Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 13, June 2022 ISSN 1531-7714

Incorporating Complex Sampling Weights in Multilevel Analyses of Education Data

Ting Shen, *Missouri University of Science and Technology*
Spyros Konstantopoulos, *Michigan State University*

Large-scale assessment survey (LSAS) data are collected via complex sampling designs with special features (e.g., clustering and unequal probability of selection). Multilevel models have been utilized to account for clustering effects whereas the probability weighting approach (PWA) has been used to deal with design informativeness derived from the unequal probability selection. However, the difficulty of applying PWA in multilevel models (MLM) has been generally underestimated and practical guidance is scarce. This study utilizes an empirical as well as a Monte Carlo simulation investigation to examine the performance of the multilevel pseudo maximum likelihood (MPML) estimation based on information derived from the Early Childhood Longitudinal Study Kindergarten cohort of 2010-2011 (ECLS-K:2011). Variance components and fixed effects estimators across four estimation methods including three MPML estimators (i.e., weighted without scaling, weighted size-scaled and weighted effective-scaled) and the unweighted estimator are provided. Practical guidance about the use of sampling weights in MLM analyses of LSAS data is also offered.

Keywords: sampling weights, multilevel models, multilevel pseudo maximum likelihood, Monte Carlo simulation, large-scale assessment surveys

Introduction

Nowadays, large-scale education data have been regularly analyzed to provide generalizable research evidence that can inform education research, policy and practice. Such data have been collected through large-scale assessment surveys (LSAS) conducted by national agencies such as the National Center for Education Statistics (NCES) and international organizations such as the International Association for the Evaluation of Educational Achievement (IEA) and the Organization for Economic Co-operation and Development (OECD). LSAS typically employ multi-stage, complex sampling designs that involve stratification and cluster sampling. For example, multi-stage sampling has been used in international LSAS such as the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) (Martin & Mullis, 2012; OECD, 2014) .

LSAS data are advantageous to education researchers. First, they provide reliable measures of students' academic achievement as well as plentiful information about students, their family backgrounds and schools that allow education researchers to investigate various research questions. Second, LSAS are designed to provide information about national probability samples of students that represent well-defined populations of interest (e.g., 4th graders in the U.S.), which facilitates researchers making explicit

projections of statistical inference from samples to populations.

The complex sampling designs used in LSAS involve some special features. One important aspect of complex sampling designs is unequal probability of selection, which can take place at different stages of the sampling design. For example, to ensure representation of minority students in the sample, American Indian students may be sampled with a higher probability than white students. Similarly, rural schools may be sampled with a higher probability than suburban schools to assure representation of schools from the countryside. When unequal probability of selection is utilized in multi-stage sampling, the ensuing sample may be informative at that stage. When the sample is informative, the distribution of a variable (e.g., the mean and the variance) may be different than that in the population. For example, because of the informativeness of the sampling design, the distribution of student achievement in mathematics in the sample could have a lower mean than that in the population. Therefore, it is important in statistical inference to take into account the informativeness of sampling designs whenever it exists (Laukaityte & Wiberg, 2018).

Another important facet of complex sampling designs is the nesting structure of the data (e.g., students nested within classrooms and schools). This grouping of individuals into larger units (e.g., students grouped into schools) creates a dependency in the data. As a result, the assumption about the independence of observations (and the residuals) which is fundamental in typical linear models such as multiple regression, becomes no longer tenable. Specifically, students in the same classroom or school are more alike compared to students in different classrooms or schools. This is typically known as the clustering effect and is a consequence of cluster sampling. LSAS in the field of education sample clusters such as schools or classrooms in which students are grouped into and the generated clustering effect needs to be addressed appropriately during data analysis. In practice, this translates to adjusting the standard errors of the regression estimates for clustering, which typically results in augmented standard errors.

To analyze data collected from LSAS that incorporate multi-stage sampling, appropriate statistical tools are needed. In particular, multi-level models (MLM) have been increasingly used to analyze LSAS data in educational research because they fit well with the multi-stage sampling scheme. MLM take into account clustering effects by design (i.e., the estimation naturally adjusts the standard errors of estimates), partition the outcome variance into components aligned with different levels that correspond to sampling stages, and define conceptually the model used at each level (e.g., student, classroom, school) (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). However, conducting MLM analysis does not necessarily account for the informativeness of the sample derived from the unequal probability of selection of units in LSAS. To tackle this issue, weighted MLM estimation has been proposed and used.

Nevertheless, the difficulties and challenges involved in applying sampling weights to MLM analysis of LSAS data have been partly neglected. First, these LSAS datasets provide many different kinds of weights that may perplex the analysts. For instance, TIMSS 2011 provides many weights variables including: (1) weighting factors and weights for nonresponse adjustment at the school, classroom and student level respectively (i.e., WGTFAC1, WGTADJ1, WGTFAC2, WGTADJ2, WGTFAC3 and WGTADJ3); (2) school and student overall weights (i.e., SCHWGT and TOTWGT); (3) senate weights (i.e., SENWGT) and house weights (i.e., HOUWGT); (4) replicate weights (i.e., Jackknife zone and replicate code). Subsequently, selecting the appropriate weights for various model-based analyses could be challenging. Although LSAS data user manuals generally recommend incorporating sampling weights in the data analysis, practical guidance regarding how to use sampling weights in MLM analysis is not available. Therefore, it is likely that analysts would encounter difficulties in selecting the appropriate weights for statistical analysis, such as which weights to use in a two-level model with students and schools. Second, unlike unweighted analysis or weighted single-level model analysis, there is some variation in MLM analyses that incorporate sampling weights. As a result, it is unclear which weighted estimation method would be preferred and under what conditions.

Therefore, providing research evidence and practical guidance to inform researchers about how to

incorporate sampling weights in MLM analysis of LSAS data is seriously needed. The purpose of the present study is timely. It focuses on a commonly used computational method, the design-based probability weighting approach (PWA), which typically involves MLM. Specifically, this study provides empirical and Monte-Carlo simulation evidence of the performance of a broadly used estimation method, namely the multilevel pseudo maximum likelihood (MPML) (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006). Data from the Early Childhood Longitudinal Study – Kindergarten Class of 2010-2011 (ECLS-K:2011) are utilized (Tourangeau et al., 2018).

Literature Review

With regard to weighted MLM methods, three main approaches have been proposed and discussed. One proposed method is the weighted ANOVA method for one-way random-effects ANOVA models (Graubard & Korn, 1996; Jia et al., 2011; Korn & Graubard, 2003). Another proposed method is the probability-weighted iterative generalized least squares (PWIGLS) based on the iterative generalized least squares (IGLS) approach (Goldstein, 1986; Pfeffermann et al., 1998). The third proposed method is the MPML estimation method (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006). MPML and PWIGLS are typically the preferred methods compared to the weighted ANOVA approach because of their applicability including software development. Specifically, based on the literature and authors' investigation on software programs, the MPML has been used in STATA, Mplus and SAS whilst the PWIGLS has been implemented in LISREL, HLM and MLwiN (Chantala & Suchindran, 2006; West & Galecki, 2011).

Overall, there is no consensus in the literature about the best weighted estimation method that involves MLM. One simulation study suggested that the PWIGLS outperformed the MPML method (Cai, 2013), whereas other studies found that the MPML worked better than the PWIGLS in terms of computational simplicity, flexibility, and applicability (Asparouhov & Muthen, 2007; Kovačević & Rai, 2003; Leite et al., 2015). Evidence seems to suggest that the MPML could be easily used in more complicated statistical models for continuous, binary and ordinal outcomes (e.g., see Asparouhov, 2006; Asparouhov & Muthén, 2007; Grilli & Pratesi, 2004; Koziol et al., 2017; Rabe-Hesketh & Skrondal, 2006), whereas the PWIGLS is mainly used to model continuous outcomes (e.g., Pfeffermann et al., 1998). The present study focuses on the MPML method because it is more flexible, versatile, and easy to implement with mainstream software programs such as STATA (the software program we used in this study).

The literature review revealed there is a lack of sufficient empirical evidence and guidance about applying sampling weights in MLM when using LSAS data. There are several practical, simulation, and methodological issues that need to be addressed concerning the incorporation of sampling weights in MLM in the context of LSAS data.

From a practical point of view, deciding whether to use weights in statistical analysis or not, especially with regard to incorporating complex sampling weights, varies across different disciplines. For example, in biostatistics and public health, researchers typically use sampling weights whereas researchers in economics and econometrics generally do not apply weights in their analyses (Bollen et al., 2016). In education, practices about involving sampling weights in data analyses have been mixed, that is, some researchers use sampling weights when analyzing LSAS data, whilst others do not (Laukaityte & Wiberg, 2018). Such disciplinary differences in dealing with sampling weights generally stem from whether a certain discipline adopts a model-based, a design-based or a combined approach to analyze data.

From a simulation point of view, prior findings obtained from simulation studies were not as applicable to LSAS data in education, because such data have special features with respect to sampling designs and data structures. To illustrate, in prior simulation research about two-level models, the small sample size of level-1 units and level-2 units was identified as a source of bias in estimation (Pfeffermann et al., 1998). However, in education, LSAS data typically have large sample sizes at both levels. Consider a typical two-level data structure where, for example, first level units (e.g., students) are nested within second level units (e.g., schools). In LSAS data, the number of level-2 units (the clusters) is rather large. For instance, the ECLS-K:2011 has more than eight hundred schools. In addition, the cluster

sample size is relatively large (e.g., on average between 20 to 30 units per cluster). For example, PISA 2015 has about 30 students per school on average across 70 countries and economies, and the ECLS-K:2011 has more than 20 students per school. As a result, potential estimation bias attributed to small cluster sample sizes (e.g., less than 10 students per school) or small number of clusters (e.g., less than 20) is rather unlikely when analyzing LSAS education data. Moreover, prior simulation studies showed that estimation bias is also linked with small intraclass correlation (ICC) values (see Asparouhov, 2006; Jia et al., 2011). Nevertheless, in LSAS education data, the average ICC value when student achievement is the outcome is not that small (e.g., less than 0.05). For example, the average ICC value in unconditional (intercept only) models in various national probability samples in the U.S. was 0.22 when student achievement was the outcome (Hedges & Hedberg, 2007).

From a methodological point of view, the informativeness of the sampling design, which implies that the sample is different than the population, as well as weights scaling at lower level need to be taken into account in MLM. First of all, the design informativeness is of utmost importance for the current study because an informative design warrants weighted estimation. LSAS education data typically adopt a two-stage sampling design where, for example, clusters such as schools are selected with unequal probabilities in the first stage, and within these selected schools, students are selected with unequal probabilities or simple random sampling in the second stage. This implies the sampling design is informative at least at the cluster (school) level.

The terms informativeness and non-informativeness have been used to describe the effect of the sampling design in model-based analysis. According to Pfeffermann (1993), a sampling design is non-informative when samples are drawn using simple random sampling. In contrast, the sampling design is informative when samples are selected using unequal probability sampling (i.e., units are selected with different probabilities) (Pfeffermann, 1993). Formally defined, the sampling design is informative when the distribution of sample units (e.g., small size schools) is different than that in the population; otherwise the sampling is non-informative (Binder et al., 2005). From a design-based perspective, design informativeness has been used to describe the difference in distribution between a sample and its population due to unequal probabilities of selection of units in the sample. However, from a model-based perspective, design informativeness is evident when the selection probabilities are contingent on the dependent variable even after conditioning on all other covariates included in the model, otherwise the sampling design is non-informative (Grilli & Pratesi, 2004; Koziol et al., 2017).

Second, to ensure weights are not too large (e.g., due to extremely small probabilities of unit selection), and to reduce bias in the estimation especially when the cluster sample size is small, scaling sampling weights at the lower level has been hypothesized as a potential solution (Pfeffermann et al., 1998; Stapleton, 2002; Asparouhov, 2006). The scaling of sampling weights refers to some normalization operation (i.e., multiplying the weights by some scaling constant) such that "the sum of the weights is equal to some kind of characteristic of the sample, for example, the total sample size" (Asparouhov, 2006, p.442). The scaling of the lower level sampling weights in particular has been used as the primary tool for bias reduction in estimation methods including the MPML (Pfeffermann et al., 1998; Stapleton, 2002). Several scaling methods have been proposed and tested. However, there is a lack of agreement in regard to which scaling method is the best (Asparouhov, 2006). For example, Pfeffermann et al. (1998) recommended the use of the size scaling method in their simulation study to reduce bias caused by an informative sampling design. Stapleton (2002) reported that the effective scaling method provided unbiased estimates in MLM estimation. Asparouhov (2006) pointed out that different scaling methods may have different effects depending on the estimation techniques employed. Nonetheless, even with rescaled weights, survey weighted estimators of variance components in MLM could still be grossly biased (Korn & Graubard, 2003). Details on the main scaling methods are provided in the methods section.

The Present Study

This study aims to provide empirical and simulation evidence as well as practical guidance about how to incorporate sampling weights in MLM. The ECLS-K:2011 is used as a typical example of LSAS

data. Specifically, the present study examines the performance of three MPML estimators: a) weighted and without scaling, b) weighted using size scaling, and c) weighted using effective scaling. These three estimators are then compared with an unweighted estimator obtained using MLM analysis of the ECLS-K:2011 data.

Both empirical and simulation investigations are conducted to address the following two research questions (RQs).

> RQ (1): Would the four estimation methods (i.e., unweighted, weighted without scaling, weighted size-scaled and weighted effective-scaled) generate similar or different estimates using two-level models to analyze the ECLS-K:2011 data?
>
> RQ (2): Which estimation method would perform better in a simulation study based on the sampling design of ECLS-K:2011?

There are three main contributions of the present study. First, the empirical and simulation analyses are grounded in realistic sampling designs and parameter estimates values, derived from the ECLS-K:2011 data. The empirical and simulation evidence produced will be directly applicable to analyses of other LSAS education data (e.g., PISA, NAEP) that have a similar sampling design to the ECLS-K:2011. Hence, the results of this study should be of interest to many education researchers.

Second, this study utilizes the STATA software to examine the performance of MPML, instead of Mplus and SAS that have been used in previous work (Cai, 2013; Koziol et al, 2017; Laukaityte & Wiberg, 2018). Because STATA is widely used in education, economics and social sciences, offering new and useful information about how to use STATA to conduct weighted MLM analysis of LSAS data should be valuable to many researchers.

Third, in the simulation component of the study, the quality of the parameter estimators including fixed-effects and variance components is appraised in terms of relative bias (RB), root mean square error (RMSE) and coverage rate (CR) within the context of LSAS data (e.g., ECLS-K:2011). These three criteria have been used to evaluate estimation quality in simulation studies conventionally (Cai, 2013), but have not been fully utilized in the field of education. For example, prior similar simulation studies had only provided point estimates (i.e., the mean) of their simulation results (see Laukaityte & Wiberg, 2018).

Methodology

Data

The ECLS-K:2011 is the latest cycle of the early childhood longitudinal program sponsored by the NCES. As a large-scale longitudinal study, it followed a sample of kindergarten students from diverse ethnic and socioeconomic backgrounds through elementary school grades (i.e., K-5). Data were collected on students, classrooms and schools. The ECLS-K:2011 provided information on children's development in early grades and their early school learning experiences. Researchers may use ECLS-K:2011 data to examine how students' cognitive, social and emotional development may be related to various family, classroom and school variables in grades K-5.

The ECLS-K:2011 adopted a multi-stage complex sampling design that involved clustering, stratification, and unequal probability of selection at different stages. Specifically, a three-stage stratified sampling strategy was employed in which 90 geographic regions served as the primary sampling units (PSUs). Then, samples of public and private schools with 5-year-old children were collected within the sampled PSUs with probabilities proportional to measures of population size (PPS). The population size refers to the total number of 5-year-old children in the population of schools in the U.S. At the third stage, on average nearly 20 students were randomly selected within each sampled school using simple random sampling (SRS) (Tourangeau et al., 2018). The first-stage (sampled PSUs) weights were not provided and prior studies showed that ignoring this level of weights does not have an impact on statistical inference (Stapleton & Kang, 2018). Therefore, the ECLS-K:2011 could be regarded as a two-stage complex sampling design that sampled schools and then students within schools, given that only school- and student-level sampling weights were provided. In the ECLS-K:2011 public data file, the PSU ID was suppressed but school- and student-level overall weights that had been adjusted for nonresponse were provided. We used data from kindergarten. The original data included 18174 students from 968 schools representing the student population of kindergarteners in the U.S. in 2010-2011.

Table 1 presents the descriptive statistics of sampling weights in which there is only one school-level final base weight (W2SCH0), whilst there are 11 student-level final base weights adjusted for nonresponse that are associated with student assessments, parents, teachers, and care-giver questionnaires and interviews in the fall and spring of kindergarten. Balanced repeated replication (BRR) weights (i.e., "W1C1" to "W1C80") and jackknife repeated replication (JRR) weights (i.e., "W1C0STR" and "W1C0PSU) were not considered because these weights were derived from the statistical estimation of the sampling variance and thus are not actual sampling weights. We focused only on sampling weights that are suitable for MLM analysis.

Based on the covariates we used, we selected the student base weight W12ACO because it adjusted for nonresponse associated with the spring kindergarten teacher-level questionnaire and the fall kindergarten child assessment. However, the student base weights W12AC0 could not be used as is, because it also included school base weights (W2SCH0). Therefore, the student specific weights was computed as the ratio of W12AC0 / W2SCH0 to get the pure student-level specific weights. The level-2 (school) base weights W2SCH0 was used as is. Notice that the mean of the school-level weights was 64.24 and the mean of the student-level weights was 3.47. The former was almost 20 times larger than the latter, suggesting that the sampling selection at the school level was the driving force in terms of design informativeness. The analytic sample in MLM included 8486 students in 631 schools for the empirical investigation.

Statistical Model

In education, a two-level model (e.g., individuals nested within clusters), typically means the first-level units are students and the second-level units are schools. We used two-level random intercept models including the null model (without covariates) and two conditional models with covariates, model I and model II. The equations (1) and (2) represent the null model and conditional model respectively:

$$y_{ij} = \beta_0 + u_j + e_{ij}$$
$$u \sim N(0, \sigma_a^2) \quad \varepsilon \sim N(0, \sigma_e^2), \qquad (1)$$
$$y_{ij} = \beta_0 + \boldsymbol{COV}_{(i)j}\boldsymbol{B} + u_j + e_{ij}$$
$$u \sim N(0, \sigma_a^2) \quad \varepsilon \sim N(0, \sigma_e^2), \qquad (2)$$

where *i, j* represent student and school respectively, y_{ij} is the dependent variable, β_0 is the intercept, $\boldsymbol{COV}$ refers to a row vector of predictors at the student and school level, $\boldsymbol{B}$ represents a column vector of regression coefficients, u is a school-level random effect and e is a student-level residual. Both u and e are assumed to follow normal distributions with zero means and variances of σ_a^2 (the between-school variance) and σ_e^2 (the within-school variance) respectively.

For the empirical analysis, the outcome was math scores in the spring of kindergarten. Model I included the following level-1 (student level) variables: prior math scores in the fall, child age in months, gender, race, language spoken at home, SES (i.e., a composite measure of the child's socioeconomic status that included information about parental education, occupation and income), class size (i.e., the actual number of students in a specific classroom), and teacher education, certificate and experience (i.e., years of teaching experience). Model II also included the following level-2 (school level) variables: school location, school sector (i.e., private or public), enrollment in kindergarten, and school SES (i.e., two variables representing the percentage of students in a school eligible for free lunch or reduced-price lunch).

The variables, prior math scores, age, SES, class size, teacher experience, school enrollment, and school SES were continuous variables. The remaining variables were categorical and were coded as follows: a) gender was coded as a dummy variable taking the value of 1 if the student is a female and 0 otherwise, b) four dummy variables were constructed for race (i.e., Black, Hispanic, Asian, and Pacific Islander and American Indian) with Whites being the reference group, c) language spoken at home was coded as a dummy variable taking the value of 1 if English was spoken at home and 0 otherwise, d) teacher education was coded as a dummy variable taking the value of 1 if the teacher had a master's or other advanced degree and 0 otherwise, e) teacher certificate was coded as a dummy taking the value of 1 if the teacher did not have a regular/standard state certificate and 0 otherwise, f) three dummy variables were created for school location (i.e., suburban, town, rural) with city being the reference group, and g) school sector was coded as a dummy variable taking the value of 1 if the school was private and zero otherwise.

Table 1. Descriptive statistics of weights

Weights	Adjustment Description	Mean	SD	Min	Max	UWE
School level (Base weight)						
W2SCH0	Adjusted for nonresponse associated with the school administrator questionnaire	64.24	47.86	0	372.03	1.56
Student level (Base weights)						
W1C0	Adjusted for nonresponse associated with the fall kindergarten child assessment	223.08	148.04	0	958.82	1.44
W1A0	Adjusted for nonresponse associated with the fall kindergarten teacher-level questionnaire	223.08	130.65	0	980.79	1.34
W1T0	Adjusted for nonresponse associated with the fall kindergarten child-level teacher questionnaire	223.08	162.36	0	1012.44	1.53
W1P0	Adjusted for nonresponse associated with the fall kindergarten parent interview	223.08	184.20	0	990.43	1.68
W2P0	Adjusted for nonresponse associated with the spring kindergarten parent interview	223.08	178.44	0	965.66	1.64
W12P0	Adjusted for nonresponse associated with the fall and spring kindergarten parent interviews	223.08	225.12	0	1026.79	2.02
W1_2P0	Adjusted for nonresponse associated with either fall or spring kindergarten parent interview	223.08	141.71	0	956.72	1.40
W12T0	Adjusted for nonresponse associated with the fall and spring kindergarten child-level teacher questionnaires	223.08	177.22	0	1109.05	1.63
W12AC0	Adjusted for nonresponse associated with the spring kindergarten teacher-level questionnaire and the fall kindergarten child assessment	223.08	170.48	0	968.62	1.58
W1PZ0	Adjusted for nonresponse associated with the fall kindergarten parent interview and the before- and after-school care provider questionnaires	223.08	241.24	0	1232.85	2.17
W12PZ0	Adjusted for nonresponse associated with both fall kindergarten and spring kindergarten parent interviews and the before- and after-school care provider questionnaires	223.08	278.74	0	1246.86	2.56

Note: SD=Standard Deviation, UWE=Unequal Weighting Effect, N=18,174

The empirical research question addressed with model I is whether teacher variables are associated with students' math scores controlling for student covariates. The empirical question addressed with model II is whether school characteristics are related to students' math scores controlling for student and teacher covariates. We compared the estimates and statistical inference among four MLM estimators produced from an unweighted model and three weighted models with and without scaling. The modified student-level weights W12AC0 / W2SCH0 and the school-level base weights W2SCH0 were used.

Suppose θ represents all parameters to be estimated, namely, the intercept (β_0), the regression coefficients $\boldsymbol{B}$ and the variances components, σ_a^2 and σ_e^2. The conditional normal likelihood for student *i* in school *j* can be expressed as:

$$L_{ij}(\theta|y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left[-\frac{(y_{ij}-\hat{\bar{y}}_j)^2}{2\sigma_e^2}\right], \quad (3)$$

where $\hat{\bar{y}}_j$ is the estimated cluster or group mean for the j^{th} school depending on the cluster-level variance σ_a^2. Thus, the marginal likelihood for school *j* is

$$L_j(\theta) = \int_{-\infty}^{+\infty} \prod_{i=1}^{N_j} L_{ij}(\theta|y_{ij})\phi(u_j)du_j, \quad (4)$$

where $\phi(u_j)$ is the density function of u_j, and the overall marginal likelihood is

$$L(\theta) = \prod_{j=1}^{M} L_j(\theta). \quad (5)$$

For computational convenience, the log-likelihood form denoted by *l* below is used:

$$l(\theta) = \sum_{j=1}^{M} \log \int_{-\infty}^{+\infty} \{\exp\left[\sum_{n=1}^{N_j} logL_{ij}\,(\theta|u_j)\right]\}\phi(u_j)d_{u_j}. \quad (6)$$

Population data are rarely available to analyze and thus typically sample data are used in empirical data analysis. Rather than using simple random sampling, large-scale data such as the ECLS-K:2011, adopt multi-stage complex sampling designs to obtain national probability samples of well-specified populations in a cost-effective way. Sampling weights are then created based on probabilities of sample selection.

Suppose the data are collected using a two-stage sampling design and the probability of selection at the first stage is p_j (e.g., probability of school selection) and the conditional probability of selection at the second stage is $p_{i|j}$ (e.g., probability of selection of students within the selected schools). The corresponding weights at the first and second stage are w_j and $w_{i|j}$ respectively. To account for potential effects arising due to unequal probabilities of selection, it is important to incorporate the level-specific sampling weights in the log-likelihood function. In a MLM, the log likelihood needs to take into account the cluster-level (e.g., school-level) variance. The MPML is then written in a multilevel model as

$$l(\theta) = \sum_{j=1}^{m} w_j log \int_{-\infty}^{+\infty} \{exp\left[\sum_{n=1}^{n_j} w_{i|j} logL_{ij}\,(\theta|u_j)\right]\}\phi(u_j)d_{u_j}. \quad (7)$$

The sampling weights can be computed as the inverse of the probability of selection at each stage: $\frac{1}{p_j}$ for j^{th} unit (e.g., school) and $\frac{1}{p_{i|j}}$ for i^{th} unit within j^{th} cluster (e.g., a student within a school). Substituting the weights with probabilities, equation (7) becomes

$$l(\theta) = \sum_{j=1}^{m} \frac{1}{p_j} log \int_{-\infty}^{+\infty} \{exp\left[\sum_{n=1}^{n_j} \frac{1}{p_{i|j}} logL_{ij}\,(\theta|u_j)\right]\}\phi(u_j)d_{u_j}. \quad (8)$$

Previous studies had suggested that some scaling procedure is necessary for the individual-level sampling weights (e.g., Stapleton, 2002). Incorporating the scaling constants in equation (7) results in

$$l(\theta) = \sum_{j=1}^{m} w_j\lambda_2 log \int_{-\infty}^{+\infty} \{exp\left[\sum_{n=1}^{n_j} w_{i|j}\lambda_1 logL_{ij}\,(\theta|u_j)\right]\}\phi(u_j)d_{u_j}. \quad (9)$$

In equation (9) above λ_1 is the scaling constant for the model level-1 weights and λ_2 is the scaling constant for the model level-2 weights. The scaling constant λ_2 does not affect the point estimates because the log-pseudolikelihood is multiplied by a scalar and as a result the log-pseudolikelihood is merely rescaled (see Rabe-Hesketh & Skrondal, 2006). Although there is no consensus about which scaling method should be the gold standard for the lower-level sampling weights, two approaches have been proposed to provide the least biased estimates (Asparouhov, 2006; Pfeffermann et al., 1998; Potthoff et al., 1992; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002). They were referred to as size and effective scaling following the language used in STATA. Specifically, the size scaling constant is defined as

$$\lambda_{1_{size}} = \frac{n_j}{\sum_{i=1}^{n_j} w_{i|j}}, \quad (10)$$

and the effective scaling constant is defined as

$$\lambda_{1effective} = \frac{\sum_{i=1}^{n_j} w_{i|j}}{\sum_{i=1}^{n_j} w_{i|j}{}^2}. \quad (11)$$

The motivation to use size scaling is to represent the number of elements in a cluster to reduce bias, but both scaling approaches are a function of the cluster sample size n_j. When the weights variable n_j is fixed, these two scaling methods are equal (Pfeffermann et al., 1998). The application of scaling varies by estimation method (e.g., PWIGLS or MPML) and by software programs (e.g., Mplus, SAS) (see Asparouhov, 2006; Cai, 2013).

With regard to computational algorithms, the MPML estimators could be obtained via any optimization algorithms such as the expectation maximization (EM) algorithm, the accelerated EM algorithm or the Quasi-Newton algorithm (see Asparouhov, 2006). There is no closed form solution for the MPML estimators in a random intercept model without predictors when the cluster sample size is unbalanced. However, in balanced data (i.e., each cluster has the same sample size), the unweighted maximum likelihood estimators (MLE) are available (McCulloch et al., 2008). Using the Laplace approximation, a well-known method for approximating the marginal densities, Asparouhov (2006) derived a closed form solution for the parameters of a random intercept model without predictors when the cluster sample size is constant across all clusters (i.e., a balanced design). Based on the scaling methods illustrated in Asparouhov (2006), we derived the analytic expressions for both the no scaling and scaling cases that corresponded to some of the analyses of our study. These results are provided in the Appendix of this manuscript.

Simulation

The simulation mimicked the ECLS-K:2011 design, a commonly used sampling design for national probability samples (Koziol et al., 2017; Stapleton & Kang, 2018). First, our simulation setup used an informative design as in the ECLS-K:2011, in which the first-sampling stage (i.e., the school) used PPS and the second-sampling stage (i.e., the student) used SRS. Second, the ICC was set as 0.20 based on the data with σ_a^2 =0.25 and σ_e^2=1. This is also the typical clustering effect in U.S. national probability samples of achievement data indicated in the What Works Clearinghouse (What Works Clearinghouse, 2020). Third, the simulation followed an unbalanced design based on the sample sizes in the ECLS-K:2011 data.

Following Pfeffermann et al. (2006), we first used the sample data including the covariates vectors to generate the population values, and then applied the sampling design in ECLS-K:2011 to select 50 schools and 12 students within each selected school. The specific simulation steps are as follows.

Step 1. Generate the random intercept for the j^{th} school. One binary (i.e., public versus private school) and one continuous (i.e., school enrolment in kindergarten) variable were included in the model in the first step:

$$\beta_{oj} = \gamma_0 + r_1 Public + r_2 SchoolEnroll + u_j$$
$$(u_j \sim N(0, \sigma_a^2), j = 1\ to\ 793). \quad (12)$$

Step 2. Generate p_j and w_j and sample 50 schools among 793 schools with p_j. To make the sampling probability p_j close to the distribution of the real data, we first truncated the random effect u_j at the lower tail by -1.5σ_a and the upper tail by 1.5σ_a (see Pfeffermann et al., 1998). Then the informative selection model at the school level was defined as

$$p_j = \frac{1}{1+\exp(4-2u_j)}. \quad (13)$$

The school sampling rate was about 0.02 on average. Schools were sampled with selection probability proportional to school size. The sampling weights w_j were computed as the inverse of the probability of selection p_j.

Step 3. Generate $p_{i|j}$ and sample 12 students with each school. The $p_{i|j}$ is defined by

$$p_{i|j} = \frac{12}{jth\ Cluster\ sample\ size}. \quad (14)$$

The student sampling rate was about 0.65 within each school. Then $w_{i|j}$ was computed as the inverse of $p_{i|j}$.

Step 4. Generate the outcome variable y_{ij} as

$$y_{ij} = \beta_{0j} + \beta_1 female + \beta_2 age + \varepsilon_{ij}$$
$$(\varepsilon_{ij} \sim N(0,1), i = 1\ to\ 12, j = 1\ to\ 50). \quad (15)$$

One binary (i.e., female versus male student) and one continuous (i.e., student age in months) variable were included in the model in the fourth step. Following Pfeffermann et al. (2006), we selected a sample of 50 schools from the original ECLS-K:2011 data, which is treated as the population, and then 12

students were selected from each sampled school (total sample size n = 600) based on the finite population model and sampling schemes. The Monte Carlo simulation process was repeated 1000 times. The true parameters values are listed in Table 6.

Following Eideh and Nathan (2009) and Cai (2013), the quality of estimates was evaluated using three criteria: empirical RB, RMSE and 95% CR. The RB indicated the degree to which the estimate deviates from the true population value including direction (negative or positive) and magnitude. The RMSE was used to measure differences between values predicted by a model (i.e., an estimator) versus the values that were observed. Small values of RB and RMSE indicate a high degree of unbiasedness and precision of the estimators respectively. The 95% CR showed the level of confidence in capturing the true parameter value based on a traditional t-test. Higher CR values indicate a higher degree of confidence in capturing the true population mean value. Muthén and Muthén (2002) suggested that when parameter bias is within 10% of the true value and coverage over 91% in the estimation is considered good (Muthén & Muthén, 2002).

Specifically, the RB is defined as

$$RB = \frac{1}{\theta}\left[\frac{1}{1000}\sum_{x=1}^{1000}\left(\widehat{\theta_x} - \theta\right)\right], \quad (16)$$

and the RMSE is defined as

$$RMSE\left(\hat{\theta}\right) = \sqrt{\left[\frac{1}{1000}\sum_{x=1}^{1000}\left(\widehat{\theta_x} - \bar{\hat{\theta}}\right)^2\right]}, \quad (17)$$

where $\bar{\hat{\theta}} = \frac{1}{1000}\sum_{x=1}^{1000}\widehat{\theta_x}$ (here we used the empirical mean theta to represent the true theta) and *x* represents each of the 1000 iterations. The CR in this study was set at 95%, which is the percentage that a true parameter value falls within the t-test based 95% confidence region of estimates (Cai, 2013).

Results

The design effect (DE) has been widely used to determine the efficiency of survey designs (see Kish, 1965; Kish, 1992; Lohr, 2019). Therefore, it is crucial to consider the DE when evaluating the quality of the MPML estimation in the context of LSAS data. There are two types of design effects: a) one that captures the unequal probability of selection and b) one that captures the clustering of the multi-stage sampling and MLM (Gabler et al., 1999). The unequal weighting effect (UWE) in equation (18) below captures the DE due to disproportional weighting (Chatrchi & Brisebois, 2015), namely

$$UWE(\bar{y}) = \frac{n\sum_i w_i^2}{(\sum_i w_i)^2} = 1 + cv_{w_i}^2,$$

(i=1, 2, …, n), (18)

where $\bar{y}$ is a sample mean, w_i is the final sample weights for the i^{th} individual, and $cv_{w_i}^2$ is the coefficient of variation (CV) of the weights squared, which denotes the relative variance of the sample weights. The UWE provides information about the variability of the weights to better understand how the precision of the estimation might be affected due to the unequal weighting.

In the last column of Table 1, we provided the UWE information for all 12 sampling weights. For student base weights, the UWE values ranged from 1.34 to 2.56. The UWE value for the weight we used (W12ACO) was 1.58. In our simulation study, the mean and the standard deviation of the UWE across students base weights in 1000 simulation iterations was 1.96 and 0.31 respectively. The range of UWE was between 1.28 and 3.01. It appears that the UWE values in our simulation were qualitatively similar to the empirical UWE values of students base weights in the ECLS-K:2011 data.

The other DE captures the clustering effect: $DEFF_{cluster}$= 1 + ($\overline{n_j}$ - 1)*ICC, where $\overline{n_j}$ is the average cluster sample size, and ICC is computed as the ratio of the between-cluster variance to the sum of the between- and within-cluster variance in a two-level model. One rule of thumb is that if $DEFF_{cluster}$ is greater than 2.00, it is necessary to take into account the design effect due to clustering effects in model analysis (Kish, 1965). In our study, the ICC was 0.19 and the average cluster sample size $\overline{n_j}$ was 22.87. Plugging in these values to the $DEFF_{cluster}$ formula yields a value of about 5.15, which is much larger than 2.00. Therefore, in this case it was necessary to incorporate clustering into account in the MLM analyses.

Table 2 displays the descriptive statistics of the variables used in the empirical analyses. Weighted and unweighted means were provided for all variables. The original unmodified student base weight (W12AC0)

was used in the weighted analysis. The last two columns of Table 2 report the difference between the weighted and the unweighted means and the ratio of the weighted to the unweighted means.

A difference of zero or close to zero indicates that the weighted and the unweighted means are equal. The difference between the weighted and the unweighted mean was large for class size and school enrollment in kindergarten. This implies that researchers may need to check the estimates of the weighted versus the unweighted analyses for these two variables, assuming they were important predictors. Also, researchers could include these two variables in the model to further control for the sampling design effect.

With respect to the ration index, a value of one indicates the two means are equal and weighting did not make a different. However, departures from one that were greater than 30% for example (i.e., a ratio greater than 1.30 or less than 0.70) were found in race, SES, and school location and sector. Researchers could include these variables in the model to further control for the sampling design effect.

Tables 3 to 5 report results of the null (intercept only) model as well as models I and II. The null (intercept only) model estimates displayed in Table 3 indicated that the mean values of the intercept are similar across the four estimation approaches (i.e., unweighted, weighted without scaling, weighted size-scaled, weighted effective-scaled). The standard error of the unweighted mean was slightly smaller than the standard errors of the three remaining weighted means. In addition, the weighted unscaled estimate of the level-2 variance was larger than the estimates from the remaining three estimation approaches (i.e., unweighted, size-scaled and effective-scaled estimation). On the contrary, the weighted unscaled estimate of the level-1 variance was the smallest compared with the estimates from the other three estimation methods. The standard errors of the variance estimates were overall similar, but the unweighted approach had the smallest standard error for level 1 variance and the weighted unscaled approach generated slightly larger standard errors for level 1 and level 2 variance than the other three approaches.

The results of model I are summarized in Table 4. For the dichotomous variables, the effect sizes were standardized mean differences and for the continuous variables the effect sizes were standardized regression coefficients. The effect sizes reported in Table 5 were also computed the same way.

Results were different in statistical significance across four estimation methods for three variables. Age reached statistical significance at the 0.05 level when weights were used either with or without scaling, but the unweighted age estimate was not significant. The Hispanic-White achievement gap in math was statistically significant only in the unweighted estimation. The effect size estimates were all small. Teacher experience was statistically significant only when the unweighted estimation was used.

For the remaining variables, statistical significance was same across four estimation methods. The female estimates were consistently non-significant at the 0.05 level across the four estimation approaches. The effect sizes were close to zero and were smaller in magnitude when scaling was used. In the same vein, the p-values were larger when scaling was used. The estimates for Black students were consistently statistically significant across estimation methods with small p-values. The effect sizes indicated a Black-White achievement gap in math of nearly one-sixth of a standard deviation favoring White students. The coefficients of Asian students were non-significant across all four estimation methods and the effect sizes were small. The estimates of English language were consistently non-significant at the 0.05 level across the four estimation approaches. For Native Islander & American Indian, although results were not significant, estimates of unweighted and weighted with and without scaling were different. The SES coefficient was consistently significant and positively related with math scores and the p-values were very small. The class size effects were consistently non-significant and the effect size estimates were close to zero, especially when scaling was used. The p-values were large when scaling was used. Teacher education and certification were also consistently non-significant

. Lastly, the level-1 variance was constantly significant across estimation methods and the estimates and standard errors were similar. Nevertheless, the estimate obtained from the weighted unscaled approach was the smallest. The level-2 variance was also constantly significant at the 0.05 level. However, the variance estimate obtained from

Table 2. Means of variables used in the analyses

Variables	N (Unweighted)	(A) Mean (Unweighted)	(B) Mean (Weighted)	(B) - (A)	(B)/(A)
Outcome					
Math scores (spring 2011)	17143	48.51	48.16	-0.35	0.99
Student variables					
Math scores (fall 2010)	15595	34.49	34.23	-0.27	0.99
Age	15775	67.45	67.53	0.08	1.00
Male	9288	0.51	0.51	0.00	1.01
Female	8847	0.49	0.49	0.00	0.99
White	8489	0.47	0.56	0.10	1.21
Black	2397	0.13	0.14	0.01	1.04
Hispanic	4590	0.25	0.25	0.00	0.98
Asian	1543	0.08	0.04	-0.05	0.46
Native Islander & American Indian	285	0.01	0.01	0.01	2.25
Other language spoken at home	2941	0.16	0.16	0.00	0.99
English spoken at home	12926	0.71	0.84	0.13	1.18
SES	15977	-0.05	-0.09	-0.04	1.74
Class size	13369	20.08	15.22	-4.86	0.76
Teacher has a bachelor degree or less	8403	0.55	0.54	-0.01	0.98
Teacher has a master's degree or more	6852	0.45	0.46	0.02	1.04
Teacher has regular or standard certificate	13455	0.74	0.93	0.19	1.25
Teacher has other certificate	1110	0.06	0.07	0.01	1.22
Teacher experience	15241	14.61	14.34	-0.28	0.98
School covariates					
City school	5963	0.33	0.30	-0.03	0.92
Suburban school	6340	0.35	0.34	-0.01	0.96
Town school	1337	0.07	0.11	0.03	1.44
Rural school	3885	0.21	0.26	0.04	1.19
Public school	15602	0.86	0.92	0.06	1.07
Private school	2189	0.12	0.08	-0.04	0.66
School enrollment in kindergarten	17758	87.24	94.45	7.21	1.08
Percentage of students eligible for free lunch	17791	43.28	43.16	-0.11	1.00
Percentage of students eligible for reduced lunch	17791	7.68	7.47	-0.20	0.97

Note: Student base weight was used in this analysis: W12AC0

Table 3. Estimates of MLM analysis: Null model

	Unweighted		Weighted unscaled		Weighted (size scaling)		Weighted (effective scaling)	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Intercept	48.82 *	0.26	48.93 *	0.30	49.02 *	0.30	49.02 *	0.30
Variance of school residual	32.53 *	2.44	39.66 *	2.49	32.27 *	2.42	32.07 *	2.42
Variance of student residual	124.42 *	1.98	118.89 *	2.61	121.62 *	2.44	121.56 *	2.44

Note: Number of students = 8486, number of schools = 631, Est. = estimate, SE = standard error, *p < 0.05

the weighted unscaled approach was the largest. By and large the estimates and their standard errors produced by the two scaling estimation methods were very similar. Also, when weighting was used, the standard errors of the estimates were typically larger than those obtained from the unweighted estimation approach for variance estimators.

Table 5 provides the results of model II that also included school variables. Generally, the estimates, standard errors, effect sizes and p-values of the student variables reported in Table 5 were similar to those reported in Table 4.

With respect to the school variables, across four estimation methods, two variables had different results whereas other remaining variables had the same results (i.e., rural school and private school) in statistical significance. The rural school estimate was statistically significant only when the weighted unscaled approach was used, and the remaining three estimates were not significant. The suburban or town school estimates were continuously non-significant. In addition, school enrollment in kindergarten and school SES were also consistently non-significant. The sector estimates reached statistical significance only in the two weighted and scaled estimation approaches. The coefficients indicated larger means in mathematics for students in public schools, compared to private schools, net of the effects of the other predictors in the model. The magnitude of the corresponding effect sizes was approximately one-tenth of a standard deviation favoring public schools. All other effect size estimates of school variables were small and close to zero.

Table 6 presents the simulation results for four regression coefficients, the intercept, and level-1 and level-2 variances. The unbiasedness of the seven parameter estimators was appraised using three commonly used criteria, namely RB, RMSE and 95% CR. The RB values were universally low and very close to zero, and the weighted and unweighted estimation provided similar results except for the level-2 variance estimator, which had negative bias especially for the unweighted estimator.

The small RMSE values for the female, age and school enrollment estimators as well as the first and second level variances estimators indicated better estimation compared with the large RMSE values for the intercept and public school estimators which indicated poorer estimation. The unweighted estimators were advantageous to the weighted estimators overall. Specifically, the unweighted estimation resulted in slightly smaller RMSE values for the intercept and the public school estimates compared to the values obtained from the three weighted estimation methods (with and without scaling). It appears that the mean estimators of the binary variables (i.e., female and public school) and the intercept had lower estimation quality than those of the continuous variables (i.e., age and school enrollment).

In regard to the 95% CR, all values were greater than 91%, the lower bound that suggests good estimation (see Muthén & Muthén, 2002), except for the CR values of the second level variance that were smaller than 91%. In particular, the CR value of the second level variance estimator using the unweighted estimation was 19% only. The CR value of the second level variance estimator using weighted estimation without scaling was much higher, namely 86%. Overall, the second level variance estiamtors were underestimated. Nevertheless, the estimation of the first level variance was good.

In summary, by and large, there were no noticeable differences between the weighted and the unweighted estimation methods with respect to bias. However, the level of bias with respect to the estimation of the second level variance was higher when the unweighted estimation was used and much lower when the

Table 4. Estimates of MLM analysis: Model I

	Unweighted				Weighted unscaled				Weighted (size scaling)				Weighted (effective scaling)			
	Est.	SE	P	ES	Est.	SE	P	ES	Est.	SE	P	ES	Est.	SE	P	ES
Intercept	21.07 *	1.32	<0.001	-	21.07 *	1.68	<0.001	-	22.40 *	1.73	<0.001	-	22.42 *	1.74	<0.001	-
Math fall 2010	0.87 *	0.01	<0.001	0.80	0.89 *	0.01	<0.001	0.81	0.88 *	0.01	<0.001	0.80	0.88 *	0.01	<0.001	0.80
Age	-0.03	0.02	0.08	-0.01	-0.05 *	0.02	0.03	-0.02	-0.05 *	0.02	0.04	-0.02	-0.05 *	0.02	0.04	-0.02
Female	0.23	0.15	0.11	0.02	0.31	0.17	0.06	0.02	0.16	0.17	0.33	0.01	0.16	0.17	0.34	0.01
Black	-2.16 *	0.27	<0.001	-0.17	-1.90 *	0.33	<0.001	-0.15	-1.97 *	0.29	<0.001	-0.16	-1.97 *	0.29	<0.001	-0.16
Hispanic	-0.51 *	0.24	0.04	-0.04	-0.35	0.29	0.22	-0.03	-0.37	0.27	0.17	-0.03	-0.37	0.27	0.16	-0.03
Asian	-0.51	0.37	0.17	-0.04	-0.36	0.46	0.43	-0.03	-0.61	0.46	0.19	-0.05	-0.61	0.47	0.20	-0.05
Native Islander & American Indian	-0.46	0.70	0.51	-0.04	-0.01	0.57	0.98	0.00	0.50	0.51	0.33	0.04	0.51	0.51	0.32	0.04
English language	-0.39	0.26	0.13	-0.03	-0.34	0.29	0.25	-0.03	-0.41	0.30	0.17	-0.03	-0.39	0.30	0.19	-0.03
SES	0.96 *	0.12	<0.001	0.06	1.16 *	0.14	<0.001	0.07	1.12 *	0.14	<0.001	0.07	1.11 *	0.14	<0.001	0.07
Class size	0.05	0.03	0.10	0.02	0.07	0.05	0.12	0.02	0.02	0.04	0.60	0.01	0.02	0.04	0.61	0.01
Teacher has a master's degree or more	-0.02	0.19	0.90	0.00	0.11	0.22	0.62	0.01	-0.18	0.22	0.40	-0.01	-0.17	0.22	0.42	-0.01
Teacher has other certificate	-0.09	0.34	0.78	-0.01	-0.40	0.41	0.32	-0.03	-0.14	0.40	0.73	-0.01	-0.13	0.41	0.75	-0.01
Teacher experience	-0.02 *	0.01	0.01	-0.02	-0.02	0.01	0.05	-0.02	-0.02	0.01	0.11	-0.01	-0.02	0.01	0.12	-0.01
Variance of school residual	8.91 *	0.70			12.03 *	0.92			9.34 *	0.87			9.28 *	0.87		
Variance of student residual	42.12 *	0.67			39.66 *	0.86			41.56 *	0.88			41.61 *	0.89		

Note: Number of students = 8486, number of schools = 631, Est. = estimate, SE = standard error, P = P-value, ES = effect size, *p < 0.05

Table 5. Estimates of MLM analysis: Model II

	Unweighted				Weighted Unscaled				Weighted (size scaling)				Weighted (effective scaling)			
	Est.	SE	P	ES	Est.	SE	P	ES	Est.	SE	P	ES	Est.	SE	P	ES
Intercept	21.12 *	1.45	<0.001		21.28 *	1.83	<0.001		23.13 *	1.86	<0.001		23.16 *	1.87	<0.001	
Math fall 2010	0.87 *	0.01	<0.001	0.80	0.89 *	0.01	<0.001	0.81	0.88 *	0.01	<0.001	0.80	0.88 *	0.01	<0.001	0.80
Age	-0.03	0.02	0.06	-0.01	-0.05 *	0.02	0.02	-0.02	-0.05 *	0.02	0.02	-0.02	-0.05 *	0.02	0.02	-0.02
Female	0.23	0.15	0.12	0.02	0.31	0.17	0.06	0.02	0.15	0.17	0.35	0.01	0.15	0.17	0.36	0.01
Black	-2.19 *	0.28	<0.001	-0.17	-1.90 *	0.34	<0.001	-0.15	-2.01 *	0.31	<0.001	-0.16	-2.01 *	0.31	<0.001	-0.16
Hispanic	-0.50 *	0.25	0.04	-0.04	-0.34	0.29	0.25	-0.03	-0.34	0.27	0.20	-0.03	-0.35	0.27	0.20	-0.03
Asian	-0.46	0.38	0.22	-0.04	-0.33	0.46	0.47	-0.03	-0.53	0.47	0.26	-0.04	-0.53	0.48	0.27	-0.04
Native Islander & American Indian	-0.52	0.70	0.46	-0.04	-0.03	0.56	0.95	0.00	0.43	0.51	0.40	0.03	0.44	0.52	0.39	0.04
English language	-0.43	0.26	0.10	-0.03	-0.36	0.29	0.23	-0.03	-0.46	0.30	0.12	-0.04	-0.44	0.30	0.14	-0.04
SES	1.03 *	0.12	<0.001	0.07	1.18 *	0.15	<0.001	0.08	1.20 *	0.14	<0.001	0.08	1.20 *	0.14	<0.001	0.08
Class size	0.05	0.03	0.07	0.02	0.08	0.05	0.11	0.03	0.02	0.04	0.58	0.01	0.02	0.04	0.59	0.01
Teacher has a master's degree or more	-0.03	0.19	0.89	0.00	0.11	0.22	0.62	0.01	-0.21	0.22	0.33	-0.02	-0.21	0.22	0.34	-0.02
Teacher has other certificate	-0.04	0.34	0.90	0.00	-0.37	0.41	0.36	-0.03	-0.04	0.40	0.93	0.00	-0.03	0.41	0.95	0.00
Teacher experience	-0.02 *	0.01	0.02	-0.02	-0.02	0.01	0.05	-0.02	-0.02	0.01	0.16	-0.01	-0.02	0.01	0.16	-0.01
School covariates																
Suburban school	0.07	0.37	0.86	0.01	0.11	0.42	0.80	0.01	-0.01	0.43	0.99	0.00	-0.01	0.43	0.97	0.00
Town school	0.42	0.54	0.43	0.03	0.04	0.56	0.94	0.00	-0.12	0.56	0.84	-0.01	-0.12	0.57	0.83	-0.01
Rural school	0.72	0.40	0.07	0.06	1.03 *	0.48	0.03	0.08	0.71	0.46	0.13	0.06	0.70	0.46	0.13	0.06
Private school	-0.78	0.52	0.14	-0.06	-1.06	0.58	0.07	-0.08	-1.38 *	0.58	0.02	-0.11	-1.39 *	0.58	0.02	-0.11
School enrollment in kindergarten	-0.01	0.00	0.10	-0.02	-0.01	0.00	0.07	-0.03	-0.01	0.00	0.11	-0.02	-0.01	0.00	0.11	-0.02
Percentage of students eligible for free lunch	0.01	0.01	0.13	0.02	0.01	0.01	0.34	0.02	0.01	0.01	0.44	0.01	0.01	0.01	0.45	0.01
Percentage of students eligible for reduced lunch	-0.02	0.02	0.24	-0.01	-0.02	0.02	0.26	-0.01	-0.02	0.02	0.16	-0.02	-0.02	0.02	0.16	-0.02
Variance of school residual	8.71 *	0.69			11.66 *	0.90			8.97 *	0.86			8.92 *	0.87		
Variance of student residual	42.10 *	0.67			39.65 *	0.86			41.54 *	0.88			41.59 *	0.89		

Note: Number of students = 8486, number of schools = 631, Est. = estimate, SE = standard error, P = P-value, ES = effect size, *p < 0.05

Table 6. Simulation Results

	True Value	Mean	RB	RMSE	95%CR
Female	**0.16**				
UNW		0.16	-0.02	0.08	95%
WNS		0.15	-0.04	0.12	96%
WSZ		0.15	-0.04	0.11	96%
WEF		0.15	-0.04	0.11	96%
Age	**0.50**				
UNW		0.50	0.00	0.01	95%
WNS		0.50	0.00	0.01	94%
WSZ		0.50	0.00	0.01	94%
WEF		0.50	0.00	0.01	94%
Public school	**4.46**				
UNW		4.47	0.00	0.27	97%
WNS		4.37	-0.02	0.39	96%
WSZ		4.37	-0.02	0.39	96%
WEF		4.37	-0.02	0.39	96%
School enrollment	**0.01**				
UNW		0.01	0.00	0.00	95%
WNS		0.01	0.03	0.00	94%
WSZ		0.01	0.03	0.00	95%
WEF		0.01	0.03	0.00	95%
Intercept	**10.02**				
UNW		10.36	0.03	0.70	92%
WNS		10.15	0.01	0.98	93%
WSZ		10.14	0.01	0.97	94%
WEF		10.14	0.01	0.97	94%
Second level variance	**0.25**				
UNW		0.13	-0.49	0.04	19%
WNS		0.19	-0.25	0.07	86%
WSZ		0.16	-0.38	0.07	68%
WEF		0.16	-0.38	0.07	68%
First level variance	**1.00**				
UNW		0.99	-0.01	0.06	95%
WNS		0.96	-0.04	0.08	93%
WSZ		0.99	-0.01	0.08	95%
WEF		0.99	-0.01	0.08	95%

Note: UNW=Non-weighted, WNS=Weighted no scaling, WSZ=Weighted size scaling, WEF=Weighted effective scaling

weighted estimation without scaling was utilized. The results for the second level variance showed a consistent pattern with that in the empirical analyses reported in Tables 4 and 5. In particular, the weighted unscaled estimates were larger than the estimates produced by the other three methods and the unweighted estimates were the smallest. In addition, the two scaling methods produced identical results, which is consistent with the empirical finding.

Discussion

Whether and how to incorporate sampling weights in statistical analysis depends on many factors such as the convention in a discipline, the sampling design, the research questions and the statistical models. To analyze data of LSAS that have adopted a complex multi-stage sampling design such as the ECLS-K:2011, researchers may employ MLM such as random intercept two-level models to estimate the between- and within-cluster variances as well as the regression coefficients of the predictors in the model. Incorporating sampling weights in MLM analysis has been of research interest in the literature, and data user manuals of LSAS recommend the use of sampling weights in statistical analysis. However, how to incorporate weights in analyses of LSAS data remains unclear to many educational researchers, and, thus, practical guidance in this area is seriously needed.

This study filled in this literature gap. First, we demonstrated empirically how to select and apply sampling weights in statistical analysis of the ECLS-K:2011 data using two-level models. Second, we conducted a Monte Carlo simulation to appraise the performance of the MPML methodology including two scaling options and juxtaposed the results with those obtained via unweighted analysis. The findings of this study are directly applicable to the ECLS-K:2011 data and other data collected from LSAS with similar sampling designs.

The estimation of variance components is of particular interest in MLM. With respect to the second level variance, σ_a^2, the unweighted estimation produced more negative bias compared to the weighted estimation with and without scaling. In terms of the individual-level variance, σ_e^2, the weighted estimation without scaling generated slightly more negative bias compared to the other three estimation methods. These findings are in congruence with the analytic expressions displayed in the Appendix of this manuscript and with findings reported in prior studies (Cai, 2013; Pfeffermann et al., 1998).

With regard to findings on fixed effects (e.g., regression coefficients), the estimators obtained from the simulation were overall close to the corresponding true values. It appears the performance of the estimation methods was better for simulated continuous variables than for simulated binary variables. One possible explanation is that continuous variables have naturally more variability than binary variables and thus the estimation may be more precise.

Prior studies had suggested that applying scaling is essential for reducing estimation bias in weighted MLM. In this study, the performance of size and effective scaling methods is very similar in the empirical analysis of the ECLS-K data. The simulation results also indicated that size and effective scaling performed similarly. This finding is not consistent with previous findings (Pfeffermann et al., 1998, Stapleton, 2002). Specifically, Pfeffermann et al. (1998) found that size scaling was preferred, whereas Stapleton (2002) found that effective scaling provided unbiased estimators of key parameters. Our results however indicate that the type of scaling did not affect the estimation in the context of LSAS at least for ECLS-K:2011 data.

To summarize, the second and first level variance estimates of σ_a^2 and σ_e^2 from the empirical analyses showed consistent statistical significance across the four estimation approaches. The simulation results indicated that both the unweighted and the weighted estimators had negative RB values. However, the RB values were more pronounced for the second level variance. With respect to the fixed effects estimators, results from the empirical analysis demonstrated some variability across the estimation methods. However, the simulation results produced fixed effects estimators that were quite homogeneous across estimation methods, which we discussed as a limitation at the end.

Practical Considerations

Education researchers may have some practical questions about which sampling weights to use and when and how to incorporate the sampling weights in MLM analyses when using LSAS data. This section

provides a brief discussion about practical considerations researchers could follow when contemplating the use of sampling weights in MLM analyses of LSAS data.

First and foremost, researchers need to read the data user manual carefully to attain a good understanding of the complex multi-stage sampling design used in the LSAS of interest. It is important to determine whether the sampling design is informative or non-informative (i.e., whether unequal probability sampling was used or not) at each sampling stage. If simple random sampling was used to select units in all sampling stages, it would not be necessary to apply any sampling weights in the statistical analysis. Unweighted analysis would be preferrable for non-informative designs, which has the advantage of providing efficient, consistent and unbiased estimators (see Cai, 2013; Pfeffermann et al., 1998). However, if unequal probability sampling is adopted in some stages, which indicates an informative design, using sampling weights in the analysis would be imperative to make projections of statistical inference from the sample to the population (Asparouhov, 2006; Pfeffermann et al., 1998). Data user manuals of LSAS typically suggest the use of sampling weights in statistical analyses. However, it is recommended that the researcher examines all sampling weights variables that are available in the data set, and chooses appropriate sampling weights variables based on their research questions and outcome and predictor variables used from different survey questionnaires. Then, it would be informative to compute the UWE to empirically check and quantify the degree of informativeness of the sampling design to confirm the need of applying sampling weights in the analysis as we have demonstrated in this study using the ECLS-K:2011 data.

Second, it is recommended that researchers check the availability of sampling weights at different levels of the hierarchy. If, for example, only one overall sampling weights variable is available in the dataset, a weighted single-level statistical model should perhaps be used. However, when sampling weights are available at different levels, applying weights at the appropriate levels is recommended (Asparouhov, 2006). If sampling weights are missing at some levels but not at other levels, applying weights at one level but not at the other levels may produce more biased estimates compared to estimates obtained from unweighted analyses (Grilli & Pratesi, 2004). This means if weights are missing at certain levels, one should conduct unweighted MLM analysis instead of a weighted analysis. Sampling weights are typically provided at different levels in LSAS, but researchers still need to select the appropriate sampling weights to use based on their model covariates and outcomes. It is because there is difference in non-response adjustment for child assessment outcome variables in spring or fall as well as for predictor variables from parent, teacher, and before- and after-school provider questionnaires as showed in Table 1 for the ECLS-K: 2011.

In addition, it is essential to ensure that when conducting MLM analysis level-specific weights should be used at each level instead of overall sampling weights. This is because there might be an overlap between the final weights at different levels. For example, as we showed in the empirical analysis of the ECLS-K:2011 data, the student weights incorporated a school weights component, which needed to be purged from the student weights. In our case we divided the student-level final sampling weights by the school-level final sampling weights to get the non-overlapped student-level specific sampling weights. Researchers may have to do similar modifications of sampling weights variables as needed.

Third, it is manageable to implement the MPML estimation method in STATA and the two scaling options are easy to use. Specifically, researchers would simply need to incorporate the "pwscale (size)" or "pwscale (effective)" in STATA "mixed" command. To illustrate, one simple syntax code for a two-level random intercept model is: mixed Y Xs [pw=student-level specific weights] || Cluster ID: , pweight (school-level weights) pwscale(size). The empirical results of the present study showed that researchers could use either the size or the effective scaling in MLM analysis of LSAS data and the generated results would be very similar.

Fourth, when the variance estimators are of key interest, the weighted estimation method without scaling performed better in estimating the second level variance, compared to the unweighted or the weighted scaling methods. However, the weighted estimation method without scaling did not perform as well as the other estimation methods in estimating the first level variance. When the fixed effects estimators are the

main focus, weighted analyses need to be conducted if sampling design is informative. By and large, unweighted estimation method generated the lowest standard errors in empirical models and lower RMSE values in simulation investigation compared with the three weighted estimation approaches. This is a disadvantage of weighted estimation methods (Pfeffermann et al., 2006; Shen & Konstantopoulos, 2022).

Limitation

One potential limitation of this study is that in our simulation, we did not include simulation evaluations with regard to the bias for fixed effects estimates across four estimation approaches. Future research may add a simulation component that associates the covariates and the error term that is due to unequal probabilities of selection. In that way, it would provide clear evidence about which estimation method would be preferred for fixed effects under the informative design in LSAS data.

References

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

Asparouhov, T., & Muthen, B. (2007). Testing for informative weights and weights trimming in multivariate modeling with survey data. *Proceedings of the 2007 JSM meeting, Section on Survey Research Methods*, Salt Lake City, Utah.

Binder, Kovacevic, & Roberts. (2005). How important is the informativeness of the sample design. *Proceedings of the Survey Methods Section,* Canada Ottawa.

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., & Berzofsky, M. E. (2016). Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Annual Review of Statistics and Its Application*, 3(1), 375–392.

Cai, T. (2013). Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses. *Sociological Methodology, 43*(1), 178-219.

Chantala, K., & Suchindran, C. (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.

Chatrchi, G., & Brisebois, F. (2015). Survey weighting adjustments and the design effect: A case study. *Proceedings of the American Statistical Association's Section on Survey Section on Survey Research Methods–JSM.*

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.

Gabler, S., Häder, S., & Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 105-106.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43-56.

Graubard, B. I., & Korn, E. L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research, 5*(3), 263-281.

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.

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.

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.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

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

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 multilevel

modelling 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.

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

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284.

Lohr, S. L. (2019). *Sampling: design and analysis*. Chapman and Hall/CRC.

Martin, M. O., & Mullis, I. V. (2012). Methods and procedures in TIMSS and PIRLS 2011. *Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College*.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear and mixed models*. Wiley.

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: A Multidisciplinary Journal*, 9(4), 599–620.

OECD. (2014). *Survey weighting and the calculation of sampling variance*. PISA 2012 technical report.https://www.oecd.org/pisa/pisaproducts/PISA%202012%20Technical%20Report_Chapter%208.pdf

Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review,* 61(2), 317-337.

Pfeffermann, D., Moura, F. A. D. S., & Silva, P. L. D. N. (2006). Multi-level modelling under informative sampling. *Biometrika, 93*(4), 943-959.

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

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 the American statistical Association, 87*(418), 383-396.

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.

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

Shen, T., & Konstantopoulos, S. (2022). Complex sampling designs in large-scale education surveys: A two-level sample distribution approach. *The Journal of Experimental Education*, *90*(2), 469-485.

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

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475-502.

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., K., W.-A., Vaden-Kiernan, N., Blaker, L., & Najarian, M. (2018). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLSK: 2011): User's manual for the ECLS-K: 2011 kindergarten– Fourth Grade data file and electronic codebook (No. NCES 2018-032)*.

West, B. T., & Galecki, A. T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician, 65*(4), 274-282.

What Works Clearinghouse. (2020). What Works Clearinghouse standards handbook, version 4.1. US Department of Education, Institute of Education Sciences. *National Center for Education Evaluation and Regional Assistance.* https://ies.ed.gov/ncee/wwc/Docs/referencere

sources/WWC-Standards-Handbook-v4-1-508.pdf

Citation:
Shen, T., & Konstantopoulos, S. (2022). Incorporating Complex Sampling Weights in Multilevel Analyses of Education Data. *Practical Assessment, Research & Evaluation*, 27(13). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/13/

Corresponding Author:
Ting Shen
Missouri University of Science and Technology
Rolla, MO USA

Email: tingshen [at] mst.edu

Appendix

This appendix provides analytic expressions for the intercept in a null (intercept only) two-level model using weighted and unweighted estimation methods. A balanced design is assumed (i.e., the cluster size is the same for all clusters in the sample) and the second level cluster variance σ_a^2 is assumed to be positive (see McCulloch et al., 2008).

Then, the analytic expression of the unweighted estimators:

$$\widehat{\beta_0} = \bar{y}_{..}, \widehat{\sigma_e{}^2} = \frac{\sum_j \sum_i (y_{ij} - \bar{y}_j)^2}{m(n-1)}, \widehat{\sigma_a^2} = \frac{\sum_j (\bar{y}_j - \bar{y}_{..})^2}{m} - \frac{\widehat{\sigma_e{}^2}}{n}, \quad \text{(A-1)}$$

where $\bar{y}_{..}$ is the grand mean, $\bar{y}_j$ is the cluster mean, m is the number of clusters (e.g., schools), and n is the cluster size, which is the same for each cluster when data are balanced and m is the number of clusters.

We followed Asparouhov (2006) to derive the analytic expressions of the weighted estimators without scaling

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e{}^2} = \frac{\sum_j w_j \sum_i w_{i|j} (y_{ij} - \bar{y}_j)^2}{\sum_j w_j (n-1)}, \widehat{\sigma_a^2} = \frac{\sum_j w_j (\bar{y}_j - \bar{y}_{..})^2}{\sum_j w_j (n-1)} - \frac{\widehat{\sigma_e{}^2}}{n}, \quad \text{(A-2)}$$

the weighted estimators with size scaling

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e{}^2} = \frac{\sum_j n\, w_j \sum_i w_{i|j} (y_{ij} - \bar{y}_j)^2}{\sum_j w_j (n-1) \sum_i w_{i|j}}, \widehat{\sigma_a^2} = \frac{\sum_j w_j (\bar{y}_j - \bar{y}_{..})^2}{\sum_j w_j (n-1)} - \frac{\widehat{\sigma_e{}^2}}{n}, \quad \text{(A-3)}$$

and the weighted estimators with effective scaling

$$\widehat{\beta_0} = \frac{\sum_j w_j \bar{y}_j}{\sum_j w_j}, \widehat{\sigma_e{}^2} = \frac{\sum_j (\sum_i w_{i|j})\, w_j \sum_i w_{i|j} (y_{ij} - \bar{y}_j)^2}{\sum_j w_j (n-1) (\sum_i w_{i|j}{}^2)}, \widehat{\sigma_a^2} = \frac{\sum_j w_j (\bar{y}_j - \bar{y}_{..})^2}{\sum_j w_j (n-1)} - \frac{\widehat{\sigma_e{}^2}}{(\sum_i w_{i|j})^2 / (\sum_i w_{i|j}^2)}. \quad \text{(A-4)}$$