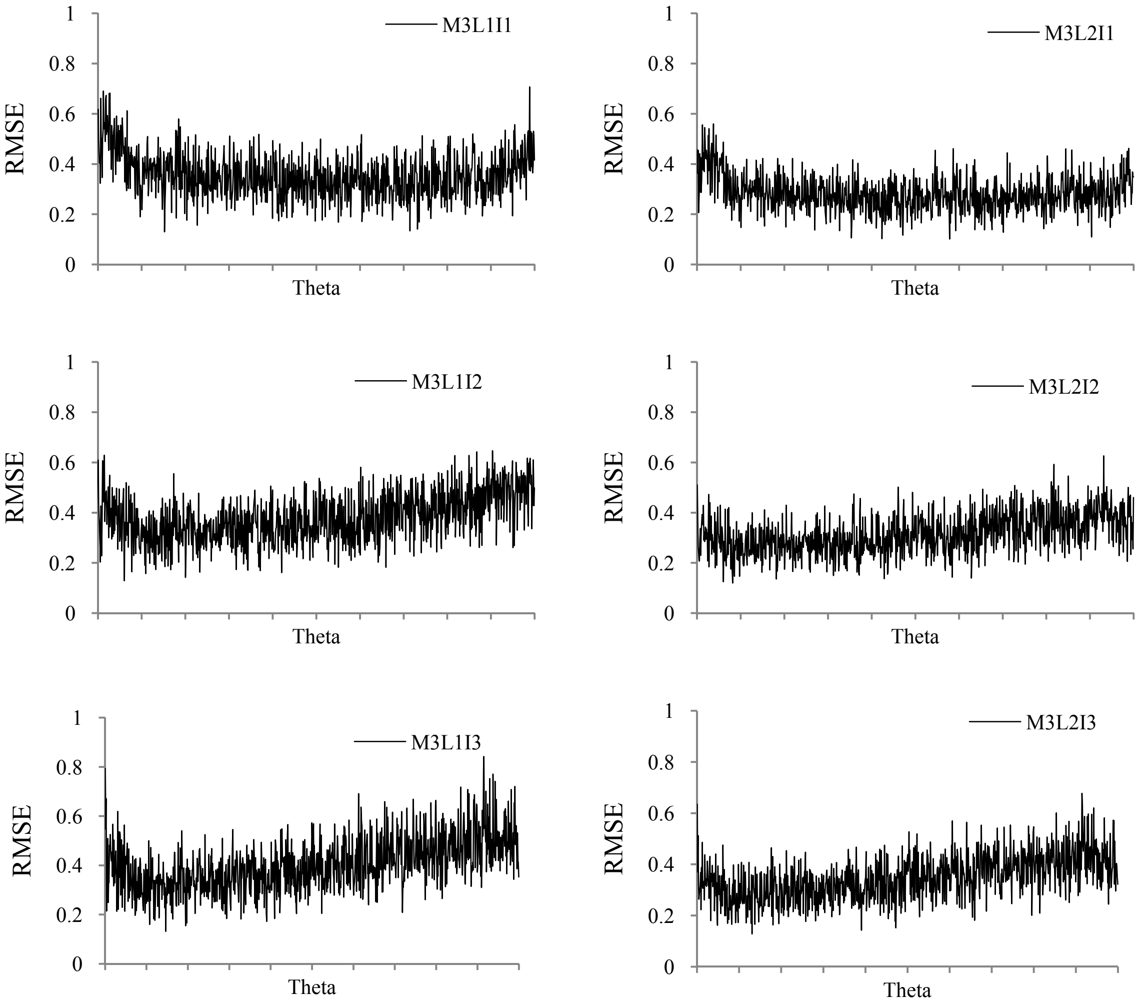


Figure 7. Distribution of Conditional RMSE Along Theta Scale for Different Test Length (L1=30 and L2=50) and ICC$_{\text{level-3}}$ (I=0, 0.2, 0.4) using MLE (M=2) with One Replication