Journal of Behavioral Data Science, 2021, *1* (2), 89–118.
DOI: https://doi.org/10.35566/jbds/v1n2/p6

A Weighted Residual Bootstrap Method for Multilevel Modeling with Sampling Weights

Wen Luo[1] and Hok Chio Lai[2]

[1] Texas A&M University, College Station, TX 77843, USA
wluo@tamu.edu
[2] University of South California, Los Angeles, CA 90089, USA
hokchiol@usc.edu

Abstract. Multilevel modeling is often used to analyze survey data collected with a multi-stage sampling design. When the selection is informative, sampling weights need to be incorporated into the estimation. We propose a weighted residual bootstrap method as an alternative to the multilevel pseudo-maximum likelihood (MPML) estimators. In a Monte Carlo simulation using two-level linear mixed-effects models, the bootstrap method showed advantages over MPML for the estimates and the statistical inferences of the intercept, the slope of the level-2 predictor, and the variance components at level-2. The impact of sample size, selection mechanism, intraclass correlation (ICC), and distributional assumptions on the performance of the methods was examined. The performance of MPML was suboptimal when sample size and ICC were small and when the normality assumption was violated. The bootstrap estimates generally performed well across all the simulation conditions but had notably suboptimal performance in estimating the covariance component in a random slopes model when sample size and ICCs were large. As an illustration, the bootstrap method is applied to the American data of the OECD's Program for International Students Assessment (PISA) survey on math achievement using the R package *bootmlm*.

Keywords: Bootstrap · Informative Selection · Multilevel Modeling · Sampling Weights · Pseudo-maximum Likelihood

1 Introduction

Multi-stage sampling design is often used in survey data collection. For example, in order to obtain a nationally representative sample of kindergartners, a two-stage sample design may be used in which a representative set of schools are sampled in the first stage and students within schools are sampled in the second stage. Besides the advantage of cost-effectiveness and convenience, data obtained by multi-stage sampling allow researchers to answer multilevel research questions. For example, researchers could examine how students' achievement is related to individual student socioeconomic status (SES) on average, how this association varies across schools, how school socioeconomic composition (i.e., school SES) affects student achievement, and how school SES affects the association between student achievement and their SES. One challenge in analyzing complex survey data is the non-independence of observations (or clustering effect) because individuals in the same cluster usually share the same environment and tend to be more alike. Another challenge arises when there are unequal selection probabilities at one or more stages of the sampling process, which is often the case due to the necessity of oversampling certain underrepresented groups or accounting for non-response.

To answer multilevel research questions and to handle the nested data structures, multilevel modeling (MLM) is frequently used. MLM allows researchers to decompose the variance into the between-cluster and within-cluster components and investigate the variability of within-cluster effects across clusters. For example, using MLM researchers could examine not only the average association between individual student achievement and their SES, but also how this association may vary across schools. Established estimation methods for MLM include maximum likelihood (ML) and iterative generalized least squares (IGLS), which are equivalent under normality (Goldstein, 1986). When there are unequal selection probabilities in the stage of selecting schools and/or the stage of selecting students within schools, in order to obtain accurate estimate of the mean outcome and/or the average association between a predictor and the outcome in the population of students, methods were developed to incorporate sampling weights in estimation, such as multilevel pseudo-maximum-likelihood (MPML)(e.g., Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006) and probability-weighted IGLS (PWIGLS; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). It has been shown that PWIGLS could result in biased standard error estimates for weighted multilevel data (Asparouhov, 2005). Hence we only considered MPML in our study.

MPML has two crucial underlying assumptions. First, it assumes that the sample size is sufficiently large at both the within-cluster (e.g., number of students per school) and the between-cluster level (e.g., number of schools), especially the latter. In practical research, even if it is possible to obtain a large number of clusters, the sample size within each cluster is often small. To reduce bias in the estimates of the standard errors of fixed effects and the estimates of variance components due to small cluster sizes, scaling of level-1 weights has been used as the major tool. However, the performances of the various scaling methods depend on a host of factors such as cluster size, intraclass correlation (ICC), the degree of informativeness of the selection mechanism, and so forth (Asparouhov, 2006). Applied researchers should select the appropriate scaling method based on the specific sampling design of a study, which could be challenging due to the lack of information. Second, MPML assumes that the error term and random effects follow a distribution of a specified class. In multilevel models, each level has its own error term and random effects; therefore the distributional assumptions should be met at each level. For example, in a two-level linear model, the level-1 errors are assumed to follow a univariate normal distribution, and the level-2 random effects are assumed to follow a multivariate normal distribution. It has been documented that although ML estimators for fixed effects and variance components are consistent even when the random-effects distribution is not normal, the standard error estimated by the inverse Fisher information matrix may be biased, especially for variance components (Verbeke & Lesaffre, 1997). The more sophisticated Huber-White robust standard errors are more accurate for the variance component estimates, but require at least 100 clusters (Maas & Hox, 2004). To our knowledge, the performance of MPML with robust standard errors under distributional misspecification has not been studied yet.

Bootstrap resampling methods for multilevel data have been developed as an alternative to ML estimation in the case where the general assumptions mentioned above are violated. In general, there are three main approaches to bootstrap: (1) the parametric bootstrap, (2) the nonparametric residual bootstrap, and (3) the case bootstrap. The parametric bootstrap has the strongest assumptions, which require that the specifications of the functional form and the distributions of the residuals are both correct. The residual bootstrap only requires the correct specification of the functional form. Finally, the case bootstrap has minimum assumptions and only requires the hierarchical structure to be correctly specified. Van der Leeden, Meijer, and Busing (2008) provided a detailed discussion of the systematic development of bootstrap resampling methods for multilevel models. It has been shown that bootstrap methods could provide accurate confidence intervals for fixed effect estimates when the distribution of the residuals are highly skewed at all levels (Carpenter, Goldstein, & Rasbash, 2003). In addition, applications to small area estimation showed that the bootstrap method could produce sensible estimates for standard errors for shrinkage estimates of small area means based on generalized linear mixed models (e.g., Booth, 1995; Hall & Maiti, 2006; Lahiri, 2003).

Given the advantages of multilevel bootstrap resampling under conditions with distributional assumption violation and small sample sizes, it is useful to extend the method to accommodate multilevel data with sampling weights. Research in this area is limited and existing methods only use the case bootstrap approach (Grilli & Pratesi, 2004; Kovacevic, Huang, & You, 2006; Wang & Thompson, 2012) . Although the case bootstrap is more robust to assumption violations than residual bootstrap, it is typically less efficient. Some studies have shown that case bootstrap performed worse than residual bootstrap even when the assumptions were violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). Hence the purpose of this paper is to propose a weighted nonparametric residual bootstrap procedure for multilevel modeling with sampling weights. The proposed procedure is an extension of the nonparametric residual bootstrap procedure developed by Carpenter et al. (2003). With a Monte Carlo simulation, we examined the performance of the proposed bootstrap method in terms of parameter estimates and statistical inferences under a variety of conditions.

The outline of the paper is as follows. First, we briefly discuss sampling weights for multilevel models, followed by a review of existing bootstrap methods for multilevel data. Next, we provide details of the proposed procedure followed by a demonstration of the method using real data. Then we present the simulation study to examine the performance of the proposed bootstrap method. Finally, the findings are summarized and discussed.

2 Sampling Weights and Pseudo-Maximum-Likelihood Estimation for Multilevel Models

Multilevel data are often collected using a multi-stage sampling design which involves sampling clusters in the first stage and then sampling units within selected clusters in the subsequent stages. Due to the clustering, observations in multilevel data often have some degree of dependence among them, which makes the traditional methods based on a simple random sample design inappropriate. Therefore, MLM is often used to account for the dependency among the observations. More importantly, MLM not only allows researchers to examine the average association between a predictor and an outcome, but also to address questions on how the associations among variables within clusters vary across clusters, such as how the association between individual student achievement and their SES varies across schools. In this section, we consider a two-level model with students nested within schools to provide a background for sampling weights in multilevel models.

Let Y_{ij} be the achievement scores, $\mathbf{X}_{ij}$ be the scores on the level-1 predictors (e.g., individual student SES, gender, etc.) associated with student $i(i = 1, \ldots, n_j)$ within school $j(j = 1, \ldots, J)$, and $\mathbf{X}_j$ be the scores on the level-2 predictors (e.g., school SES, school sector, etc.) associated with school j. A two-level model can be specified as

$$Y_{ij} = \boldsymbol{\beta}_1\mathbf{X}_{ij} + \boldsymbol{\beta}_2\mathbf{X}_j + \boldsymbol{\mu}_j\mathbf{Z}_{ij} + \varepsilon_{ij} \tag{1}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are row vectors of regression coefficients associated with student-level and school-level predictors respectively, which represent the average effects of the predictors in the population of students. The row vector $\boldsymbol{\mu}_j$ contains random effects associated with school j, which could be a random intercept, or a random slope of a student-level predictor, or both. The design vector $\mathbf{Z}_{ij}$ usually includes the constant 1 (for the random intercept) and the student-level predictors that have random slopes across schools. Finally, ε_{ij} is the level-1 error. The main parameters of interest in MLM are usually the fixed effects (i.e., $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$) and the variance and covariance components (i.e., the variances and covariances of the random effects $\boldsymbol{\mu}_j$). The conventional maximum likelihood estimates of the parameters are obtained by maximizing the likelihood function $L(\theta) = \prod_{j=1}^{J}[\int \prod_{i=1}^{n_j} f(Y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1)q(\boldsymbol{\mu}_j|\mathbf{X}_j, \boldsymbol{\beta}_2)d\boldsymbol{\mu}_j]$ where $f\left(Y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1\right)$ is the density function of Y_{ij} and $q(\boldsymbol{\mu}_j|\mathbf{X}_j, \boldsymbol{\beta}_2)$ is the density function of $\boldsymbol{\mu}_j$.

Suppose that schools and students within schools are selected with unequal probabilities. Let the probability of selecting school j be p_j and the probability of selecting student i given that school j is sampled be $p_{i|j}$. The sampling weight for school j is $w_j = 1/p_j$. The conditional sampling weight for student i within school j is $w_{i|j} = 1/p_{i|j}$. The unconditional sampling weight for an individual student is $w_{ij} = w_j \times w_{i|j}$. If the sampling weights are related to the dependent variable after conditioning on the covariates in the model, they are called informative weights (Pfeffermann, 1993). For example, if students with lower achievement have a higher probability of being sampled controlling for the predictors $\mathbf{X}_{ij}$ and $\mathbf{X}_j$, then the sampling weights are informative. Informative sampling weights should be incorporated in statistical inferences to avoid bias in estimates or poor performance of test statistics and confidence intervals. For multilevel models, the sampling weights at each level need to be taken into account when they are informative, to ensure that the average association between the predictors and the outcome in the population of students as well as the variance and covariance components of school random effects can be accurately estimated. One approach to incorporate the sampling weights is to use multilevel pseudo maximum likelihood estimation (MPML), which defines the likelihood function as $l\left(\theta\right) = \prod_{j=1}^{J}(\int \prod_{i=1}^{n_j} f\left(Y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\mu}_j, \boldsymbol{\beta}_1\right)^{w_{i|j}} q(\boldsymbol{\mu}_j|\mathbf{X}_j, \boldsymbol{\beta}_2)d\boldsymbol{\mu}_j)^{w_j}$.

Extant literature has shown that the level-1 weights should be scaled in order to reduce the bias of variance component estimates and standard error estimates of fixed effects when cluster sizes are not large (e.g., Pfeffermann et al., 1998; Potthoff, Woodbury, & Manton, 1992; Stapleton, 2002). There are two commonly used scaling methods: relative vs. effective sample size scaling. In relative sample size rescaling, the level-1 weights $w_{i|j}$ are multiplied by a scaling factor $s_{1j} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}$ so that the sum of the rescaled level-1 weights within a cluster equals the actual cluster size. In effective sample size rescaling, the scaling factor $s_{1j} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} w_{i|j}^2}$ is used such that the sum of the rescaled level-1 weights within a cluster equals the effective cluster size which is defined as $\frac{\left(\sum_{i=1}^{n_j} w_{i|j}\right)^2}{\sum_{i=1}^{n_j} w_{i|j}^2}$. Some simulation studies showed that relative sample size rescaling works better for informative weights, whereas effective sample size rescaling is more appropriate for non-informative weights (Pfeffermann et al., 1998). Some researchers argue that non-informative weights should not be used in multilevel analyses because they tend to result in a loss of efficiency and even bias in parameter estimates under some conditions. For example, Asparouhov (2006) found bias in the estimation of multilevel models when cluster sample size is small and non-informative within-cluster weights are used.

However, in practical applications, choosing the right scaling method may be challenging. Pfeffermann (1993) described a general method for testing the informativeness of the weights. Asparouhov (2006) proposed a simpler method based on the informative index, and recommended to consider both the value of the informative index and Pfeffermann's test, the invariance of selection mechanism across clusters, and the average cluster size when determining weighting in multilevel modeling.

3 Bootstrap for Multilevel Data

Depending on whether and what parametric assumptions are involved, there are multiple approaches to do bootstrapping (Davison & Hinkley, 1997), and additional care is needed to address the dependencies in the data when resampling with multilevel data (Van der Leeden et al., 2008). Below we first provide a brief summary of the common bootstrap procedures for multilevel data in general (i.e., the parametric bootstrap, the residual bootstrap, and the case bootstrap) and then focus on the bootstrap method for multilevel data with sampling weights. Readers should consult Davison and Hinkley (1997), Goldstein (2011), and Van der Leeden et al. (2008) for more detailed reviews of the statistical theory of multilevel bootstrapping methods.

3.1 Parametric Bootstrap

As described in Goldstein (2011), with parametric bootstrap, researchers first fit a multilevel model to obtain fixed effect estimates, and the random effect variance estimates, $\hat{\boldsymbol{\tau}}$ and $\hat{\sigma}$. Then, for each bootstrap sample, a new set of N level-1 errors, ε^*_{ij}, and a new set of J level-2 random effects, $\boldsymbol{\mu}^*_j$, are drawn from independent $N(0, \hat{\boldsymbol{\tau}})$ and $N(0, \hat{\sigma})$ distributions to form a new set of responses, y^*_{ij}. The multilevel model is then refitted to the new bootstrap data, and the target statistics (e.g., fixed effects) are computed. The resampling process is repeated for a large number of B bootstrap samples (e.g., $B = 1,999$) to obtain bootstrap sampling distributions of the target statistics.

3.2 Non-parametric Residual Bootstrap

The (nonparametric) residual bootstrap is similar to the parametric bootstrap except that, when forming new responses, the new errors and random effects were obtained by sampling with replacement the residuals of the multilevel fitted model. In this paper, the resampled residuals were denoted as $\tilde{\boldsymbol{\mu}}_j$ and $\tilde{\varepsilon}_{ij}$ to distinguish them from the counterparts in the parametric bootstrap. In addition, because the sampling variance of $\tilde{\boldsymbol{\mu}}_j$ is generally smaller than $\hat{\boldsymbol{\tau}}$, and so is the sampling variance of $\tilde{\varepsilon}_{ij}$ smaller than $\hat{\sigma}$ (albeit to a lesser extent). Carpenter et al. (2003) and Goldstein (2011) recommended to first “reflate” the residuals so that the sample variances of the reflated residuals were exactly $\hat{\boldsymbol{\tau}}$ and $\hat{\sigma}$, respectively. Finally, as in parametric bootstrap, a new set of response $\tilde{y}_{ij}$ is formed, and the target statistics are computed, and then the process is repeated B times to obtain a bootstrap sampling distribution of the target statistics.

3.3 Case Bootstrap

With the case bootstrap, each bootstrap sample consisted of observations (i.e., “cases”) sampled with replacement from the original data. When there are two levels in the data so that a case can mean a cluster or a unit within a cluster, there are two variants of the case bootstrap (Davison & Hinkley, 1997): (a) to resample with replacement intact clusters but no resampling within a cluster, and (b) to first resample the clusters, and within each cluster resample with replacement the units. Both Davison and Hinkley (1997) and Goldstein (2011) recommended (a) over (b).

A few previous studies have examined these three bootstrap methods for multilevel analyses. Seco, García, García, and Rojas (2013) showed that the residual bootstrap produced more precise estimates, in terms of smaller root mean squared errors, for fixed effects than restricted maximum likelihood. On the other hand, because the case bootstrap makes fewer assumptions than the parametric and the residual bootstraps, it requires more information from the data. As such, previous literature found that its performance was poor compared to the other two methods, even when the assumptions for the latter two methods were violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). On the other hand, Thai, Mentré, Holford, Veyrat-Follet, and Comets (2014) found that in longitudinal linear-mixed models where cluster size is constant, residual bootstrap and case bootstrap performed similarly when there were at least 100 individuals (i.e., $J = 100$).

3.4 Bootstrap for Multilevel Data with Sampling Weights

For multilevel data with sampling weights, the extant literature documents two types of bootstrap methods, both of which can be viewed as modifications to case bootstrap. One type involves generating a pseudo (or artificial) population that mimics the population from which the original sample is selected, and then selecting bootstrap samples from the pseudo population based on the sampling weights in the original sample (Grilli & Pratesi, 2004; Wang & Thompson, 2012). As described in Grilli and Pratesi (2004), when generating the pseudo population, the ith unit ($i = 1, \ldots, n_j$) in the jth cluster ($j = 1, \ldots, J$) is duplicated $w_{i|j}$ times, rounding the weight to the nearest integer to form J artificial clusters. Then each of the J artificial clusters is replicated w_j times, rounding the weight to the nearest integer, to obtain the artificial population. From the artificial population, bootstrap samples are obtained by first selecting J clusters with probability proportional to $1/w_j$ and then selecting n_j units with probability proportional to $1/w_{i|j}$ from the jth resampled cluster. Wang and Thompson (2012)'s procedure is similar except that they added an additional step to account for the potential biases caused by rounding the weights when generating the pseudo population.

The other type of bootstrap for multilevel data with sampling weights involves a two-stage resampling and rescaling of weights at each level. As described in Kovacevic et al. (2006), $J - 1$ clusters are first drawn from the original sample using simple random sampling with replacement (SRSWR). Then w_j is rescaled to obtain the cluster bootstrap weights $w_j^* = w_j \frac{J}{J-1} t_j$ where t_j is the number of times that cluster j is included in the bootstrap sample. From each resampled cluster, $n_j - 1$ units are drawn using SRSWR and the unadjusted conditional bootstrap weights are calculated for level-1 units as $b^*_{i|j} = w_{i|j}\left(\frac{n_j}{n_j-1}\right)\left(\frac{t_{i|j}}{t_j}\right)$ where $t_{i|j}$ is the total number of times that the ith unit is resampled. Based on the rescaled cluster bootstrap weights and the unadjusted conditional bootstrap weights, the unadjusted unconditional bootstrap weights are computed as $b^*_{ij} = b^*_{i|j}w^*_j$. The adjusted unconditional bootstrap weights (w^*_{ij}) are obtained after applying all the same adjustments done in the process of calculating the original full sample unconditional weights. If no adjustment is made, then $w^*_{ij} = b^*_{ij}$. Finally, the within-cluster conditional weights are calculated as $w^*_{i|j} = w^*_{ij}/w^*_j$.

Both Grilli and Pratesi (2004) and Kovacevic et al. (2006) noted that the steps concerning the level-1 units in their procedures can be omitted when the sampling fraction is low at the cluster level. Kovacevic et al. (2006) also showed that the accuracy and stability of variance estimation improved when using the relative within-cluster weights (i.e., the sum of the rescaled level-1 weights within a cluster equals the actual cluster size) as compared to the original unscaled within-cluster weights. However, to the best of our knowledge, these methods have not been developed into statistical packages that can be easily accessed by applied researchers.

4 The Proposed Weighted Residual Bootstrap

4.1 Algorithm

The weighted residual bootstrap method was developed based on an idea similar to the one outlined in Goldstein, Carpenter, and Kenward (2018). Without loss of generality, we present the weighted nonparametric residual bootstrap algorithm for a two-level model. An extension to a model with more levels is straightforward.

Step 1: Obtain parameter estimates for model 1 (i.e., $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$) based on sample data using unweighted maximum likelihood and restricted maximum likelihood, and compute level-1 residuals ε_{ij} and level-2 residuals $\boldsymbol{\mu}_j$.

Step 2: Obtain reflated level-1 and level-2 residuals ($\varepsilon_{ij}^{'}$and $\boldsymbol{\mu}_j^{'}$) using Carpenter et al. (2003)'s procedure.

Step 3: Sample independently with replacement from the set of reflated level-1 residuals using level-1 unconditional weights and from the set of reflated level-2 residuals using level-2 weights, obtaining two new sets of residuals $\varepsilon_{ij}^{'b}$ and $\boldsymbol{\mu}_j^{'b}$, where b is the index of bootstrap samples. It is noted that the level-1 unconditional weights are used instead of the conditional weights to resample level-1 residuals, because the new set of level-1 residuals are selected from the entire sample across clusters rather than within clusters. This approach makes it unnecessary to scale the within-cluster weights.

Step 4: The new response of the bth bootstrap sample is then obtained by $Y_{ij}^{'b} = \hat{\boldsymbol{\beta}}_1\boldsymbol{X}_{ij} + \hat{\boldsymbol{\beta}}_2\boldsymbol{X}_j + \boldsymbol{\mu}_j^{'b}\boldsymbol{Z}_{ij} + \varepsilon_{ij}^{'b}$.

Step 5: Refit the model to the bootstrap sample to obtain one set of bootstrap parameter estimates using either unweighted maximum likelihood or restricted maximum likelihood.

Step 6: Repeat steps 2-5 to obtain B set sets of bootstrap parameter estimates.

4.2 Illustration

As a demonstration, we applied the proposed procedure to examine the associations between student math achievement and student gender and school SES among 15-year-old students in the United States using the 2000 PISA data Organization for Economic Co-operation and Development (2000) . PISA used a cluster sampling design with unequal selection probabilities. Specifically, schools with more than 15% of minority students were oversampled, and minority students were oversampled within those schools. The data include weights at the school level (named WNRSCHBW) and unconditional weights at the student level (named W_FSTUWT). We used a two-level random intercept model with students' math test scores (Y_{ij}) as the dependent variable, student gender ($Gender_{ij} = 0$ for females and 1 for males) and school mean ISEI (*ISEI_m*) as the school-level predictor (Equation 2),

$$Y_{ij} = \beta_0 + \beta_1 Gender_{ij} + \beta_2 ISEI\_m_j + u_{0j} + e_{ij} \tag{2}$$

where i indexes students and j indexes schools, u_{0j} represents random effects associated with the intercept. The main parameters of interest are the average effects of gender (β_1) and school SES (β_2) on students' math achievement in the population of 15-year-old students in the United States. Although we used a random intercept model in this demonstration, researchers could further examine whether the association between student gender and achievement varies across schools by adding a random effect associated with the slope of gender that varies across schools (i.e., a random slope model).

The US sample consists of 2135 students from 145 schools. 74% students had complete data on both *ISEI* and *Math* while 26% had at least one missing value on the two variables. After removing cases with missing data, the final sample of analysis consists of 1578 students from 145 schools. The cluster size ranged from 1 to 20, with the first quartile of 8, median of 12, and the third quartile of 14. To determine the degree to which the weights were informative, we followed the recommendation by Asparouhov (2006) and computed the informative index by $|\widehat{\mu_w} - \widehat{\mu_0}| / \sqrt{\upsilon_0}$ where $\widehat{\mu_w}$ is the weighted mean of the dependent variable, $\widehat{\mu_0}$ is the unweighted mean, and υ_0 is the unweighted variance. The informative index for math was 0.03, indicating that the sampling weights were very slightly informative.

The bootstrap estimates were obtained using researcher developed R package *bootmlm* (see Appendix for the R code). As a comparison, the model was also estimated using unweighted ML, and MPML with relative and effective weights respectively. The MPML estimates were obtained using M*plus* 8.2 Muthén and Muthén (1998, see Appendix B for the Mplus code). The ML estimates were obtained using the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015). Percentile confidence intervals were computed in the bootstrap method (i.e., $\alpha/2$

and $1-\alpha/2$ quantiles of the bootstrap distribution), profile likelihood confidence intervals were computed in *lme4* for the ML estimates, and the delta method[3] was used to construct approximate confidence intervals for the MPML variance component estimates. The MPML results based on relative weights were almost identical to those based on effective weights, thus we only reported the latter.

Table 1. ML, MPML, and Bootstrap Results Based on the PISA Data

		Estimate	SE	95% CI
Unweighted ML	Intercept	74.33	2.49	[69.45, 79.20]
	Gender	-1.6	0.66	[-2.88, -0.31]
	ISEI_m	0.16	0.05	[0.06, 0.26]
	Variance			
	School	9.43	3.02	[4.35, 16.48]
	Residual	162.4	6.06	[151.07, 174.87]
	Conditional ICC	0.06		
MPML Effective Weights	Intercept	80.42	5.52	[69.59, 91.24]
	Gender	-2.43	1.16	[-4.70, -0.16]
	ISEI_m	0.06	0.12	[-0.17, 0.28]
	Variance			
	School	10.86	9.03	[2.12, 55.41]
	Residual	152.47	24.3	[111.56, 208.38]
	Conditional ICC	0.07		
Bootstrap	Intercept	74.94	2.51	[70.17, 80.18]
	Gender	-1.56	0.67	[-2.85, -0.17]
	ISEI_m	0.16	0.05	[0.05, 0.26]
	Variance			
	School	7.42	2.68	[2.23, 13.02]
	Residual	162.51	9.95	[144.5, 184.0]
	Conditional ICC	0.04		

Before looking at the parameter estimates, we examined the distribution of the residuals. The level-1 residuals based on the ML estimates were slightly non-normal with skewness of -1.45 and kurtosis of 6.77. The distribution of the level-2 residuals was close to normal with skewness of -0.46 and kurtosis of 3.39. Table 1 shows the parameter estimates, standard error estimates, 95% confidence intervals, and conditional ICCs. There was little difference between the ML estimates and the bootstrap estimates. However, the MPML results showed different point estimates and standard error estimates, especially for the slope of school mean ISEI (i.e., ISEI_m). As a result, the statistical inference also reached different conclusions regarding the slope of school mean ISEI, which

[3] The 1- α confidence interval of a variance component θ is given by $\exp\left[\ln\left(\hat{\theta}\right) \pm z_{1-\frac{\alpha}{2}}\frac{\sqrt{Var\left(\hat{\theta}\right)}}{\hat{\theta}}\right]$ where $\hat{\theta}$ is the MPML estimate of θ, $Var\left(\hat{\theta}\right)$ is the asymptotic variance of $\hat{\theta}$.

was statistically significant based on the ML and the bootstrap results, but nonsignificant based on MPML.

From this particular sample and model, we obtained inconsistent results from the bootstrap and the MPML methods. We suspected that the MPML results might not be trustworthy because the specific condition of this sample (i.e., small cluster size, low ICC, and very slight informativeness) has been shown to be unfavorable to MPML (e.g., Asparouhov, 2006). However, it is unknown whether the performance of the bootstrap method is acceptable, thus a Monte Carlo simulation is needed to assess the performance of these methods under various conditions.

5 Simulation

5.1 Data Generation

To evaluate the performance of the weighted bootstrap procedure in accounting for nonrandom sampling, we used R 3.5.0 (R Core Team, 2018) to simulate two-level data mimicking the data structure of students nested in schools. The population models were either (a) a random intercept model or (b) a random slopes model. The models include one level-1 predictor such as student SES (denoted as $X1_{ij}$) and one level-2 predictor such as school SES (denoted as $X2_j$). Because multilevel modeling is a model-based technique usually justified by a superpopulation model (Cochran, 1977; Lohr, 2010), the data generating model is treated as the superpopulation, and in each replication, we first generated a finite population with $J_{pop} = 500$ clusters and $n_{pop} = 100$ observations for each cluster.

When generating a finite population based on the random intercept model (see Equation 2), we simulated $X2_j$ from $N(0,\,1)$ distributions and the cluster-level random intercept effect u_{0j} from either normal distributions or scaled $\chi^2(df = 2)$ distributions with mean 0 and variance τ, depending on the simulation condition described in the next section. We then simulated $n_{pop} \times J_{pop}$ values of $X1_{ij}$ from $N(0,\,1)$ distributions and e_{ij} from either normal distributions or scaled $\chi^2(df = 2)$ distributions with mean 0 and variance σ, depending on the simulation condition. For all simulation conditions, we set β_0= 0.5, $\beta_1 = \beta_2 = 1$, and the total variance $\tau+\sigma = 2.5$. The outcome was computed based on Equation (2).

When generating a finite population based on the random slopes model, the following equation was used

$$Y_{ij} = \beta_0 + \beta_1 X1_{ij} + \beta_2 X2_j + u_{0j} + u_{1j} X1_{ij} + e_{ij} \tag{3}$$

where u_{0j} and u_{1j} represent the random effects associated with the intercept and the slope of $X1_{ij}$ respectively. We simulated u_{0j} and u_{1j} from a bivariate normal distribution with mean of 0 and variance-covariance of $\begin{bmatrix} \tau_{00} & \\ \tau_{01} & \tau_{11} \end{bmatrix}$ in which τ_{00} represents the variance of the random intercept, τ_{11} the variance of the random slope of $X1_{ij}$, and τ_{01} the covariance between the random intercept and the random slope. The magnitude of τ_{00} depends on the simulation condition, and the magnitude of τ_{11} is half of τ_{00} because the variance of random slopes is typically smaller than the variance of random intercepts. The covariance τ_{01} is computed as $\rho\sqrt{\tau_{00}\tau_{11}}$ where ρ denotes the correlation between the random intercepts and the random slopes and was set at 0.5 to represent a moderate correlation.

After simulating the finite populations, we first sampled J clusters with a sampling fraction f according to a certain selection mechanism depending on the simulation condition. Then in each cluster we randomly sampled n observations with the same sampling fraction f according to a certain selection mechanism depending on the simulation condition.

5.2 Design Factors

We considered 5 design factors to generate a variety of experimental conditions. First, the variance of the random intercepts: 0.125, 0.5, and 1.25. They correspond to small, medium, and large conditional ICCs (i.e., ICC = 0.05, 0.2, and 0.5) commonly seen in multilevel data. Second, sampling fraction (f): 0.1 and 0.5. Similar levels were used in previous simulations such as 0.12 in Grilli and Pratesi (2004) and 0.6 in Rabe-Hesketh and Skrondal (2006). Under the 0.1 sampling fraction condition, the cluster size was 10 and the number of clusters was 50. This was considered a small sample size condition. Under the 0.5 sampling fraction condition, the cluster size was 50 and the number of clusters was 250, which was considered a large sample size. Third, normality of random effects. For the random intercept model, we considered the normal distribution vs. the scaled $\chi^2(df = 2)$ distribution for the random effects and the level-1 errors. The $\chi^2(df = 2)$ distribution has skewness $= \sqrt{8/2} = 2$ and kurtosis $= 12/2 = 6$. For the random slopes model, we only considered normal distribution.

Fourth, between-cluster selection mechanism: non-informative vs. informative. For non-informative selection, simple random sampling (SRS) was used. For the random intercept model with informative sampling, we first divided the clusters into two strata: $\mu_{0j} > 0$ (stratum 1) and $\mu_{0j} < 0$ (stratum 2), and then sampled without replacement in each stratum such that the sampling probability of each cluster is $1.4f$ for stratum 1 and $0.6f$ for stratum 2. In other words, it was expected that for each replication, 70% of the sampled units came from stratum 1, and 30% of the sampled units came from stratum 2. For the random slopes model with informative sampling, we divided the clusters into four strata: $\mu_{0j} > 0$ and $\mu_{1j} > 0$ (stratum 1), $\mu_{0j} > 0$ and $\mu_{1j} < 0$ (stratum 2), $\mu_{0j} < 0$ and $\mu_{1j} > 0$ (stratum 3), and $\mu_{0j} < 0$ and $\mu_{1j} < 0$ (stratum 4), with sampling probabilities of 1.96*f*, 0.84*f*, 0.84*f*, and 0.36*f*, respectively. It was expected that for each replication, 49% of the sampled units came from stratum 1, 21% from stratum 2, 21% from stratum 3, and 9% from stratum 4.

Finally, within-cluster selection mechanism: non-informative vs. informative. For non-informative selection, within-cluster units were sampled using SRS. For informative selection, units in each cluster were first divided into two strata: $e_{ij} >$

0 (stratum 1) and $e_{ij} < 0$ (stratum 2), and then sampled without replacement according to the 7:3 ratio of sampling probability. The informative index was about 0.17 when informative selection occurred at level-1 only, 0.09 when at level-2 only, and 0.27 when at both levels based on the random intercept models. These values represent slight to moderate informativeness according to Asparouhov (2006).

Combining the five design factors, there are a total of 48 data conditions (3 ICCs × 2 sampling fractions × 2 distributions × 2 between-cluster selection mechanisms × 2 within-cluster selection mechanisms) for the random intercept models and 24 conditions (3 ICCs × 2 sampling fractions × 2 between-cluster selection mechanisms × 2 within-cluster selection mechanisms) for the random slopes models. We conducted 500 replications for each simulation condition. For each generated data set, three estimators were applied: the proposed bootstrap method (using the R package *bootmlm*), MPML with effective weights (using M*plus* 8.2 for the random intercept models and Stata 16 for the random slopes models), and unweighted maximum likelihood (using the R package *lme4*).

5.3 Analysis

For each parameter in the models (including both fixed effects and variance components), we examined the relative bias of the point estimate and the coverage rate of the 95% confidence intervals. For the bootstrap method, we used the 2.5 and 97.5 percentile of the empirical sampling distribution as the lower and upper boundaries of the 95% confidence interval. Following Hoogland and Boomsma (1998), relative biases of point estimates are considered acceptable if their magnitudes are less than 0.05. The coverage rate of a 95% confidence interval should be approximately equal to 95%, with a margin of error of 1.9% based on 500 replications. Hence coverage rates between 93% and 97% are acceptable.

5.4 Results

5.4.1 Random intercept models Tables 2 to 5 show the relative bias and coverage rate for parameter estimates under all conditions based on the random intercept models. The relative biases for the slope of the level-1 predictor *X1* and the slope of the level-2 predictor *X2* are not shown in the tables because they were close to zero for all conditions. In addition, the coverage rate for the slope of *X1* was close to 95% under all conditions, therefore it was not included in the tables.

Intercept. As shown by the relative biases of the ML estimates, ignoring sampling weights when the selection mechanism was informative caused moderate to large relative biases, ranging from 0.14 to 1.38 (see Table 2 and 3). As a result of biased point estimate, the coverage rates of the confidence intervals for the ML estimates were also poor under those conditions ranging from 0.00 to 0.85 (see Table 4 and 5).

MPML successfully reduced the relative biases to an acceptable level under the majority of conditions, however, there were still small to moderate relative biases under 11 conditions where the sample size was small and the selection mechanism was informative at level 1 or both levels (relative bias ranging from 0.07 to 0.13). As a result, there was slight under-coverage (ranging from 0.88 to 0.92) in about half of those conditions (6 out of 11), mainly when there was informative selection at both levels.

The bootstrap method performed the best in terms of relative biases because they were below 0.05 under all conditions. However, the advantage of the bootstrap method over MPML was less obvious in terms of the coverage rate because the bootstrap method also had slightly low coverage rate (ranging from 0.88 to 0.92) under similar conditions.

Slope of *X2*. The relative bias of the estimated slope of *X2* was acceptable for all methods under all conditions. However, the MPML confidence intervals suffered from slight under-coverage (89%-92%) in 18 conditions, mainly when sample size was small and selection was informative at level 2 or both levels.

Variance component of the random intercepts (τ). ML estimates had small relative biases under 18 conditions when there was informative sampling at level-2 or at both levels. The biases were negative ranging from -0.07 to -0.11 when the distribution was normal, and were positive ranging from 0.10 to 0.12 when the distribution was skewed. MPML suffered from small to moderate biases (-0.10 to 0.27) under 10 conditions when small sample size was combined with small to moderate ICCs. It was noted that the two moderately large relative biases (i.e., 0.25 and 0.27) both occurred when there was informative selection at level-1 or at both levels. The bootstrap method performed better with only small positive biases (0.08 to 0.11) under 5 conditions where both ICC and sample size were small. It was noted that out of the 5 conditions where relative biases were obvious, one was under the normal distribution and four under the skewed distribution, indicating that the performance of the bootstrap method might be sensitive to skewed distributions.

In general, all three methods tended to have under-coverage, with ML being the worst and bootstrap being the best. Where the distribution was normal, 15 conditions had under-coverage ranging from 0.87 to 0.92 for ML, 14 conditions ranging from 0.86 to 0.92 for MPML, and 11 conditions ranging from 0.89 to 0.92 for bootstrap. When data were skewed, 23 conditions had under-coverage ranging from 0.67 to 0.92 for ML, 22 conditions ranging from 0.76 to 0.92 for MPML, and 15 conditions ranging from 0.81 to 0.92 for bootstrap. For both MPML and bootstrap, the coverage rate tended to worsen as the sample size decreased. In addition, when data were skewed, larger ICCs led to lower coverage rate for MPML.

Level-1 residual variance (σ). Only ML estimates had small negative relative biases when there was informative selection at level-1 or at both levels. As a result, ML estimates had severe under-coverage under those conditions, especially when sample size was large. The performance of ML deteriorated when the distribution was skewed as there were severe under-coverage across all conditions.

Although MPML and bootstrap estimates had minimum relative biases, both had slight under-coverage under certain conditions. Specifically, when the distribution was normal, under-coverage mainly occurred when sample size was small combined with informative selection at both levels. When the distribution was skewed, under-coverage mainly occurred when sample size was small and when the selection was non-informative or only informative at level-2.

Table 2. Relative Bias for the Random Intercept Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	-0.01	-0.01	-0.01	0.03	**0.07**	**-0.07**	0.00	0.00	0.00
		0.5	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.93**	0.01	**0.10**	-0.04	0.00	**0.25**	**-0.08**	0.00	-0.02
		0.5	**0.70**	0.00	0.01	0.01	0.01	0.02	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.22**	-0.04	0.01	**-0.07**	0.00	**-0.10**	0.01	0.01	0.00
		0.5	**0.16**	-0.02	0.00	-0.05	-0.01	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.15**	-0.02	**0.12**	-0.10	-0.03	**0.27**	**-0.09**	0.00	-0.02
		0.5	**0.86**	-0.02	0.01	-0.05	-0.01	0.01	-0.05	0.00	0.00
0.2	Non-informative	0.1	-0.01	-0.01	-0.01	0.00	0.01	-0.05	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.85**	0.02	**0.09**	-0.01	-0.01	0.02	**-0.08**	0.00	-0.02
		0.5	**0.63**	-0.01	0.00	0.00	0.00	0.00	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.44**	-0.04	0.03	**-0.09**	-0.02	**-0.06**	0.01	0.01	0.00
		0.5	**0.32**	-0.01	0.00	-0.05	0.00	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.30**	-0.01	**0.13**	**-0.11**	-0.04	0.01	**-0.09**	0.00	-0.02
		0.5	**0.97**	-0.01	0.01	-0.05	-0.01	-0.01	-0.05	0.00	0.00
0.5	Non-informative	0.1	-0.01	-0.01	-0.01	0.00	0.01	-0.04	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	0.00	0.00	-0.01	0.00	0.00	0.00
	Informative at level-1	0.1	**0.67**	0.02	**0.07**	0.00	0.00	-0.03	**-0.08**	0.00	-0.02
		0.5	**0.49**	-0.01	-0.01	0.00	0.00	0.00	-0.05	0.00	0.00
	Informative at level-2	0.1	**0.70**	0.00	0.05	**-0.09**	-0.01	-0.05	0.01	0.01	0.00
		0.5	**0.51**	0.00	0.00	-0.05	0.00	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.38**	0.03	**0.13**	**-0.10**	-0.02	-0.04	**-0.09**	-0.01	-0.02
		0.5	**1.02**	0.00	0.01	-0.05	0.00	-0.01	-0.05	0.00	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

5.4.2 Random slopes models Tables 6 to 9 show the relative biases and coverage rates for parameter estimates under all conditions based on the random slopes models. Notably, while convergence was not an issue for ML and the bootstrap method, MPML estimation suffered from a low convergence rate (ranging between 0.59 and 0.76) when both ICC and sample size were small.

Intercept. Similar to the pattern under the random intercept models, ML estimates of the intercept suffered from moderate to large relative biases (ranging

Table 3. Relative Bias for the Random Intercept Model Under $\chi^2(2)$ Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	.00	.00	.00	0.03	**0.07**	**-0.07**	0.00	0.00	0.00
		0.5	.00	.00	.00	0.00	0.00	-0.02	0.00	0.00	0.00
	Informative at level-1	0.1	**0.81**	0.01	**0.08**	0.04	**0.09**	-0.05	**0.12**	0.01	0.02
		0.5	**0.60**	0.00	0.00	0.00	0.00	-0.04	**0.10**	0.00	0.00
	Informative at level-2	0.1	**0.18**	-0.04	0.01	**0.12**	**0.11**	**-0.09**	-0.01	-0.01	-0.01
		0.5	**0.14**	-0.02	0.00	**0.10**	0.02	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.00**	-0.02	**0.10**	**0.11**	**0.10**	**-0.08**	**0.11**	0.00	0.01
		0.5	**0.75**	-0.02	0.01	**0.10**	0.03	-0.04	**0.10**	0.00	0.00
0.2	Non-informative	0.1	-0.01	-0.01	-0.01	0.00	-0.01	**-0.06**	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	-0.01	0.01	-0.02	0.00	0.00	0.00
	Informative at level-1	0.1	**0.74**	0.01	**0.07**	0.01	0.01	-0.05	**0.12**	0.01	0.02
		0.5	**0.55**	-0.01	0.00	-0.01	-0.01	-0.02	**0.10**	0.00	0.00
	Informative at level-2	0.1	**0.37**	-0.04	0.01	**0.11**	0.03	**-0.06**	-0.01	-0.01	-0.01
		0.5	**0.28**	-0.01	0.00	**0.10**	0.01	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.12**	-0.02	0.10	**0.10**	0.03	-0.05	**0.11**	0.00	0.01
		0.5	**0.83**	-0.01	0.01	**0.10**	0.01	-0.04	**0.10**	0.00	0.00
0.5	Non-informative	0.1	-0.01	-0.01	-0.01	-0.01	0.00	-0.05	0.00	0.00	0.00
		0.5	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	0.00	0.00	0.00
	Informative at level-1	0.1	**0.57**	0.01	0.05	0.00	0.00	-0.04	**0.12**	0.01	0.02
		0.5	**0.43**	-0.01	0.00	-0.01	-0.01	-0.02	**0.10**	0.00	0.00
	Informative at level-2	0.1	**0.58**	-0.02	0.02	**0.11**	0.01	-0.04	-0.01	-0.01	-0.01
		0.5	**0.44**	0.00	0.00	**0.10**	0.00	-0.01	0.00	0.00	0.00
	Informative at both levels	0.1	**1.17**	0.00	**0.09**	**0.11**	0.02	-0.04	**0.11**	0.01	0.01
		0.5	**0.88**	-0.01	0.01	**0.10**	0.00	-0.01	**0.10**	0.00	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

Table 4. Coverage Rate for the Random Intercept Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X2 Slope			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.93	0.94	0.93	0.95	0.95	0.94	**0.91**	**0.92**	**0.91**	0.97	0.96	0.94
	informative	0.5	0.96	0.97	0.95	0.94	0.94	0.94	0.95	0.96	0.94	0.95	0.95	0.95
	Informative	0.1	**0.05**	0.96	0.97	0.95	0.96	0.93	**0.92**	0.95	0.96	**0.73**	0.93	0.93
	at level-1	0.5	**0.49**	0.96	0.96	0.93	0.93	**0.91**	0.94	0.94	0.93	**0.01**	0.95	0.96
	Informative	0.1	**0.74**	**0.92**	0.94	0.94	0.94	**0.91**	0.93	0.93	**0.92**	**0.98**	0.94	0.94
	at level-2	0.5	**0.14**	0.93	0.96	0.95	0.94	0.94	0.91	0.93	0.94	0.96	0.95	0.96
	Informative	0.1	**0.00**	**0.88**	**0.88**	0.94	0.95	**0.89**	**0.92**	0.95	0.93	**0.71**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.95	0.95	0.94	0.94	0.94	**0.90**	**0.92**	0.94	**0.02**	0.94	0.94
0.2	Non-	0.1	0.94	0.94	0.94	0.95	0.95	0.94	**0.90**	**0.91**	**0.86**	0.96	0.96	0.93
	informative	0.5	0.95	0.96	0.95	0.93	0.94	**0.92**	0.94	0.94	0.94	0.95	0.96	0.95
	Informative	0.1	**0.00**	0.94	**0.90**	0.95	0.96	0.93	**0.92**	**0.92**	**0.90**	**0.71**	**0.92**	0.93
	at level-1	0.5	**0.00**	0.95	0.95	0.93	0.93	**0.92**	0.94	0.94	0.94	**0.01**	0.96	0.96
	Informative	0.1	**0.51**	0.93	0.93	0.94	0.96	0.93	**0.91**	**0.92**	**0.88**	0.95	0.93	0.94
	at level-2	0.5	**0.06**	0.96	0.96	0.94	0.94	0.94	**0.88**	**0.92**	**0.92**	0.96	0.95	0.96
	Informative	0.1	**0.00**	**0.90**	**0.90**	0.95	0.95	**0.91**	**0.88**	**0.89**	**0.90**	**0.69**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.96	0.97	0.95	0.96	0.93	**0.87**	0.93	**0.92**	**0.02**	0.94	0.94
0.5	Non-	0.1	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.93	**0.87**	0.96	0.95	0.93
	informative	0.5	0.95	0.96	0.95	0.93	0.93	0.93	0.93	0.94	0.94	0.95	0.96	0.95
	Informative	0.1	**0.04**	0.95	0.94	0.95	0.95	0.93	**0.91**	**0.91**	**0.87**	**0.71**	**0.92**	0.93
	at level-1	0.5	**0.00**	0.95	0.95	0.93	0.93	0.93	0.94	0.95	0.94	**0.01**	0.95	0.96
	Informative	0.1	**0.41**	0.93	0.94	0.95	0.96	**0.92**	**0.89**	**0.91**	**0.87**	0.95	0.94	0.94
	at level-2	0.5	**0.05**	0.96	0.97	0.94	0.95	0.94	0.97	0.93	**0.92**	0.96	0.95	0.96
	Informative	0.1	**0.02**	**0.92**	**0.91**	0.95	0.95	**0.92**	**0.88**	**0.89**	**0.87**	**0.69**	**0.91**	**0.92**
	at both levels	0.5	**0.00**	0.96	0.97	0.94	0.95	0.94	**0.87**	**0.92**	**0.92**	**0.02**	0.93	0.94

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

Table 5. Coverage Rate for the Random Intercept Model Under $\chi^2(2)$ Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X2 Slope			TAU			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.91**	**0.90**	0.93	**0.90**	**0.78**	**0.91**	**0.91**
		0.5	0.97	0.97	0.97	0.94	0.94	0.94	**0.79**	**0.89**	**0.92**	**0.67**	0.94	0.94
	Informative at level-1	0.1	**0.00**	0.94	0.94	0.94	0.94	**0.92**	**0.92**	**0.92**	**0.92**	**0.82**	0.96	0.94
		0.5	**0.00**	0.96	0.97	0.94	0.94	0.93	**0.81**	**0.90**	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.85**	**0.92**	0.93	0.95	0.95	**0.90**	**0.92**	0.96	0.95	**0.77**	**0.89**	**0.90**
		0.5	**0.25**	0.95	0.96	0.95	0.95	0.96	**0.77**	0.94	**0.92**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.00**	0.93	**0.91**	0.95	0.94	0.93	0.93	0.93	0.95	**0.82**	0.95	0.95
		0.5	**0.00**	0.95	0.97	0.95	0.95	0.94	**0.78**	**0.91**	**0.89**	**0.00**	0.96	0.96
0.2	Non-informative	0.1	0.93	0.93	**0.92**	0.94	0.94	**0.92**	**0.77**	**0.83**	**0.78**	**0.66**	**0.91**	**0.91**
		0.5	0.97	0.97	0.97	0.93	0.93	**0.92**	**0.72**	**0.91**	**0.91**	**0.67**	0.94	0.94
	Informative at level-1	0.1	**0.10**	0.95	0.94	**0.92**	**0.92**	**0.92**	**0.77**	**0.83**	**0.79**	**0.58**	0.96	0.94
		0.5	**0.00**	0.97	0.97	0.93	0.93	0.94	**0.73**	**0.92**	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.71**	**0.92**	**0.92**	0.94	0.94	0.93	**0.83**	**0.90**	**0.83**	**0.69**	**0.89**	**0.90**
		0.5	**0.17**	0.97	0.96	0.95	0.94	0.96	**0.70**	0.94	**0.91**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.00**	0.94	**0.92**	0.95	0.95	0.93	**0.85**	**0.90**	**0.86**	**0.62**	0.95	0.94
		0.5	**0.00**	0.96	0.96	0.95	0.94	0.95	**0.71**	0.95	**0.90**	**0.00**	0.97	0.96
0.5	Non-informative	0.1	0.93	0.93	**0.92**	0.94	0.94	**0.92**	**0.71**	**0.81**	**0.76**	**0.66**	**0.91**	**0.91**
		0.5	0.97	0.97	0.96	**0.92**	0.93	**0.92**	**0.70**	**0.92**	**0.91**	**0.67**	0.93	0.94
	Informative at level-1	0.1	**0.63**	0.93	0.93	0.93	**0.92**	**0.92**	**0.70**	**0.81**	**0.77**	**0.58**	0.95	0.94
		0.5	**0.11**	0.97	0.97	0.93	0.93	**0.92**	**0.70**	0.93	**0.91**	**0.00**	**0.98**	0.95
	Informative at level-2	0.1	**0.64**	**0.92**	**0.92**	0.95	0.95	0.94	**0.76**	**0.88**	**0.82**	**0.69**	**0.88**	**0.90**
		0.5	**0.14**	0.97	0.96	0.95	0.94	0.95	**0.67**	0.94	**0.90**	**0.69**	0.93	0.94
	Informative at both levels	0.1	**0.04**	0.93	0.94	0.95	0.94	0.93	**0.77**	**0.89**	**0.81**	**0.62**	0.95	0.94
		0.5	**0.00**	0.96	0.96	0.94	0.93	0.95	**0.67**	0.94	**0.90**	**0.00**	0.97	0.96

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

from 0.23 to 1.64) when the selection mechanism was informative (see Table 6). The relative biases based on MPML estimates were acceptable under the majority of conditions, except for 6 conditions where the sample size was small and the selection mechanism was informative at level 1 or both levels (relative bias ranging from 0.12 to 0.14). The bootstrap method performed the best in terms of relative biases because there were only 3 conditions where small biases were found (ranging from -0.06 to -0.10).

Table 6. Relative Bias for Fixed Effects Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X1		
			ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	0.01	0.01	0.00	0.00	0.00	0.00
		0.5	0.00	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.93**	0.04	**0.12**	0.00	0.00	0.01
		0.5	**0.70**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.30**	**-0.06**	0.00	**0.11**	0.06	0.00
		0.5	**0.23**	-0.03	0.00	**0.08**	0.02	0.00
	Informative at both levels	0.1	**1.23**	-0.03	**0.13**	**0.11**	**0.06**	0.01
		0.5	**0.92**	-0.03	0.01	**0.08**	0.02	0.00
0.2	Non-informative	0.1	0.00	0.01	0.00	0.00	0.00	0.00
		0.5	-0.01	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.85**	0.04	**0.13**	0.00	0.00	0.01
		0.5	**0.64**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.61**	**-0.10**	0.00	**0.22**	**0.06**	0.00
		0.5	**0.45**	-0.02	0.00	**0.16**	0.01	0.00
	Informative at both levels	0.1	**1.46**	**-0.07**	**0.14**	**0.22**	**0.06**	0.01
		0.5	**1.09**	-0.02	0.01	**0.16**	0.01	0.00
0.5	Non-informative	0.1	0.00	0.01	0.01	0.00	0.00	0.00
		0.5	-0.01	0.00	0.00	0.00	0.00	0.00
	Informative at level-1	0.1	**0.66**	0.03	**0.13**	0.00	0.00	0.01
		0.5	**0.50**	0.00	0.01	0.00	0.00	0.00
	Informative at level-2	0.1	**0.96**	**-0.07**	0.00	**0.35**	0.04	0.00
		0.5	**0.71**	-0.01	0.00	**0.25**	0.01	0.00
	Informative at both levels	0.1	**1.64**	-0.04	**0.14**	**0.35**	0.04	0.01
		0.5	**1.22**	-0.01	0.01	**0.25**	0.01	0.00

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

As a result of the biased point estimate based on ML, the coverage rates of the confidence intervals for the ML estimates were also poor (ranging from 0.00 to 0.61) under informative selection mechanisms (see Table 6). On the other hand, both MPML and the bootstrap method had the issue of over-coverage (coverage rate above 0.98) in the majority of the conditions, indicating that the estimated confidence intervals were wider than expected.

Table 7. Coverage Rate for Fixed Effects Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	Intercept			X1			X2		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-informative	0.1	0.96	**0.99**	**0.98**	0.96	0.96	0.96	0.95	0.95	0.93
		0.5	0.96	**1.00**	**1.00**	0.97	**0.99**	**0.99**	0.95	0.95	0.96
	Informative at level-1	0.1	**0.00**	0.97	**0.89**	0.96	0.97	0.95	0.94	0.95	**0.91**
		0.5	**0.00**	**1.00**	**1.00**	0.96	**0.99**	**0.98**	0.95	0.95	0.94
	Informative at level-2	0.1	**0.61**	0.95	**0.98**	**0.72**	**0.85**	0.95	0.95	0.95	**0.86**
		0.5	**0.01**	0.97	**0.99**	**0.02**	**0.86**	**0.98**	0.96	0.96	0.95
	Informative at both levels	0.1	**0.00**	**0.90**	**0.89**	**0.70**	**0.84**	0.96	0.95	0.96	**0.89**
		0.5	**0.00**	0.97	0.99	**0.02**	**0.88**	**0.99**	0.96	0.96	0.95
0.2	Non-informative	0.1	0.95	**0.99**	**0.99**	0.96	0.97	0.97	0.94	0.94	**0.92**
		0.5	0.96	**1.00**	**1.00**	0.97	**1.00**	**1.00**	0.96	0.96	0.95
	Informative at level-1	0.1	**0.06**	**0.99**	0.96	0.95	**0.99**	0.98	0.94	0.94	**0.91**
		0.5	**0.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.95	0.95	0.95
	Informative at level-2	0.1	**0.23**	0.97	**0.99**	**0.36**	**0.88**	0.96	0.94	0.95	**0.88**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**0.98**	**1.00**	0.97	0.97	0.95
	Informative at both levels	0.1	**0.00**	0.94	0.93	0.33	0.86	0.96	0.94	0.96	**0.88**
		0.5	**0.00**	**1.00**	**0.99**	**0.00**	**0.98**	**1.00**	0.96	0.96	0.95
0.5	Non-informative	0.1	0.94	**0.99**	**0.99**	0.97	**1.00**	**1.00**	0.94	0.94	**0.91**
		0.5	0.95	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.96	0.96	0.95
	Informative at level-1	0.1	**0.49**	**1.00**	**0.99**	0.94	**1.00**	**1.00**	0.94	0.94	0.91
		0.5	**0.07**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.95	0.97	0.95
	Informative at level-2	0.1	**0.14**	**0.98**	**0.99**	**0.17**	0.96	**0.98**	0.94	0.94	**0.87**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	0.96	0.97	0.95
	Informative at both levels	0.1	**0.00**	0.96	0.97	**0.18**	0.95	**0.98**	0.95	0.96	**0.88**
		0.5	**0.00**	**1.00**	**1.00**	**0.00**	**1.00**	**1.00**	0.97	0.97	0.96

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate $<$ 0.93 or $>$ 0.97)

Table 8. Relative Bias for Variance Components Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	TAU00			TAU11			TAU01			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.02	0.01	**-0.87**	**0.16**	**0.20**	**0.55**	**-0.09**	**-0.35**	**-0.57**	-0.01	0.03	0.01
	informative	0.5	0.01	**-0.10**	**-0.35**	0.01	**-0.07**	**-0.09**	0.01	**-0.26**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	-0.01	-0.02	**-0.61**	**0.17**	**0.25**	**1.24**	-0.04	**-0.31**	**-0.48**	**-0.09**	0.02	-0.01
	at level-1	0.5	0.00	**-0.10**	**-0.34**	0.01	**-0.07**	-0.01	0.01	**-0.27**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.11**	-0.05	**-0.83**	-0.01	**0.19**	**0.69**	**-0.34**	**-0.47**	**-0.50**	-0.01	0.03	0.01
	at level-2	0.5	**-0.09**	**-0.11**	**-0.35**	**-0.08**	**-0.10**	**-0.09**	**-0.15**	**-0.30**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	**-0.18**	**-0.10**	**-0.62**	**0.06**	**0.26**	**1.61**	**-0.36**	**-0.47**	**-0.61**	**-0.09**	0.03	-0.02
	at both levels	0.5	**-0.09**	**-0.11**	**-0.34**	**-0.09**	**-0.10**	-0.01	**-0.16**	**-0.30**	**-0.41**	-0.05	0.02	0.01
0.2	Non-	0.1	0.00	**-0.08**	**-0.38**	0.01	**-0.25**	-0.01	0.01	-0.05	**-0.39**	0.00	**0.08**	0.05
	informative	0.5	0.00	**-0.11**	**-0.16**	0.01	**-0.36**	**-0.10**	0.00	**-0.09**	**-0.40**	0.00	0.02	0.01
	Informative	0.1	0.00	**-0.09**	**-0.27**	0.04	**-0.23**	**0.17**	0.02	-0.04	**-0.39**	**-0.09**	**0.07**	0.00
	at level-1	0.5	0.00	**-0.11**	**-0.16**	0.01	**-0.37**	**-0.08**	0.00	**-0.09**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.15**	**-0.12**	**-0.38**	**-0.24**	**-0.30**	-0.01	**-0.16**	**-0.12**	**-0.38**	0.00	**0.09**	0.05
	at level-2	0.5	**-0.09**	**-0.11**	**-0.16**	**-0.16**	**-0.37**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	0.00	0.02	0.01
	Informative	0.1	**-0.17**	**-0.13**	**-0.31**	**-0.24**	**-0.32**	**0.23**	**-0.12**	**-0.07**	**-0.41**	**-0.09**	**0.08**	0.00
	at both levels	0.5	**-0.09**	**-0.11**	**-0.15**	**-0.16**	**-0.37**	**-0.08**	**-0.09**	**-0.10**	**-0.41**	-0.05	0.02	0.01
0.5	Non-	0.1	0.01	**-0.09**	**-0.18**	0.01	**-0.33**	**-0.07**	0.01	**-0.07**	**-0.40**	0.00	**0.11**	0.04
	informative	0.5	0.00	**-0.10**	**-0.12**	0.01	**-0.39**	**-0.10**	0.00	**-0.10**	**-0.41**	0.00	0.03	0.01
	Informative	0.1	0.00	**-0.09**	**-0.16**	0.02	**-0.32**	-0.03	0.01	**-0.07**	**-0.39**	**-0.09**	**0.09**	0.00
	at level-1	0.5	0.00	**-0.11**	**-0.12**	0.01	**-0.40**	**-0.10**	0.00	**-0.10**	**-0.41**	-0.05	0.02	0.01
	Informative	0.1	**-0.15**	**-0.10**	**-0.20**	**-0.23**	**-0.33**	**-0.09**	**-0.15**	**-0.10**	**-0.37**	0.00	**0.12**	0.05
	at level-2	0.5	**-0.09**	**-0.10**	**-0.12**	**-0.16**	**-0.40**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	0.00	0.03	0.01
	Informative	0.1	**-0.16**	**-0.10**	**-0.18**	**-0.23**	**-0.34**	-0.01	**-0.14**	**-0.08**	**-0.39**	**-0.09**	**0.10**	0.00
	at both levels	0.5	**-0.09**	**-0.10**	**-0.12**	**-0.16**	**-0.40**	**-0.10**	**-0.09**	**-0.10**	**-0.41**	-0.05	0.02	0.01

Note. Values in bold represent unacceptably large relative bias (i.e., absolute value > 0.05)

Table 9. Coverage rate for Variance Components Estimates from the Random Slopes Model Under Normal Distribution

ICC	Selection Mechanism	Sampling Fraction	TAU00			TAU11			TAU01			SIGMA		
			ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML	ML	BOOT	MPML
0.05	Non-	0.1	0.95	0.93	**0.71**	**0.98**	**1.00**	**0.71**	0.97	0.95	**0.87**	0.94	0.93	0.94
	informative	0.5	0.96	**0.85**	**0.09**	0.94	**0.92**	**0.41**	0.95	**0.76**	**0.92**	0.95	**0.70**	**0.84**
	Informative	0.1	0.95	**0.95**	**0.86**	**0.99**	**1.00**	**0.81**	0.96	0.94	**0.66**	**0.68**	**0.88**	0.95
	at level-1	0.5	0.96	**0.85**	**0.11**	0.95	**0.92**	**0.39**	0.96	**0.75**	0.95	**0.02**	**0.71**	**0.91**
	Informative	0.1	0.96	0.95	**0.68**	**0.98**	**0.99**	**0.70**	0.94	**0.92**	**0.86**	0.96	**0.83**	0.94
	at level-2	0.5	**0.85**	**0.80**	**0.16**	**0.90**	**0.87**	**0.44**	**0.87**	**0.68**	0.93	0.94	**0.68**	**0.86**
	Informative	0.1	0.94	0.93	**0.84**	**0.98**	**0.99**	**0.77**	0.96	**0.91**	**0.66**	**0.68**	**0.76**	**0.91**
	at both levels	0.5	**0.85**	**0.83**	**0.20**	**0.90**	**0.86**	**0.46**	**0.85**	**0.65**	0.96	**0.03**	**0.67**	**0.89**
0.2	Non-	0.1	0.94	0.93	**0.78**	0.96	0.94	**0.85**	0.93	0.94	0.97	0.95	**0.79**	**0.91**
	informative	0.5	0.96	**0.77**	**0.56**	0.94	**0.83**	**0.16**	0.97	**0.31**	**0.82**	0.95	**0.54**	0.84
	Informative	0.1	0.95	0.93	**0.84**	0.96	0.96	**0.85**	0.93	0.93	**0.92**	**0.73**	**0.78**	0.94
	at level-1	0.5	0.96	0.77	**0.57**	0.94	**0.82**	**0.15**	0.96	**0.28**	**0.86**	**0.02**	**0.60**	**0.91**
	Informative	0.1	**0.86**	**0.87**	**0.77**	**0.90**	**0.88**	**0.83**	**0.92**	**0.87**	0.95	0.96	**0.70**	**0.91**
	at level-2	0.5	**0.80**	**0.75**	**0.64**	**0.84**	**0.81**	**0.24**	**0.80**	**0.29**	**0.87**	0.94	**0.55**	**0.86**
	Informative	0.1	**0.84**	**0.86**	**0.85**	**0.91**	**0.91**	**0.82**	**0.90**	**0.87**	**0.88**	**0.72**	**0.70**	**0.92**
	at both levels	0.5	**0.80**	**0.75**	**0.65**	**0.84**	**0.79**	**0.25**	**0.80**	**0.30**	**0.89**	**0.03**	**0.57**	**0.89**
0.5	Non-	0.1	0.95	0.90	**0.85**	0.96	0.93	**0.75**	0.94	**0.86**	0.94	0.95	**0.66**	**0.91**
	informative	0.5	0.96	**0.75**	**0.72**	0.93	**0.77**	**0.11**	0.97	**0.17**	**0.75**	0.95	**0.50**	**0.84**
	Informative	0.1	0.96	**0.89**	**0.86**	0.95	**0.92**	**0.76**	0.93	**0.86**	0.94	**0.73**	**0.71**	0.94
	at level-1	0.5	**0.85**	**0.75**	**0.70**	0.93	**0.77**	**0.10**	0.95	**0.17**	**0.78**	**0.02**	**0.57**	**0.91**
	Informative	0.1	**0.83**	**0.82**	**0.78**	**0.87**	**0.85**	**0.75**	**0.88**	**0.78**	**0.90**	0.97	**0.62**	**0.91**
	at level-2	0.5	**0.79**	**0.75**	**0.75**	**0.82**	**0.77**	**0.19**	**0.77**	**0.19**	**0.83**	0.94	**0.52**	**0.85**
	Informative	0.1	**0.80**	**0.83**	**0.82**	**0.88**	**0.85**	**0.75**	**0.84**	**0.76**	**0.91**	**0.72**	**0.64**	**0.92**
	at both levels	0.5	**0.79**	**0.75**	**0.76**	**0.81**	**0.75**	**0.20**	**0.77**	**0.20**	**0.82**	**0.02**	**0.54**	**0.89**

Note. Values in bold represent under-coverage or over-coverage (i.e., coverage rate < 0.93 or > 0.97)

Slope of *X1*. As expected, the ML estimates of the slope of *X1* were biased when the selection mechanism was informative at level 2 or both levels (relative bias ranging between 0.08 and 0.35). The magnitude of the biases increased as ICC increased. On the other hand, both the MPML and the bootstrap estimation methods successfully reduced the biases to an acceptable level, although the MPML method performed slightly better than the bootstrap method when sample size was small and the selection mechanism was informative at level 2 or both levels.

Similarly, due to the biased point estimates, the coverage rates of the confidence intervals for the ML estimates were also poor (ranging from 0.00 to 0.72) under informative selection mechanisms. The MPML confidence intervals demonstrated over-coverage, especially when sample size and ICC were large. The bootstrap confidence intervals demonstrated slight under-coverage (ranging between 0.84 and 0.88) when informative selection occurred at level 2 or both levels, but showed a similar over-coverage pattern as the MPML confidence intervals in the other conditions.

Slope of *X2*. The relative bias of the estimated slope of *X2* was acceptable for all methods under all conditions. However, the MPML confidence intervals suffered from slight under-coverage (0.86 to 0.92) in about half of the conditions, mainly when sample size was small. The performance of the ML and the bootstrap confidence intervals was acceptable under all conditions.

Variance of the random intercepts (τ_{00}). ML estimates had small relative bias (-0.09 to -0.18), mainly when there was informative sampling at level-2 or at both levels. MPML suffered from moderate to large biases (-0.34 to -0.87) when ICC was small. The magnitude of the relative biases decreased as ICC or sample size increased, but was still more than 0.12 when ICC and sample size were large. The bootstrap method showed small negative biases (-0.08 to -0.13) across all conditions consistently and had the greatest advantages over MPML when ICC was small.

The coverage rates of the confidence intervals based on the three methods showed similar patterns. The ML-based confidence intervals had slight under-coverage (0.79 to 0.85) when there was informative sampling at level-2 or at both levels. The MPML-based confidence intervals suffered from severe under-coverage (0.09 to 0.20) under conditions where small ICCs were combined with large sample sizes. The bootstrap confidence intervals had slight under-coverage (0.75 to 0.90) in the majority of the conditions. It is noted that when ICC was large, MPML and bootstrap confidence intervals performed similarly.

Variance of the random slopes (τ_{11}). Similar to τ_{00}, ML estimate of τ_{11} showed small to moderate relative bias (-0.16 to 0.17), mainly when there was informative sampling at level-2 or at both levels. The MPML estimates had large positive biases (0.55 to 1.61) when ICC and sample size were both small, and mostly small negative biases (-0.08 to -0.10) under the other conditions. The bootstrap estimates had small positive biases (0.19 to 0.26) when ICC and sample size were both small, and moderate negative biases (-0.23 to -0.40) when

ICC was moderate and large. Comparing the three methods, ML showed the least amount of bias across all conditions.

In terms of the confidence intervals, MPML had the worst performance because of the severe under-coverage (0.10-0.46) when sample size was large. The bootstrap confidence intervals had somewhat under-coverage (0.77-0.92) across the conditions. The ML confidence intervals had the best performance, showing slight under-coverage (0.81 to 0.91) when there was informative sampling at level-2 or at both levels.

Covariance of the random intercepts and the random slopes (τ_{01}). The ML estimate of τ_{01} showed small to moderate negative biases (-0.09 to -0.36) when there was informative sampling at level-2 or at both levels. The MPML estimates showed moderate negative biases across all conditions, ranging from -0.37 to -0.61. The bootstrap estimates showed small to moderate negative biases, with the magnitude decreasing from -0.34 to -0.09 as ICC increased from 0.05 to 0.5.

The ML confidence intervals had slight under-coverage (0.77 to 0.92) when there was informative sampling at level-2 or at both levels. Despite the moderate negative biases in the point estimates, MPML confidence intervals only showed slight under-coverage in most of the conditions (0.66 to 0.92). In general, the bootstrap confidence intervals suffered from under-coverage (0.17 to 0.92), and the degree of under-coverage was severe (0.17 to 0.31) when sample sizes were large and ICCs were moderate to large.

Level-1 residual variance (σ^2). ML estimates had small negative relative biases (-0.09) when there was informative selection at level-1 or at both levels. The bootstrap estimates showed small positive relative biases (0.07 to 0.12) when sample size was small and ICC was moderate to large. MPML estimates had the best performance with little bias across all conditions.

The ML-based confidence intervals showed under-coverage when there was informative selection at level-1 or at both levels. The degree of under-coverage was severe (0.02 to 0.03) when sample size was large. The bootstrap confidence interval had moderate under-coverage across all conditions, ranging from 0.50 to 0.88. The MPML confidence intervals had slight under-coverage across all conditions, ranging from 0.84 to 0.91.

6 Discussion and Conclusion

We proposed a weighted residual bootstrap method for multilevel modeling of data from complex sampling designs. Unlike previously proposed bootstrap methods (e.g., Grilli & Pratesi, 2004; Kovacevic et al., 2006; Wang & Thompson, 2012), our method does not require generating a pseudo population or rescaling weights. The performance of the proposed bootstrap method for linear two-level models was investigated under various conditions, and compared with the multilevel pseudo maximum likelihood (MPML) approach and the unweighted ML approach using Monte Carlo simulations.

In general, the proposed weighted bootstrap method performed similar to or better than the MPML method in random intercept models and had mixed results in random slopes models. As expected, for the random intercept model, unweighted ML resulted in biased intercept estimate when there were informative selections. Both the bootstrap and the MPML estimates of the slopes for the level-1 and level-2 predictors (*X1* and *X2*) had acceptable performance. However, the bootstrap showed advantages over MPML for the estimate of the level-2 variance component when sample size is small (i.e., 50 clusters and 10 units per cluster), selection mechanism is informative, and ICC is low (i.e., 0.05). As a result, the confidence interval of the slope of the level-2 predictor (*X2*) based on the bootstrap method also had a better coverage rate compared to MPML under those conditions. It has been demonstrated in the literature that MPML estimates have increased biases as ICC decreases (Asparouhov, 2006; Kovacevic & Rai, 2003). As Asparouhov (2006) explained, the weakness of MPML is in the estimation on the individual level, therefore as ICC decreases the individual level becomes more influential, which exacerbates the problem.

For the random slopes model, the ML estimates of both the intercept and the slope of the level-1 predictor (*X1*) showed moderate to severe biases when there are informative selections. The bootstrap and the MPML approaches performed similarly in terms of the estimates of the fixed effects, with the bootstrap method slightly better for the estimate of the intercept and the slope of the level-2 predictor, while the MPML slightly better for the slope of the level-1 predictor. While convergence was not an issue for the bootstrap method, MPML suffered from a high rate of non-convergence when ICC is low. As a result, MPML had severe biases in the estimates of the level-2 variance components when ICC is low. The performance of the bootstrap estimate of the variance components was not ideal either as small to moderate biases existed across the conditions. However, the bootstrap confidence intervals performed much better than the MPML approach, especially when sample size is large. The only drawback of the bootstrap method is in the estimation of the covariance between the random intercept and the random slope, which showed severe under-coverage when sample size is large.

Another advantage of the bootstrap method is that it is more robust to the distributional violation. Previous simulation studies on MPML for linear models only considered normally distributed random effects and residuals. Our findings showed that when the normality assumption was violated, the coverage rate of the MPML confidence interval for the level-2 variance component in a random intercept model became much worse with 8 more conditions showing under-coverage. The bootstrap method was also affected by the distributional violation, but to a lesser degree because only 4 more conditions showed under-coverage when the distributions were skewed.

As a demonstration, the weighted residual bootstrap method was applied to the American 2000 PISA data on math achievement. Based on the random intercept model, the bootstrap and the MPML results showed some inconsistency, especially for the slope of the level-2 predictor. We believe that the bootstrap results were more trustworthy in this case because conditions in the simulation study that were similar to the specific condition of this sample (i.e., small cluster size, low ICC, very slightly informative, and slight distributional violation) have shown favorable results in the bootstrap than the MPML method.

6.1 Implications

The weighted residual bootstrap method provides a robust alternative to MPML. Applied researchers can use the bootstrap approach when the traditional MPML estimation fails to converge or when there is severe violation of the normality assumption. In analyses of random intercept models, the weighted residual bootstrap method is preferred to MPML when the effect of level-2 predictors (e.g., school SES), or the variance of the random intercept (e.g., variance of school mean achievement) are of interest and when both sample sizes and ICCs are small. In random slopes models, the bootstrap method has advantages over MPML in the point estimates and the confidence interval estiamtes of the slopes of level-2 predictors, as well as the variance component estimates associated with the random intercept and the random slopes (e.g., variance of the association of student SES and student achievement across schools). However, the statistical inferences for the covariance component (e.g., the covariance between school mean achievement and the slope of student SES and student achievement) based on the bootstrap method might not be trustworthy.

It is recommended that researchers conduct sensitivity analyses using different methods. Discrepancies among the results may indicate that the conditions for MPML to work properly are not satisfied. The weighted residual bootstrap method is implemented in the developmental version of the R package *bootmlm*, which has the capacity to analyze two-level linear random intercept and random coefficients models with sampling weights.

6.2 Limitations and Future Directions

The findings of the study should be interpreted in light of the limitations. First, there is still room for improvement in terms of the bootstrap confidence interval for level-2 variance and covariance components. We used percentile confidence interval for its simplicity. Future research may be conducted to investigate whether more sophisticated methods such as bias-corrected and accelerated confidence intervals and studentized intervals could further improve the performance. Second, the proposed bootstrap method was only applied to multilevel linear models. Although it is possible to extend it to generalized multilevel models (Goldstein et al., 2018), Monte Carlo experiments should be conducted to examine the performance of the method for generalized multilevel models such as multilevel ordinal and binary models. Third, this study only compared the performance of the proposed method with MPML. Future studies could compare the proposed method with other bootstrap methods for multilevel data with sampling weights.

References

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*(3), 411–434. doi: https://doi.org/10.1207/s15328007sem1203_4

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, *35*(3), 439–460. doi: https://doi.org/10.1080/03610920500476598

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: https://doi.org/10.18637/jss.v067.i01

Booth, J. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation. In G. Seeber, J. Francis, R. Hatzinger, & G. Steckel-Berger (Eds.), *Statistical modelling* (pp. 43–51). New York, NY: Springer. doi: https://doi.org/10.1007/978-1-4612-0789-4_6

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 431–443. doi: https://doi.org/10.1111/1467-9876.00415

Cochran, W. G. (1977). *Sampling techniques (3rd ed.)*. New York, NY: Wiley.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge, UK: Cambridge University. doi: https://doi.org/10.1017/cbo9780511802843

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman and Hall. doi: https://doi.org/10.1201/9780429246593

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*(1), 43–56. doi: https://doi.org/10.1093/biomet/73.1.43

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (p. 163–171). New York, NY: Routledge.

Goldstein, H., Carpenter, J., & Kenward, M. G. (2018). Bayesian models for weighted data with missing values: a bootstrap approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(4), 1071–1081. doi: https://doi.org/10.1111/rssc.12259

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

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 221–238. doi: https://doi.org/10.1111/j.1467-9868.2006.00541.x

Hoogland, J. J., & Boomsma, A. (1998). *Robustness studies in covariance structure modeling.* Sociological Methods and Research, 26:329–367. doi: https://doi.org/10.1177/0049124198026003003

Kovacevic, M. S., Huang, R., & You, Y. (2006). *Bootstrapping for variance estimation in multi-level models fitted to survey data.* ASA Proceedings of the Survey Research Methods Section.

Kovacevic, M. S., & Rai, S. N. (2003). *A pseudo maximum likelihood approach to multi-level modeling of survey data.* Communications in Statistics—Theory and Methods, 32:103–121. doi: https://doi.org/10.1081/sta-120017802

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, *18*(2), 199–210. doi: https://doi.org/10.1214/ss/1063994975

Lohr, S. L. (2010). *Sampling: Design and analysis (2nd ed.).* Boston, MA: Cengage.

Maas, C. J., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427–440. doi: https://doi.org/10.1016/j.csda.2003.08.006

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide (8th ed.).* Los Angeles, CA: Muthén & Muthén.

Organization for Economic Co-operation and Development. (2000). Manual for the pisa 2000 database [Computer software manual]. Retrieved from `http://www.pisa.oecd.org/dataoecd/53/18/33688135.pdf`

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistics Review.* doi: https://doi.org/10.2307/1403631

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistics Society: Series B (Statistical Methodology).* doi: https://doi.org/10.1111/1467-9868.00106

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of American Statistical Association.* doi: https://doi.org/10.2307/2290269

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria.

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

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema*, *25*(4), 520–528.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling.* doi: https://doi.org/10.1207/s15328007sem0904_2

Thai, H. T., Mentré, F., Holford, N. H. G., Veyrat-Follet, C., & Comets, E. (2014). Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *Journal of Pharmacokinetics and Pharmacodynamics*, *41*(1).

Van der Leeden, R., Meijer, E., & Busing, F. M. (2008). Resampling multilevel models. In *Handbook of multilevel analysis* (pp. 401–433). New York, NY: Springer.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*. doi: https://doi.org/10.1016/s0167-9473(96)00047-3

Wang, Z., & Thompson, M. E. (2012). A resampling approach to estimate variance components of multilevel models. *Canadian Journal of Statistics*, *40*(1), 150–171. doi: https://doi.org/10.1002/cjs.10136

Appendix A. R Code for the Analysis of PISA Data using Weighted Residual Bootstrap

```
# Check if devtools were installed
if (!require("devtools")) {
   install.packages("devtools")
}
# Install developmental version of the bootmlm package
devtools::install_github("marklhc/bootmlm",
  ref = "weighted_boot")

# Load required packages
library(bootmlm)
library(boot)
library(lme4)

# Unweighted ML
m1 <- lmer(SC17Q01 ~ ISEI_m + male + (1 | Sch_ID),
           data = PISA, REML = FALSE)

# Weighted semi-parameteric bootstrap
boo <- bootstrap_mer(
  m1,
  FUN = function(x) {
    c(x@beta,
      c(x@theta ^ 2, 1) * sigma(x) ^ 2)
  },
  nsim = 999L,
  type = "residual_cgr",
w1 = PISA$ W_FSTUWT,
w2 = unique(PISA[c("Sch_ID", "WNRSCHBW")])$WNRSCHBW
)

# Print the output
boo  # bootstrap results
colMeans(boo$t)  # parameter estimates
apply(boo$t, 2, sd)  # bootstrap SE

# Percentile intervals for the six parameters
boot.ci(boo, type = "perc", index = 1L)
boot.ci(boo, type = "perc", index = 2L)
boot.ci(boo, type = "perc", index = 3L)
boot.ci(boo, type = "perc", index = 4L)
boot.ci(boo, type = "perc", index = 5L)
boot.ci(boo, type = "perc", index = 6L)
```

Appendix B. Mplus Code for the Analysis of PISA Data using MPML

```
Data:         File=pisa.csv;
Variable:     Names are math ISEI_m male Sch_ID
              W_FSTUWT WNRSCHBW lv1_con_wt;
              Usevariables are math ISEI_m male;
              Between = ISEI_m;
              Within = male;
              Cluster = Sch_ID;
              Weight = lv1_con_wt;  !lv1_con_wt=
                 W_FSTUWT/WNRSCHBW;
              Bweight = WNRSCHBW;
Analysis:     Type = twolevel;
Model:        %within%
              math on male;
              %between%
              math on ISEI_m;
Output:       Cinterval;
```