WEIGHTING IN MULTILEVEL MODELS

By

Bing Tong

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods -- Doctor of Philosophy

2019

ABSTRACT

WEIGHTING IN MULTILEVEL MODELS

By

Bing Tong

Large-scale survey programs usually use complex sampling designs such as unequal probabilities of selection, stratifications, and/or clustering to collect data to save time and money. This leads to the necessity to incorporate sampling weights into multilevel models in order to obtain accurate estimates and valid inferences. However, the weighted multilevel estimators have been lately developed and minimal guidance is left on how to use sampling weights in multilevel models and which estimator is most appropriate.

The goal of this study is to examine the performance of multilevel pseudo maximum likelihood (MPML) estimation methods using different scaling techniques under the informative and non-informative condition in the context of a two-stage sampling design with unequal probabilities of selection. Monte Carlo simulation methods are used to evaluate the impacts of three factors, including informativeness of the sampling design, intraclass correlation coefficient (ICC), and estimation methods. Simulation results indicate that including sampling weights in the model still produce biased estimates for the school-level variance. In general, the weighted methods outperform the unweighted method in estimating intercept and student-level variance while the unweighted method outperforms the weighted methods for school-level variance estimation in the informative condition. In general, the cluster scaling estimation method is recommended in the informative sampling design. Under the non-informative condition, the unweighted method can be considered a better choice than the weighted methods for all the parameter estimates. Besides, the ICC has obvious effects on school-level variance estimates in

the informative condition, but in the non-informative condition, it also affects intercept estimates. An empirical study is included to illustrate the model.

This dissertation is dedicated to my family.

ACKNOWLEDGEMENTS

I have received a great deal of support and assistance throughout the writing of my dissertation. This dissertation could not be completed without their help.

I am especially indebted to my advisor and dissertation chair, Dr. Kimberly S. Kelly. With her encouragement, I chose MQM program. During my PhD career, she gave me tremendous help in my academic studies, and my spiritual life as well. Her expertise was invaluable in formulating the research topic.

I would like to acknowledge my committee members, Dr. Yuehua Cui, Dr. Richard Houang, and Dr. William Schmidt. I am grateful to them and appreciate them offering me enlightening feedback, enormous support and guidance with great patience.

My special thanks go to my CSTAT colleagues, including Dr. Frank Lawrence, Dr. Steven Pierce, Dr. Dhruv Sharman, Dr. Wenjuan Ma, Dr. Sarah Hession. In the last four and half years here, they have become my family members and I love to work with them. They have shared with me tremendous resources and insightful ideas. More importantly, they never hesitate to help me whenever I encounter any problems. I would never forget them and would miss every single of them in the future.

Nobody has been more important to me in the pursuit of this dissertation than my family members. I would like to thank my mom and my sisters, who always love me and support me without conditions. Most importantly, I would thank my beloved daughter, Shiyuan, who provides unending support. She is always there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ECLS-K: 2011	Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011
PML	Pseudo Maximum Likelihood
MPML	Multilevel Pseudo Maximum Likelihood
PML	Pseudo Maximum Likelihood
PWIGLS	Probability Weighted Iterative Generalized Least Squares
ICC	Intraclass Correlation Coefficient
RB	Relative Bias
RMSE	Root Mean Square Error
CR	Coverage Rate
UW	Unweighted Estimation Method
RW	Estimation Method with Raw Weights
CS	Estimation Method with Cluster Scaling
ES	Estimation Method with Effective Scaling
NAEP	National Assessment of Educational Progress
NCES	National Center for Education Statistics
NSF	National Science Foundation

CHAPTER 1 INTRODUCTION

A survey is defined as a data collection tool and is commonly used in social science to collect self-report data from study participants. It allows researchers to collect a large amount of data quickly and less expensively. Besides, the samples in survey research are often large, and a wide variety of variables can be examined (Boslaugh, 2007; Koziol, Bovaird, & Suarez, 2017), including personal facts, attitudes, previous behaviors, and opinions. Also, a survey can be often quickly created and easily administered. Thus, secondary data analysis is becoming increasingly popular (Stapleton, 2006). Many large-scale survey programs in social science use complex sampling designs to collect data, such as unequal probabilities of selection, stratification, and/or cluster sampling due to the impracticality of simple random sampling. In educational research, large scale data collection efforts such as National Assessment of Educational Progress (NAEP afterwards), Early Childhood Longitudinal Study-Kindergarten Class of 1998-1999 (ECLS-K afterwards), Early Childhood Longitudinal Study-Kindergarten Class of 2010-2011 (ECLS-K afterwards), available through National Center for Education Statistics (NCES) or National Science Foundation (NSF) use complex sampling plans. These three-stage surveys first involve sampling geographic areas with different probabilities of selection according to characteristics. These areas are often termed primary sampling units (PSUs). Then schools are sampled with different probabilities from the selected areas and lastly students are sampled from each of the selected schools, resulting in a cluster sampling design. Students chosen from the same school tend to be more alike than students chosen from other schools, and these groups of students show some degree of dependence (Hox & Kreft, 1994; Kish, 1965; Skinner, Holt & Smith, 1989) when compared to students from other schools. This type of sampling design brings challenges when

performing statistical analyses. If we disaggregate higher order variables to individual variables, ignoring the nested structure of the data and assuming each observation is independent, the assumption of independence of observations is not tenable. Conventional parametric analytic methods (e.g., regression, analysis of variance, *t*-tests) do not work well because they violate the assumption of observation independence (Cohen, West, & Aiken, 2003). The standard errors for the point estimates are estimated incorrectly, which could lead to erroneous conclusions arising from increased Type-I errors due to the violation of this assumption (Arceneaux & Nickerson, 2009; Clarke, 2008; Hahs-Vaughn, 2005; Heck & Mahoe, 2004; Judd, McClelland, & Ryan, 2009; Musca et al., 2011). However, if all the individual level variables are aggregated to the higher level, then important information could be lost. Multilevel models or Hierarchical linear models (HLM) were proposed and have been widely used in education, because they can be used to account for clustering, and allow the variance of the dependent variable to be partitioned explicitly into within- and between-variance (Lee & Fish, 2010; Lubienski & Lubienski, 2006; Palardy, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). They are an alternative to some of the approaches used by survey analysis for dealing with nested data structures.

Furthermore, some groups of the population are oversampled for various reasons. Units with higher data collection costs may be drawn with lower selection probabilities and individuals from small subpopulations of particular interest may be sampled with higher probabilities. For example, both ECLS-K and ECLS-K:2011 oversampled Asian, Native Hawaiians, and other Pacific islanders with the rate of 2.5 compared with other racial groups. This feature suggests applying sampling weights in the model to reflect the unequal probabilities of selection whenever selection probabilities are related to the outcome variable after conditioning on covariates in the model. The sampling design is said to be informative in this case (Fuller, 2009; Grilli & Pratesi,

2004). Ignoring this feature and without using weights, parameter estimates would be severely biased (Korn & Graubard, 1995; Pfeffermann, Skinner & Goldstein, 1998; Rodriguez & Goldman, 1995, 2001; Zaccarin & Donati, 2008).

But, appropriately using weights is not an easy task. For large-scale data sets, for example, ECLS-K:2011, there are many sampling weight variables, including school-level and student-level weights. For student level, this includes weights generated for the child assessments, teacher-level questionnaire, student-level questionnaire, parent interview, and care provider questionnaire. Appropriate use of complex sampling weights is of great importance because ignoring them may produce erroneous standard errors and consequently, inaccurate statistical inference. What's more, there is not much guidance on how to incorporate sampling weights in the multilevel models. It can be dated back from the late 1980s (e.g., Pfeffermann & LaVange, 1989). The pseudo maximum likelihood (PML) method, developed by Skinner (1989) and following the thoughts of Binder (1983), is a well-established estimation procedure for any weighted single-level models. However, flexible techniques for estimating weighted multilevel models have only newly been developed (cf., Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006; Koziol et al., 2017). One possible reason for this is multilevel weights are not available, which is often the case for public-released data file (Kovačević & Rai, 2003; Stapleton, 2012). The second reason might be that weighted multilevel modeling requires scaling of the lower level sampling weights (Pfeffermann et al., 1998). Currently, there is no well-established general multilevel consistent estimation method incorporating weights.

It is controversial whether to weight or not (Bertolet, 2008; Kish, 1992; Skinner, 1994; Smith, 1988; Xia & Torian, 2013). For example, on the one hand, some researchers (e.g., Graubard & Korn, 1996; Korn & Graubard, 1995, 2003; Lohr & Liu, 1994) suggested using sampling

weights in the model, as mentioned above to take into account for the complex sampling scheme. On the other hand, Winship and Radbill (1994) preferred unweighted estimators because estimates were unbiased, and consistent because they produced smaller standard errors. However, although the use of sampling weights will result in the increase of variance from unequal inclusion probabilities, it is still required and necessary because it prevents producing biased parameter estimates under informative sampling in multilevel models (Pfeffermann et al., 1998; Kim & Skinner, 2013), protects against misspecification, and makes full use of population-level information (Kim & Skinner, 2013).

The estimation quality can be affected by a number of factors and some of them have been investigated in the past research across different conditions, such as cluster size, distribution of the response variable, estimator/software program, informativeness of the sampling design, intraclass correlation coefficient (ICC), model type, invariance of selection across clusters, number of clusters, relative variance of weights, sample design features, and weight approximation method. In this study, I focus on the multilevel pseudo maximum likelihood (MPML) estimation method. First of all, although various conditions have been examined, conclusions are not inconclusive and rely on the particular model or sampling mechanism. Second, there are limited number of studies evaluating MPML (i.e., Asparouhov, 2006; Asparouhov & Muthén, 2006; Cai, 2013; Grilli & Pratesi, 2004; Koziol et al., 2017; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012). Third, MPML, compared with other estimators, are more flexible. Therefore, more studies are needed to evaluate MPML.

The purpose of the present study is to evaluate the performance of MPML using different scaling procedures in the context of a two-stage sampling design with unequal probabilities of selection in the informative and non-informative conditions across different levels of ICC using a

linear random-intercept model with covariates at both levels. Monte Carlo simulation methods are used to estimate the relative bias (RB), root mean square error (RMSE) and coverage rate/probability (CR) of the corresponding 95% confidence interval estimators. The following factors are manipulated: (a) informativeness; (b) ICC of the unconditional model; and (c) estimation method. All factors are fully crossed.

Cai (2013) conducted Monte Carlo simulations and found that the unweighted estimator produces biased estimates for the intercept and school-level variance, while the estimates for fixed effects and student-level variance are nearly unbiased within 10% of the true value in terms of Muthén and Muthén (2002). Generally speaking, the MPML estimators have higher coverage rates than the unweighted estimator in the informative condition. Including sampling weights increases MSE substantially and produces biased estimates for the intercept and school-level variance in the informative sampling design. Furthermore, ignoring informative sampling design could produce biased estimates. Pfeffermann et al. (1998) pointed out that the unweighted method only produced biased estimates for the intercept and school-level variance, not for student-level variance when the design is informative at school-level variance. Prior studies (e.g., Asparouhov & Muthén, 2006; Kovačević & Rai, 2003) show that as the ICC increases the bias decreases for all the parameters using an unconditional model. Asparouhov and Muthén (2007) also found that the MPML estimator outperforms substantially the other estimators.

The plan of this study is as follows. Chapter 2 discusses theoretical background and reviews the related literature. We briefly review multistage design and general multilevel models. Pseudo maximum likelihood estimation (MPML) method is presented, followed by two scaling methods. Intraclass correlation coefficient (ICC) and informativeness are also described in this section. In Chapter 3, I introduce the empirical data set I use in this study: ECLS-K:2011, and procedures of

simulation for the present study. Chapter 4 presents the results of the empirical data analysis and simulation analysis. Chapter 5 provides a discussion of overall findings, limitations, and topics for future research.

CHAPTER 2 THEORETICAL BACKGROUND AND LITERATURE REVIEW

2.1 Research Goal

Using empirical and simulated data, the present study focuses on examining the performance of MPML in the context of a two-stage sampling design with unequal probability of selection. Since MPML is newly developed compared to PML, there are far fewer studies examining MPML. And no consensus has been achieved on which one performs best and under which condition for the existing weighted multilevel estimators. MPML is considered the most flexible and popular method if the consistency of estimates and computation intensity are considered for multilevel data. But it is also obvious that weighted estimators produce larger standard errors than unweighted methods do. Therefore, it is controversial whether to use weight to not. More studies are needed to compare them and examine the performance of MPML. What's more, the scaling effect used in the multilevel estimation method is inconclusive based on the previous literature. Lastly, to my knowledge, except one study (c.f., Koziol et al., 2017), all other previous simulation studies manipulating ICC values use only an unconditional random intercept model.

Therefore, the main goal for this study is to examine the impact of sampling weights and to evaluate the performance of the MPML methods with different scaling techniques in the context of two-stage informative and non-informative sampling designs across different values of ICC with unequal probability of selection using random intercept model with covariates at both levels. Monte Carlo simulation methods are used to evaluate several factors, including: (a) informativeness of the sample design (non-informativeness vs. informativeness at both stages); (b) ICC with five different values; (c) estimation methods (unweighted, raw/unscaled weighted,

cluster scaling, effective scaling). All the factors are fully crossed. This gives rise to $2 \times 5 \times 4 = 40$ combination of conditions.

This study makes several contributions to the complex survey data literature. First, it provides a comparison between unweighted and weighted multilevel approaches in the context of unequal probability of selection. Second, it provides a comparison of estimation methods between informative and non-informative sampling design. Third, it provides a comparison of estimation methods under different levels of ICC values.

In order to cover the gaps of the current body of literature, the following research questions are addressed:

1. How do MPML estimators differ from unweighted estimator in multilevel models in the informative and non-informative sampling designs in terms of relative bias, root mean square error and 95% confidence interval coverage rate?
2. How does intraclass correlation influence the performance of estimators under the informative and non-informative condition in terms of relative bias, root mean square error and 95% confidence interval coverage rate?

Large-scale surveys in social studies usually use complex sampling designs based on the characteristics of the population to glean information in order to address various research questions. This feature brings challenges to the analysis. This chapter includes several topics which are central to understanding weighted multilevel analysis of survey data.

2.2 Multistage Sampling

Multistage designs are commonly used in many practical cases. For a two-stage sampling in the educational setting, for example, clusters or PSUs such as schools are selected in the first

stage. In the second stage, individual units, such as students are then sampled from the clusters. Each sampling stage corresponds to a multilevel model level. In this case, second stage corresponds to Level 1, first stage to Level 2.

At the first stage, cluster j is sampled with probability p_j, j =1, …, m, where m is the number of clusters to be sampled from the total number of clusters in the population, M. At the second stage, individual i is sampled from the cluster selected at the first stage with conditional probability $p_{i|j}$, i = 1, …, n, where n is the cluster sample size. Usually, clusters are sampled with probabilities that are proportional to their sizes, that is, the number of individual units in their *clusters,* S_j,

$$p_j = \frac{mS_j}{\sum_j^M S_j} \tag{2.1}$$

and the weight at cluster level is the inverse of the probability p_j, that is, $w_j = 1/p_j$. Each unit is sampled from cluster j with conditional probability (assuming that equal number of units are sampled from each cluster)

$$p_{i|j} = \frac{n}{S_j} \tag{2.2}$$

and the weight for individual unit i given cluster j is the inverse of the conditional probability $p_{i|j}$, that is, $w_{i|j} = 1/p_{i|j}$. Then the unconditional probability is defined as

$$p_{ij} = p_{i|j} * p_j = \frac{n\,m}{\sum_j^M S_j} \tag{2.3}$$

2.3 Multilevel Model

A typical two-level linear model can be specified with two equations. The first equation is used to describe the relationship between dependent variables and the covariates at the student level, within each group. Some or all of the parameters of the student-level equation are viewed as

varying randomly across the groups. The second equation, school-level equation, defines these parameters as dependent variables with the school-level variables as covariates. If we combine them together, a two-level linear mixed model can be specified in matrix vector form as follows, based on Laird and Ware (1982),

$$Y_j = X_j\alpha + Z_j b_j + e_j. \quad (2.4)$$

In the above equation, j indexes the cluster, with j = 1, …, m, where m is the number of clusters. For the jth cluster with size $i = 1, ..., n_j$, $\boldsymbol{Y_j}$ is an n_j x 1 vector of observed response, $\boldsymbol{X_j}$ is an n_j x p observed matrix for fixed effects, $\boldsymbol{\alpha}$ is a p x 1 vector of unknown coefficients, $\boldsymbol{Z_j}$ denotes an n_j x q random-effect design matrix, $\boldsymbol{b_j}$ is a q x 1 vector of cluster-specified random effects, and $\boldsymbol{e_j}$ is an n_j x 1 vector of random residual errors, where p is the number of unknown coefficients including the intercept and q is the number of random effects. Since random intercept model is used in the current study, q equals 1.

Either full maximum likelihood (ML/FIML) or the restricted maximum likelihood (REML) estimation method is often used to estimate the unknown model parameters in a general linear mixed model, such as fixed regression coefficients and variance components. Searle, Casella, and McCulloch (1992) defines the likelihood function for a linear mixed model as follows,

$$L\,(Y|X, Z, \alpha, D, \sigma_e^2) = \frac{exp(-\frac{1}{2}(Y-X\alpha)'\,V^{-1}(Y-X\alpha))}{(2\pi)^{\frac{N}{2}}|V|^{\frac{1}{2}}}, \quad (2.5)$$

where V is the covariance matrix of vector Y, $V = ZDZ' + \sigma_e^2 I$, D denotes covariance matrix for the random effect vector $\boldsymbol{b_i}$, and in our case, it is a scalar $\boldsymbol{\sigma_u^2}$, and $\boldsymbol{\sigma_e^2}$ is the variance of the error term. For computational convenience, the log likelihood function is more often used instead of likelihood function. It is specified in mathematical form as

$$l = log\left(L\left(Y|X, Z, \alpha, D, \sigma_e^2\right)\right) = -\frac{1}{2} Nlog(2\pi) - \frac{1}{2} log|V| - \frac{1}{2}(Y - X\alpha)' V^{-1}(Y - X\alpha) \quad (2.6)$$

where N is the total number of observations, $N = \sum_{j=1}^{m} n_j$.

2.4 Multilevel Pseudo-Maximum Likelihood (MPML) Estimation Methods

In order to achieve valid inference for the population, sampling weights must be used for all the levels of the data. But the literature does not obviously describe when and how to use sampling weights properly in the multilevel models. Using single-level weights to replace multilevel weights, is not always appropriate for the following reasons. First, sampling weights are placed into sum of squares and cross-products in a single-level regression. Final-level weights are the product of multilevel weights. Based on Christ, Biemer, & Wiesen (2007), if we use final-level weights, it might lead to biased estimates in multilevel models. Second, Pfeffermann et al., (1998) noted that single final-level weights or overall inclusion probabilities may not contain sufficient information to correct for unequal sampling probabilities at higher levels, because units at either level can be selected with differential probabilities. Therefore, multilevel weights need to be used in multilevel models. We use sample data and the sampling weights to estimate unknown parameters by maximizing the weighted sample likelihood.

So far, researchers have explored different estimation methods incorporating sampling weights for complex surveys, such as multilevel pseudo maximum likelihood (MPML) (Asparouhov, 2004, 2006; Grilli & Pratesi, 2004; Rabe-Hesketh & Skrondal, 2006), probability-weighted iterative generalized least squares (PWIGLS) (Pfeffermann et al., 1998), sample distribution methods (Eideh & Nathan, 2009; Pfeffermann, Moura, & Silva, 2006), weighted composite likelihood (WCL) estimation (Rao, Verret, & Hidiroglou, 2013), and pseudo empirical

likelihoods (Chaudhuri, Handcock, & Rendall, 2010; Chen & Sitter, 1999; Francisco & Fuller, 1991; Fuller, 1984; Lin, Steel, & Chambers, 2004; Rao & Wu, 2010; Scott & Holt, 1982). As Asparouhov & Muthén (2006) stated that there is no best estimation method for multilevel models if sampling weights are used. MPML method and PWIGLS method are the two most widely used estimation methods in multilevel models incorporating sampling weights. Compared with PWIGLS, MPML is more flexible and more widely applied, from the perspective of software implementation. Currently, MPML has been applied in the software of Stata, M*plus*, and SAS while PWIGLS has been used in LISRAEL, HLM and MLwiN. Different software would generate different output (Chantala, Blanchette, & Suchindran, 2011; Chantala & Suchindran, 2006). The application of MPML, compared with PWIGLS, requires less computational intensity and is much more flexible (Kovačević & Rai, 2003; Rabe-Hesketh & Skrondal, 2006). Besides, MPML can be applied to any general multilevel model (Rabe-Hesketh & Skrondal, 2006) just as the PML method can be used in any single-level models. The third advantage is that MPML is versatile and it can be modified for different estimation issues (Asparouhov, 2004; Asparouhov & Muthén, 2006). In addition, MPML can account for stratification and extra non-substantive clustering levels in the estimation of standard errors without having to incorporate such design features into the parameterization of the model (Asparouhov & Muthén, 2006; Koziol et al., 2017; Rabe-Hesketh & Skrondal, 2006). Because of these advantages, only the MPML with different scaling techniques is considered in the present study.

Let the estimates $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ be the parameters and the likelihood function for a general multilevel model can be expressed as

$$L(\theta_1, \theta_2) = \prod_{j=1}^{m} \left(\int \left(\prod_{i=1}^{n_j} f\left(y_{ij} \middle| x_{ij}, u_j, \theta_1\right) \phi(u_j | z_j, \theta_2) du_j \right) \right. \tag{2.7}$$

where y_{ij} is the response variable in cluster j = 1, …, m of individual i = 1, …, n and u_j the cluster-specific random effect; x_{ij} is student-level covariates and z_j the cluster level covariates; $f(y_{ij}|x_{ij}, u_j, \theta_1)$ is the density function of y_{ij} and $\phi(u_j|z_j, \theta_2)$ the density function of u_j, where θ_1 and θ_2 are the parameters to be estimated for the fixed effects for the student level and school level, respectively.

If weighting is incorporated into the analysis, and scaling procedures are also applied in order to reduce the bias arising from unequal probabilities of selection for complex survey data, the population likelihood function is directly estimated by weighting the sampling likelihood function,

$$L(\theta_1, \theta_2) = \prod_{j=1}^{m} \left(\int (\prod_{i=1}^{n_j} f(y_{ij}|x_{ij}, u_j, \theta_1)^{w_{i|j}\lambda_{2j}}) \phi(u_j|z_j, \theta_2) du_j \right)^{w_j \lambda_{1j}}, \qquad (2.8)$$

where $w_{i|j}$ = 1/ $p_{i|j}$ is student-level weights where $p_{i|j}$ is the conditional inclusion probability for the *i*th unit in the *j*th cluster, given that the *j*th cluster is sampled; w_j = 1/p_j is the school-level weights where p_j is the inclusion probability for the *j*th cluster; λ_{1j} and λ_{2j} are the scaling factors for the school-level and individual level sampling weights, respectively.

Numerical techniques are needed to integrate out the unobserved school-level random effect u_j to approximate the weighted likelihood.

Sandwich variance estimator is employed to obtain standard errors because some researchers (e.g., Huber, 1967; White, 1980) claimed that they are robust to nonnormality and heterogeneity. The asymptotic covariance matrix of the parameter θ using this method is defined by

$$(l'')^{-1}\mathrm{Var}(\mathrm{l}')\ (l'')^{-1} \qquad (2.9)$$

where ' and " refer to the first and second derivative of the log-likelihoods with respect to the parameters θ. M*plus* (Muthén & Muthén 1998-2017) implements this method using a robust variance estimator having the following form:

$$(\frac{\partial^2 log\, L}{\partial \theta^2})^{-1} (\sum_{j=1}^{m}\big((\lambda_{1j} w_j)^2\big) \frac{\partial log L}{\partial \theta} (\frac{\partial log L}{\partial \theta})')(\frac{\partial^2 log\, L}{\partial \theta^2})^{-1}. \quad (2.10)$$

2. 5 Scaling Sampling Weights for Multilevel Models

In multilevel weighted estimation literature, one of the main problems is the fact that the parameter estimates are usually only approximately unbiased. There are many factors that have substantial influence on the quality of the estimation, such as sample size of cluster, informativeness of selection, variability of sampling weights, intraclass correlation and scaling methods (Asparouhov, 2006; Asparouhov & Muthén, 2006; Bertolet, 2008; Cai, 2013; Grilli & Pratesi, 2004; Jia, Stokes, Harris, & Wang, 2011; Kovačević & Rai, 2003; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). For instance, parameter estimation would be severely biased when the cluster sample size is not sufficiently large enough (Asparouhov, 2006; Rabe-Hesketh and Skondal, 2006). In order to correct this, two scaling methods were proposed by Pfeffermann et al. (1998).

The scaling method is an indicator of how the weights are normalized at each level (Asparouhov, 2006). The first method, assuming individual level weights are approximately non-informative, may produce approximately unbiased estimator for both variance components. This approach produces a scaling factor so that the individual level weights equal the 'effective' cluster size (Longford, 1995, 1996; Pfeffermann et al., 1998). The scalar factor, which was referred to as "Method 1" in Pfeffermann et al. (1998), is specified as follows

$$\lambda_j = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} w_{i|j}^2}. \quad (2.11)$$

Method 2 in Pfeffermann et al. (1998) is used when both levels of sampling design are assumed to be informative. The scaling factor is defined as

$$\lambda_j = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}, \quad (2.12)$$

where n_j is the number of sample units in the *j*th cluster. The scaling factor is set so that the individual level weights equal the actual cluster size. These two scaling methods are termed as effective cluster scaling (ES) and cluster scaling (CS) respectively in the current study.

Currently, there is no consensus about which scaling method works better and under what conditions. For example, Pfeffermann et al. (1998) pointed out Method 2 (cluster scaling) works better in reducing bias in simulation in the informative sampling design while Stapleton (2002) found that Method 1 (effective cluster scaling) produces unbiased estimates in multilevel SEM analysis. Asparouhov (2006) noted that the different scaling methods may have different effects on different estimation techniques. If a scaling method performs well with the MPML approach, it does not necessarily mean that it performs well with other estimation techniques, for example, PWIGLS. Sometimes, which scaling method to use depends on the purpose of the research. If the main interest is point estimates, cluster scaling method is recommended. If cluster variance estimates are, then effective scaling method might be used (Asparouhov, 2006; Carle, 2009).

2.6 Intraclass Correlation Coefficient (ICC)

Besides sample size of cluster, informativeness of selection, variability of sampling weights, and scaling methods, ICC also affects estimation quality (Asparouhov, 2006; Asparouhov and Muthén, 2006; Bertolet, 2008; Cai, 2013; Grilli & Pratesi, 2004; Jia et al., 2011; Kovačević

& Rai, 2003; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). Prior studies have found that the larger the ICC values are, the less biased the estimates are in simulation studies manipulating ICCs using random intercept models without any covariates at both levels (Asparouhov, 2006; Jia et al., 2011; Kovačević & Rai, 2003).

ICC is one of the factors that is examined in this study. It can be used for model construction because it helps to determine the predictors which are most important to account for the outcome variable (Raudenbush & Bryk, 2002). It is also used as an index for including cluster level in multilevel modeling if ICC is not close to zero. Larger ICC values usually represent larger variations in cluster level, indicating larger proportion of total variance in the response variable that is accounted for by the clustering and thus larger clustering effect. In addition, the ICC value is informative for planning group-randomized experiments in education (Hedges & Hedberg, 2007, 2013).

To estimate the ICC for a given outcome, *y*, a multilevel model is fit for the *i*th student in the *j*th school

$$y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij}, \tag{2.13}$$

and the REML estimates of the variance of u_{0j}, (labeled as $\hat{\sigma}_u^2$), which is the variation between schools, and the variance of ε_{ij} (labeled as $\hat{\sigma}_e^2$), which represents variation at student level are used to compute ICC. The estimate of the ICC, $\hat{\rho}$, is then defined as

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2 + \hat{\sigma}_u^2}, \tag{2.14}$$

which is the proportion of total variability in scores due to the school-to-school differences.

Moreover, the ICC is used to calculate the design effect, which shows how much standard errors are underestimated. The design effect is defined as follows

$$\text{Designeffect} = 1 + (\text{averageclsutersize} - 1) * \hat{\rho}. \tag{2.15}$$

Based on Kish (1965), a design effect which is greater than 2 indicates that we need to take into account the clustering effect of the data during estimation.

2.7 Informativeness of Selection

The informativeness of selection, according to Asparouhov (2006), is an indicator of how biased the selection is. If the sampling design is informative, the inclusion probabilities are related to the response variable after conditioning on the variables in the model (Fuller, 2009; Grilli & Pratesi, 2004). Otherwise, it is non-informative. Pfeffermann (1993) and Cai (2013) pointed out that if weights are informative, they are quite influential on the results and therefore, should be considered in the multilevel analysis. However, if the sampling designs or weights are not informative, the effect of weights could be negligible and it is not necessary to include weights in the analysis. Therefore, to check whether the sampling design/weight is informative or not is necessary. Following Laukaityte and Wiberg (2018), weights are informative if the effective sample size is smaller than the real sample size. Effective sample size for two-level models can be defined as follows. Effective sample size at level 2 (between schools) is calculated using the following formulas:

$$N^{eff} = \frac{(\sum_j w_j)^2}{\sum_j (w_j^2)} \tag{2.16}$$

and effective sample size at level 1 (within schools) for school j is obtained by

$$n_j^{eff} = \frac{(\sum_j w_{i|j})^2}{\sum_j (w_{i|j}^2)}. \tag{2.17}$$

Pfeffermann (1993) developed a model to evaluate whether the sampling design is informative or not. The informativeness of sampling design is examined by the x^2 test, which is defined as follows

$$I = (\hat{\theta}_w - \hat{\theta}_0)'[\hat{V}(\hat{\theta}_w) - \hat{V}(\hat{\theta}_0)]^{-1}(\hat{\theta}_w - \hat{\theta}_0) \sim x_p^2 \quad (2.18)$$

where $\hat{\theta}_w$ and $\hat{\theta}_0$ are the estimates of weighted and unweighted analyses, respectively, and $\hat{V}(\hat{\theta}_w)$ and $\hat{V}(\hat{\theta}_0)$ are their variance estimates. The informativeness statistic follows a x_p^2 distribution with $p = dim(\theta)$ degrees of freedom.

CHAPTER 3 METHODS

Two primary sections are included in this chapter: one introduces methods for empirical data; one introduces simulation design.

3.1 Empirical Data

3.1.1 Data and Variables

This study uses data from the public-use the Early Childhood Longitudinal Study, Kindergarten Class of 2010–2011 (ECLS-K: 2011, see Mulligan, Hastedt, & McCarroll, 2012, for an overview) data set, which is sponsored by the National Center for Education Statistics (NCES). It is a latest study in early childhood longitudinal study that follows a U.S. nationally representative sample of students entering Kindergarten in 2011-2012 to the spring of 2016, fifth grade. ECLS-K:2011 provides descriptive information about children's school experience. Data have been collected related to family, classroom and school environment. Individual variables are available as well, studying how cognitive, social and emotional development is related to them.

The ECLS-K: 2011 data are not a simple random sample of individuals or clusters. The study employed a 3-stage cluster sampling design. 90 geographic areas (counties or groups of counties) as the primary sampling units (PSUs) were first sampled at stage 1. Then samples of public and private schools were selected at stage 2 from the selected PSUs. Lastly, five-year-old children were randomly sampled within selected schools at stage 3. Stratification and probability proportional to size sampling were used at the first two stages of selection; stratification and unequal sampling were used at the final stage. In the base year, Asian, Native Hawaiians, and other Pacific islanders were oversampled. The user's manual for the ECLS-K: 2011 kindergarten data

file and electronic codebook, public version (Tourangeau et al., 2015) offers an excellent overview of the characteristics of complex sample designs including clustering, stratification, unequal probabilities of selection, and non-response and poststratification.

The analytic samples in this paper only include kids in kindergarten, and data collected in both the fall and the spring semesters. Approximately 18,200 children enrolled in 970 schools during the 2010-11 school year participated during their kindergarten year.

Although the use of sampling weights will result in the increase of variance due to unequal inclusion probabilities, it is still required and necessary because it prevents producing biased parameter estimates under informative sampling in multilevel models (Pfeffermann et al., 1998; Kim & Skinner, 2013), protects against misspecification, and makes full use of population-level information (Kim & Skinner, 2013). The supplied sampling weights adjusted for school-level nonresponse and inverses of estimated student-level response probability are used. Weights for first sampling stage are not available. For student level, I use composite variables based on the parent survey as the primary independent variables of interest, as well as controlling for the student's fall test score in order to predict the spring score. The parent is used as a primary component to adjust for non-response, suggesting that child base weight adjusting for non-response associated with either fall or spring kindergarten parent interviews (W1_2P0) would be a good choice of weight. For school-level weight, school base weight adjusted for non-response associated with the school administrator questionnaire (W2SCH0) are used.

The academic outcome variables in this study are reading and mathematics scale scores calibrated using Item Response Theory (IRT) procedures. The reading assessment (User's Manual for the ECLS-K:2011, Mulligan et al., 2012) measures basic skills (print familiarity, letter recognition, beginning and ending sounds, rhyming words, word recognition), vocabulary

knowledge, and reading comprehension. Reading comprehension consists of questions identifying information specifically in text, making complex inferences within and across texts, and considering the text objectively to judge its appropriateness and quality. The mathematics assessment measures skills in conceptual knowledge, procedural knowledge, and problem solving.

The construct validity has been established for ECLS-K:2011 assessments as the assessment, national and state performance standards in each of the domains were examined and specifications for reading and mathematics were established based on NAEP framework. Furthermore, curriculum specialists in the subject areas were recruited and the pool of items created were examined for content and framework strand design, accuracy, on-ambiguity of response options, and appropriate formatting.

The reliability of the reading score for Fall and Spring Kindergarten is 0.95, and the reliability of the mathematics score is 0.92 for Fall Kindergarten, 0.94 for Spring. The kindergarten mathematics mean score for this study's sample was 45.28 (*SD* = 12.19). For reading, the sample's kindergarten mean score was 61.26 (*SD* = 13.56). To model mathematics and reading achievements, we use three student-level covariates and two school-level covariates. Descriptive statistics of these variables are presented in Table 3.1.

Table 3.1. *ECLS-K: 2011 Variable Descriptive Statistics*

Variables	Variable_in_the_data	Mean	SD	MIN	MAX
Math	X2MSCALK2	45.28	12.19	7.19	88.76
Reading	X2RSCALK2	61.26	13.56	25.68	109.92
Pre_Math	X1MSCALK2	31.67	11.37	7.19	111.58
Pre_Reading	X1RSCALK2	46.92	11.5	25.45	109.92
SES	X12SESL	-0.05	0.81	-2.33	2.6
Female	X_CHSEX_R	0.49	0.5	0	1
School_Locale	X2LOCALE				
Suburban		0.36	0.48	0	1
Rural		0.22	0.42	0	1
Student_Weight	W1_2P0	223.08	141.71	0	956.72
School_Weight	W2SCH0	64.24	47.86	0	372.02

Note: SD=standard deviation; MIN=minimum; MAX=maximum.

3.1.2 Statistical Models

The unexplained variance among randomly sampled clusters (e.g., schools) in outcomes of interest could be inferred by using multilevel models. The effects of covariates at each level could also be estimated. Researchers could use models with random intercepts to account for the correlations within clusters caused by longitudinal or clustered design (West et al., 2015). In a survey with multistage samples, there are always various levels of cluster. But only the lowest level of clustering usually has the greatest impact on individual outcome (Asparouhov & Muthén, 2006). Furthermore, Stapleton and Kang (2016) found minor impacts could be found on inference and no difference could be detected even if we disregard the first stage sampling design which is beyond the levels in the model. For large-scale data sets, the first-stage weights are usually not provided, for example, ECLS-K: 2011. Hence this first stage sampling design is not considered in this study. Therefore, for simplicity, two-level random intercept regression models are used in this study to fit multilevel models in which individual students are nested in schools to two academic

dependent variables, reading IRT scale score, and mathematics IRT scale scores. But I would not take account of IRT measurement errors in the analysis. Three different two-level models are examined with different sets of covariates. Model 1 is an unconditional model without any covariates at both levels, model 2 includes all the student level predictors and model 3 is a full model consisting of all the student level and school level predictors.

Model 1: unconditional model

$$\text{Level 1: } y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{3.1}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j} \tag{3.2}$$

$$\text{Combined: } y_{ij} = \gamma_{00} + u_{0j} + \varepsilon_{ij} \tag{3.3}$$

Model 2: student model with three student-level predictors

$$\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}\text{*Female} + \beta_{2j}\text{*SES} + \beta_{3j}\text{*Pretest} + \varepsilon_{ij} \tag{3.4}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j} \tag{3.5}$$

$$\text{Combined: } y_{ij} = \gamma_{00} + \beta_{1j}\text{*Female} + \beta_{2j}\text{*SES} + \beta_{3j}\text{*Pretest} + u_{0j} + \varepsilon_{ij} \tag{3.6}$$

Model 3: full model including two level covariates

$$\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}\text{*Female} + \beta_{2j}\text{*SES} + \beta_{3j}\text{*Pretest} + \varepsilon_{ij} \tag{3.7}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\text{*Suburb} + \gamma_{02}\text{*Rural} + \gamma_{02}\text{*Suburban} + u_{0j} \tag{3.8}$$

$$\text{Combined: } y_{ij} = \gamma_{00} + \beta_{1j}\text{*Female} + \beta_{2j}\text{*SES} + \beta_{3j}\text{*Pretest} + \gamma_{01}\text{*Suburb} + \gamma_{02}\text{*Rural} + \gamma_{02}\text{*Suburban} + u_{0j} + \varepsilon_{ij} \tag{3.9}$$

Since there are many factors affecting the quality of estimation in complex sampling design, it is noteworthy to investigate both unweighted and weighted models. In this study, all the three multilevel models above are explored using the following four estimation methods:

(a) maximum likelihood estimation method with no weights (UW),

(b) MPML using raw /unscaled weights (RW),

(c) MPML using cluster scaling (CS),

(d) MPML using effective cluster scaling (ES).

The missing data at level 1 ranges from 0.2% for female to 14.2% for math pretest. Listwise deletion is used for handling missing data for the empirical study. Multiple imputation can be used in this case, but the exact models for real data is less important here. So listwise deletion is used to simplify the problem. Missing data at level 2 is 3.6% for rural and suburban. Level 2 missing values cannot be simply removed because they have impact on the lower level. Schafter and Graham (2002) mentioned if the probabilities of missingness only depended on observed items, missing data could be assumed to be missing at random (MAR afterwards). Therefore, I assume missingness at level 2 here is MAR. Two methods are recommended for handling MAR data. One method is multiple imputation method (Robin, 1987; Enders, 2010; Howell, 2008), and the other is the full-information maximum likelihood (FIML) method (Danielsen, Wiium, Wilhelmsen, & Wold, 2010; Enders, 2010; Laukaityte & Weibert, 2018). I use FIML for handling missing data in this study.

3.2 Simulations

3.2.1 Simulation Design

The informativeness (Asparouhov, 2006; Cai, 2013) and the intraclass correlation were found to be influential factors on the performance of weighted estimation in multilevel models (Asparouhov, 2006; Jia et al., 2011; Kovačević & Rai, 2003). Monte Carlo simulation methods are applied to evaluate the effect of ICC and examine the performance of MPML using different scaling techniques in the context of two-stage informative and non-informative sampling design

(please see Table 3.2 *Simulation Design*). All the conditions are fully crossed. The full study design results in a total of 2 × 5 × 4= 40 simulation settings.

Table 3.2. *Simulation Design*

design	ICC	UW	RW	CS	ES
Informative	ICC=0.5				
	ICC=0.3				
	ICC=0.2				
	ICC=0.1				
	ICC=0.01				
Non-Informative	ICC=0.5				
	ICC=0.3				
	ICC=0.2				
	ICC=0.1				
	ICC=0.01				

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES= estimation method with effective cluster scaling.

Five different ICC values are used in this simulation: 0.5, 0.3, 0.2, 0.1, and 0.01. The unconditional ICCs that may typically be found in educational and psychological research in the United States are in the range of 0.15 and 0.25 for academic large-scale assessments (Bloom, Bos, & Lee, 1999; Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007, 2013; Kreft & Yoon, 1994; Schochet, 2008). Accordingly, the values of 0.1, 0.2, and 0.3 are chosen for this study. The lowest ICC value found in Hedges and Hedberg (2013) is 0.02, in which students were nested in grades for each state. Raykov (2015) showed that the lower bound of 95% confidence interval of ICC could be as low as 0.014. Murry and short (1995) found that in a school-based intervention design, ICC values were generally smaller, in the range of 0.01 to 0.05. The current study considers students may be nested in school district, or even larger geographic areas, which may result in a lower ICC value. Therefore 0.01, a very small non-zero value, is chosen, because small ICC still

affects estimates of standard errors if we ignore the dependency. Musca et al. (2011) said small ICC would impact Type-I error dramatically.

Different values for $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ are used while the total variance of *y* is kept fixed, $\hat{\sigma}_u^2 + \hat{\sigma}_e^2$ = 60. This value is determined based on the empirical data results (See Table 3.5). Five different ICC values 0.5, 0.3, 0.2, 0.1, and 0.01 are obtained by setting $\hat{\sigma}_u^2$ to be 30, 18, 12, 6 and 0.6 respectively, while the value of $\hat{\sigma}_e^2$ is 60 - $\hat{\sigma}_u^2$, i.e., 30, 42, 48, 54, and 59.4 correspondingly.

3.2.2 Model

To evaluate the performance of MPML approach for a linear two-level regression model under informative and non-informative sampling condition, the Monte Carlo simulation mimics the sampling design in ECLS-K:2011. Specifically, about 18,200 kindergarteners from 970 schools were sampled. Overall, about 19 students were selected on average from each school. Mulligan et al. (2012) indicated that the school and student selection probability (i.e., sampling rate) is 0.02 and 0.25 respectively and the overall student selection probability is 0.02 × 0.25 = 0.005. All the school population are categorized into six groups based on the percentages of public schools in ECLS-K:2011: 5.69% of schools have students ranging from 16 to 24; 11.49% of schools have students ranging from 25 to 49, 43.53% of schools have students varying from 50 to 99, 25.3% of schools varying students from 100 to 149, 8.59% of schools have students ranging from 150 to 199, then 5.22% of schools have more than 200 students. Then finally 150 schools and 3915 students are drawn from the population with the expected sampling rate for schools and students in ECLS-K:2011. The true values for the parameters are all obtained using the empirical data set ECLS-K: 2011 with maximum likelihood estimation method (see Table 3.5). Thus, the data are generated using the following model:

$$y_{ij} = 17.43 + 0.91*\text{Female} + 1.06*\text{SES} + 0.92*\text{Pretest} + 1.04*\text{Rural} + u_j + \varepsilon_{ij} \qquad (3.10)$$

where u_j is school-level random effect and ε_{ij} is student-level error term, u_j and ε_{ij} are normally distributed with mean of 0 and variance 30, 18, 12, 6, and 0.6 for $\hat{\sigma}_u^2$, and corresponding variance of 60 - $\hat{\sigma}_u^2$ for $\hat{\sigma}_e^2$. Explanatory variables (e.g., female, social economic status (SES), pretest, rural and suburban) are determined because they contribute significantly to the model and are also variables other researchers are also interested in (e.g., Hedberg, 2016; Hedges & Hedberg, 2007). Female follows Bernoulli distribution with probability of 0.49. Social economic status (SES) follows normal distribution with mean -0.05 and variance 0.66 (*SD* = 0.81). Pretest score follows normal distribution with mean 46.92 and variance 132.22 (*SD* = 11.50). Suburban follows Bernoulli distribution with probability of 0.36. Rural follows Bernoulli distribution with probability of 0.22.

3.2.3 Sampling Selection

Finite population are generated according to the model described above. The expected sampling rate used in this study is still 0.02 for schools and 0.25 for students as in ECLS-K: 2011, which results in the overall sampling rate of 0.005.

Sampling selection is determined by whether the sampling design is informative or non-informative. In order to introduce unequal probability sampling at both levels and make our sampling design informative, the present study uses the similar plan used by Asparouhov (2006), Cai (2013) and Koziol et al. (2017). Poisson sampling is used to select the *j*th school with probability:

$$prob\,(I_j = 1) = \frac{1}{1 + exp(-\frac{\tilde{u}_{0j}}{2} + 4.02)} \tag{3.11}$$

where the $\tilde{u}_{0j}$ is equal to u_{0j} (the random intercept effect for the *j*th cluster) but rescaled to have a variance of 2. For the selected school, Poisson sampling is used to select the *i*th student within the *j*th school with probability:

$$prob\,(I_{i|j} = 1) = \frac{1}{1 + exp(-\frac{\tilde{e}_{ij}}{2} + 1.23)}. \quad (3.12)$$

The $\tilde{e}_{0j}$ is equal to e_{0j} (the residual effect for the *i*th student in the *j*th cluster) but rescaled to have a variance of 2. This sampling plan results in a design which is informative at both levels, because at both levels, the inclusion probabilities are linked to the response variable, according to the definition of sampling design informativeness (c.f., Fuller, 2009; Grilli & Pratesi, 2004).

The random variable variance is rescaled in order to keep a constant level of informativeness across different levels of the ICC. A variance of 2 for both random variables and the slope coefficients (1/2) are selected to have approximately 0.3 of informativeness for both the school level and student level, which Asparouhov (2006) used as a moderate level of informativeness in his simulations. The intercept values (4.12 and 1.23 for school level and student level, respectively) are determined using expected sampling rates (0.02 and 0.25 for the school level and the student level, respectively) and the formulas above (equation 3.11 and 3.12) to obtain desired sample sizes.

Under the non-informative sampling condition, $\tilde{u}_{0j}$ and $\tilde{e}_{ij}$ are replaced by other variables that are not part of the population model. Still Poisson sampling is used to select the *j*th school with probability

$$prob\,(I_j = 1) = \frac{1}{1+exp(-\frac{\beta_{0j}}{2}+4.02)} \quad (3.13)$$

where β_{0j} ~ N (0, 2) and is not related to any variables in the model. Conditional on the selected school, Poisson sampling is used to select the *i*th student in the *j*th school with probability of

$$prob\,(I_{i|j} = 1) = \frac{1}{1+exp(-\frac{r_{ij}}{2}+1.23)}. \quad (3.14)$$

where r_{ij} ~ N (0, 2) and is not related to any variables in the model. Although this design uses unequal probability of selection, it is not informative, because the selection probability is not related to the response variable.

Data are generated using the software Stata. The syntax for data generation is provided in APPENDIX A and APPENDIX B.

3.2.4 M*plus* and Data Analysis

Each simulation is replicated 1000 times for each study condition. Each 1000 replications are analyzed in M*plus* Version 8 (Muthén & Muthén, 1998-2017) using the TYPE = MONTECARLO option under the M*plus* DATA command. The M*plus* user's manual (Muthén & Muthén, 1998-2017) provides guidance on how to incorporate sampling weights and how to use scaling methods in a two-level model.

The two scaling methods that are used are referred to ECLUSTER and CLUSTER respectively in M*plus* documentation, which correspond to effective cluster scaling and clustering scaling respectively in this study.

Altogether, four estimation methods are considered: (a) unweighted estimation method (UW); (b) MPML method using raw/unscaled weights (RW); (c) MPML method using cluster scaled (CS) weights, and (d) MPML method using effective cluster scaled (ES) weights.

Then Sandwich variance estimators (ESTIMATOR = MLR) are used in all instances. The TYPE option is set to TWOLEVEL, and appropriate variables are identified for the CLUSTER, WEIGHT, and BWEIGHT options. For MPML models, WTSCALE and BWTSCALE are also specified based on different scaling methods: UNSCALED and UNSCALED are used respectively for raw scaling method, CLUSTER and SAMPLE for cluster scaling method, and ECLUSTER and SAMPLE for effective scaling method for three weighted methods respectively. For a general

multilevel model ignoring weighting in the present study, WTSCALE and BWTSCALE are not used under the VARIABLE command.

3.2.5 Evaluation Criteria

Empirical (absolute) Relative Bias, Root Mean Square Error (RMSE), and 95% Confidence Interval Coverage Rate are used as the primary criteria to estimate the quality of the performance of the estimators as previous simulation studies (e.g., Cai, 2013; Eideh & Nathan, 2009). In measurement or sampling situations, bias is defined as "the difference between a population mean of the measurements or test results and an accepted reference or true value" (Bainbridge, 1985). Then the true value can be under- or overestimated. Since large number of replications are applied in this study, even small values of bias may be deemed significantly different from 0. As such, the relative bias instead of bias is used. The relative bias is defined as

$$RBias(\hat{\theta}) = \frac{1}{\theta}\left(\frac{1}{1000}\sum_{1}^{1000}(\hat{\theta}_i - \theta)\right). \quad (3.15)$$

where θ is the true value set, and $\hat{\theta}_i$ is the estimated value in each iteration. It is noted in Muthén and Muthén (2002) that, if the absolute relative bias is less than 10% of the true value, then the parameter estimates can be considered unbiased.

A common accuracy measure called mean square error (MSE) is the mean of the squared differences. It indicates how close the estimate is to the true value. This measure incorporates concepts of bias and precision because it equals to the sum of the variance of the estimates and the squared mean error. The root MSE (RMSE) tells us how far the approximation will be from the true value on average. RMSE is used because it can penalize large values. It is computed using the following formula

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{1000}\sum_{1}^{1000}(\hat{\theta}_i - \overline{\hat{\theta}})^2}, \quad (3.16)$$

where $\overline{\hat{\theta}} = \frac{1}{1000} \sum_{1}^{1000} \hat{\theta}_i$. The smaller the RMSE is, the better the estimate is.

The coverage rate/probability (CR) in this study is set at 95%. It is utilized to evaluate the proportion of replication in each parameter estimate that the interval estimator contains the population parameter value (Muthén & Muthén, 1998-2017). It is recommended that the coverage rate should be at least 0.91 by Muthén & Muthén (2002). That is, at least 91% of replications having true parameter values within the 95% confidence interval.

M*plus* syntax for analysis is provided in APPENDIX C.

CHAPTER 4 RESULTS

This chapter consists of two primary sections: one for simulation results, the other for empirical study results.

4.1 Simulation Results

The primary evaluation criteria are (absolute) relative bias, root mean square error (RMSE) and coverage rate of the interval estimators. Simulation results are depicted in Table 4.1-4.6 and Figure 4.1-4.16. Table 4.1-4.2 illustrate the Monte Carlo estimates of relative bias, RMSE and 95% confidence interval coverage rate for the fixed effects, intercept and variance components in the informative condition, Table 4.3-4.4 for those in the non-informative condition. Table 4.5-4.6 display the average standard errors of the estimates and the standard deviations in the informative and non-informative design respectively. Figure 4.1-4.2, and Figure 4.13-4.14 plot relative bias for the four covariates, intercept and variance components in the informative condition, and Figure 4.3-4.4 and Figure 4.15-4.16 for those in the non-informative condition. Dashed horizontal lines indicate bounds for acceptable levels of relative bias ($|RB\%| \leq 10$; Muthén & Muthén, 2002). Figure 4.5-4.6 plot RMSE for the four covariates and intercept and variance components in the informative design and Figure 4.7-4.8 for those in the non-informative design. Figure 4.9-4.10 plot coverage rate for the four covariates and intercept and variance components in the informative design and Figure 4.11-4.12 for those in the non-informative design. Dashed horizontal lines indicate the nominal coverage rate of 95%.

Results are organized by research questions and evaluation criteria. Under each evaluation criteria, the results are illustrated by informative and non-informative condition respectively.

4.1.1 Research Question One

Research question one allows me to evaluate the performance of weighted and unweighted estimators under the informative and non-informative condition in terms of (absolute) relative bias, RMSE, and 95% confidence interval coverage rate. Comparison between unweighted and weighted estimators can give us a picture understanding whether differences among them are due to sampling weights application and which estimator performs best.

4.1.1.1 (Absolute) Relative Bias

In general, all the fixed effects are estimated somewhat unbiasedly in both informative and non-informative conditions if the criterion of Muthén and Muthén (2002) is applied. However, a different story can be told for the intercept and variance components estimates. On average, the absolute relative bias is comparatively larger in magnitude under the informative condition than that in the non-informative design. The most variability in the absolute relative bias occurs for the school-level variance estimators in both conditions.

4.1.1.1.1 Informative Design

From the presented simulation results in Table 4.1 and Figure 4.1, it is evident that all the estimates of absolute relative bias for the four fixed effects are less than 10% of the true value and can be considered unbiased if the criterion of Muthén and Muthén (2002) is used across the four estimators. The absolute relative biases for the three student-level covariates (i.e., female, SES and pretest) are less than or close to 1%. Although the relative bias for the school-level covariate (i.e., rural) is higher than those of student-level covariates, it is still within 10% of the true value.

Table 4.2 and Figure 4.2 show that the intercept and student-level variance are unbiasedly estimated (in terms of Muthén & Muthén, 2002) except for the intercept estimate in the unweighted

Table 4.1. *RB(%), RMSE, 95% CI CR for Covariates in the Informative Design*

	Informative																			
	ICC=0.5				ICC=0.3				ICC=0.2				ICC=0.1				ICC=0.01			
Covariate	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR
female	0.910																			
UW	0.904	-0.714	0.198	0.944	0.903	-0.802	0.234	0.944	0.903	-0.824	0.250	0.943	0.903	-0.813	0.264	0.943	0.903	-0.725	0.274	0.941
RW	0.906	-0.462	0.318	0.932	0.905	-0.549	0.376	0.932	0.905	-0.593	0.402	0.932	0.904	-0.626	0.426	0.932	0.904	-0.626	0.446	0.932
CS	0.905	-0.527	0.310	0.936	0.904	-0.626	0.367	0.937	0.904	-0.681	0.393	0.936	0.904	-0.714	0.416	0.937	0.904	-0.648	0.433	0.935
ES	0.905	-0.560	0.290	0.942	0.904	-0.692	0.344	0.940	0.903	-0.747	0.368	0.941	0.903	-0.813	0.390	0.944	0.903	-0.780	0.404	0.944
SES	1.060																			
UW	1.055	-0.481	0.125	0.954	1.054	-0.575	0.148	0.955	1.054	-0.613	0.157	0.954	1.053	-0.651	0.167	0.954	1.053	-0.651	0.173	0.959
RW	1.050	-0.925	0.21	0.939	1.048	-1.094	0.244	0.939	1.048	-1.160	0.261	0.938	1.047	-1.226	0.277	0.938	1.046	-1.283	0.290	0.938
CS	1.051	-0.830	0.200	0.939	1.050	-0.972	0.237	0.939	1.049	-1.038	0.255	0.939	1.048	-1.104	0.271	0.938	1.047	-1.217	0.283	0.945
ES	1.051	-0.821	0.188	0.938	1.050	-0.962	0.223	0.939	1.049	-1.028	0.239	0.940	1.048	-1.094	0.254	0.936	1.047	-1.226	0.265	0.943
pretest	0.920																			
UW	0.921	0.054	0.010	0.948	0.921	0.065	0.010	0.948	0.921	0.076	0.010	0.948	0.921	0.076	0.014	0.947	0.921	0.087	0.014	0.949
RW	0.921	0.087	0.014	0.938	0.921	0.109	0.017	0.937	0.921	0.109	0.017	0.938	0.921	0.120	0.020	0.939	0.921	0.130	0.020	0.939
CS	0.921	0.087	0.014	0.941	0.921	0.098	0.017	0.938	0.921	0.109	0.017	0.938	0.921	0.120	0.020	0.934	0.921	0.130	0.020	0.934
ES	0.921	0.087	0.014	0.941	0.921	0.098	0.014	0.940	0.921	0.109	0.017	0.939	0.921	0.120	0.017	0.938	0.921	0.141	0.017	0.939
rural	1.040																			
UW	1.040	0.019	1.125	0.942	1.039	-0.077	0.903	0.941	1.038	-0.154	0.766	0.943	1.038	-0.221	0.596	0.945	1.036	-0.433	0.368	0.943
RW	1.140	9.577	1.698	0.904	1.122	7.885	1.368	0.904	1.111	6.779	1.169	0.908	1.095	5.288	0.929	0.910	1.068	2.654	0.638	0.936
CS	1.141	9.731	1.697	0.902	1.124	8.077	1.365	0.909	1.113	6.990	1.163	0.913	1.097	5.481	0.915	0.908	1.067	2.558	0.593	0.932
ES	1.140	9.615	1.697	0.904	1.122	7.894	1.364	0.909	1.110	6.731	1.159	0.911	1.094	5.154	0.902	0.912	1.062	2.125	0.571	0.937

Note: RB=relative bias; RMSE=root mean square error; CR=95% confidence interval coverage rate; UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling.

Table 4.2. *RB(%), RMSE, 95% CI CR for Intercept and Variance Components in the Informative Design*

	Informative																			
	ICC=0.5				ICC=0.3				ICC=0.2				ICC=0.1				ICC=0.01			
Covariates	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR
intercept	17.430																			
UW	23.809	36.599	6.415	0.000	23.446	34.512	6.052	0.000	23.129	32.694	5.737	0.000	22.640	29.889	5.250	0.000	21.676	24.362	4.294	0.000
RW	17.702	1.562	1.105	0.910	17.713	1.624	1.092	0.912	17.710	1.604	1.081	0.917	17.694	1.516	1.065	0.925	17.621	1.267	1.037	0.924
CS	17.708	1.594	1.091	0.910	17.723	1.679	1.073	0.916	17.723	1.678	1.061	0.917	17.713	1.623	1.043	0.919	17.675	1.403	1.007	0.925
ES	17.718	1.653	1.071	0.904	17.745	1.806	1.049	0.908	17.756	1.869	1.035	0.914	17.765	1.921	1.018	0.912	17.761	1.900	0.982	0.919
lv1_var	30.000				42.000				48.000				54.000				59.400			
UW	26.948	-10.173	3.136	0.016	37.729	-10.169	4.388	0.016	43.122	-10.163	5.012	0.016	48.521	-10.146	5.630	0.018	53.437	-10.040	6.131	0.025
RW	27.911	-6.962	2.424	0.506	39.077	-6.959	3.392	0.506	44.661	-6.956	3.875	0.506	50.248	-6.948	4.356	0.506	55.288	-6.922	4.778	0.506
CS	28.346	-5.512	2.042	0.603	39.689	-5.503	2.855	0.602	45.365	-5.489	3.256	0.606	51.055	-5.454	3.646	0.611	56.250	-5.303	3.936	0.626
ES	27.579	-8.069	2.649	0.367	38.617	-8.054	3.702	0.368	44.143	-8.036	4.223	0.371	49.687	-7.987	4.726	0.376	54.778	-7.780	5.088	0.390
lv2_var	30.000				18.000				12.000				6.000				0.600			
UW	29.735	-0.884	3.767	0.920	18.189	1.049	2.423	0.936	12.394	3.283	1.783	0.946	6.565	9.418	1.195	0.943	1.191	81.750	0.717	0.758
RW	31.707	5.691	6.065	0.894	21.202	17.791	5.051	0.886	15.904	32.531	4.899	0.713	10.539	75.648	4.898	0.164	4.563	660.550	4.111	0.001
CS	30.308	1.028	5.796	0.865	19.252	6.957	4.020	0.902	13.682	14.015	3.289	0.909	8.042	34.037	2.737	0.815	2.763	360.433	2.341	0.170
ES	29.924	-0.255	5.768	0.854	18.701	3.896	3.837	0.893	13.044	8.699	2.952	0.907	7.313	21.877	2.167	0.899	1.940	223.383	1.538	0.564

Note: RB=relative bias; RMSE=root mean square error; CR=95% confidence interval coverage rate; UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling.

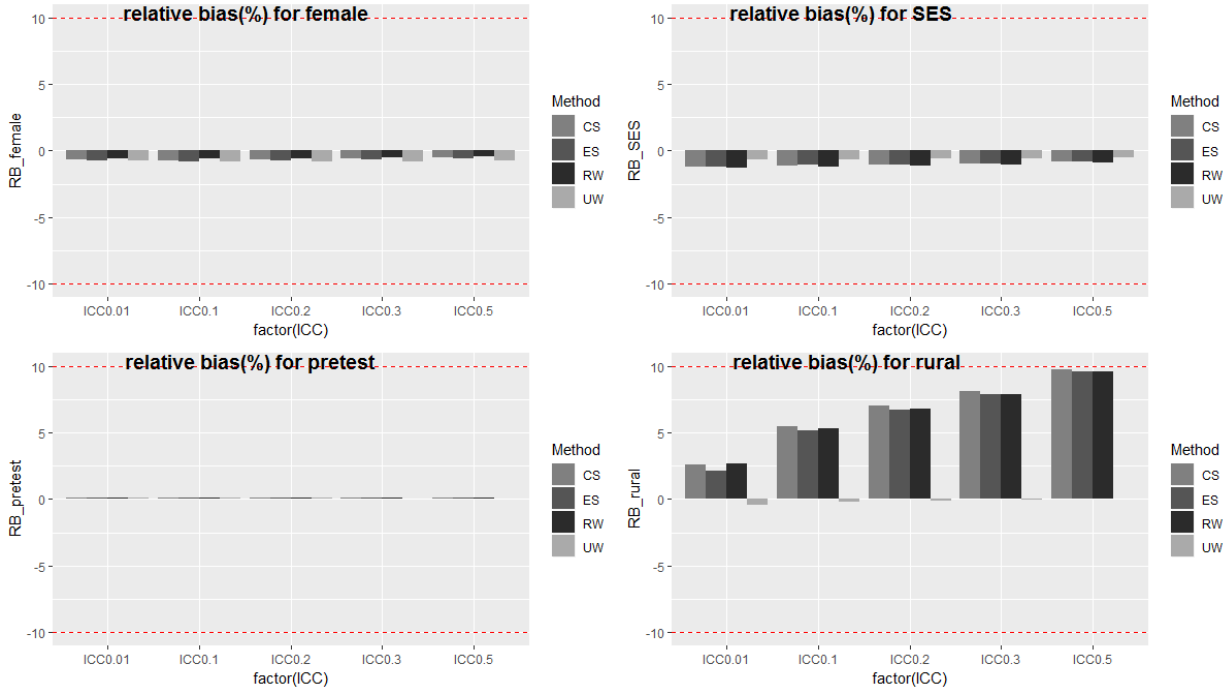

Figure 4.1. Relative bias (%) for covariates in the informative design

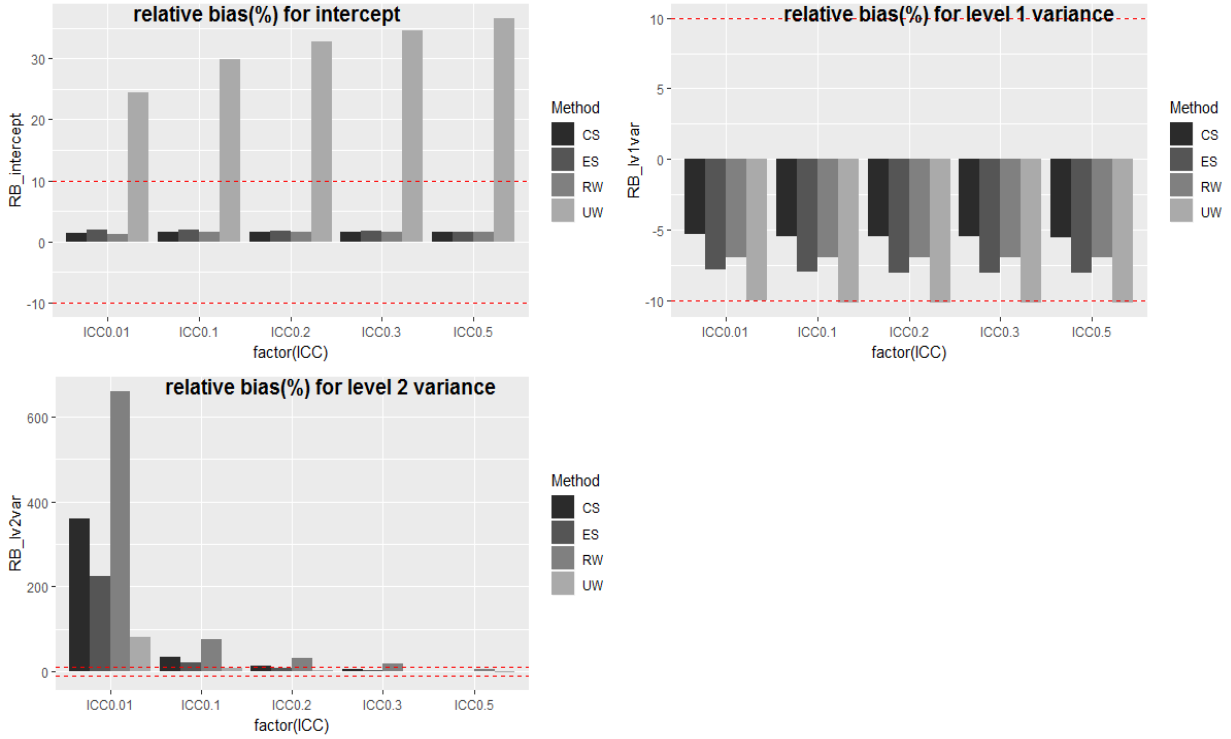

Figure 4.2. Relative bias (%) for intercept and variance components in the informative design

case. The three weighted estimators perform almost equally well since all the relative biases of the intercept estimates produced by them are less than 2. The unweighted estimator performs the worst and produces substantially larger relative bias than the weighted estimators do. As for the student-level variance, the absolute relative biases are all less than or close to 10%. Among the four estimators, the unweighted method produces larger absolute relative bias than the weighted methods do. The cluster scaling method has the smallest values of absolute relative bias. Therefore, the cluster scaling method works the best and the unweighted method works the worst for the student-level variance in terms of (absolute) relative bias. As for the estimates of school-level variance, all four estimators do not perform well and have very large relative biases when the ICC is extremely small. To be more specific, the relative bias is as large as over 600 with the raw weighted method. Even for the best estimator, the unweighted one, has the relative bias of over 80, which is much larger than the standard used in the present study. In general, the raw weighted estimator performs the worst and the unweighted estimator performs the best for the school-level variance across all the ICC levels.

In all, the weighted models perform quite similarly with each other and outperform the unweighted estimator for the intercept and student-level variance while the unweighted model has smaller relative bias and outperform the weighted estimators for the school-level variance. The intercept is always overestimated and the student-level variance is underestimated. The school-level variance, in most cases is overestimated, except with the unweighted method and effective scaling method when ICC equals 0.5. The student-level variables Female and SES are underestimated and pretest is overestimated. School-level variable, rural, is overestimated in the weighted case, while underestimated in the unweighted case.

4.1.1.1.2 Non-Informative Design

Table 4.3 and Figure 4.3 show that the absolute relative biases of the four covariate estimates are all smaller than 10% in the non-informative condition. It means that these four covariates are considered to be estimated unbiasedly in terms of Muthén & Muthén (2002). Also, the two continuous covariates have smaller absolute relative biases than the two dichotomous covariates do. At the same time, the unweighted method produces lower or equal absolute relative bias for the four fixed effects than or as the other three weighted estimators do. So, the unweighted estimator performs the best for all the fixed effects among the four estimators.

The intercept is precisely estimated since all the absolute relative biases are no more than 0.205 (see Table 4.4 and Figure 4.4). The unweighted method outperforms the other estimators when the ICC equals 0.01, 0.1, and 0.2, while it performs the worst when the ICC equals 0.5. Results also show that the student-level variance is estimated unbiasedly since the absolute relative biases are all less than 5% across all the estimators. Among them, the raw weighted method has the largest relative bias, indicating it works the worst. The effective scaling and unweighted method outperform the other two. As for the school-level variance, all the four estimators produce substantially large relative bias when the ICC is extremely small and all the estimators do not work well when the ICC is 0.01. Comparatively, the raw weighted method works the worst while the unweighted method performs the best across different levels of the ICC for the school-level variance estimates.

Table 4.3. *RB(%), RMSE, 95% CI CR for Covariates in the Non-Informative Design*

	Noninformative																			
	ICC=0.5				ICC=0.3				ICC=0.2				ICC=0.1				ICC=0.01			
Covariate	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR
female	0.910																			
UW	0.922	1.264	0.203	0.950	0.924	1.505	0.240	0.951	0.925	1.626	0.256	0.950	0.926	1.736	0.271	0.949	0.927	1.824	0.281	0.950
RW	0.924	1.484	0.278	0.948	0.926	1.758	0.329	0.949	0.927	1.879	0.352	0.949	0.928	2.011	0.373	0.949	0.930	2.176	0.390	0.949
CS	0.922	1.297	0.275	0.947	0.924	1.560	0.325	0.945	0.926	1.703	0.347	0.946	0.927	1.868	0.367	0.945	0.930	2.187	0.383	0.942
ES	0.921	1.242	0.275	0.944	0.924	1.505	0.325	0.944	0.925	1.648	0.347	0.943	0.927	1.813	0.367	0.941	0.930	2.176	0.382	0.941
SES																				
UW	1.054	-0.528	0.136	0.951	1.053	-0.642	0.161	0.949	1.053	-0.708	0.172	0.949	1.052	-0.774	0.182	0.949	1.051	-0.840	0.189	0.950
RW	1.049	-1.038	0.188	0.942	1.047	-1.236	0.222	0.941	1.046	-1.330	0.238	0.941	1.045	-1.425	0.252	0.940	1.044	-1.509	0.264	0.941
CS	1.047	-1.208	0.187	0.943	1.044	-1.472	0.220	0.941	1.043	-1.613	0.236	0.940	1.041	-1.755	0.249	0.942	1.040	-1.906	0.259	0.943
ES	1.047	-1.236	0.187	0.941	1.044	-1.519	0.221	0.941	1.042	-1.660	0.236	0.941	1.041	-1.811	0.249	0.944	1.039	-1.972	0.260	0.935
pretest	0.920																			
UW	0.920	0.000	0.010	0.943	0.920	0.000	0.010	0.944	0.920	0.000	0.014	0.945	0.920	0.000	0.014	0.945	0.920	0.000	0.014	0.949
RW	0.920	0.043	0.014	0.923	0.921	0.054	0.017	0.923	0.921	0.054	0.017	0.923	0.921	0.054	0.017	0.923	0.921	0.065	0.020	0.924
CS	0.920	0.043	0.014	0.927	0.921	0.054	0.017	0.927	0.921	0.054	0.017	0.926	0.921	0.065	0.017	0.926	0.921	0.065	0.017	0.929
ES	0.920	0.043	0.014	0.933	0.921	0.054	0.017	0.932	0.921	0.054	0.017	0.929	0.921	0.065	0.017	0.931	0.921	0.065	0.017	0.937
rural	1.040																			
UW	1.032	-0.760	1.119	0.944	1.030	-0.962	0.902	0.945	1.029	-1.038	0.768	0.947	1.029	-1.077	0.603	0.947	1.031	-0.856	0.382	0.944
RW	0.967	-6.990	1.416	0.936	0.982	-5.615	1.159	0.937	0.991	-4.740	1.002	0.932	1.003	-3.635	0.811	0.934	1.024	-1.587	0.570	0.938
CS	0.966	-7.135	1.411	0.937	0.980	-5.788	1.147	0.939	0.989	-4.865	0.986	0.939	1.002	-3.654	0.785	0.936	1.026	-1.346	0.519	0.941
ES	0.966	-7.106	1.410	0.936	0.980	-5.798	1.146	0.935	0.989	-4.885	0.984	0.936	1.002	-3.644	0.782	0.937	1.026	-1.346	0.510	0.939

Note: RB=relative bias; RMSE=root mean square error; CR=95% confidence interval coverage rate; UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling.

Table 4.4. *RB(%), RMSE, 95% CI CR for Intercept and Variance Components in the Non-Informative Design*

	Noninformative																			
	ICC=0.5				ICC=0.3				ICC=0.2				ICC=0.1				ICC=0.01			
Covariates	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR	Mean	RB(%)	RMSE	CR
intercept	17.430																			
UW	17.415	-0.085	0.705	0.944	17.418	-0.072	0.698	0.942	17.419	-0.064	0.693	0.941	17.420	-0.056	0.686	0.939	17.422	-0.046	0.668	0.944
RW	17.424	-0.035	0.946	0.935	17.417	-0.074	0.955	0.933	17.413	-0.096	0.958	0.929	17.409	-0.121	0.958	0.930	17.403	-0.157	0.949	0.934
CS	17.426	-0.025	0.936	0.935	17.418	-0.069	0.942	0.934	17.413	-0.098	0.942	0.935	17.407	-0.132	0.937	0.940	17.397	-0.189	0.916	0.936
ES	17.425	-0.028	0.935	0.933	17.417	-0.075	0.941	0.930	17.412	-0.105	0.941	0.934	17.406	-0.139	0.936	0.937	17.394	-0.205	0.908	0.938
lv1_var	30.000				42.000				48.000				54.000				59.400			
UW	29.629	-1.238	0.866	0.906	41.482	-1.234	1.211	0.907	47.411	-1.227	1.383	0.909	53.349	-1.206	1.551	0.911	58.756	-1.084	1.676	0.918
RW	28.485	-5.049	1.813	0.646	39.880	-5.047	2.538	0.647	45.579	-5.044	2.899	0.647	51.281	-5.036	3.258	0.647	56.430	-5.001	3.567	0.650
CS	29.359	-2.138	1.196	0.867	41.105	-2.131	1.673	0.869	46.982	-2.120	1.909	0.869	52.871	-2.091	2.140	0.870	58.247	-1.941	2.313	0.883
ES	29.681	-1.064	1.067	0.921	41.555	-1.059	1.493	0.921	47.495	-1.052	1.705	0.921	53.443	-1.032	1.916	0.920	58.845	-0.934	2.095	0.920
lv2_var	30.000				18.000				12.000				6.000				0.600			
UW	30.145	0.485	3.697	0.936	18.424	2.354	2.447	0.941	12.544	4.537	1.847	0.942	6.632	10.533	1.275	0.932	1.188	97.950	0.737	0.779
RW	31.873	6.242	5.070	0.934	20.857	15.870	4.250	0.891	15.304	27.534	4.063	0.765	9.679	61.310	4.005	0.251	4.415	635.883	3.914	0.002
CS	30.423	1.409	4.677	0.922	18.848	4.713	3.173	0.932	13.026	8.553	2.471	0.934	7.152	19.192	1.832	0.914	1.688	181.250	1.258	0.615
ES	29.987	-0.142	4.645	0.909	18.223	1.241	3.047	0.915	12.338	2.818	2.252	0.923	6.428	7.130	1.459	0.931	1.022	70.350	0.703	0.943

Note: RB=relative bias; RMSE=root mean square error; CR=95% confidence interval coverage rate; UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling.

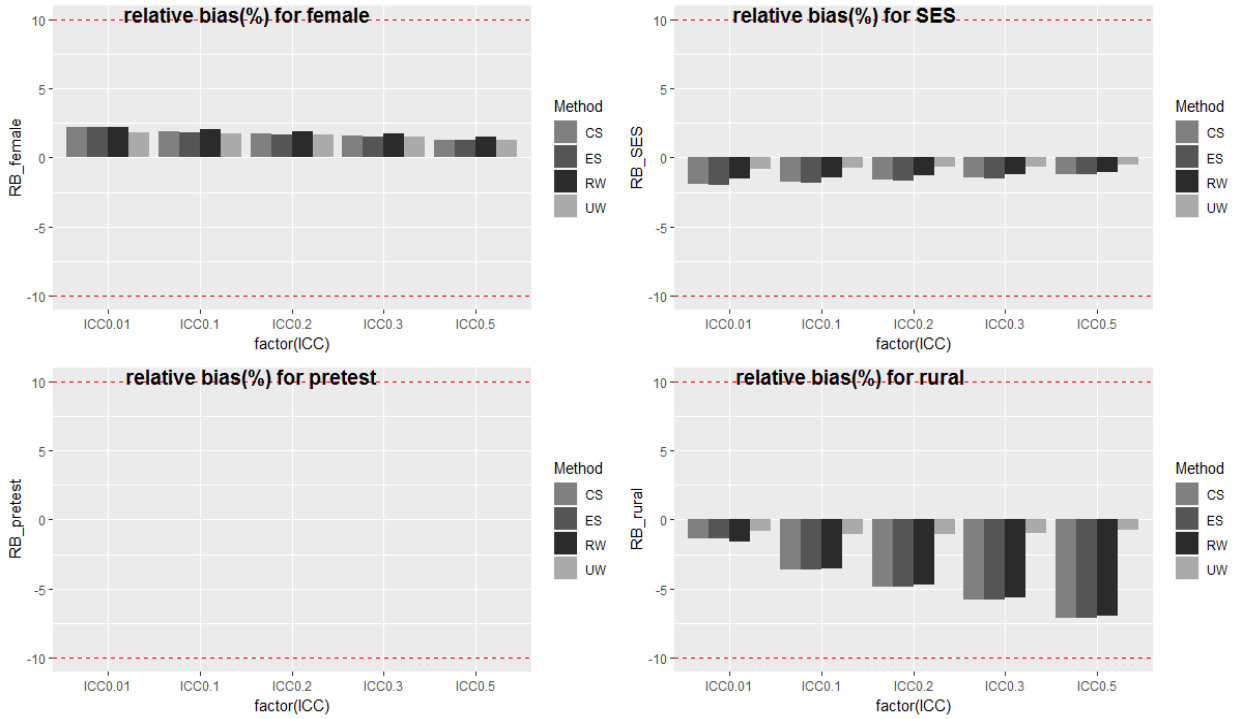

Figure 4.3. Relative bias (%) for covariates in the non-informative design

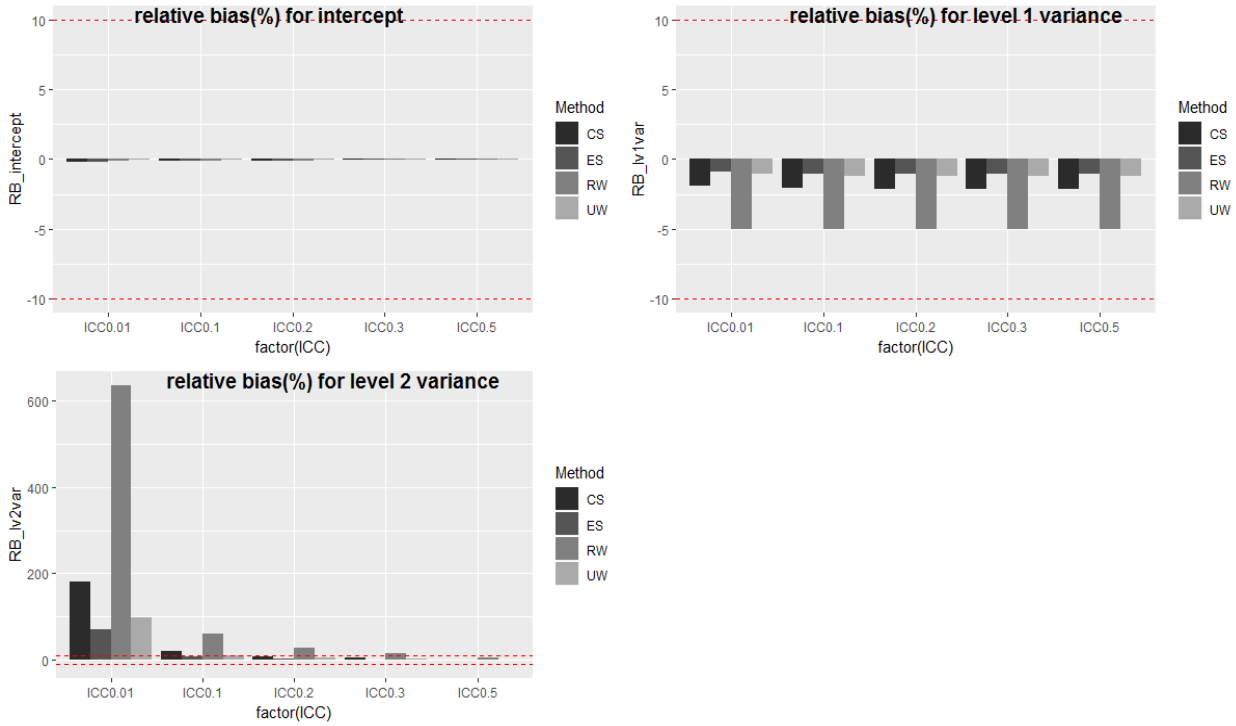

Figure 4.4. Relative bias (%) for intercept and variance components in the non-informative design

4.1.1.2 RMSE

An overview of the RSME of the fixed effect point estimators and the intercept and variance component estimators across informativeness and ICCs is provided in Table 4.1-4.4, Figure 4.5-4.8. There is not much difference on the RMSE for the fixed effects between the informative and non-informative condition. However, on average, the RMSE is comparatively larger under the informative condition than those in the non-informative condition.

4.1.1.2.1 Informative Design

Compared with weighted estimators, the unweighted estimator has smaller RMSE value for the four covariates under the informative condition (see Table 4.1 and Figure 4.5). The weighted estimates of the RMSE show almost the same patterns for the four covariates. The unweighted estimator performs the most efficiently among the four estimators.

As the relative biases of the intercept and variance components, similar results are obtained for the RMSE. For example, the unweighted method has comparatively much larger RMSE for the intercept than the weighted estimators do and the three weighted estimators perform very much similarly to each other (see Table 4.2 and Figure 4.6). The unweighted estimator produces the largest RMSE for the student-level variance and performs the least efficiently among the four. The cluster scaling method performs the most efficiently. As for the school-level variance, the unweighted estimator has the smallest RMSE and performs the most efficiently among the four. The raw weighted estimator has the least efficiency.

In all, the unweighted estimator performs the worst for the intercept and student-level variance estimates, but performs the best for school-level variance estimates in terms of RMSE in the informative design.

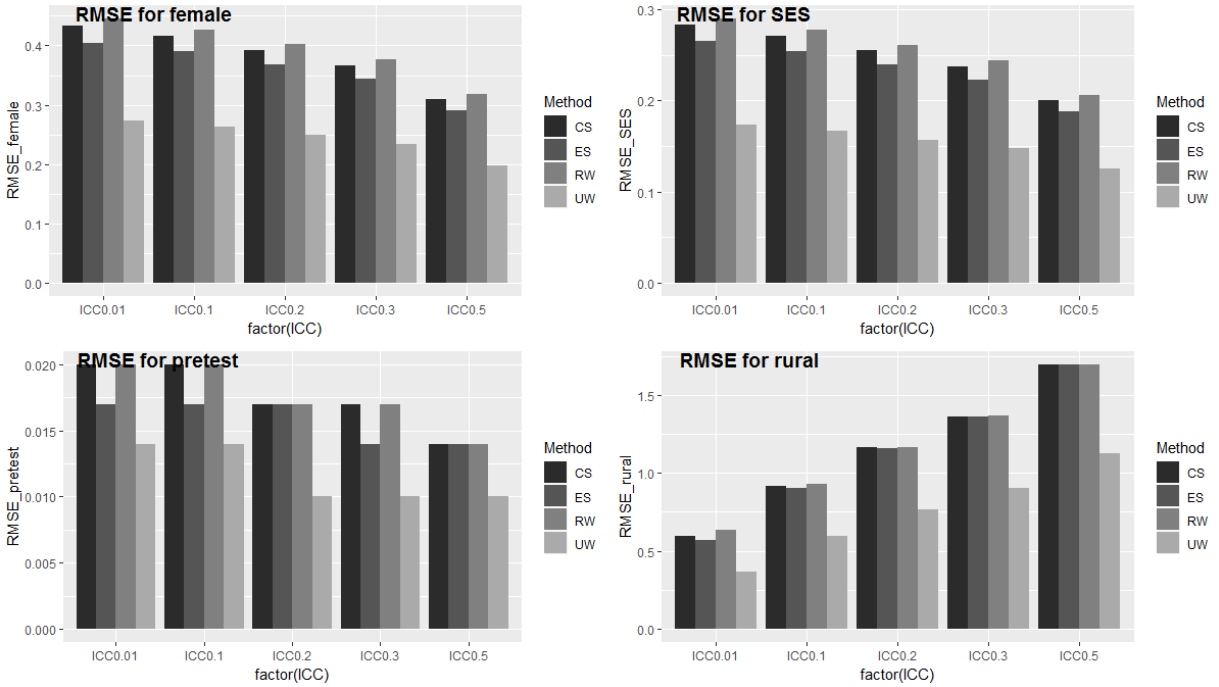

Figure 4.5. RMSE for covariates in the informative design

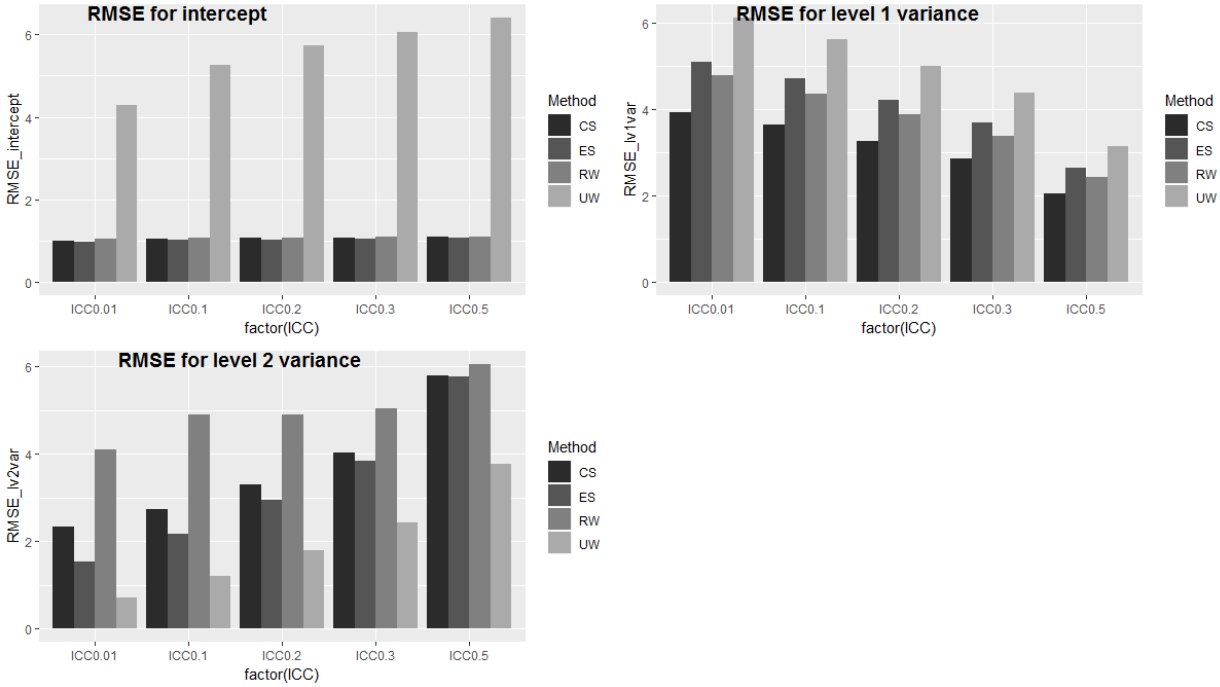

Figure 4.6. RMSE for intercept and variance components in the informative design

4.1.1.2.2 Non-Informative Design

Table 4.3 and Figure 4.7 show that the unweighted method has the smaller RMSE for the four covariates than the weighted methods do, and in most cases, there is not much difference across the weighted methods for the four covariates at different levels of the ICC. Therefore, the unweighted method performs the best among the four estimators for all the fixed effects.

Apparently, the unweighted method has the smallest RMSE for the intercept and the two variance components across all the conditions in the non-informative condition (see Table 4.4 and Figure 4.8) and performs the most efficiently among the four estimators across all the levels of the ICC. And the raw weighted method produces the largest RMSE among the four estimators for the intercept and the two variance component estimates.

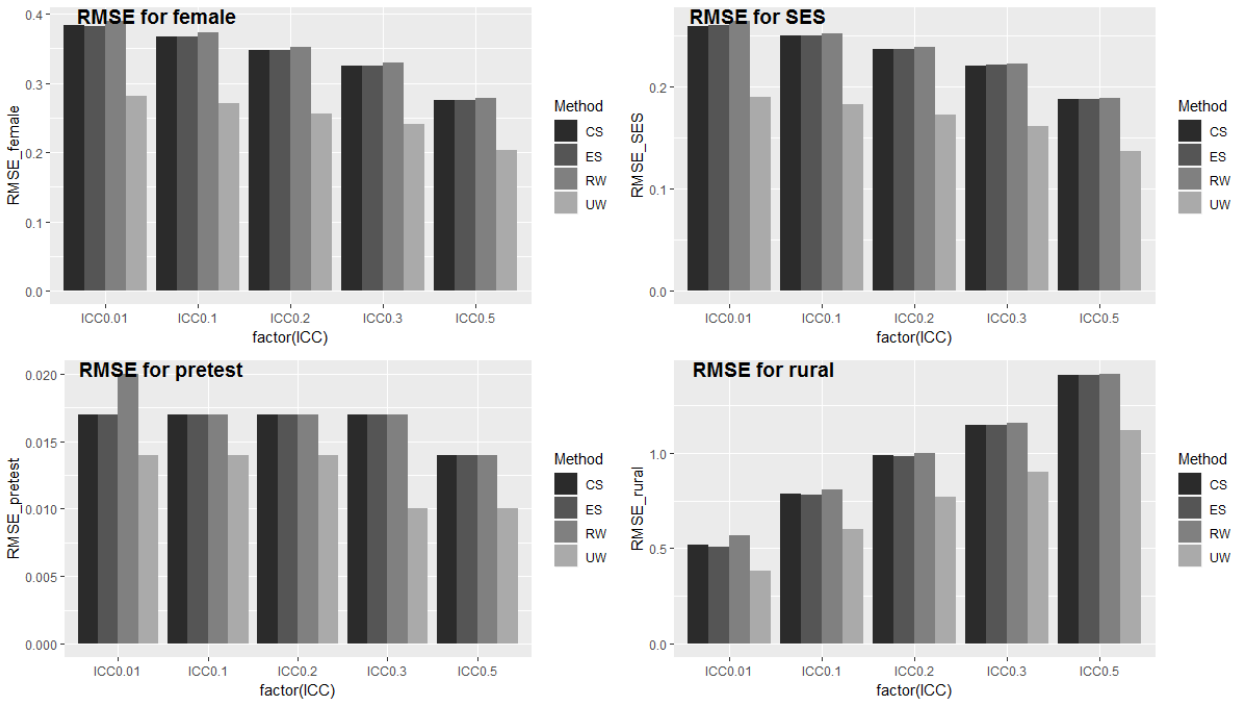

Figure 4.7. RMSE for covariates in the non-informative design

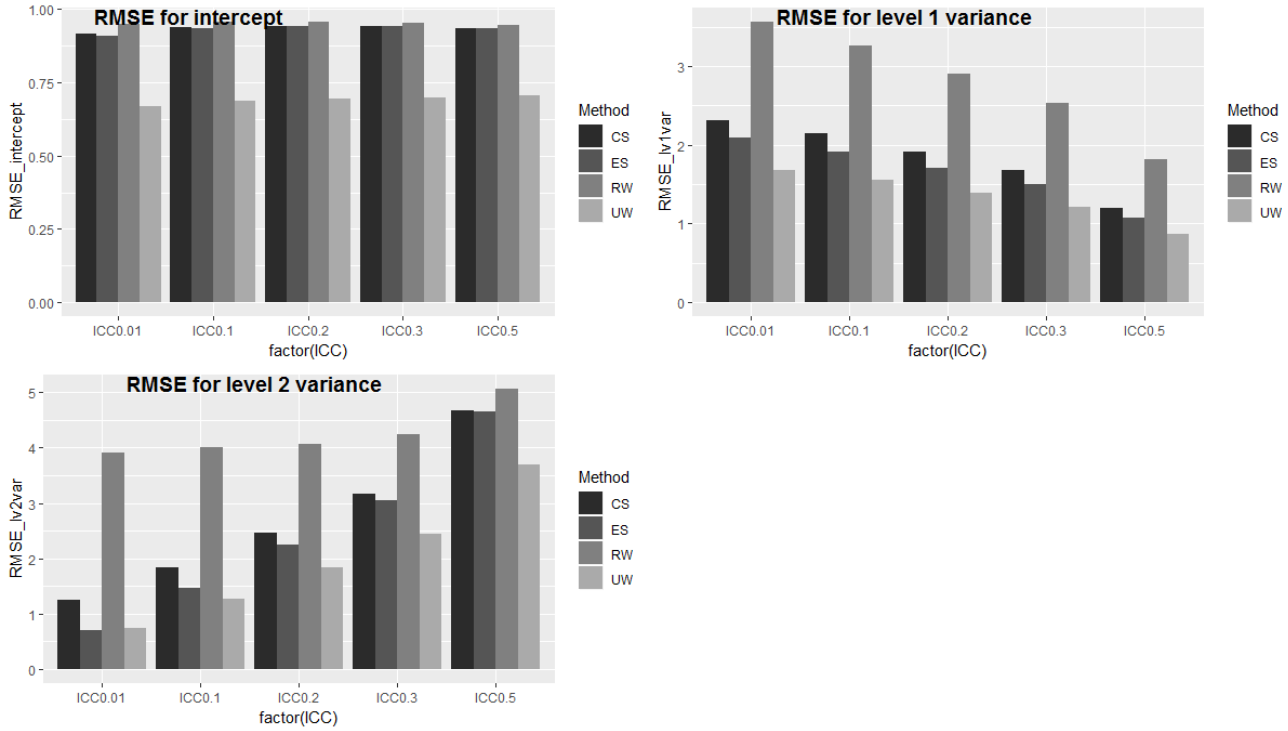

Figure 4.8. RMSE for intercept and variance components in the non-informative design

4.1.1.3 Coverage Rate

An overview of coverage of the fixed effects, intercept and variance component estimators across informativeness and ICCs is provided in Table 4.1-4.4 and Figure 4.9-4.12. All the fixed effects are estimated without much bias (<10%) in both the informative and non-informative conditions if the criterion of Muthén & Muthén (2002) is applied. The corresponding coverage rates for them are good and not much difference can be found among them. For the intercept and variance components, on average, their coverage rates are much lower under the informative condition than those under the non-informative condition. Under the informative condition, the most variability in coverage occurs for the intercept estimators, whereas under the non-informative condition, the most variability in coverage occurs for the school-level estimators.

4.1.1.3.1 Informative Design

Because the four covariates are precisely or slightly biasedly estimated, the coverage rates for them are all above or close to 0.91, especially for the three level-one predictors (see Table 4.1 and Figure 4.9).

Because the unweighted method produces substantially larger biases for the intercept and student-level variance estimates, this leads to very poor coverage rates for both of them (see Table 4.2 and Figure 4.10): with the coverage rate of 0 for the intercept and less than 3% for the student-level variance. The three weighted methods perform almost equally well and have the coverage rates of around or over 0.91 for the intercept. However, even the best student-level variance estimator, the cluster scaling estimator, has the coverage rates no more than 0.63. For the school- level variance estimates, raw weighted method performs the worst while the unweighted estimator performs the best.

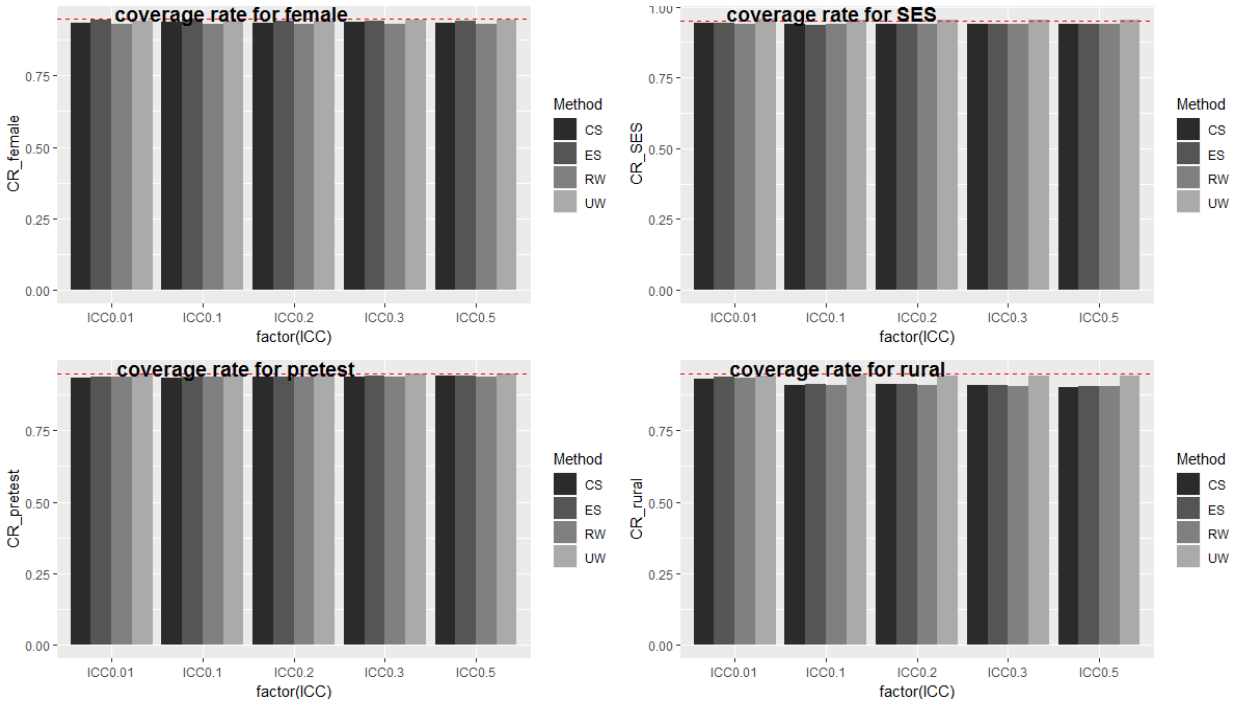

Figure 4.9. Coverage rate for covariates in the informative design

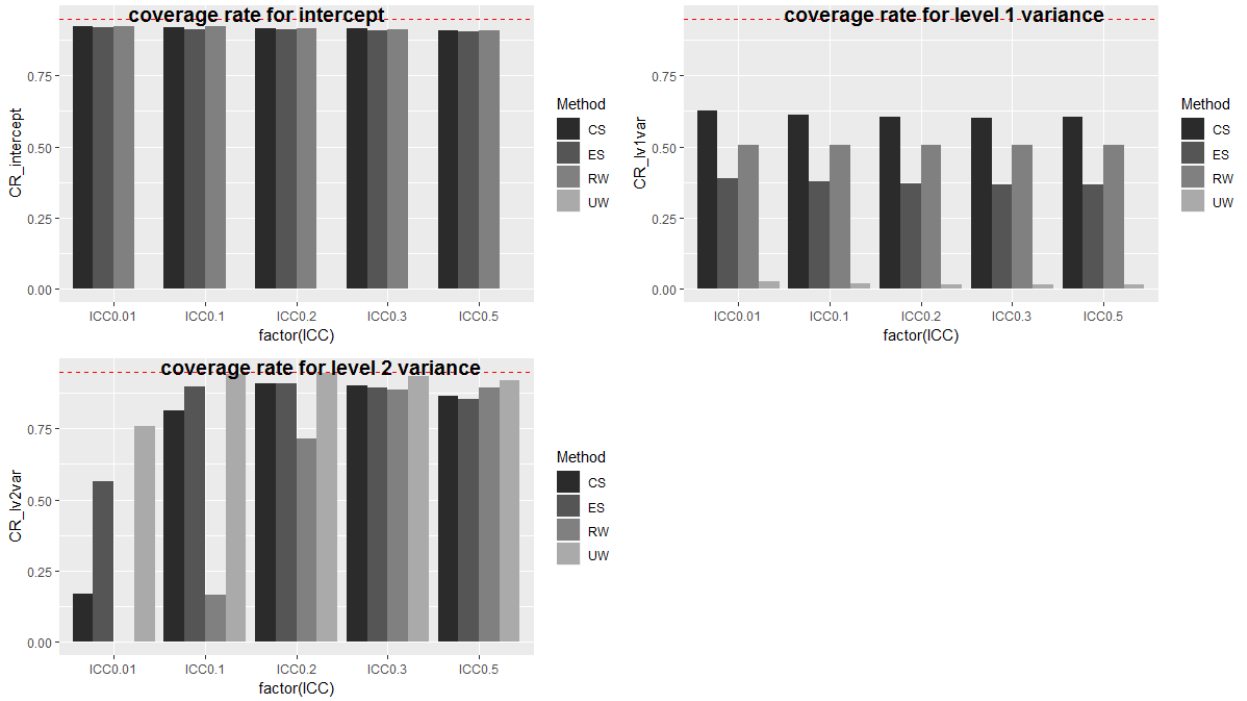

Figure 4.10. Coverage rate for intercept and variance components in the informative design

4.1.1.3.2 Non-Informative Design

The coverage rates for the four covariate estimates in the non-informative condition are all above or close to 0.95. Among the four estimators, the unweighted method performs the best.

The unweighted method has the highest coverage rates for the intercept among the four estimators as well and they are all above or around 0.94. As for the student-level variance, the effective scaling method has the highest coverage rates, which are around 0.92 whereas the raw weighted method has the lowest coverage rates, which are around 0.65. The unweighted estimator has very similar coverage rate to the effective scaling method. The coverage rates for the school-level variance with unweighted method are the highest among all the estimators and are all larger than 0.93 except when the ICC is 0.01, while the raw weighted has the smallest one.

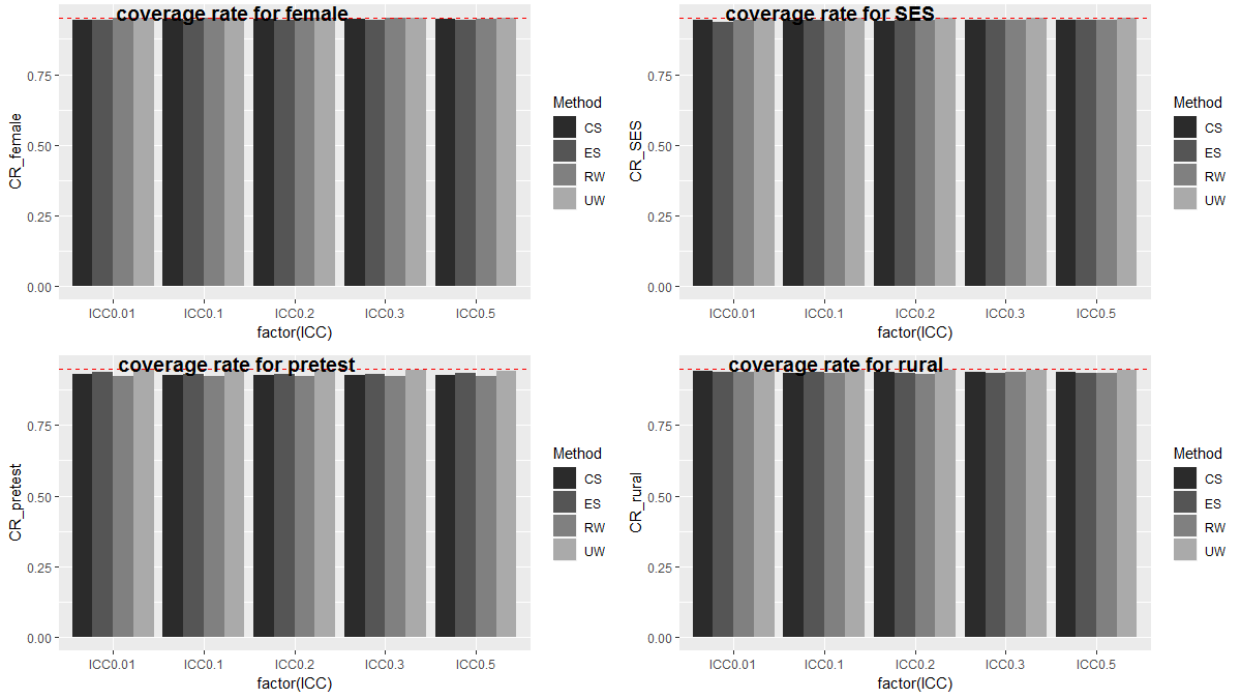

Figure 4.11. Coverage rate for covariates in the non-informative design

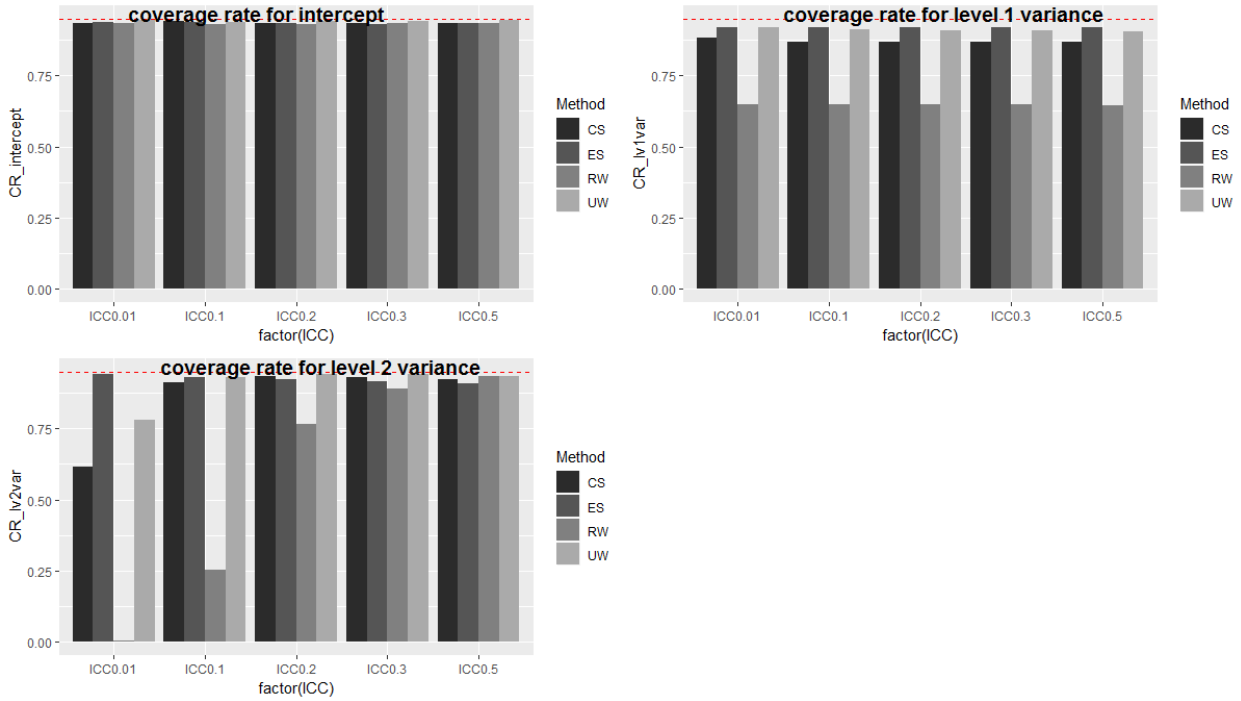

Figure 4.12. Coverage rate for intercept and variance components in the non-informative design

4.1.2 Research Question Two

Research question two addresses the ICC effect on the different estimation methods in the informative and non-informative design.

4.1.2.1 (Absolute) Relative Bias

4.1.2.1.1 Informative Design

Table 4.1 and Figure 4.13 show that, as the ICC increases, the absolute relative biases for the two continuous covariates (e.g., SES and pretest) decrease. For the covariate female, there is no monotonous pattern for its relative bias. As the ICC increases, it increases first and then starts to decrease. For the covariate rural, the relative bias increases as the ICC increases in the weighted case, while the absolute relative bias decreases in the unweighted case. Therefore, for all the fixed effects, there is no overall consistent pattern.

It is evident (see Figure 4.14) that the absolute relative bias for the intercept estimate with unweighted method increases as the increase of ICC, but no consistent monotonous pattern can be found for the relative biases for the intercept estimate with the weighted methods and they do not vary much across the weighted methods at each different levels of the ICC (see Table 4.2). The absolute relative biases for the student-level variance estimates decrease as the ICC decreases with all the four estimators, but the decrease rate is very tiny and hard to find from Figure 4.14. There is an obvious increase pattern in the relative bias of the school-level variance estimates as the ICC decreases with the four estimators (see Table 4.2 and Figure 4.14).

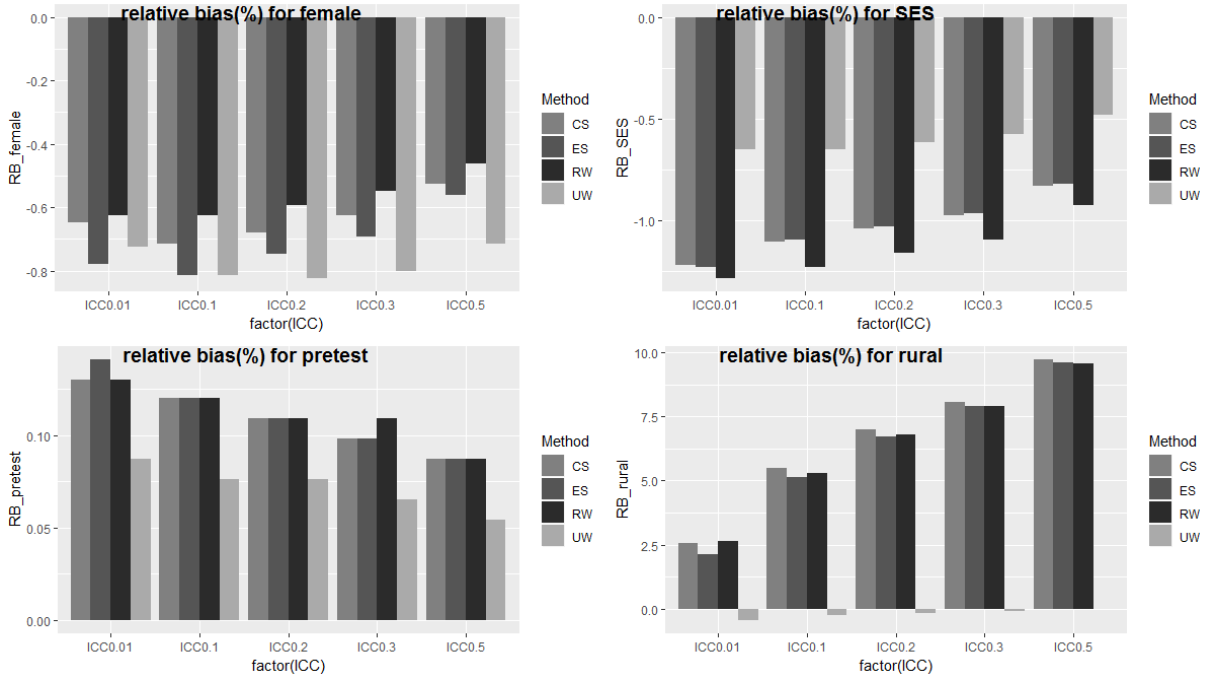

Figure 4.13. Relative bias (%) for covariates in the informative design

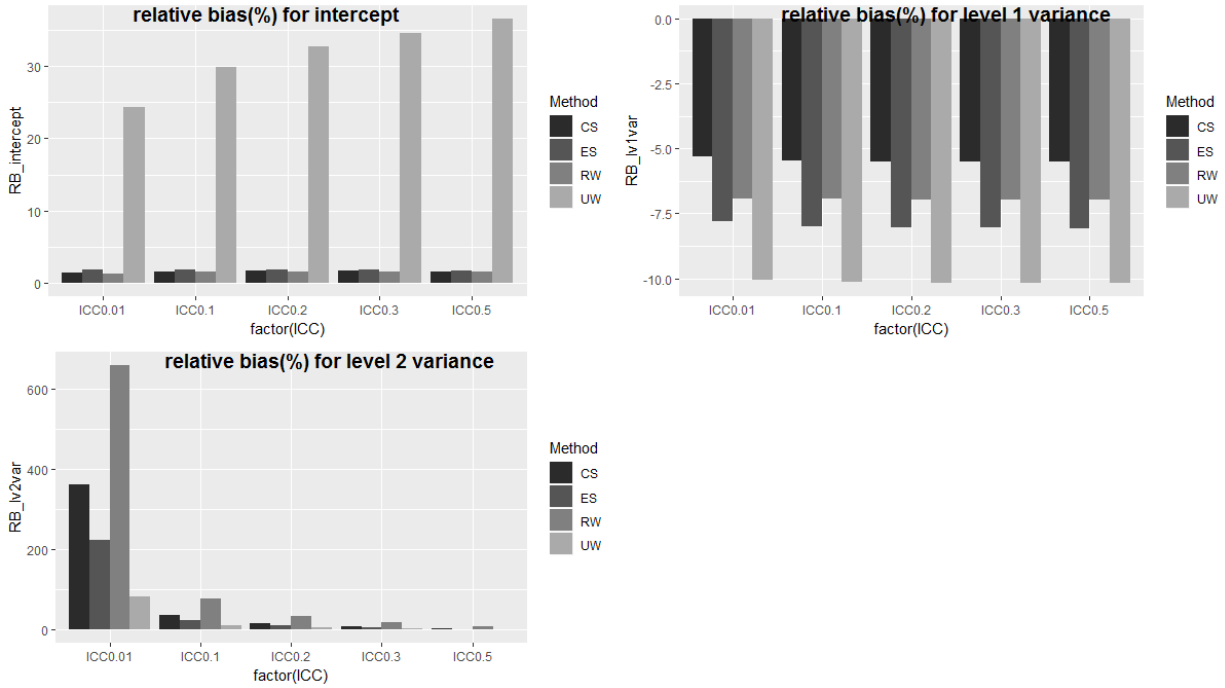

Figure 4.14. Relative bias (%) for intercept and variance components in the informative design

4.1.2.1.2 Non-Informative Design

Clear patterns can be found under the non-informative sampling design. Table 4.3 and Figure 4.15 indicate as the ICC increases, the absolute relative bias decreases for the three student-level covariates, and increases for rural, the school-level covariate.

Simulation results show that as the increase of the ICC, the absolute relative bias for the intercept decreases with the three weighted methods whereas it increases with the unweighted model (see Table 4.4 and Figure 4.16). As for the relative bias of student-level variance, it decreases as the ICC decreases, but the decrease rate is so small that similar patterns hold for the estimators across different ICC values. The relative bias for the school-level variance increases as the decreases of the ICC.

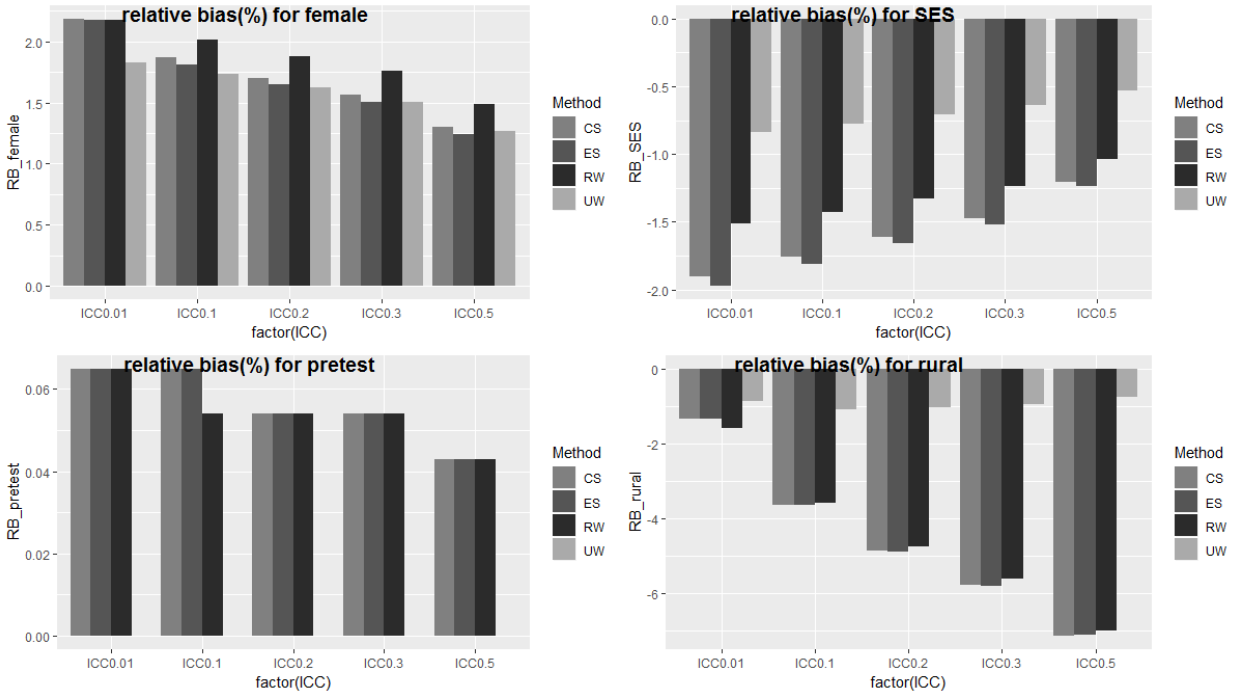

Figure 4.15. Relative bias (%) for covariates in the non-informative design

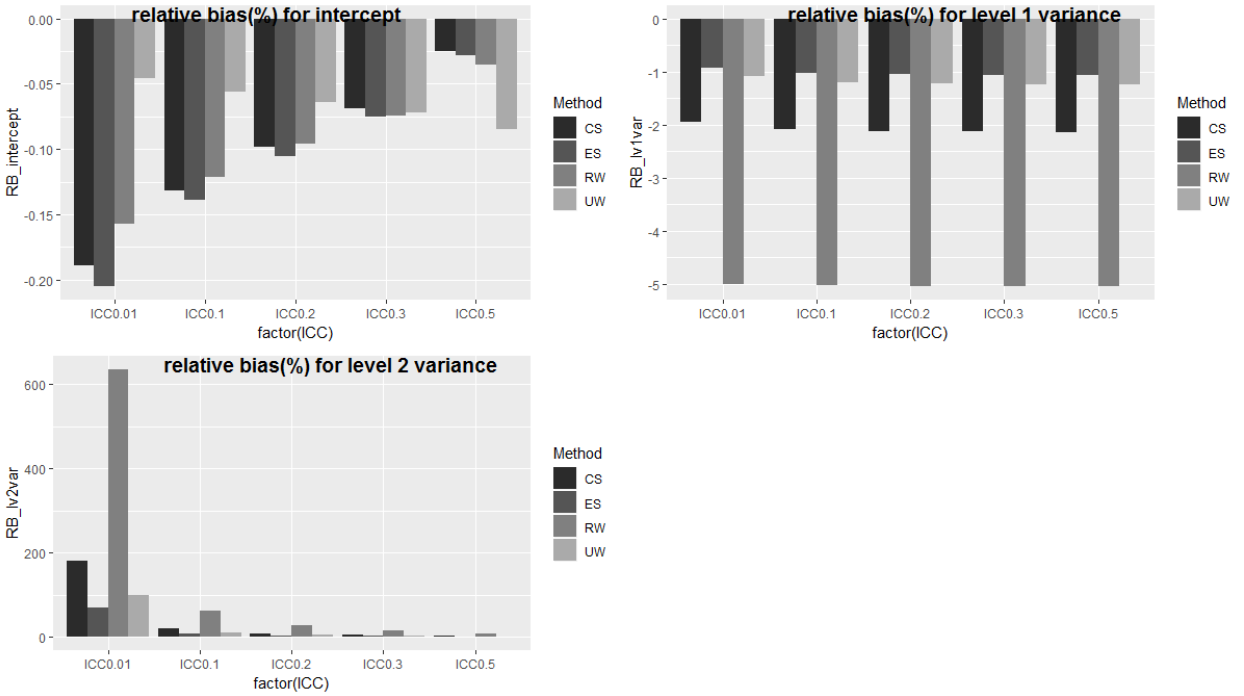

Figure 4.16. Relative bias for intercept and variance components in the non-informative design

4.1.2.2 RMSE

4.1.2.2.1 Informative Design

Contrary to the relative bias, there are clear patterns of RMSE for all the fixed effects (see Table 4.1 and Figure 4.5). As the ICC increases, the RMSE decreases for all the student-level fixed effects, and increases for the school-level fixed effect with all the estimators.

Table 4.2 and Figure 4.6 show that the RMSE for the intercept is increasing as the ICC increases. The increase rate is quite obvious with the unweighted method but it is so small with the three weighted methods that not much variation can be found across different ICC values. As for the variance components, there are clear patterns for both of them. As the ICC increases, the RMSE of the student-level variance decreases whereas the RMSE of the school-level variance increases.

4.1.2.2.2 Non-Informative Design

The RMSE decreases for the three student-level covariates, and increases for rural, the school-level covariate as the ICC increases with all the four estimators (see Table 4.3 and Figure 4.7).

Figure 4.8 shows that the RMSE for the intercept remains almost unchanged across different levels of the ICC with the four estimators, but Table 4.4 shows that the RMSE does increase as the increase of the ICC consistently. It is clear that as the ICC increases, the RMSE for the student-level variance decreases whereas the RMSE for the school-level variance increases with all the estimators.

4.1.2.3 Coverage Rate

4.1.2.3.1 Informative Design

Table 4.1 and Figure 4.9 show as the increase of the ICC, there is not much variation on the coverage rates for all the fixed effects.

The coverage rate for the intercept and student-level variance remains almost the same as the ICC increases (see Table 4.2 and Figure 4.10). For the school-level variance, although the coverage rate changes as the increase of ICC, no consistent pattern can be seen for the estimators except for with the raw weighted method. Overall, coverage rate is not sensitive to the change of the ICC in the current case.

4.1.2.3.2 Non-Informative Design

No obvious ICC effect can be found in terms of the coverage rate for all the parameter estimates except for school-level variance (see Table 4.3-4.4, and Figure 4.11-4.12). The coverage rates for the fixed effects, intercept, and student-level variance remain almost unchanged as the increase of the ICC. Although there are some variations of the coverage rates for the school-level

variance, there is no clear pattern with the four estimators. For example, the coverage rates with the unweighted model and cluster scaling method increase first and then decreases later as the increase of the ICC. The coverage rate keeps on increasing with the effective scaling method and decreasing with the raw weighted method as the decrease of the ICC. In sum, the effect of ICC cannot be found for the all the parameters in terms of the coverage rate in the non-informative condition.

4.1.3 Simulated Standard Errors and Standard Deviations

If we tend to repeat the Monte Carlo simulation and tally the sample mean each time, a normal distribution (based on Central Limit Theorem) would result in the distribution of the sample mean. To assess how well the standard errors of the estimates approximate the true sampling variation, the sample standard deviation of each replicate, that is, the Monte Carlo standard deviation, can be compared to the average of the estimated standard errors. We might expect the sample standard deviation, an approximation to the true sampling variation, to be "close" to and the average of standard errors. It means that the standard error is a good estimate of the standard deviation of the normal distribution if the sample size is sufficiently large. The differences are calculated between the standard deviations and averaged standard errors of 1000 point estimates for all the seven parameters: four regression coefficients for female, SES, pretest and rural, intercept, and two random effects (the student-level variance and school-level variance). Table 4.5 presents the results of standard deviations of simulation and standard errors of estimates in the informative sampling design. The differences between them for the four fixed effects and intercept are on the second or even third decimal place. The differences for student-level variance and school-level variance are a little bit larger, but still they are less than or close to 1. Clearly, the

unweighted method produces the smallest standard errors and works best compared with the weighted estimators.

Table 4.6 contains the results of standard deviations of simulation and standard errors of estimates in the non-informative sampling design. It tells us the same story as in the informative setting. The differences for all the parameter estimates are even smaller, and the largest absolute difference is 0.273, indicating the estimation performs quite well. Still, the unweighted method has the smallest standard errors and performs best compared with the three weighted models.

Table 4.5. *Simulation Standard Deviations and Standard Errors of Estimates in the Informative Design*

	ICC0.5			ICC0.3			ICC0.2			ICC0.1			ICC0.01		
Covariates	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff
female															
UW	0.198	0.193	0.005	0.234	0.228	0.006	0.250	0.244	0.006	0.264	0.258	0.006	0.274	0.269	0.006
RW	0.318	0.300	0.018	0.376	0.355	0.021	0.402	0.380	0.023	0.427	0.403	0.024	0.446	0.422	0.025
CS	0.310	0.293	0.017	0.367	0.348	0.020	0.393	0.372	0.021	0.417	0.395	0.022	0.433	0.411	0.022
ES	0.290	0.277	0.013	0.344	0.329	0.015	0.368	0.352	0.016	0.390	0.374	0.017	0.404	0.389	0.016
SES															
UW	0.125	0.131	-0.006	0.148	0.155	-0.007	0.158	0.165	-0.008	0.167	0.175	-0.008	0.173	0.182	-0.008
RW	0.206	0.201	0.005	0.244	0.238	0.006	0.261	0.254	0.007	0.277	0.270	0.007	0.290	0.282	0.008
CS	0.200	0.196	0.004	0.237	0.232	0.005	0.255	0.249	0.006	0.271	0.264	0.007	0.283	0.275	0.008
ES	0.188	0.186	0.002	0.223	0.220	0.003	0.239	0.235	0.004	0.254	0.250	0.004	0.265	0.260	0.005
pretest															
UW	0.009	0.009	0.000	0.011	0.011	0.000	0.012	0.012	0.000	0.012	0.012	0.000	0.013	0.013	0.000
RW	0.014	0.014	0.000	0.017	0.017	0.000	0.018	0.018	0.000	0.019	0.019	0.000	0.020	0.020	0.000
CS	0.014	0.014	0.000	0.017	0.016	0.000	0.018	0.018	0.000	0.019	0.019	0.000	0.020	0.019	0.000
ES	0.013	0.013	0.000	0.016	0.016	0.000	0.017	0.017	0.000	0.018	0.018	0.000	0.019	0.018	0.000
rural															
UW	1.125	1.096	0.030	0.903	0.884	0.020	0.767	0.754	0.013	0.596	0.592	0.004	0.368	0.373	-0.005
RW	1.696	1.485	0.211	1.366	1.207	0.160	1.168	1.040	0.128	0.928	0.838	0.089	0.637	0.598	0.040
CS	1.695	1.482	0.213	1.364	1.200	0.163	1.162	1.029	0.132	0.913	0.821	0.093	0.593	0.557	0.036
ES	1.695	1.481	0.213	1.362	1.198	0.164	1.158	1.025	0.133	0.905	0.813	0.092	0.571	0.540	0.031
Intercept															
UW	0.676	0.682	-0.006	0.666	0.669	-0.004	0.660	0.662	-0.001	0.654	0.652	0.001	0.640	0.637	0.003
RW	1.072	1.009	0.063	1.055	1.007	0.048	1.045	1.005	0.039	1.032	1.002	0.030	1.014	0.997	0.017
CS	1.055	0.996	0.059	1.033	0.991	0.042	1.020	0.987	0.033	1.005	0.981	0.024	0.978	0.968	0.010
ES	1.032	0.972	0.060	1.001	0.957	0.044	0.983	0.948	0.035	0.962	0.937	0.025	0.925	0.916	0.008
lv1_var															
UW	0.720	0.712	0.008	1.008	0.996	0.011	1.151	1.139	0.013	1.295	1.281	0.014	1.425	1.409	0.015
RW	1.230	1.159	0.071	1.722	1.623	0.099	1.968	1.855	0.113	2.213	2.087	0.127	2.434	2.295	0.139
CS	1.198	1.137	0.061	1.676	1.592	0.084	1.915	1.819	0.096	2.151	2.046	0.105	2.361	2.251	0.110
ES	1.076	1.044	0.032	1.505	1.462	0.043	1.720	1.671	0.049	1.934	1.880	0.054	2.128	2.072	0.057
lv2_var															
UW	3.759	3.556	0.203	2.417	2.314	0.103	1.740	1.683	0.057	1.054	1.036	0.017	0.407	0.406	0.001
RW	5.823	4.738	1.085	3.908	3.225	0.682	2.962	2.474	0.488	2.033	1.730	0.303	1.220	1.081	0.138
CS	5.791	4.703	1.088	3.822	3.149	0.673	2.828	2.363	0.466	1.823	1.564	0.260	0.896	0.823	0.073
ES	5.770	4.687	1.083	3.774	3.112	0.661	2.763	2.310	0.452	1.725	1.487	0.238	0.756	0.709	0.047

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling; SD=standard deviation; SE=standard error; Diff=difference.

Table 4.6. *Simulation Standard Deviations and Standard Errors of Estimates in the Non-Informative Design*

	ICC0.5			ICC0.3			ICC0.2			ICC0.1			ICC0.01		
Covariates	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff	SD	SE	Diff
female															
UW	0.203	0.203	0.001	0.240	0.239	0.001	0.256	0.256	0.001	0.271	0.270	0.000	0.281	0.281	-0.001
RW	0.278	0.272	0.006	0.329	0.322	0.007	0.352	0.344	0.008	0.373	0.364	0.008	0.390	0.381	0.009
CS	0.275	0.270	0.006	0.325	0.318	0.007	0.347	0.340	0.007	0.367	0.360	0.008	0.383	0.375	0.008
ES	0.275	0.269	0.006	0.325	0.318	0.007	0.347	0.340	0.007	0.367	0.360	0.007	0.382	0.374	0.007
SES															
UW	0.136	0.136	0.000	0.161	0.161	0.000	0.172	0.172	0.000	0.182	0.182	0.000	0.189	0.189	0.000
RW	0.188	0.181	0.007	0.222	0.214	0.008	0.238	0.229	0.009	0.252	0.243	0.009	0.263	0.254	0.010
CS	0.186	0.180	0.007	0.220	0.212	0.008	0.235	0.227	0.008	0.249	0.240	0.009	0.259	0.250	0.009
ES	0.186	0.180	0.007	0.220	0.212	0.008	0.235	0.227	0.009	0.249	0.240	0.009	0.259	0.250	0.009
pretest															
UW	0.010	0.010	0.000	0.012	0.011	0.000	0.012	0.012	0.000	0.013	0.013	0.000	0.014	0.013	0.000
RW	0.014	0.013	0.001	0.016	0.015	0.001	0.017	0.016	0.001	0.018	0.017	0.001	0.019	0.018	0.001
CS	0.014	0.013	0.001	0.016	0.015	0.001	0.017	0.016	0.001	0.018	0.017	0.001	0.019	0.018	0.001
ES	0.014	0.013	0.001	0.016	0.015	0.001	0.017	0.016	0.001	0.018	0.017	0.001	0.019	0.018	0.001
rural															
UW	1.119	1.109	0.011	0.902	0.896	0.006	0.769	0.766	0.003	0.603	0.604	-0.001	0.382	0.384	-0.002
RW	1.415	1.351	0.064	1.158	1.100	0.058	1.002	0.948	0.053	0.811	0.764	0.047	0.570	0.535	0.035
CS	1.409	1.350	0.060	1.147	1.096	0.051	0.985	0.941	0.044	0.785	0.749	0.036	0.519	0.018	0.501
ES	1.409	1.350	0.060	1.145	1.095	0.050	0.983	0.940	0.043	0.781	0.746	0.035	0.510	0.486	0.025
Intercept															
UW	0.705	0.699	0.006	0.699	0.691	0.008	0.694	0.686	0.008	0.686	0.678	0.008	0.668	0.664	0.004
RW	0.946	0.896	0.050	0.956	0.899	0.057	0.958	0.900	0.058	0.958	0.900	0.058	0.949	0.896	0.053
CS	0.937	0.891	0.045	0.942	0.892	0.051	0.942	0.891	0.052	0.937	0.887	0.050	0.916	0.876	0.039
ES	0.936	0.892	0.044	0.942	0.892	0.050	0.942	0.891	0.051	0.936	0.887	0.049	0.907	0.875	0.032
lv1_var															
UW	0.783	0.772	0.011	1.095	1.081	0.015	1.252	1.235	0.017	1.408	1.389	0.019	1.548	1.528	0.020
RW	0.997	0.991	0.006	1.396	1.388	0.008	1.595	1.586	0.009	1.795	1.785	0.011	1.976	1.964	0.012
CS	1.010	1.008	0.002	1.414	1.411	0.003	1.616	1.613	0.003	1.819	1.815	0.004	2.006	1.999	0.007
ES	1.019	1.018	0.001	1.426	1.426	0.001	1.630	1.629	0.001	1.834	1.833	0.001	2.021	2.016	0.006
lv2_var															
UW	3.696	3.626	0.070	2.412	2.366	0.046	1.766	1.726	0.040	1.107	1.071	0.037	0.445	0.430	0.015
RW	4.714	4.441	0.273	3.148	2.974	0.174	2.366	2.242	0.124	1.585	1.511	0.073	0.875	0.859	0.016
CS	4.660	4.413	0.247	3.059	2.912	0.147	2.249	2.148	0.101	1.426	1.365	0.061	0.633	0.615	0.018
ES	4.648	4.408	0.240	3.040	2.898	0.142	2.227	2.128	0.100	1.396	1.335	0.061	0.563	0.556	0.007

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling; SD=standard deviation; SE=standard error; Diff=difference.

4.2 Results for ECLS-K:2011

First, the informativeness of the weights is examined following Laukaityte and Wiberg (2018). The student-level effective sample sizes n_j^{eff} are all smaller than the actual sample sizes except those schools which have only one student. The school-level effective sample size N^{eff} is 614, which is smaller than the actual number of schools. Therefore, both level weights are informative and both level weights would affect the results of the multilevel analysis.

Three two-level HLM models with different sets of covariates are used to fit two dependent variables: reading achievement scores and mathematics achievement scores. The first model is a null model, the second is the model with student-level predictors (I label it as student model), and the third model is a full model with student-level and school-level predictors included. Table 4.7 presents the results of the unweighted and weighted null models. Even this simple model shows there are important differences in the estimates of the variance components. Having no weights produces the largest estimates of student-level variance, whereas using raw weights produces the largest estimates of school-level variance. The estimates of intercept are found to be in the same direction and have similar sizes to each other across the four estimators in reading and mathematics. Still, the weighted intercept estimates are consistently larger than unweighted estimate. Overall, the unweighted method has the smallest standard errors and largest test statistics consistently among the four estimators. In addition, the two scaling methods perform more similarly with much more similar results of point estimates, standard errors, and consequently the test statistics.

The ICC (see Table 4.7) shows 19.6% and 16.2% of the total variance in mathematics and reading achievement are attributable to schools. Based on Equation 2.15, the design effects are 13.61 and 13.65 for mathematics and reading respectively. They are greater than 2, indicating that using multilevel model to analyzed data here is reasonable.

Table 4.7. *Null Model for ECLS-K: 2011 Mathematics and Reading*

	math				reading			
parameter	UW	RW	CS	ES	UW	RW	CS	ES
Intercept	45.100***	45.616***	45.712***	45.738***	61.127***	51.296***	51.410***	51.427***
SE	0.190	0.268	0.251	0.252	0.205	0.316	0.273	0.273
Statistic	237.616	170.419	182.311	181.764	297.933	170.419	182.311	224.707
$\hat{\sigma}_e^2$	119.447**	10.235**	15.811**	116.203***	150.821**	35.376**	44.198**	45.070 ***
SE	1.615	2.171	1.896	1.917	2.407	2.998	2.719	2.746
Statistic	73.956	50.787	61.073	60.619	62.659	50.787	61.073	52.833
$\hat{\sigma}_u^2$	29.049***	39.612***	30.189***	29.257***	33.370***	51.268***	34.361***	33.087***
SE	1.544	2.369	2.033	2.033	1.937	3.663	2.593	2.609
Statistic	18.814	16.724	14.852	14.393	17.224	16.724	14.852	12.680
ICC	0.196	0.264	0.207	0.201	0.162	0.275	0.192	0.186

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling; SE=standard error.
*$p < .05$; **$p < .01$; ***$p < .001$.

The results of the model with student-level predictors are depicted in Table 4.8. Contrary to the null model, the intercept estimates in the weighted models are smaller than those in the unweighted model. As in the null model, the unweighted model produces the largest estimate of student-level variance and the raw weighted model produces the largest estimate of school-level variance. Furthermore, the indices of goodness of fit AIC, BIC, and deviance are substantially larger when raw weighted estimation method is applied. Compared with the null model, the standard errors for the intercept increase, while the standard errors for student-level and school-level variance decrease. The within-school variance decreases by 67% for both mathematics and reading, the between-school variance decrease varies from by 68% to 72% for mathematics, and from by 61% to 64% for reading. Similar results are obtained when both student-level and school-level weights are used in the model. The standard errors of all the parameters with the unweighted method are consistently smaller than those of weighted methods, and the test statistics of the

unweighted estimator consistently larger than those of the weighted estimators, as expected. The significance is stable for all the parameters as well.

Table 4.8. *Model with Student-Level Predictors for ECLS-K: 2011 Mathematics and Reading*

	Mathematics				Reading			
parameter	UW	RW	CS	ES	UW	RW	CS	ES
Intercept	18.211***	17.635***	18.035***	18.038***	17.650***	16.806***	16.834 ***	16.817***
SE	0.257	0.366	0.315	0.317	0.366	0.485	0.450	0.454
Statistic	70.736	48.214	57.217	56.881	48.162	34.677	37.390	37.057
female	0.094	0.092	0.046	0.040	0.901***	1.001***	0.982***	0.979***
SE	0.110	0.163	0.133	0.134	0.128	0.194	0.144	0.144
Statistic	0.856	0.566	0.344	0.299	7.058	5.159	6.837	6.797
SES	0.876***	1.178***	0.990***	0.989***	1.072***	1.300***	0.986***	0.985***
SE	0.088	0.138	0.104	0.104	0.104	0.180	0.126	0.127
Statistic	9.924	8.554	9.520	9.487	10.297	7.208	7.837	7.775
pretest	0.853***	0.869***	0.860***	0.860***	0.922***	0.932***	0.933***	0.934***
SE	0.006	0.009	0.008	0.008	0.007	0.009	0.008	0.009
Statistic	132.816	93.282	113.064	112.305	131.044	101.76	109.992	109.004
$\hat{\sigma}_e^2$	39.459***	35.760***	38.439***	38.503***	49.660***	44.265***	47.230***	47.259 ***
SE	0.577	0.816	0.664	0.669	0.985	1.242	0.968	0.976
Statistic	68.373	43.827	57.913	57.554	50.405	35.654	48.789	48.427
$\hat{\sigma}_u^2$	8.132***	12.827***	8.631***	8.595***	12.038***	19.791***	12.760 ***	12.722***
SE	0.587	0.921	0.759	0.762	0.912	1.812	1.182	1.185
Statistic	13.843	13.926	11.373	11.281	13.202	10.919	10.795	10.733
deviance	89979	1630285364	89670.5	89701.1	93596.1	1689150119	92972.82	92993.4
AIC	89991	1630285376	89682.5	89713.1	93608.1	1689150131	92984.82	93005.4
BIC	90036.1	1630285480	89727.6	89758.2	93653.2	1689150236	93029.96	93050.5

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling; SE=standard error.
AIC: Akaike Information Criteria; BIC=Bayesian Information Criteria.
**p < .05; **p < .01; ***p < .001.*

Table 4.9 reports the results of the full model. The covariate suburban is found not to contribute significantly to the model for both reading and mathematics data. Another model excluding suburban is also run. These two models are then compared using likelihood ratio test:

Table 4.9. *Full Model for ECLS-K: 2011 Mathematics and Reading*

	Mathematics				Reading			
parameter	UW	RW	CS	ES	UW	RW	CS	ES
Intercept	18.042***	17.42***	17.773***	17.775***	17.427 ***	16.565***	16.571***	16.554***
SE	0.267	0.365	0.328	0.330	0.377	0.502	0.471	0.475
Statistic	67.620	47.673	54.260	53.930	46.188	33.026	35.196	34.876
female	0.114	0.077	0.053	0.048	0.909***	0.934***	0.980***	0.978***
SE	0.111	0.161	0.133	0.134	0.129	0.166	0.143	0.144
Statistic	1.026	0.475	0.401	0.358	7.052	5.635	6.835	6.796
SES	0.864***	1.181***	0.994***	0.993***	1.056***	1.203***	0.968***	0.967 ***
SE	0.089	0.137	0.105	0.105	0.105	0.151	0.124	0.125
Statistic	9.736	8.604	9.457	9.428	10.055	7.957	7.789	7.730
pretest	0.853***	0.866***	0.859***	0.859***	0.922***	0.931***	0.933***	0.933***
SE	0.007	0.009	0.008	0.008	0.007	0.009	0.009	0.009
Statistic	131.011	97.867	113.103	112.373	129.325	102.808	107.969	106.949
rural	0.762**	1.201***	0.988**	0.988**	1.043***	1.265**	1.152**	1.153**
SE	0.271	0.368	0.346	0.346	0.304	0.432	0.385	0.385
Statistic	2.817	3.264	2.854	2.853	3.434	2.924	2.994	2.992
$\hat{\sigma}_e^2$	39.457***	36.118***	38.628***	38.695***	49.657***	44.021 ***	47.333***	47.371***
SE	0.579	0.779	0.662	0.667	0.996	0.982	0.961	0.969
Statistic	68.093	46.386	58.354	58.017	49.880	44.824	49.271	48.895
$\hat{\sigma}_u^2$	8.082***	12.820***	8.700***	8.664***	11.963***	19.871***	12.803***	12.765***
SE	0.590	0.908	0.765	0.769	0.924	1.833	1.194	1.198
Statistic	13.696	14.118	11.367	11.274	12.943	10.839	10.722	10.659
deviance	88334.6	1532438229	88104.4	88135.8	91891.7	1584270571	91316.2	91337.2
AIC	88348.6	1532438243	88118.4	88149.8	91905.7	1584270585	91330.2	91351.1
BIC	88401.1	1532438364	88170.9	88202.3	91958.3	1584270706	91382.8	91403.7

Note: UW=unweighted estimation method; RW=estimation method with raw weights; CS=estimation method with cluster scaling; ES=estimation method with effective cluster scaling; SE=standard error.
AIC: Akaike Information Criteria; BIC=Bayesian Information Criteria.
**p < .05; **p < .01; ***p < .001.*

one with suburban and one without. No significant result is found. Therefore, I simplify the model and include the three student-level predictors and only one school-level predictor rural in the model as full model in this study. The findings from comparison of weighted and unweighted analyses are similar to the those obtained from the model with only student level predictors. The estimates,

standard errors, and consequently the test statistics do not show much differences between the full model and student model. However, one can see that the significance remains unchanged for all the parameters except for school-level covariate rural. It changes, from being significant at 0.01 with raw weighted model to being significant at 0.001 with the other three models for mathematics data. For reading data, the estimate for rural is significant at 0.001 with unweighted model, but it changes to be significant at 0.01 with other three weighted models.

In general, for both reading and mathematics data in ECLS-K:2011, using weighted approaches produce larger standard errors and smaller test statistics than unweighted model. Hah-Vaughn (2005) pointed out that "the larger standard errors and resulting smaller test statistic values generated suggest that, given a different model, the chance of committing Type I error will increase substantially when weights are used, although rejection of the hypotheses remain the same across all the models". Among the weighted approaches, the raw weighted method produces larger standard errors than the other two weighted methods do. The two scaling methods perform quite similarly for all the parameters in all models.

CHATPER 5 SUMMARY AND DISCUSSION

This chapter provides a summary, a discussion and limitations of the results. It consists of four sections. The first section summarizes the research objectives, and results. The second section presents the discussion of major findings, followed by the implications. Limitations of this study and directions for future research are discussed in the final section.

5.1 Summary of This Study

The primary aim for this study is to examine the performance of the four estimators and analyze the impact of sampling weights in multilevel models in the context of two-stage informative and non-informative sampling designs. Large-scale data in social science usually adopt complex sampling designs, such as clustering and unequal probability of selection, which bring challenges in statistical analysis. Using multilevel models to analyze complex large-scale assessment data accounting for clustering is becoming more and more popular, but it is still a question in when and how to use sampling weights in such models, to correct for unequal probability of selection. For example, there is controversy whether to use weight or not. It has long history arguing this issue between model-based and design-based schools. Even if we have determined to use weights, for instance, in a two-level model, using single-level weight derived from the product of the weights from each level, or using multilevel weights is debatable. I use multilevel weights in this study because single-level weight may not carry adequate information to correct for unequal probability of selection. The analysis with real data shows that incorporating sampling weights in the model does produce different parameter estimates, standard errors, test statistics and even sometimes the significance of a certain variable from those obtained when both

levels are informative. Weighted models have larger standard errors and smaller test statistics than unweighted model does. And the cluster scaling and effective scaling method produce more similar results compared with the unweighted and raw weighted model. Therefore, caution should be exercised while weights are applied in the multilevel analysis.

In this study, Monte Carlo simulations are conducted to evaluate the performance of the four estimation methods in the informative and non-informative sampling design in a linear random-intercept model, because prior studies (e.g., Cai, 2013) found that the estimates were biased if the informativeness was ignored. Summary of the comparisons of the estimators are depicted in Table 5.1. Substantial differences are found among these four estimation methods while estimating the intercept and variance components. In the informative design, in terms of bias, the weighted estimators outperform the unweighted for the intercept and student-level variance estimation, whereas the unweighted estimator works the best for school-level variance estimation. Although the three weighted estimators produce almost unbiased estimates for the intercept and student-level variance, they perform quite differently. The three weighted perform almost equally well for intercept estimation, while the cluster scaling estimator performs the best for student-level variance estimation. Raw weighted method works the worst and should be used with caution when estimating school-level variance. The weighted methods give better coverage rates for the intercept and student-level variance, but unweighted method does for school-level variance in the informative design. In the non-informative setting, the unweighted method gives the better coverage rate for all the parameter estimates. The unweighted estimator performs the best or the second best in terms of relative bias in the non-informative condition. Furthermore, including sampling weights decreases the RMSE for the intercept and student-level variance and increase

Table 5.1. *Summary of Comparisons of the Estimation Methods*

Criterion	Estimate	Informativeness	Noninformativeness
(absolute) RB	covariates	They are all nearly unbiasedly estimated.	They are all nearly unbiasedly estimated.
	intercept	The three weighted estimators performs equally well and better than the unweighted.	The unweighted estimator performs the best in most cases and the effective scaling estimator the worst in most cases.
	level-1 variance	The cluster scaling estimator performs the best and the unweighted estimator the worst.	The effective scaling and the unweighted estimator perform the best and the raw weighted the worst.
	level-2 variance	The unweighted estimator works the best and the raw weighted estimator the worst.	The effective scaling and the unweighted estimator perform the best and the raw weighted the worst.
RMSE	covariates	The unweighted performs the most efficiently the weighted.	The unweighted estimator performs the most efficiently.
	intercept	The three weighted estimators performs equally well and more efficiently than the unweighted.	The unweighted estimator performs the most efficiently.
	level-1 variance	The cluster scaling estimator performs the most efficiently and the unweighted estimator the least.	The unweighted estimator performs the most efficiently.
	level-2 variance	The unweighted estimator performs the most efficiently and the raw weighted estimator the least.	The unweighted estimator performs the most efficiently.
95%CR	covariates	Most of the coverage rates are over 0.94.	Most of the coverage rates are over 0.94.
	intercept	Thre weighted estiamtors have almost equal coverage rates (>0.91) and the unweighted has the coverage rate of 0.	The four estimators have similar coverage rates and they are all over 0.93.
	level-1 variance	The cluster scaling estimator has the highest coverage rate and the unweighted has the smallest coverage rate.	The effective scaling and the unweighted estimator have the highest coverage rates and the raw weighted the smallest.
	level-2 variance	The unweighted estimator has the highest coverage rate and the raw weighted has the smallest in most cases.	The unweighted estimator has higher coverage rates in most cases and the raw weighted has the smallest in most cases.

Note: RB=relative bias; RMSE=root mean square error; 95%CR=95% confidence interval coverage rate.

the RMSE for the school-level variance in the informative design. However, it increases the RMSE for the intercept, student-level variance and school-level variance in the non-informative design. Therefore, the unweighted method works the most efficiently for all the parameter estimates across different levels of the ICC in the non-informative design. Tentatively, the cluster scaling estimator and effective scaling estimator might be preferred in the informative condition.

ICC is one of the factors that influences the quality of estimation (e.g., Asparouhov & Muthén, 2006; Kovačević & Rai, 2003). Therefore, it is manipulated in this study. Simulation results are summarized in Table 5.2 and it shows, the effect of the ICC is related to relative bias and RMSE, but not sensitive to coverage rate. As the ICC increases, the bias for student-level variance increases and the bias for school-level variance decreases in both conditions. These changes are quite obvious for school-level variance, but hard to see for student-level variance. No monotonic patterns for the relative bias can be found as the ICC increases for fixed effects and intercept in the informative condition, but clear patterns can be seen for fixed effects and intercepts as the increase of the ICC.

RMSE shows the similar patterns in both conditions for all the parameters. As the ICC increases, the RMSE decreases for the three student-level fixed effects and variance, and increases for the school-level fixed effect and variance with all the four estimators.

Take the following scenario when ICC = 0.3 for example. In the informative condition, when ICC = 0.3, the simulation results show that the cluster scaling estimator works best for the intercept and student-level variance in terms of relative bias, RMSE and coverage rate. Although it is not the best estimator for the school-level variance estimates among the weighted estimators, it gives the best coverage rate and just slightly higher RMSE compared with the best weighted estimator, the effective scaling estimator. In addition, it produces unbiased estimates for the

Table 5.2. *ICC Effect*

Criterion	Estimate	Informativeness	Noninformativeness
(absolute) RB	covariates	No consistent pattern found.	As the ICC increases, the absolute relative bias for the three level-1 covariates decreases, whereas the relative bias for level-2 covariate increases except with unweighted estimator.
	intercept	as the ICC increases, the unweighted estimates increases, whereas the weighted estimates remain almost the same.	As the ICC increases, the weighted estimates decreases while the unweighted estimates increases.
	level-1 variance	The increasing rate is so tiny that almost same patterns are found for all the estiamtors.	The increasing rate is so tiny that almost same patterns are found for all the estiamtors.
	level-2 variance	As the ICC increases, the relative bias decreases.	As the ICC increases, the relative bias decreases.
RMSE	covariates	As the ICC increases, the RMSE for the level-1 covariates decreases and for the level-2 covariate increases.	As the ICC increases, the RMSE for the level-1 covariates decreases and for the level-2 covariate increases.
	intercept	As the ICC increases, the unweighted estimates increases as well, while the weighted estimates remain almost the same.	As the ICC increases, the RMSE for the intercept remains almost the same.
	level-1 variance	As the ICC increases, the RMSE for level-1 variance decreases.	As the ICC increases, the RMSE for the level-1 variance decreases.
	level-2 variance	As the ICC increases, the RMSE for level-2 variance increases as well.	As the ICC increases, the RMSE for the level-2 variance increases as well.
95%CR	covariates	No obvious monotonic increasing or decreasing pattern	No obvious monotonic increasing or decreasing pattern
	intercept	No obvious monotonic increasing or decreasing pattern	No obvious monotonic increasing or decreasing pattern
	level-1 variance	No obvious monotonic increasing or decreasing pattern	No obvious monotonic increasing or decreasing pattern
	level-2 variance	No obvious monotonic increasing or decreasing pattern	No obvious monotonic increasing or decreasing pattern

Note: RB=relative bias; RMSE=root mean square error; 95%CR=95% confidence interval coverage rate.

school-level variance. Therefore, in the informative setting, cluster scaling estimator is preferred in most cases. In the non-informative condition, when ICC is 0.3, the unweighted estimator has the least (absolute) relative bias, RMSE and highest coverage rate in almost all the cases. Therefore, the unweighted estimator is preferred in the non-informative condition.

5.2 Discussion of Results

The design of current simulation captures the general features of large-scale data sets available in social studies, for example, large number of clusters with different sizes, unequal probability of selection, and moderate informativeness values. Some of the findings from the previous studies are confirmed, and some are not in this study. For example, prior studies showed that the unweighted method produces biased estimate for the intercept and school-level variance when the sampling design is informative at both levels (Cai, 2013; Pfeffermann et al., 1998). Pfeffermann et al. (1998) pointed out that when the design is informative at the cluster level, the unweighted method only produces biased estimates for intercept and school-level variance, not for student-level variance. However, the current study shows that the unweighted method works quite well most of the time for school-level variance estimation, and it only does not work well when the ICC is extremely small in the informative design. None of the estimators works well when the ICC is extremely small. This is expected because, based on the equation 3.15, we have a very small denominator, 0.6, which results in a very large relative bias compared with relative bias when ICC is comparatively larger. As for student-level variance, although the unweighted estimator works the worst in the informative condition, it produces unbiased estimates. In addition, Cai (2013) pointed out that including the sampling weights substantially increases the MSE. This is only confirmed in the non-informative setting, but not in the informative setting in the current study.

All the fixed effects are nearly unbiased estimated in terms of Muthén & Muthén (2002). This is confirmed in both studies. In general, including sampling weights still produces biased estimates. This is confirmed by all the studies. Asparouhov and Muthén (2007) reported that the MPML estimator outperforms substantially the other estimators. This is partially confirmed in the present study, since cluster scaling estimator performs better than others in the informative condition, while raw weighted estimator needs to be used with caution, especially when we estimate variance components in the informative condition.

Previous studies (e.g., Asparouhov & Muthén, 2006; Kovačević & Rai, 2003) found that the bias increases for all the parameters as the ICC decreases. This is only partially confirmed in the current study. Current results do not show monotonic patterns of the relative bias for the fixed effects and intercept, but bias increases for student-level variance and decreases for school-level variance as the ICC increases in the informative condition. In the non-informative condition, the increase of the ICC decreases the bias for student-level fixed effects and variance, and increase the bias for school-level fixed effect and variance. Therefore, the tentative conclusion is that weighted estimators with cluster scaling and effective scaling weights are preferred when the ICC is not extremely small in the informative design and unweighted method could be used in the non-informative design.

The differences above might be due to the different settings of simulation. For example, either the estimators are examined using random-intercept model with no covariates at both levels (cf., Asparouhov & Muthén, 2006; Kovačević & Rai, 2003) or the linear random-intercept model is used with no school-level predictors (cf., Cai, 2013). Therefore, it is possible that our results might not be replicated in different settings.

5.3 Implications

The major finding from this study confirms that including sampling weights in the analysis produce different estimates in the informative sampling design and the unweighted method works best in the non-informative sampling design. The fair comparison between the weighted and unweighted, and between the informative and non-informative design might indicate to use sampling weights in the informative design and use unweighted estimation method in the non-informative sampling design. Calculation of informativeness is necessary since it gives us the extent to which the design is informative and indicate whether it is necessary to include sampling weights. Second, researchers should examine the ICC and evaluate the magnitude and significance of variance components to determine whether multilevel modeling is necessary. Lastly but not the least, caution should be taken in using sampling weights when ICC is extremely small.

5.4 Limitations and Future Studies

There are several limitations in this study. The primary limitation is that only a simple linear random-intercept model is applied. It is more real if the slopes are random and different types of outcome variables, such as Poisson or nominal, may be used. This may provide us with a clearer picture which estimator works best. Second, besides scaling the sampling weights, trimming weights can be an alternative, which is not considered in this study. Third, I just roughly divide the situation into two: informative or non-informative. It might be better idea if different levels of informativeness, for example, low, medium and high levels of informativeness are all included in the analysis. This might tell us under which condition of informativeness, the parameter estimates can be estimated unbiasedly. Fourth, multistage sample selection is more complicated in real life. Therefore, the simulation design may not well reflect the reality.

Not all the findings in the prior studies are confirmed in this study. Therefore, more studies are needed to evaluate MPML performance in different settings. For example, different types of outcome variable, such as discrete response or count data can be used. There are more and more research focusing on them (Chaudhuri, Handcock, & Rendall, 2008; Natarajan, Lipsitz, Fitzmaurice, Moore, & Gonin, 2008; Nordberg, 1989; Rodriguez & Goldman, 1995, 2001), or higher level HLM models (e.g., three-level model) can be used. Furthermore, as is true with any simulation, conclusions from this study are restricted to a particular sampling design and modeling context. In order to see if comparable findings happen in alternative situations, future research is necessary. In this study, the simulation is conducted on the basis of a large of number of clusters. Small samples are possible in practice. The performance of estimators might suffer from the small number of clusters (Asparouhov & Muthén, 2005; Li & Redden, 2015; Mass & Hox, 2005). Research to examine the performance of different estimation methods in unideal conditions is necessary. Above all, future research is needed to enhance weighted multilevel models. Asparouhov & Muthén (2010) stated that Bayesian estimation method could be an alternative with maximum likelihood estimation methods when sample sizes are small if we have informative priors, but few comparisons were made in the context of informative sampling designs.

APPENDICES

APPENDIX A. Stata Simulation Syntax in the Informative Sampling Design

```
/*******************************************************************************/
set more off
local info 30 18 12 6 0.6 /*level 2 variance*/
forvalues i = 1/1000 {/*to repeat the process 1000 times*/
	display "iteration `i'"
	foreach j in `info' {
		clear
		display "l2var `j'"
		*generate school level data
		quietly: set seed 1`i'1
		quietly: set obs 75000
		quietly: gen uj = rnormal(0, sqrt(`j')) /*need sd here, so need to square root j*/
		*uj recaled
		quietly: egen ujmean = mean(uj)
		quietly: egen ujsd = sd(uj)
		quietly: gen uj_scaled = ((uj-ujmean)/ujsd)*sqrt(2)
		quietly: gen pj = 1/(1+exp(4.12-uj_scaled/2))
		quietly: gen wj = 1/pj
		quietly: gsample 150 [aw=pj] /*draws a unequal probability sample with sampling probabilities pj.*/
		quietly: gen index = 1
		quietly: gen school = _n

		*school covariates
		quietly: gen rand = runiform()
		quietly: gen locale = cond(rand < 0.22, 1, cond(rand < 0.58, 2, 3))
		quietly: gen rural = locale==1
		quietly: gen suburb = locale==2
		quietly: gen urban = locale==3
		*expand students based on percentages of different types of schools
		quietly: expand 16+int((24-10+1)*runiform()) if school<=8  /*5.69% of 150 schools: 8*/
		quietly: expand 25+int((49-25+1)*runiform()) if school>=9 &school<=25  /*11.49% of 150 schools: 17*/
		quietly: expand 50+int((99-50+1)*runiform()) if school>=26 & school<=91 /*43.53% of 150 schools:66*/
		quietly: expand 100+int((149-100+1)*runiform()) if school>=92 & school<=129 /*25.48% of 150 schools:38*/
		quietly: expand 150+int((199-150+1)*runiform()) if school>=130 & school<=142 /*8.59% of 150 schools:13*/
```

```
                quietly: expand 200+int((600-200+1)*runiform()) if school>=143 & school<=150
/*5.22% of 150 schools:8*/
                quietly: bysort school: generate student = _n
                *generate student data
                quietly: gen eij = rnormal(0, sqrt(60-`j'))
                *eij recaled
                quietly: egen eijmean = mean(eij)
                quietly: egen eijsd = sd(eij)
                quietly: gen eij_scaled = ((eij-eijmean)/eijsd)*sqrt(2)
                quietly: gen pi_j = 1/(1+exp(1.23-eij_scaled/2))
                quietly: gen wi_j = 1/pi_j
                quietly: gen pij = pi_j*pj
                quietly: gen wij = 1/pij
                *generate correlated data for female, SES and pretest
                quietly: local p = 0.49
                quietly: matrix m = (0, -0.05, 46.92)
                quietly: matrix sd = (0.5, 0.81, 11.50)
                quietly: matrix input c = (1, 0.005, 1, 0.07, 0.409, 1)
               quietly: corr2data female SES pretest, corr(c) means(m) sds(sd) cstorage(lower)
                /* Steps 2-3 for the one Bernoulli variable */
               quietly: replace female = cond(normal(female)>=(1-`p'),1,0)
                /*merge two level data*/
                quietly: gen yij = 17.43+0.91*female+1.06*SES + 0.92*pretest+1.04*rural+uj+eij
                quietly: rename yij achieve
                quietly: rename wj schwgt
                quietly: rename wi_j stdwgt
                *select final sample
                quietly: keep if index == 1
                quietly: gsample 3915 [aw=pi_j]
                if `j' == 30 local r = 1
                if `j' == 18 local r = 2
                if `j' == 12 local r = 3
                if `j' == 6 local r = 4
                if `j' == 0.6 local r = 5
                quietly: keep student schwgt school locale rural suburb urban stdwgt female SES
pretest achieve
                gen iteration = `i'
**************************************************************************/
```

APPENDIX B. Stata Simulation Syntax in the Non-Informative Sampling Design

```
/*********************************
set more off
local info 30 18 12 6 0.6 /*level 2 variance*/
forvalues i = 1/1000 {/*to repeat the process 1000 times*/
        display "iteration `i'"
        foreach j in `info' {
                clear
                display "l2var `j'"
                *generate school level data
                quietly: set seed 1`i'1
                quietly: set obs 75000
                quietly: gen uj = rnormal (0, sqrt(`j'))
                *betaj recaled
                quietly: gen betaj = rnormal (0, sqrt(2))
                quietly: egen betajmean = mean(betaj)
                quietly: egen betajsd = sd(betaj)
                quietly: gen betaj_scaled = ((betaj-betajmean)/betajsd)*sqrt(2)
                quietly: gen pj = 1/(1+exp(4.12-betaj_scaled/2))
                quietly: gen wj = 1/pj
                quietly: gsample 150 [aw=pj] /*draws a unequal probability sample with sampling
probabilities pj.*/
                quietly: gen index = 1
                quietly: gen school = _n
                *school covariates
                quietly: gen rand = runiform()
                quietly: gen locale = cond(rand < 0.22, 1, cond(rand < 0.58, 2, 3))
                quietly: gen rural = locale==1
                quietly: gen suburb = locale==2
                quietly: gen urban = locale==3
                *expand students based on percentages of different types of schools
                quietly: expand 16+int((24-10+1)*runiform()) if school<=8  /*5.69% of 150
schools: 8*/
              quietly: expand 25+int((49-25+1)*runiform()) if school>=9 &school<=25  /*11.49%
of 150 schools: 17*/
                quietly: expand 50+int((99-50+1)*runiform()) if school>=26 & school<=91
/*43.53% of 150 schools:66*/
                quietly: expand 100+int((149-100+1)*runiform()) if school>=92 & school<=129
/*25.48% of 150 schools:38*/
                quietly: expand 150+int((199-150+1)*runiform()) if school>=130 & school<=142
/*8.59% of 150 schools:13*/
```

```
                quietly: expand 200+int((600-200+1)*runiform()) if school>=143 & school<=150
/*5.22% of 150 schools:8*/
                quietly: bysort school: generate student = _n
                *generate student data
                quietly: gen eij = rnormal(0,sqrt(60-`j'))
                *rij recaled
                quietly: gen eij = rnormal(0,sqrt(60-`j'))
                quietly: gen rij = rnormal(0,sqrt(2))
                quietly: egen rijmean = mean(rij)
                quietly: egen rijsd = sd(rij)
                quietly: gen rij_scaled = ((rij-rijmean)/rijsd)*sqrt(2)
                quietly: gen pi_j = 1/(1+exp(1.23-rij_scaled/2))
                quietly: gen wi_j = 1/pi_j
                quietly: gen pij = pi_j*pj
                quietly: gen wij = 1/pij
                *generate correlated data for female, SES and pretest
               quietly: local p = 0.49
               quietly: matrix m = (0, -0.05,46.92)
               quietly: matrix sd = (0.5,0.81,11.50)
                quietly: matrix input c = (1, 0.006, 1, 0.07, 0.409, 1)
                quietly: corr2data female SES pretest, corr(c) means(m) sds(sd) cstorage(lower)
                /* Steps 2-3 for the one Bernoulli variable */
                quietly: replace female = cond(normal(female)>=(1-`p'),1,0)
                /*merge two level data*/
                quietly: gen yij = 17.43+0.91*female+1.06*SES + 0.92*pretest+1.04*rural+uj+eij
                quietly: rename yij achieve
                quietly: rename wj schwgt
                quietly: rename wi_j stdwgt
                *select final sample
                quietly: keep if index == 1
                quietly: gsample 3915 [aw=pi_j]
                if `j' == 30 local r = 1
                if `j' == 18 local r = 2
                if `j' == 12 local r = 3
                if `j' == 6 local r = 4
                if `j' == 0.6 local r = 5
                quietly: keep student schwgt school locale rural suburb urban stdwgt female SES
pretest achieve
                gen iteration = `i'
        }
}

**********************************************************************/
```

APPENDIX C. M*plus* Syntax

```
/*************************Mplus VERSION 8*****************************/
/***************** *******Unweighted estimation method*********************/
Title: READING with NO weights;
Data: File is iteration_list.csv;
      Type = MONTECARLO;
Variable: Names are
      schwgt school locale rural suburb urban student stdwgt female
      SES pretest achieve iteration;
    USEVARIABLES are achieve school female SES pretest rural;
    CLUSTER = school;
    WITHIN = female SES pretest;
    BETWEEN = rural;
MODEL: %WITHIN%
      achieve on female*.91 SES*1.06 pretest*.92;
      achieve*30; !variance at level1
      %BETWEEN%
      achieve on rural*1.04;
      [achieve*17.43]; ![gamma00]
      achieve*30; !variance at level2
ANALYSIS:
        TYPE = TWOLEVEL;

/***************** Estimating method with raw weights *********************/
Title: READING with raw weights (unscaled);
Data: File is iteration_list.csv;
      Type = MONTECARLO;
Variable: Names are
          schwgt school locale rural suburb urban student stdwgt female
          SES pretest achieve iteration;
    USEVARIABLES are achieve school female SES pretest rural;
    CLUSTER = school;
    WITHIN = female SES pretest;
    BETWEEN = rural;
    Weight is stdwgt;
    Bweight = schwgt;
    Wtscale = UNSCALED;
    Bwtscale = UNSCALED;
MODEL: %WITHIN%
      achieve on female*.91 SES*1.06 pretest*.92;
      achieve*30; !variance at level1
```

```
      %BETWEEN%
      achieve on rural*1.04;
      [achieve*17.43]; ![gamma00]
      achieve*30; !variance at level2
ANALYSIS:
       TYPE = TWOLEVEL;
       algorithm = integration;
       estimator = MLR;

/*********************Estimation method with cluster scaling************************/
Title: READING with scaling1;
Data: File is iteration_list.csv;
     Type = MONTECARLO;
Variable: Names are
          schwgt school locale rural suburb urban student stdwgt female
        SES pretest achieve iteration;
    USEVARIABLES are achieve school female SES pretest rural;
    CLUSTER = school;
    WITHIN = female SES pretest;
    BETWEEN = rural;
    Weight is stdwgt;
    Bweight = schwgt;
    Wtscale = cluster;
    Bwtscale = sample;
MODEL: %WITHIN%
     achieve on female*.91 SES*1.06 pretest*.92;
     achieve*30; !variance at level1
     %BETWEEN%
     achieve on rural*1.04;
     [achieve*17.43]; ![gamma00]
     achieve*30; !variance at level2
ANALYSIS:
       TYPE = TWOLEVEL;
       algorithm = integration;
       estimator = MLR;

/************* Estimation method with effective scaling (ecluster scaling)***************/
Title: READING with scaling2;
Data: File is iteration_list.csv;
     Type = MONTECARLO;
Variable: Names are
          schwgt school locale rural suburb urban student stdwgt female
          SES pretest achieve iteration;
```

```
    USEVARIABLES are achieve school female SES pretest rural;
    CLUSTER = school;
    WITHIN = female SES pretest;
    BETWEEN = rural;
    Weight is stdwgt;
    Bweight = schwgt;
    Wtscale = ecluster;
    Bwtscale = sample;
MODEL: %WITHIN%
      achieve on female*.91 SES*1.06 pretest*.92;
      achieve*30; !variance at level1
      %BETWEEN%
      achieve on rural*1.04;
      [achieve*17.43]; ![gamma00]
      achieve*30; !variance at level2
ANALYSIS:
       TYPE = TWOLEVEL;
        algorithm = integration;
        estimator = MLR;
```

REFERENCES

REFERENCES

Arceneaux, K., & Nickerson, D. W. (2009). Modeling certainty with clustered data: A comparison of methods. *Political Analysis,* 17, 177-190. doi: 10. 1093/pan/mpp004

Asparouhov, T. (2004). *Weighting for unequal probability of selection in multilevel modeling*, M*plus* Web Notes: No. 8, available from http://www.statmodel.com/

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods, 35*(3), 439-460.

Asparouhov, T., & Muthén, B. (2005). *Multivariate statistical modeling with survey data* (M*plus* Web Notes). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2007). *Testing for informative weights and weights trimming in multivariate modeling with survey data*. Retrieved August 21, 2012 from http://www.statmodel.com/download/JSM2007000745.pdf

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using Mplus* (M*plus* Technical Report Version 4). Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/Bayes-Advantages18.pdf

Asparouhov, T., & Muthén, B. (2006). *Multilevel modeling of complex survey data.* Paper presented at the Proceedings of the Joint Statistical Meeting in Seattle.

Bainbridge, T. R. (1985). The Committee on standards: precision and bias. – ASTM Standardization News 13, 44-46.

Bertolet, M. (2008). *To weight or not to weight? Incorporating sampling designs into model-based analyses.* (Ph. D.), Carnegie Mellon University, Ann Arbor.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*(3), 279-292.

Bloom, H. S., Bos, J. M., & Lee, S. (1999). Using cluster random assignment to measure program impacts: statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445-469.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59. doi: 10.3102/0162373707299550Schochet, 2008

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide.* New York, NY: Cambridge University Press.

Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology, 43*(1), 178-219.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*. doi:10.1186/1471-2288-9-49

Chantala, K., & Suchindran, C. M. (2006). Adjusting for unequal selection probability in multilevel models: a comparison of software packages. *Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association,* 2815-2824.

Chantala, K., Blanchette, D., & Suchindran, C. M. (2011). Software to compute sampling weights for multilevel analysis. Available from http://www.cpc.unc.edu/research/tools/data_analysis/ml_sampling_weights/Compute%20Weights%20for%20Multilevel%20Analysis.pdf.

Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70*(2), 311-328.

Chaudhuri, S., Handcock, M. S., & Rendall, M. S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistical Sinica, 9*(2), 385-406.

Christ, S., Biemer, P., & Wiesen, C. (2007). *Guidelines for applying multilevel modeling to the NSCAW data*. Ithaca, NY: National Data Archive on Child Abuse and Neglect.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health, 62,* 752-758. doi: 10. 1136/jech.2007.060798

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Danielsen, A. G., Wiium, N., Wilhelmsen, B. U., & Wold, B. (2010). Perceived support provided by teachers and classmates and students' self-reported academic initiative. *Journal of School Psychology, 48*(3), 247-67. doi:10.1016/j.jsp.2010.02.002

Eideh, A., & Nathan, G. (2009). Two-stage informative cluster sampling with application in small area estimation. *Journal of Statistical Planning and Inference, 139*, 3088-3101.

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Francisco, C. A., & Fuller, W. A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics, 19*(1), 454-469.

Fuller, W. (1984). Least squares and related analyses for complex survey design. *The Annals of Statistics, 10*(1), 99-118.

Fuller, W. (2009). *Sampling Statistics.* Hoboken: Wiley.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Graubard, B. I., & Korn, E. L. (1996). Modeling the sampling design in the analysis of health surveys. *Statistical methods in medical research, 5*(3), 43-56.

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology, 30*(1), 93-103.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education, 73*(3), 221-248. doi: 10.3200/JEXE.73.3.221-248

Heck, R. H., & Mahoe, R. (2004). *An example of the impact of sample weights and centering on multilevel SEM models.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. doi: 10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445-489.

Howell, D. C. (2008). The analysis of missing data. In *Handbook of social science methodology*, ed. W. Outhwaite and S. Turner, (208-224). London, GB: Sage.

Hox, J. J., & Kreft, I. G. (1994). Multilevel analysis methods. *Sociological Methods & Research, 22*(3), 283-299.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 221-233). Berkeley, CA: University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512988

Jia, Y., Stokes, L., Harris, I., & Wang, Y. (2011). Performance of random effects model estimators under complex sampling designs. *Journal of Educational and Behavioral Statistics, 36*(1), 6-32.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: A model comparison approach.* New York, NY: Routledge.

Kim, J. K., & Skinner, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika, 100*(2), 385-398. https://www.jstor.org/stable/43304565

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

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*(2), 183-200.

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*(3), 291-295.

Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(1), 175-190.

Kovačević, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multi-level modeling of survey data. *Communications in Statistics-Theory and Methods, 32*(1), 103-121.

Koziol, N. A., Bovaird, J. A., & Suarez, S. (2017). A comparison of population-averaged and cluster-specific approaches in the context of unequal probabilities of selection. *Multivariate Behavioral Research, 52*(3), 325-349. doi: 10.1080/00273171.2-17.1292115

Kreft, I. G. G., & Yoon, B. (1994). *Are multilevel techniques necessary*? *An attempt at demystification*. Retrieved from http://eric.ed.gov/?id=ED371033

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics-Theory and Methods, 47*(20), 4991-5012. https://doi.org/10.1080/03610926.2017.1383429

Lee, J., & Fish, R. M. (2010). International and interstate gaps in value-added math-achievement: multilevel instrumental variable analysis of age effect and grade effect. *American Journal of Education, 117*(1), 109-137.

Li, P., & Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine, 34,* 281-296. http://dx.doi.org/10.1002/sim.6344

Lin, Y. X., Steel, D., & Chambers, R. L. (2004). Restricted quasi-score estimating functions for sample survey data. *Journal of Applied Probability, 41*, 119-130.

Longford, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment.* Washington, DC.

Longford, N. T. (1995). *Random coefficient models*. Handbook of Statistical Modeling for the Social and Behavioral Sciences, 519-570.

Lubienski, S. T., & Lubienski, C. (2006). School sector and academic achievement: a multilevel analysis of NAEP mathematics data. *American Educational Research Journal, 43*(4), 651-698.

Mass, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92. http://dx.doi.org/10.1027/1614-1881.1.3.86

Mels, G. (2006). *LISREL for windows: getting started guide.* Lincolnwood, IL: Scientific Software International.

Mulligan, G. M., Hastedt, S., & McCarroll, J. C. (2012). *First-Time Kindergarteners in 2010-2011: First Findings From the Kindergarten Rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 20101-11 (ECLS-K:2011) (NCES 2012-049).* U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Murray, D. M., & Short, B. (1995). Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates, and applications in intervention studies. *Journal of Studies on Alcohol, 56*(6), 681-694.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: Impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology, 2*(74). doi: 10.3389/fpsyg.2011.00074

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide.* 8th ed. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599-620.

Natarajan, S., Lipsitz, S. R., Fitzmaurice, G., Moore, C. G., & Gonin, R. (2008). Variance estimation in complex survey sampling for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 57*(1), 75-87.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *journal of Official Statistics, 5*(3), 223.

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement, 70*(3), 401-419.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*(2), 317-337. doi: 10.2307/1403631.

Pfeffermann, D., & LaVange, L. (1989). Regression models for stratified multi-stage cluster samples. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds), *Analysis of complex surveys* (237-260). New York, NY: John Wiley & Sons.

Pfeffermann, D., Krieger, A. M., & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica, 8*(4), 1087-1114.

Pfeffermann, D., Skinner, C. J., Holmes D. J., Goldstein, H. & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of Royal Statistical Society*: *Series B, 60*(1), 23-40.

Rabe-Hesketh, S. & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of Royal Statistical Society: Series A, 169*(4), 805-827. https://doi.org/10.1111/j.1467-985X.2006.00426.x

Rao, J. N. K., & Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 533-544.

Rao, J. N. K., Verret, F., & Hidiroglou, M. A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology, 39*(2), 263-282.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear modes (2nd ed.).* Tousand Oaks, CA: SAGE.

Raykov, T. (2011). Intraclass correlation coefficients in hierarchical designs: Evaluation using latent variable modeling. *Structural Equation Modeling*, *18*(1), 73-90. doi: 10.1080/10705511.2011.534319

Raykov, T., & Marcoulides, G. A. (2015). Intraclass correlation coefficient in hierarchical design studies with discrete response variables: a note on a direct interval estimation procedure. *Educational and Psychological Measurement, 75*(6), 1063-1071.

Robin, D. B. (1987). *Multiple imputations for non-response in surveys.* New York, NY: Wiley.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 73-79.

Rodriguez, G., & Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: a case study. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 164*(2), 339-355.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi: 10.1037//1082-989X.7.2.147

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of educational programs. *Journal of Educational and Behavioral Statistics, 22*(1), 62-87. doi: 10.3102/1076998607302714

Scientific Software International, 2005-2012. Multilevel Models. LISREL Documentation. Retrieved July 22, 2011 from http://www.ssicentral.com/lisrel/complexdocs/chapter4_web.pdf

Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association, 77*(380), 848-854.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components.* New York: Wiley.

Skinner, C. J. (1994). *Sample models and weights*. Paper presented at the Proceedings of the Section on Survey Research Methods.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys.* Chichester, UK: Wiley.

Snijder, T. A., & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling, 2nd edition.* London: Sage Publication Ltd.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 28-58. doi: 10.1207/s15328007sem1301_2

Stapleton, L. M. (2012). Evaluation of conditional weight approximations for two-level models. *Communications in Statistics – Simulation and Computation, 41,* 182-204. doi: 10.1080/03610918.2011.579700

Stapleton, L. M., & Kang, Y. (2018). Design effects of multilevel estimates from national probability samples. *Sociological Methods & Research, 47*(3), 430-457.

Tourangeau, K., Nord, C., Lê., T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011). User's Manual for the ECLS-K:2011 Kindergarten Data File and Electronic Cdebook, Public Version. NCES 2015-074. *National Center for Education Statistics.*

West, B. T., Beer, L., Gremel, G. W., Weiser, J., Johnson, C. H., Garg, S., & Skarbinski, J. (2015). Weighted multilevel models: a case study. *American Journal of Public Health, 105*(11), 2214-2215.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48,* 817-830. doi: 10.2307/1912934

Winship, C., & Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research, 23*(2), 230-257.

Xia, Q., & Torian, L. V. (2013). To weight or not to weight in time-location sampling: why not do both? *AIDS and Behavior, 17*(9), 3120-3123.

Zaccarin, S., & Donati, C. (2008). *The effects of sampling weights in multilevel analysis of PISA data* (Working Paper No. 119). Universita Degli Studi di Trieste: Departimento di Scienze Economiche e Statistiche. Retrieved from: http://www2.units.it/nirdses/sito_inglese/working%20papers/files%20for%20wp/wp119.pdf.