Listening. Learning. Leading.®

Research Report
ETS RR–11-13

The Evaluation of Bias of the Weighted Random Effects Model Estimators

Yue Jia

Lynne Stokes

Ian Harris

Yan Wang

April 2011

The Evaluation of Bias of the Weighted Random Effects Model Estimators

Yue Jia
ETS, Princeton, New Jersey

Lynne Stokes, Ian Harris, and Yan Wang
Southern Methodist University, Dallas, Texas

April 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Dan Eignor

Technical Reviewers: Sandip Sinharay Jiahe Qian

Abstract

Estimation of parameters of random effects models from samples collected via complex multistage designs is considered. One way to reduce estimation bias due to unequal probabilities of selection is to incorporate sampling weights. Many researchers have been proposed various weighting methods (Korn, & Graubard, 2003; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998) in estimating the parameters of hierarchical models, including random effects models as a special case. In this paper, the bias of the weighted analysis of variance (ANOVA) estimators of the variance components for a two-level, one-way random effects model is evaluated. For these estimators, analytic bias expressions are first developed, the expressions are then used to examine the impact of sample size, intraclass correlation coefficient (ICC), and the sampling design on the bias of the estimators. In addition, two-stage sampling designs are considered, with a general probability design at the first stage (Level 2) and simple random sampling without replacement (SRS) at the second stage (Level 1). The study shows that first-order weighted variance component estimators perform well when for moderate cluster sizes and ICC values. However, noticeable estimation bias can be found with this weighting method for small cluster sizes (less than 20), particularly when ICC is small (less than 0.2). In such scenarios, scaled first-order weighted estimators can be an alternative. This paper is discussed in the context of National Assessment of Educational Progress (NAEP) 2003 4th Grade Reading National and State Assessment data, with Level 1 being the student level and Level 2 being the school level.

Key words: random effects model, variance components, estimation bias, ANOVA estimators, complex sampling designs, selection probability, sampling weights, ICC, NAEP

Acknowledgments

This research for the first author was partially supported by a grant from the American Educational Research Association, which receives funds for its AERA Grants Program from the National Science Foundation and the National Center for Education and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant #REC-0310268. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

The authors would like to thank Sandip Sinharay, Jiahe Qian, Daniel Eignor and two external reviewers for their invaluable comments on a draft of this manuscript. We also gratefully acknowledge Kim Fryer for her editorial assistance. In addition, the first author would like to thank American Educational Research Association for supporting the early development of this research.

Table of Contents

List of Tables

Page

List of Figures

Page

1. Introduction

The National Assessment of Educational Progress (NAEP) is a large-scale educational assessment designed to give information on what U.S. students know and can do. Data for the NAEP are collected from a complex multistage sample of schools and students, therefore sampling weights are required for proper analysis of these data. Online documentation from the National Center for Education Statistics (NCES) provides secondary data analysts with information on how to use weights on the NAEP data file when estimating means, population totals, and regression coefficients but nothing on how to use weights when fitting hierarchical models. Because these models are increasingly popular in educational research and several different weighting methods have been proposed for estimating the model parameters, guidance for data analysts is needed. The motivation for the research reported here was to offer such guidance for secondary analysts of NAEP data.

Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash (1998) and Graubard and Korn (1996) presented two methods for incorporating sampling weights in estimation of hierarchical models. The former used only first-order weights and the latter used both first- and second-order weights. First-order weights are (before adjustments for nonsampling errors) reciprocals of the inclusion probabilities of sampling units, while second-order weights are reciprocals of the joint inclusion probabilities of pairs of units. Estimates for parameters of hierarchical models that use only first-order weights are currently available in commercial software (e.g., HLM 6.0, MLWIN, LISREL, and Stata GLLAMM), but those using second-order weights are not available. Further, second-order weights are not typically provided on data files, so users have to produce them from knowledge of the sampling design, which is difficult for all but the most expert users.

Estimators that are linear in the data (such as estimators of totals) are design-unbiased if they incorporate the appropriate first-order weights. However, weighting might not reduce design bias for those that are nonlinear in the data (such as estimators of variance components). In fact, Korn and Graubard (2003) noted that estimators of variance components that used only first-order weights could be substantially biased, even for designs with simple random sampling without replacement (SRS) at each stage. The goal of the current study is to determine when first-order weighted estimators of variance components are adequate and when they are not by focusing on data and designs related to those found in NAEP.

Section 2 reviews the background of sampling weights and hierarchical models. Section 3 presents analytical expressions for bias of the first-order weighted ANOVA estimators under the random effects model. Section 4 characterizes the conditions under which the first-order weighted estimators studied in section 3 have an unacceptably high bias. In section 5, first- and second-order weighted ANOVA estimators are computed for a random effects model fit to the NAEP 2003 fourth-grade reading data. First-order weighted estimators adjusted by scaling are evaluated in section 6. Finally, a summary and recommendations for users of NAEP data follows in section 7.

2. Hierarchical Models and Sampling Weights

When the purpose of an educational assessment program is to make valid inferences from a sample to a population of students, the students must be chosen according to a probability design; that is, the probability of selection of each sampled student must be known. Sampling designs for educational assessments often have a two-stage structure because it is cost-efficient to test groups of students from the same school. The selection probabilities for different schools and different students within a school may be unequal, and if they are, the estimation procedure must take this into account by weighting in order to assure approximately design unbiased estimation. One estimator that is design unbiased for the total for any probability design is the Horvitz-Thompson (H-T) estimator. It weights each student's score by the inverse of his or her selection probability and can be written for the two-stage design as

$$\hat{T} = \sum_{i=1}^{k} \sum_{s=1}^{m_i} y_{is} / \pi_i \pi_{s|i} ,$$

where k is the number of schools in the sample, m_i is the number of students sampled from each selected school, y_{is} is the score of the sth student in the ith school, $\pi_i = P(\text{school } i \text{ in sample})$, and $\pi_{s|i} = P(\text{student } s \text{ in sample} \mid \text{school } i \text{ in sample})$. The first-order weights, defined as $w_i = 1/\pi_i$ and $w_{s|i} = 1/\pi_{s|i}$, are needed to prevent bias if the design is informative; that is, if the model that holds for the sample is different from the model for the population (Pfeffermann & Smith, 1985). See Binder, Kovacevic, and Roberts (2005) and Binder and Roberts (2001) for more detailed discussion on the informativeness of the sampling design.

For assessments such as NAEP, which collect a rich amount of background information, educational researchers may also be interested in fitting models designed to examine

relationships between a student's performance and his or her personal or school characteristics. Because of the multistage sampling design, models accommodating the hierarchical structure are more appropriate for analysis. A simple hierarchical model (Raudenbush & Bryk, 2002) having two levels can be written as

$$\text{Level 1: } y_{is} = \beta_{0i} + x_{is}\beta_{1i} + \varepsilon_{is} \quad , \tag{1}$$

$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01} z_i + a_{0i} \quad ,$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} z_i + a_{1i} \quad ,$$

for $i = 1,\ldots,k$ and $s = 1, \ldots, m_i$, where x_{is} are covariates corresponding to the student, z_i are covariates corresponding to the school, $\underline{\beta} = [\beta_{0i}, \beta_{1i}]^T$ is a vector of unknown regression parameters, and $\underline{a_i} = [a_{0i}, a_{1i}]^T$ and ε_{is} are random effects, which are mutually independent and normally distributed with zero means and constant variances, $Var(\underline{a_i}) = \Omega$ and $Var(\varepsilon_{is}) = \sigma_e^2$.

This paper considered a simple special case of this model, the one-way random effects model, in which $\beta_{0i} = \mu$ was the grand mean and $\beta_{1i} = 0$. Thus our model is

$$y_{is} = \mu + a_i + \varepsilon_{is}, \tag{2}$$

for $i = 1,\ldots,k$ and $s = 1,\ldots,m_i$, where $a_i \sim N(0, \sigma_a^2)$ and $\varepsilon_{is} \sim N(0, \sigma_e^2)$, and a_i and ε_{is} are all mutually independent. Besides estimating the mean, or the variance components themselves, researchers may also be interested in estimating the intraclass correlation coefficient (ICC),

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, \tag{3}$$

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

Korn and Graubard (2003) showed in a simulation study that the estimators of variance components that used only first-order weights were biased, even when the design was noninformative at both school and student levels. Their proposed estimators, which used the second-order weights, were nearly unbiased.

Second-order weights are needed for an approximately unbiased estimation of variance components because the full-population functions of the data being estimated are nonlinear, specifically involving squares of sums of the individual scores. However, the estimation method

incorporating second-order weights is difficult to employ in practice, both because no commercial software is yet available and because second-order weights are not routinely included on data files.

The next section develops analytical expressions for the bias of Graubard and Korn's first-order weighted estimators of the variance components (Graubard & Korn, 1996) for the one-way random effects model. This process allows examination of the estimation bias for a larger range of sampling designs and population scenarios than simulation does. Most of the available commercial multilevel software packages use maximum likelihood based estimation methods (Chantala & Suchindran, 2006). However, any theoretical evaluation of the weighted estimators becomes rapidly intractable when the computation involves iterative methods and complex sampling structures. The focus of this paper is the analysis of variance (ANOVA) estimators (Searle, Casella, & McCulloch, 1992, p. 59), also known as method of moments estimators (Korn & Graubard, 2003) because they are easier to examine analytically.

3. Bias of First-Order Weighted Analysis of Variance (ANOVA) Estimators

3.1 First-Order Weighted ANOVA Estimators

In a super-population view (Binder & Roberts, 2001), it is assumed that the data in a population have arisen from Equation 2 and we are interested in estimating its parameters μ, σ_e^2, and σ_a^2. If all students from all schools in the population are observed, the parameters μ, σ_e^2, and σ_a^2 in Equation 2 can be estimated by (Searle et al., 1992):

$$\bar{Y} = \frac{\sum_{i=1}^{K}\sum_{s=1}^{M_i} Y_{is}}{\sum_{i=1}^{K} M_i}, \tag{4}$$

$$S_e^2 = \frac{1}{\sum_{i=1}^{K}(M_i - 1)}\sum_{i=1}^{K}\sum_{s=1}^{M_i}\left(Y_{is} - \bar{Y}_{i.}\right)^2, \tag{5}$$

$$S_a^2 = \frac{1}{(K-1)M_0}\sum_{i=1}^{K} M_i\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2 - \frac{S_e^2}{M_0}, \tag{6}$$

where K is the total number of schools in the population, M_i is the total number of students within each school, $\bar{Y}_{i.}$ is the *i*th school average, $\bar{Y}$ is the overall average, and

$$M_0 = \frac{1}{K-1}\left(\sum_{i=1}^{K} M_i - \frac{1}{\sum_{i=1}^{K} M_i} \sum_{i=1}^{K} M_i^2 \right). \quad (7)$$

Equations 4 to 6 are model consistent for the parameter values. Of course, access to data from all students in the population is usually not available. Instead, the parameters in Equation 2 must be estimated from a sample. If a sample from a two-stage probability sampling design of students chosen within schools is available, and if the sample units have equal selection probabilities at each of the two stages, then estimators of these expressions can be obtained by replacing the sums over all population units with the analogous sums over all sample units in Equations 4 to 7. But this estimation method can lead to biased results even asymptotically if either the students or the schools are unequally weighted (see Jia, 2007, for detailed discussion).

Graubard and Korn (1996) suggested the first-order weighted ANOVA estimators:

$$\bar{y}_{..FW} = \frac{\sum_{i=1}^{k} \sum_{s=1}^{m_i} w_i w_{s|i} y_{is}}{\sum_{i=1}^{k} \sum_{s=1}^{m_i} w_i w_{s|i}}, \quad (8)$$

$$s_{eFW}^2 = \frac{1}{\sum_{i=1}^{k} w_i \sum_{s=1}^{m_i} (w_{s|i} - 1)} \sum_{i=1}^{k} w_i \sum_{s=1}^{m_i} w_{s|i} \left(y_{is} - \bar{y}_{i.FW} \right)^2 , \quad (9)$$

$$s_{aFW}^2 = \frac{1}{m_{0FW}\left(\sum_{i=1}^{k} w_i - 1 \right)} \sum_{i=1}^{k} w_i \left(\sum_{s=1}^{m_i} w_{j|i} \right) \left(\bar{y}_{i.FW} - \bar{y}_{..FW} \right)^2 - \frac{s_{eFW}^2}{m_{0FW}}, \quad (10)$$

where

$$m_{0FW} = \frac{1}{\sum_{i=1}^{k} w_i - 1}\left(\sum_{i=1}^{k} w_i \sum_{s=1}^{m_i} w_{s|i} - \frac{1}{\sum_{i=1}^{k} w_i \sum_{s=1}^{m_i} w_{s|i}} \sum_{i=1}^{k} w_i \left(\sum_{s=1}^{m_i} w_{s|i} \right)^2 \right),$$

$$\bar{y}_{i.FW} = \frac{\sum_{s=1}^{m_i} w_{s|i} y_{is}}{\sum_{s=1}^{m_i} w_{s|i}}.$$

These statistics estimate μ, σ_e^2, and σ_a^2 by replacing all population sums in Equations 4 to 7 with weighted sample sums. The weighted estimator $\bar{y}_{..FW}$ is (for fixed sample sizes) unbiased for μ, but s_{eFW}^2 and s_{aFW}^2 require large sample sizes at both levels of the design for approximate unbiasedness for σ_e^2 and σ_a^2. The sample size within the school is often not large, so there can

be substantial bias in the estimators. In the next subsection, expressions for their approximate biases are derived.

3.2 Bias Expressions for the First-Order Weighted ANOVA Estimators

Expressions of the approximate estimation bias for fairly general sample designs were developed to evaluate the performance of s^2_{eFW} and s^2_{aFW}. The designs considered were two-stage, with a general probability design at the school level and SRS at the student level, which are common in educational surveys, including NAEP. The school level selection probability π_i was allowed to be related to both the school level random effect a_i and the school population size M_i. Then $\pi_i = \pi(M_i, a_i)$, so that π_i was also a random variable in this framework.

The expectation of the estimators was approximated by taking the expectation of the first term of their Taylor expansion, first with respect to the sampling design and then to the model (see the appendix). This yielded an approximate relative bias for s^2_{eFW} of

$$RB_{I,a,e}(s^2_{eFW}) = \frac{E_{I,a,e}(s^2_{eFW}) - \sigma^2_e}{\sigma^2_e} \approx -\frac{\sum_{i=1}^{K}(M_i/m_i) - K}{N - K} = -\frac{avg(M/m) - 1}{\bar{M} - 1}, \qquad (11)$$

where $N = \sum_{i=1}^{K} M_i$, $\bar{M} = N/K$, and $avg(M/m) = (1/K)\sum_{i=1}^{K} M_i/m_i$. Equation 11 shows that s^2_{eFW} was negatively biased, with larger relative bias for small school sample size (unless M_i is also small) and bounded below by -1. A complex design at the school level did not affect its approximate relative bias.

The bias and relative bias of s^2_{aFW} were approximated using similar methods (see the appendix). The resulting bias expression (A20) was too complicated to be helpful for drawing general conclusions, so a simpler balanced case was considered in which $M_i = M$ and $m_i = m$ for all *i*. Then

$$RB_{I,a,e}\left(s^2_{aFW}\right) \approx \frac{1}{m}\frac{1 - ICC}{ICC}\left(\frac{K - E(w_i)}{K-1} - \frac{m-1}{M-1}\right) - \frac{E(w_i) - 1}{K-1} - \rho\left(\pi_{ij} w_i w_j, a_i a_j\right) sd\left(\pi_{ij} w_i w_j\right) - \frac{\rho\left(w_i, a_i^2\right) sd(w_i)}{K-1}, \qquad (12)$$

where $E(\ )$ and $\rho(\ )$ were defined as the expectation and the correlation of the random variables with respect to the school level random effect a_i.

Note that if the schools were censused, all terms but the first in Equation 12 would have been equal to zero and the bias would have been positive unless the students were also censused ($m = M$). The relative bias in this case could have been large if the ICC and m were both small. The second term,

$$-\frac{E(w_i)-1}{K-1},$$

was negative for a given sample but can be substantial only if a small proportion of schools in the population are selected in the sample. The next two terms were related to the informativeness of the sample. The third term rarely made an important contribution to the relative bias unless for designs where π_{ij} is considerably different from $\pi_i\pi_j$, for example, if a small school level sampling rate was present. Otherwise, $\pi_{ij} \approx \pi_i\pi_j = 1/w_i w_j$. If extreme schools (those with either high or low scores) were oversampled, then the last term in Equation 12,

$$-\frac{\rho\left(w_i, a_i^2\right) sd\left(w_i\right)}{K-1},$$

would have contributed a positive component to the relative bias.

Since the bias expressions reported in this section are approximations, a simulation study was conducted to check how accurate they were in reflecting the true bias of the estimators. In the simulation, we assumed a population of K = 1,500 schools, each of size M = 56 students (which was the estimated average population size of schools in the NAEP 2003 fourth-grade reading sample). A two-stage stratified design was selected with two strata at the school level and SRS at the student level. Three experimental factors (denoted as Factors A, B, and C) were considered. Factor A varied the nature of the informativeness of the stratification design: Level A_1 indicated oversampling schools with large values of $|a_i|$ (extreme schools, symmetric strata), and Level A_2 indicated oversampling schools with large values of a_i (high-performing schools, asymmetric strata). Factor B denoted the sample size assignment at the school level. Defining Stratum 1 as the oversampled stratum and Stratum 2 the remainder, Level B_1 denoted selecting

all the units from Stratum 1 and half of units from Stratum 2 $\left(k_1 = K_1; k_2 = K_2 / 2\right)$ and Level B_2 denoted selecting 90 schools from Stratum 1 and nine schools from Stratum 2 $\left(k_1 = 90; k_2 = 9\right)$. Factor C was the student-level sample size, with C_1 denoting a large sample (m = 23, which was the average school sample size for the NAEP 2003 fourth-grade reading sample) and C_2 denoting a small sample (m = 5).

The population data (K = 1,500, M = 56 for all schools) was simulated using Equation 2, with $\sigma_e^2 = 1$ and ICC = 0.23. Then 5,000 samples were simulated from the data for each of the $2 \times 2 \times 2 = 8$ conditions just described. The first-order weighted estimators s_{eFW}^2 and s_{aFW}^2 from Equations 9 and 10 were computed for each sample, the bias for each estimator was computed by averaging the estimates, and the relative bias was computed. The results are reported in Table 1. Note that $\sigma_a^2 = ICC \cdot \sigma_e^2$, and for a given ICC value, further simulation results suggest that for any given σ_e^2 value, the relative biases of s_{eFW}^2 and s_{aFW}^2 were almost identical to the ones presented in Table 1, and the differences were mostly due to the simulation error. Expressions for relative bias were then computed from Equations 11 and 12 for each of the eight designs. The table shows that the simulated and analytically derived approximate biases are very similar in all cases considered. Based on this result, the analytic expressions were used to investigate the conditions under which the bias of the first-order weighted estimators of variance components would be problematic.

4. Examination of Bias of the First-Order Variance and Weighted Analysis of Variance (ANOVA) Estimators

The bias expressions derived in section 3 provided a systematic way to examine estimation bias for a variety of models and sampling designs. Equations 11 and 12 show that the relative bias of the first-order weighted estimators of the variance components was affected by sample sizes, sampling rates, ICC, and the informativeness of the design. This section uses these expressions to examine how much these factors affect the bias and to determine how important that bias is. The examples of the previous section and its results in Table 1 show that the relative bias of the variance components estimators could vary tremendously and that cases could exist at both extremes; that is, when the effect on bias was negligible (as in the upper half of Table 1) and when it was unacceptably high (as in the lower half of Table 1).

Table 1

Comparison of Simulated and Approximate Relative Bias (RB) of First-Order Weighted Estimators From a One-Way Random Effects Model With Informative Designs

		A1 (symmetric strata)		A2 (asymmetric strata)	
		$RB(s^2_{ew})$	$RB(s^2_{aw})$	$RB(s^2_{ew})$	$RB(s^2_{aw})$
		C_1 (m = 23)			
B1	Simulated	-2.6%	8.7%	-2.6%	8.8%
	Analytic	-2.6%	8.7%	-2.6%	8.8%
B2	Simulated	-2.6%	2.4%	-2.6%	8.1%
	Analytic	-2.6%	3.2%	-2.6%	7.3%
		C_2 (m = 5)			
B1	Simulated	-18.5%	62.1%	-18.6%	62.2%
	Analytic	-18.6%	62.3%	-18.6%	62.3%
B2	Simulated	-18.8%	55.2%	-18.8%	59.2%
	Analytic	-18.6%	55.2%	-18.6%	59.2%

Note. Simulation results are based on 5,000 iterations. Analytic results were calculated from Equations 11 and 12.

The goal in this section is to characterize the situations in which the first-order weighted estimators of variance components are adequate and when they are not. This was done by systematically varying features of the model parameters and sampling design and using the analytic expressions of bias for evaluation.

4.1 Effect of Sample Size Under Balanced Noninformative Designs

Section 3 noted that the first-order weighted estimators of the variance components could be substantially biased even if the sampling design was noninformative. In the first example, the bias in the first-order weighted estimator of the between- and within-school variance components was examined. The simple case of a single-stage sample from a population of equal-sized schools was assumed; that is, all schools and a simple random sample of m students within each school were selected. From Equations 11 and 12,

$$RB_{I,a,e}(s^2_{eFW}) = -\frac{M-m}{(M-1)m} \tag{13}$$

$$RB_{I,a,e}(s^2_{aFW}) = \frac{M-m}{(M-1)m}\frac{1-ICC}{ICC} \tag{14}$$

Figure 1 shows these relative biases for a range of school population sizes (M) and school sample sizes (m) when ICC = 0.2. If a relative bias of 10% or greater in magnitude was considered to be unacceptably large, then s^2_{eFW} had too large of a bias if $m < 10$ for M ranging from about 40 to 140. The estimator s^2_{aFW} also required larger values of m to have an acceptably small bias. For example, m needed to be at least 20 when $M = 40$ and at least 30 when $M = 100$.

4.2 Effect of Varying Population and Sample Sizes Under Unbalanced Noninformative Design

The second example was designed to examine whether varying school population sizes or varying school sample sizes affected the bias of the first-order weighted variance component estimators. It was assumed that the school population size M_i followed a specified distribution. It was also assumed that all schools and a simple random sample of m_i students per school were selected. Equation A20 (see the appendix) could then be simplified to

$$RB_{I,a,e}(s^2_{aFW}) = \frac{\sum_{i=1}^{K}\frac{M_i}{m_i}\sum_{i=1}^{K}M_i - \sum_{i=1}^{K}\frac{M_i^2}{m_i}}{\sum_{i\neq j=1}^{k}M_iM_j}\frac{1-ICC}{ICC} - \frac{(K-1)\sum_{i=1}^{K}\left(\frac{M_i(m_i-1)}{m_i}\right)\sum_{i=1}^{K}M_i}{\sum_{i\neq j=1}^{k}M_iM_j\sum_{i=1}^{K}(M_i-1)}\frac{1-ICC}{ICC}. \tag{15}$$

Again ICC = 0.2 as in the first example. In order to examine a realistic range of distributions of school population size, we first fitted a gamma distribution to the empirical distribution of estimated school population sizes from the NAEP 2003 fourth-grade reading assessment by matching the first two moments ($\bar{M}_{weighted} = 56,\ \mathrm{S}_{weighted}(M) = 44$). The corresponding coefficient of variation (CV) is 0.78. Figure 2 plots the histogram of the estimated school population size along with the gamma density approximation. Then K (= 1,500) units was generated from that gamma distribution. To have varying school sample sizes, $m_i = M_i/2$ was set. In addition, cases were considered for which the school population sizes were generated from three other gamma distributions with approximately the

same mean value (= 56) but varying CVs, both smaller and larger than those observed in the NAEP data. The corresponding histograms are displayed in Figure 3.

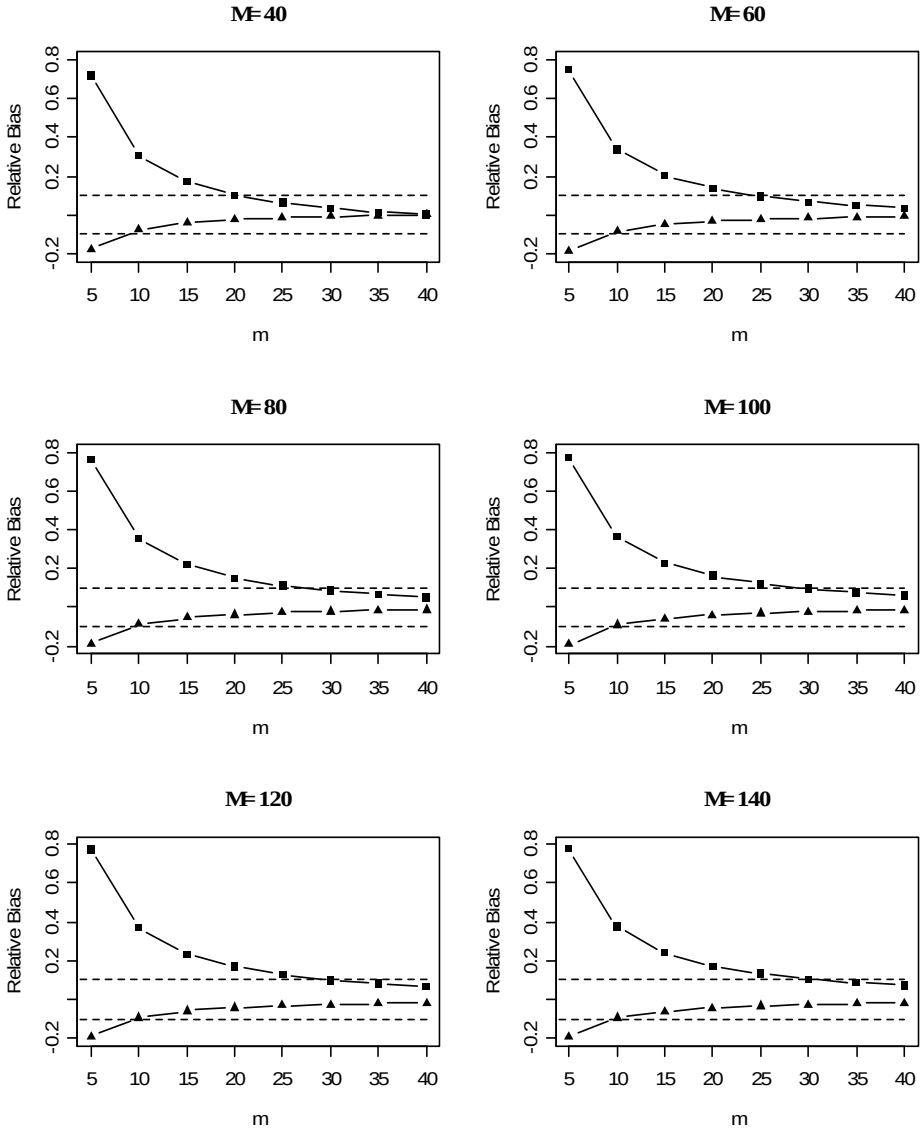

Figure 1. **Relative bias of first-order weighted variance estimators as a function of school population and sample sizes for a noninformative design in which all schools are sampled and a simple sample of *m* students are selected within each school.**

Note. The dashed lines are the bench marks for -10% and 10% relative bias (▪–relative bias of the estimators of the between-school variance; ▲– relative bias of the estimator of the within-school variance.) *M* = school population size; *m* = school sample size.

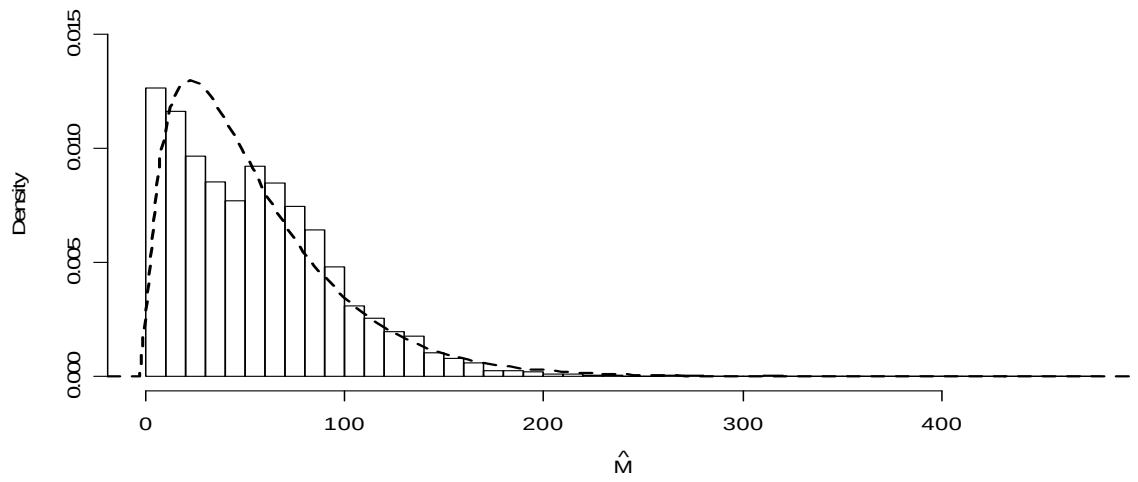

Figure 2. **Histogram of the estimated school population size for National Assessment of Educational Progress (NAEP) 2003 fourth-grade national assessment.**

Note. $\hat{M}$ = estimated school population size.

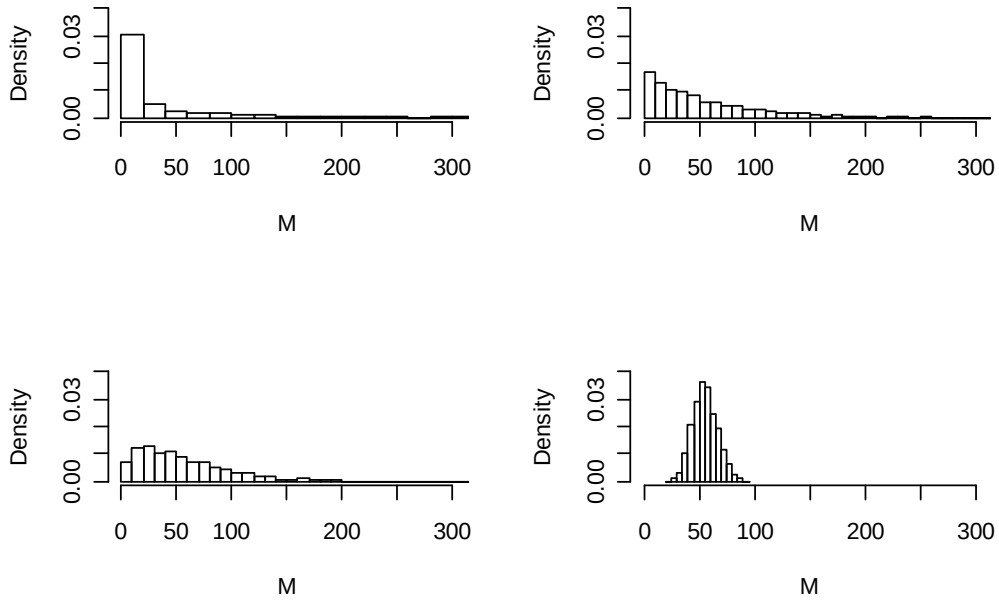

Figure 3. **Histogram of the simulated school population size.**

Note. The distributions from which the finite population of school were generated from top left to the bottom right: gamma(0.25,0.004), gamma(1,0.018), gamma(1.70,0.030), and gamma(25, 0.448). *M* = school population size.

Table 2 shows the relative biases computed from Equations 12 and 15. Note that gamma (1.70, 0.030) was chosen to approximate the school population size distribution for the above given NAEP assessment. It can be seen that the estimators underestimated the within-school variability and overestimated the between-school variability, as in the equal school size case. In addition, even though the CV of the school sizes varied from 0.2 to 2.0, the relative biases calculated were all similar to the one with the constant school population size of 56 ($RB_{I,a,e}(s^2_{eFW}) = -1.8\%$ and $RB_{I,a,e}(s^2_{aFW}) = 7.3\%$). The result suggested that varying school population sizes and varying school sample sizes did not have a substantial effect on the relative bias of s^2_{eFW} and s^2_{aFW}.

Table 2

Relative Bias (RB) of the First-Order Weighted Estimators of Within-School and Between-School Variance Components for Variable School Population Size and School Sample Size

Model	$CV(M)$	$RB_{I,a,e}\left(s^2_{eFW}\right)$	$RB_{I,a,e}\left(s^2_{aFW}\right)$
Gamma(0.25,0.004)	2	-1.9%	7.6%
Gamma(1.00,0.018)	1	-1.8%	7.1%
Gamma(1.70,0.030)	0.78	-1.8%	7.2%
Gamma(25, 0.448)	0.2	-1.8%	7.3%

Note. The *RB*s for comparable constant school sample size cases for within-school and between-school variance components are -1.8% and 7.3%, respectively. *CV* = coefficient of variation; *M* = school population size.

4.3 Joint Effect of School Sample Sizes and Interclass Correlation Coefficient (ICC) Level

The joint effect of the school sample sizes and ICC on the bias of the estimators of the between-school variance component was examined next. Kovacevic and Rai (2003) observed from a simulation study that the relative bias of their proposed weighted estimators increased as the ICC level decreased. Similar results were found in the simulation study conducted by Asparouhov (2006). The analytic bias expression and Table 1 show that the effect of ICC on $RB_{I,a,e}(s^2_{aFW})$ was mitigated by large school sample size (*m*). This example looked systematically

at the joint effect of these factors for both informative and noninformative designs. The analysis was restricted to the equal school and sample size case for simplicity.

In this example, the number of schools in the population was fixed as 1,500, and the population was assumed to follow the model in Equation 2. Four different school level designs were considered. The first three were informative and the last was noninformative (SRS at the school level). The three informative designs were all stratified, with strata defined by varying cut-points on the school random effect. In a real application, the stratification design would likely be less informative than these, so in some sense, this example was the worst case. Design 1 oversampled high-performing schools (that is, a school belonged to Stratum 1 if $a_i \geq \sigma_a$ and to Stratum 2 otherwise); Design 2 oversampled above-average schools (strata defined by $a_i \geq 0$ and $a_i < 0$); and Design 3 oversampled extreme-performing schools (strata defined by $|a_i| \geq 0.6745 \cdot \sigma_a$ and $|a_i| < 0.6745 \cdot \sigma_a$). Design 4 selected schools by SRS. For the first three designs, 90 schools were sampled from the oversampled stratum and nine from the other one; 99 schools were selected for the fourth design. At the student level, a sample was randomly selected without replacement from each selected school. The school population size was 56, and the school sample sizes ranged from 5 to 30. We investigated bias for ICC from 0.05 to 0.30.

The relative biases of s^2_{aFW} were calculated using Equation 12, where w_i and π_{ij} were all functions of normally distributed random variable a_i. Figure 4 plots $RB_{I,a,e}(s^2_{aFW})$ as a function of ICC and *m* under the four given designs. The trends were similar for the four designs, showing that the relative bias increased as ICC decreased and as school sample size decreased. A design having small school sample sizes could make the relative bias unacceptable. The informative designs showed similar magnitudes of bias as the noninformative design, so it appeared that the relative bias of the first-order weighted estimators of the between-school variance components was mainly due to the school sample size and ICC effect.

4.4. Summary

The purpose of this section was to examine whether the first-order weighted estimators had an acceptably small bias for estimation of variance components in the random effects model. Our examples showed that the first-order weighted variance components estimators were biased under both informative and noninformative designs. However, the degree of informativeness of the

school sampling design was not the main factor contributing to the bias. The first-order weights appeared to remove most of the bias due to this source. Rather, the relative bias was large when the ICC and school sample size were both small. In any particular case, when a data analyst has an idea about the size of ICC, *m*, and *M*, he can investigate the magnitude of the relative bias by using the simplified expressions in Equations 13 and 14 when *K* is relatively large.

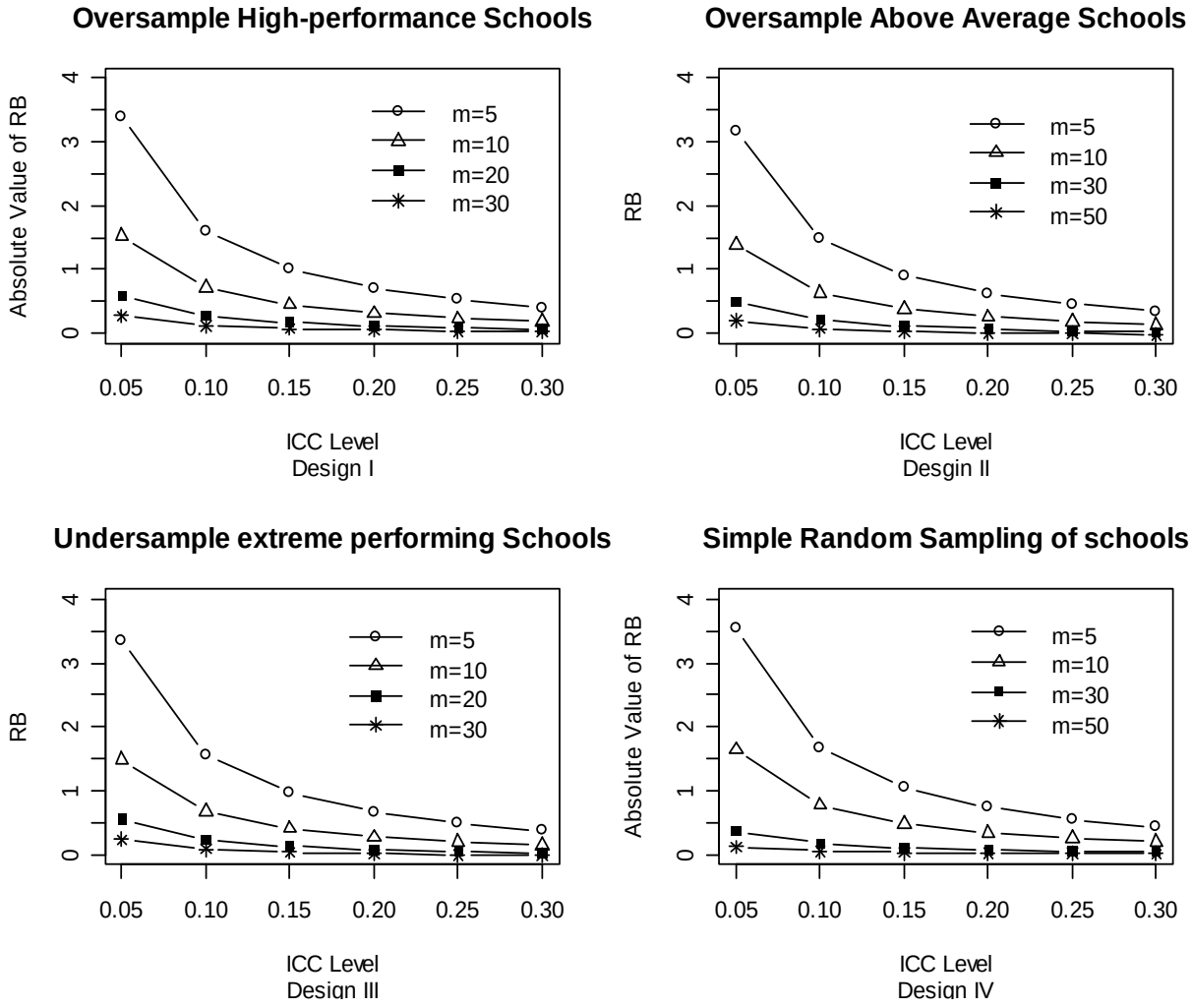

Figure 4. **Effect of interclass correlation coefficient (ICC), school sample size (*m*), and sampling design on the magnitude of the relative bias of the first-order weighted estimator of the between-school variance component.**

5. Application—National Assessment of Educational Progress (NAEP) 2003 Fourth-Grade Reading Assessment

In the previous section, we examined the size of the bias of the first-order weighted estimators of variance components in the random effects model for a variety of parameter settings and design features. In this section, we calculate first-order and second-order weighted estimates (Korn & Graubard, 2003) of the variance components from a random effects model fitted to the NAEP 2003 fourth-grade reading assessment data for the nation as a whole and for two jurisdictions. Although the true values of the variance components weren't known, it was known that the second-order weighted estimators were approximately unbiased (Korn & Graubard, 2003). Hence, the appropriateness of the first-order weighted estimators was evaluated and compared to results based on second-order weights.

More than 187,000 students from 54 jurisdictions were assessed in the NAEP 2003 fourth-grade reading assessment. Jurisdictions included states, the District of Columbia, U.S. territories, and Department of Defense schools. The sampling design is described briefly as follows: Schools were stratified with one stratum per state for public schools and several region-based strata for private schools. Within each stratum, schools were selected using a stratified systematic probability proportional to size design so as to oversample minority, nonpublic, and relatively large schools. This step was followed by a simple random sample of students drawn from each school. The average school sample size for the national sample was 23; the estimated average school population size was 56. First-order weights for both stages of the sample design were available from the restricted use data file.

We fitted a one-way random effects model to the NAEP national data, using one of the plausible values (Mislevy, 1991) for the assessment score as the response variable. Estimation of the model was conducted twice: once computing first-order weighted estimators as given in Equations 8 through 10 and once computing second-order weighted estimators as specified in Korn and Graubard (2003). Because second-order weights were not provided on the NAEP file, they had to be inferred from the first-order weights and from knowledge about the sample design. As all the details about the school level design were not known, the simplifying assumption was made that the selection of schools was independent; that is, $\pi_{ij} = \pi_i \pi_j$. At the student level, we calculated second-order selection probabilities for students from school *i* as $\pi_{st|i} = m_i(m_i - 1)/M_i(M_i - 1)$, as it would be for SRS within school. Based on this analysis, the

ICC was estimated by the second-order weighted estimators to be around 0.24. Both Figure 4 and Equation 11 suggested that bias of the first-order weighted estimators of variance components would not likely be a problem for this combination of ICC and sample size.

In addition, the one-way random effects models were fitted using both first-order and second-order weighted estimation methods to data from two jurisdictions. The jurisdictions were chosen to exemplify different kinds of weight structures. All the schools for Jurisdiction 1 were selected so the design was noninformative. The sample consisted of 24 schools with an average school sample size of 30. The estimated average school population size was 64, and the ICC value was estimated at around 0.08 from the second-order weighted estimators. Jurisdiction 2 had a design for which several extreme-performing schools (those with high and low performance) had large weights. The sample consisted of about 120 schools. The average school sample size was 16; the estimated average school population size was 32. The ICC for reading assessment score was estimated to be 0.34 based on the second-order weighted estimators. Equation 11 suggested that bias of estimators of the within-school variance component was not likely to be a problem for either jurisdiction. Figure 4 suggested that the first-order weighted estimator of the between-school variance for Jurisdiction 2 was also likely to have acceptable bias, but that we should be cautious when using it for Jurisdiction 1 due to the small value of ICC, even for the design's relatively large school sample size.

Table 3 shows the estimates of variance components as well as ICC calculated using first- and second-order weights for the national data and the two jurisdictions. In parentheses below each first-order weighted estimator is the estimated relative bias, calculated as the difference between the first- and second-order weighted estimators divided by the value of the second-order weighted estimators. This assessment of the actual bias of the first-order weighted estimator is reasonable if our approximated second-order weights are accurate. The results show, as expected, that the estimated relative bias was negative for all estimates of within-school variance and positive for estimates of between-school variances. The estimated relative biases were less than 10% for all variance component estimators except the between-school component for Jurisdiction 1. This result was predicted due to the small ICC value in that jurisdiction. However, in cases like Jurisdiction 1, where less than 10% of total variance contributes to the differences among schools before introducing any regression models, multilevel modeling might not be

necessary. This study shows that the analytic expressions can accurately predict which estimators will perform better based on our knowledge of the design and population characteristics.

Table 3

First- and Second-Order Weighted Estimators of Variance Components and Intraclass Correlations Coefficients (ICC) for 2003 National Assessment of Educational Progress (NAEP) Fourth-Grade Reading Assessment Data

Estimators using…	Estimates of σ_e^2	Estimates of σ_a^2	Estimates of ICC
	NAEP national data		
First-order weights	1026.5 (-2.3%)	355.9 (7.2%)	0.26 (8.3%)
Second-order weights	1050.6	331.9	0.24
	NAEP Jurisdiction 1 data		
First-order weights	1616.3 (-1.7%)	175.1 (19.6%)	0.10 (25%)
Second-order weights	1644.8	146.4	0.08
	NAEP Jurisdiction 2 data		
First-order weights	1111.8 (-2.8%)	573.9 (4.7%)	0.34 (3.0%)
Second-order weights	1144.4	571.2	0.33

Note: The estimated relative bias, calculated as the difference between the first- and second-order weighted estimators divided by the second-order weighted estimators, is in parentheses.

6. Weight Scaling

It was noted that the first-order weighted estimators of the variance components were biased regardless of whether the sampling design was informative. One approach to reduce the bias of the first-order weighted variance component estimators was to scale the weights. Recent statistical literature provided several scaling methods (Asparouhov, 2006; Korn & Graubard, 2003; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006; and Stapleton, 2002). Pfeffermann et al. proposed two scaling procedures that only scaled the student within-school

conditional weight ($w_{s|i}$). To be more specific, the scaled student conditional weight under their Scaling Method 1 was

$$w_{s|i}^{(1)} = w_{s|i} \frac{\sum_{s=1}^{m_i} w_{s|i}}{\sum_{s=1}^{m_i} w_{s|i}^2} \tag{16}$$

and the sum of $w_{s|i}^{(1)}$ over *s* was equal to the effective sample size

$$\frac{\left(\sum_{s=1}^{m_i} w_{s|i}\right)^2}{\sum_{s=1}^{m_i} w_{s|i}^2}.$$

Under Pfeffermann's Scaling Method 2, the scaled student conditional weight was

$$w_{s|i}^{(2)} = w_{s|i} \frac{m_i}{\sum_{s=1}^{m_i} w_{s|i}}. \tag{17}$$

For this method, the sum of $w_{s|i}^{(2)}$ over *s* was equal to the sample size for school i.

For designs that were SRS at the student level, Pfeffermann's Scaling Method 2 was appropriate to produce an approximately unbiased estimator of the within-school variance. For such designs, the scaled student conditional weight in Equation 17 was equal to

$$w_{s|i}^{(2)} = \frac{\sum_{s=1}^{m_i} w_{s|i}}{m_i} \frac{m_i}{\sum_{s=1}^{m_i} w_{s|i}} = 1,$$

and the scaled first-order weighted (SFW) estimator (s_{eSFW}^2) reduced to the unweighted one (with weight of 1), which was approximately unbiased, so that

$$RB_{I,a,e}\left(s_{eSFW}^2\right) \approx 0. \tag{18}$$

However, the SFW estimator $\left(s_{aSFW}^2\right)$ of the between-school variance was still biased. For the same sampling design assumed before with constant $M_i's$ and $m_i's$, when scaled weights were used,

$$RB_{I,a,e}(s^2_{aSFW}) \approx \left(\frac{1-E(w_i)}{(K-1)m}\right)\frac{1-ICC}{ICC} - \rho\left(\pi_{ij}a_i a_j, z_i z_j\right) sd\left(\pi_{ij} w_i w_j\right) + \frac{1-E(w_i)}{K-1} - \frac{\rho\left(w_i, a_i^2\right) sd\left(w_i\right)}{\left(K-1\right)} \quad , \tag{19}$$

where $\rho(\ ), E(\ )$, and $sd(\ )$ were all taken with respect to a. Note that Equation 19 was approximately zero for large *K* while the first two moments of w_i were finite or if a large fraction of schools was selected.

To examine the accuracy of the bias expressions for the SFW estimators, the simulation study in section 3.2 was revisited. The scaled weighted estimators were calculated for each simulated sample, averaged over 5,000 replications to obtain the relative biases, and compared with values computed from Equations 18 and 19. Table 4 shows that the simulated and calculated relative biases were similar for all parameters in all four scenarios. Thus the SFW estimators of within-school variance were approximately unbiased and those of between-school variance were negatively biased. The relative bias of s^2_{aSFW} was trivial for $k \approx 750$ (Condition B_1) and increased a bit for *k* = 99 (Condition B_2). Compared to the first-order weighted estimators whose relative biases are shown in Table 3 for the same sample designs, those of the SFW estimators were much smaller.

In summary, scaling of the first-order weighted estimator using Scaling Method 2 (Pfeffermann et al., 1998) eliminated most of the bias from estimators of the variance components for designs that were SRS at the student level, along with a large number of schools in the population or a large fraction of schools being selected.

7. Summary and Discussion

The analytic bias expressions derived in this paper are based on one-way random effects models and ANOVA estimators. Such models commonly serve as the preliminary step in the hierarchical model fitting in providing information about the outcome variability at each of level of the model (Raudenbush & Bryk, 2002).

The research results suggest that incorporating first-order weights can help to reduce bias due to the informativeness of sampling designs. However, large relative bias still exists when both school sample size and ICC values are small, regardless of the design informativeness. The

Table 4

Comparison of Simulated and Approximate Relative Bias (RB) of the Scaled First-Order Weighted Estimators From a One-Way Random Effects Model With Informative Designs at Level 2

		A1 (asymmetric strata)		A2 (symmetric strata)	
		$RB(s^2_{eSFW})$	$RB(s^2_{aSFW})$	$RB(s^2_{eSFW})$	$RB(s^2_{aSFW})$
		C_1 (m = 23)			
B1	Simulated	0.02%	-0.03%	0.00%	0.01%
	Analytic	0.00%	-0.07%	0.00%	0.02%
B2	Simulated	-0.03%	-6.35%	0.01%	-0.67%
	Analytic	0.00%	-5.57%	0.00%	-1.52%
		C_2 (m = 5)			
B1	Simulated	0.00%	-0.23%	0.00%	0.09%
	Analytic	0.00%	-0.08%	0.00%	-0.03%
B2	Simulated	-0.26%	-6.92%	-0.31%	-2.90%
	Analytic	0.00%	-7.15%	0.00%	-3.10%

Note. Simulation results are based on 5,000 iterations. Analytic results were calculated from Equations 18 and 19.

study also found that with small sample sizes (less than 20) and small ICC values (less than 0.2), if the weights are relatively constant at both student and school levels, then the unweighted estimators of variance components will be less biased than the first-order weighted estimator. On the other hand, if the weights vary at either level, then the second-order weighted estimators are needed for estimating variance components. This difference presents a dilemma for data users as second-order weights typically do not exist in the database, and constructing those weights accurately requires a level of knowledge about the design that is not likely to be available either, not to mention the unavailability of commercial software to compute these second-order weighted estimators. In that case, scaled first-order weighted estimators that were discussed in section 6 provide an alternative to the difficult-to-use second-order weighted estimators for designs in which SRS is used at the student level, given a large number of schools in the population or a large fraction of schools being selected. But until some method of making the

second-order weights available to users is implemented in publicly available software programs, an adequate and unique solution does not appear to be available.

As a limitation of the analytic approach, the obtained bias expressions only apply to the sampling designs described in this study. The bias expressions will become much more difficult to tackle if the SRS assumption at the student level is violated. Simulation studies might be a practical approach for future study of various sampling schemes at lower levels of hierarchical models.

References

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

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

Binder, D. A., Kovacevic M. S., & Roberts G. (2005). How important is the informativeness of the sample design? In *2005 Proceedings of the survey methods section, annual meeting of the statistical society of Canada*. Saskatoon, Saskatchewan, Canada: Statistical Society of Canada.

Binder, D. A., & Roberts, G. (2001, January). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section, American Statistical Association, 12,* 1, 4–6.

Chantala, K., & Suchindran, C. (2006). Adjusting for unequal selection probability in multilevel models: A comparison of software packages. In *2006 Proceedings of the survey research methods section, joint statistical meeting* (pp. 2815–2824). Alexandria, VA: American Statistical Association.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample weights in multiple regression analyses of stratified samples. *Journal of American Statistical Association, 78,* 535–543.

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

Jia, Y. (2007). *Using sampling weights in the estimation of random effects model.* Unpublished doctoral dissertation, Southern Methodist University, Dallas, TX.

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

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61,* 317–337.

Pfeffermann, D., & Holmes, D. J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A, 148,* 268–278.

Pfeffermann, D., & Lavange, L. (1989). Regression models for stratified multistage samples. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys*. Chichester, England: John Wiley & Sons Ltd.

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, 60,* 23–40.

Pfeffermann, D., & Smith, T. M. F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review, 53*(1), 37–59.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A, 169*(4), 805–827.

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

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

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

Appendix

Bias Expression of First-Order Weighted Estimators

Bias Expression of the First-Order Weighted Estimator of the Within-School Variance

The first-order weighted ANOVA estimator of the within-school variance is given as

$$s^2_{eFW} = \frac{sse_{FW}}{\sum_{i=1}^{K} I_i w_i (\sum_{s=1}^{M_i} I_{s|i} w_{s|i} - 1)}, \tag{A1}$$

with

$$sse_{FW} = \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_i y_{is}^2 - \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \overline{y}^2_{i.FW} \quad . \tag{A2}$$

where I_i and $I_{s|i}$ are indicator functions with

$$I_i = \begin{cases} 1 & \text{if unit i is in the sample} \\ 0 & \text{if unit i is not in the sample} \end{cases},$$

$$I_{s|i} = \begin{cases} 1 & \text{if unit } s \text{ within } i \text{ is in the sample, given that unit } i \text{ is in the sample} \\ 0 & \text{Otherwise} \end{cases},$$

and

$$\overline{y}_{i.FW} = \frac{\sum_{s=1}^{M_i} I_{s|i} w_{s|i} y_{is}}{\sum_{s=1}^{M_i} I_{s|i} w_{s|i}} \quad .$$

The expectations of I_i and $I_{s|i}$ with respect to the sampling design are

$$E_p(I_i) = \pi_i = 1/w_i \qquad \text{and} \qquad E_p(I_{s|i}) = \pi_{s|i} = 1/w_{s|i}\,.$$

We first take the expectation of each quantity on the right side of Equation A1 with respect to the design, then to the model

$$E_{\xi p}(\theta) = E_{\xi} E_{p|\xi}(\theta) = E_{\xi I} E_{\xi II} E_{pI|\xi I} E_{pII|\xi II}(\theta) \tag{A3}$$

Given SRS at Level 1, the student selection probability is independent of the student level random effect ε_{is}, and with the property of

$$E_p\left(I_{s|i}\right)=E_p\left(I_{s|i}^2\right)=\pi_{s|i}=\frac{m_i}{M_i} \quad . \tag{A4}$$

Given the designs, Expression A3 can be further simplified as

$$E_{\xi I}E_{\xi II}E_{pI|\xi I}E_{pII|\xi II}\left(\theta\right)=E_{\xi I}E_{\xi II}E_{pI|\xi I}E_{pII}\left(\theta\right).$$

Therefore,

$$\begin{aligned}E_{\xi p}\left(\sum_{i=1}^{K}I_i w_i\sum_{s=1}^{M_i}I_{s|i}w_{s|i}y_{is}^2\right)&=E_{\xi I}E_{\xi II}E_{pI|\xi I}E_{pII}\left(\sum_{i=1}^{K}I_i w_i\sum_{s=1}^{M_i}I_{s|i}w_{s|i}y_{is}^2\right)\\&=E_{\xi I}E_{\xi II}\left[\sum_{\mathrm{i}=1}^{K}\sum_{s=1}^{M_i}\left(\mu+a_i+\varepsilon_{is}\right)^2\right]\\&=E_{\xi I}\left[\sum_{\mathrm{i}=1}^{K}\left(\mu^2+a_i^2+\sigma_{\mathrm{e}}^2+2\mu a_i\right)M_i\ \right]\end{aligned} \tag{A5}$$

and

$$\begin{aligned}E_{\xi p}\left(\sum_{i=1}^{K}I_i w_i\sum_{s=1}^{M_i}I_{s|i}w_{s|i}\bar{y}_{i.FW}^2\right)&=E_{\xi I}E_{pI|\xi I}E_{\xi II}E_{pII}\left(\sum_{i=1}^{K}I_i w_i\sum_{s=1}^{M_i}I_{s|i}w_{s|i}\bar{y}_{i.FW}^2\right)\\&=E_{\xi I}\left[\sum_{i=1}^{K}\pi_i w_i\left(\mu^2M_i+a_i^2M_i+\frac{\sum_{s=1}^{M_i}\pi_{s|i}w_{s|i}^2}{M_i}\sigma_{\mathrm{e}}^2+2\mu a_iM_i\right)\right]\\&=E_{\xi I}\left[\sum_{i=1}^{K}\left(\mu^2+a_i^2+\frac{1}{m_i}\sigma_{\mathrm{e}}^2+2\mu a_i\right)M_i\right]\ .\end{aligned} \tag{A6}$$

As a result,

$$\begin{aligned}E_{\xi p}(sse_{FW})&=E_{\xi I}\left[\sum_{\mathrm{i}=1}^{K}\left(\mu^2+a_i^2+\sigma_{\mathrm{e}}^2+2\mu a_i\right)M_i-\sum_{i=1}^{K}\left(\mu^2+a_i^2+\frac{1}{m_i}\sigma_{\mathrm{e}}^2+2\mu a_i\right)M_i\right]\\&=E_{\xi I}\left[\sigma_{\mathrm{e}}^2\sum_{\mathrm{i}=1}^{K}\left(\frac{M_i\left(m_i-1\right)}{m_i}\right)\right]\\&=\sigma_{\mathrm{e}}^2\sum_{\mathrm{i}=1}^{K}\left(\frac{M_i\left(m_i-1\right)}{m_i}\right)\end{aligned} \quad . \tag{A7}$$

Meanwhile,

$$E_{\xi p}\left[\sum_{i=1}^{K} I_i w_i \left(\sum_{s=1}^{M_i} I_{s|i} w_{s|i} - 1\right)\right] = E_{\xi I} E_{pI|\xi I} E_{\xi II} E_{pII}\left[\sum_{i=1}^{K} I_i w_i \left(\sum_{s=1}^{M_i} I_{s|i} w_{s|i} - 1\right)\right] \quad . \tag{A8}$$

The right side of Expression A7 can be written as

$$\begin{aligned} E_{\xi I} E_{pI|\xi I} E_{\xi II} E_{pII}\left[\sum_{i=1}^{K} I_i w_i \left(\sum_{s=1}^{M_i} I_{s|i} w_{s|i} - 1\right)\right] &= E_{\xi I} E_{pI|\xi I}\left(\sum_{i=1}^{K} I_i w_i \left(M_i - 1\right)\right) \\ &= E_{\xi I}\left(\sum_{i=1}^{K} \pi_i w_i \left(M_i - 1\right)\right) \\ &= \sum_{i=1}^{K} (M_i - 1) \quad . \end{aligned} \tag{A9}$$

Equations A6 and A8 together yield

$$E_{\xi p}\left(s_{eFW}^2\right) \approx \frac{\sum_{i=1}^{K}\left(\frac{M_i\left(m_i - 1\right)}{m_i}\right)}{\sum_{i=1}^{K}\left(M_i - 1\right)} \sigma_e^2 \quad , \tag{A10}$$

and

$$RB_{\xi p}\left(s_{eFW}^2\right) \approx \frac{\sum_{i=1}^{K}\left(\frac{m_i - M_i}{m_i}\right)}{\sum_{i=1}^{K}\left(M_i - 1\right)} . \tag{A11}$$

Bias Expression of the First-Order Weighted Estimator of the Between-School Variance

The first-order weighted ANOVA estimator of the between-school variance is given as

$$s_{aFW}^2 = \frac{ssa_{FW}}{(\sum_{i=1}^{K} I_i w_i - 1) m_{0FW}} - \frac{s_{eFW}^2}{m_{0FW}} \tag{A12}$$

with

$$ssa_{FW} = \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \bar{y}_{i.FW}^2 - \bar{y}_{..FW}^2 \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \tag{A13}$$

$$\bar{y}_{..FW} = \frac{\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} y_{is}}{\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i}} \tag{A14}$$

$$m_{0FW} = \frac{1}{\left(\sum_{i=1}^{K} I_i w_i - 1\right)} \left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} - \frac{\sum_{i=1}^{K} I_i w_i \left(\sum_{s=1}^{M_i} I_{s|i} w_{s|i} \right)^2}{\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i}} \right). \tag{A15}$$

Note that

$$\begin{aligned}
&\bar{y}_{..FW}^2 \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \\
&= \frac{\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} y_{is} \right)^2}{\left(\sum_{i=1}^{K} I_i w_i M_i \right)^2} \sum_{i=1}^{K} I_i w_i M_i \\
&= \frac{\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \left(\mu + a_i + \varepsilon_{is} \right) \right)^2}{\sum_{i=1}^{K} I_i w_i M_i} \\
&= \mu^2 \sum_{i=1}^{K} I_i w_i M_i + \frac{\left(\sum_{i=1}^{K} I_i w_i a_i M_i \right)^2}{\sum_{i=1}^{K} I_i w_i M_i} + \frac{\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} \right)^2}{\sum_{i=1}^{K} I_i w_i M_i} \\
&+ 2\mu \sum_{i=1}^{K} I_i w_i a_i M_i + 2\mu \sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} + 2 \frac{\left(\sum_{i=1}^{K} I_i w_i a_i M_i \right) \left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} \right)}{\sum_{i=1}^{K} I_i w_i M_i}
\end{aligned} \tag{A16}$$

Since

$$E_{\xi p} \left[\frac{\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} \right)^2}{\sum_{i=1}^{K} I_i w_i M_i} \right] = E_{\xi I} E_{\xi II} E_{pI|\xi I} E_{pII} \left[\frac{\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} \right)^2}{\sum_{i=1}^{K} I_i w_i M_i} \right],$$

$$\approx \frac{\sum_{i=1}^{K} \frac{M_i^2}{m_i} E_{\xi I}\left(w_i \right)}{\sum_{i=1}^{K} M_i} \sigma_e^2$$

$$E_{\xi p} \left(\sum_{i=1}^{K} I_i w_i a_i M_i \right) = E_{\xi p} \left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \varepsilon_{is} \right) = 0 \quad ,$$

$$E_{\xi p}\left(\sum_{i=1}^{K} I_i w_i a_i M_i\right)^2 = \sum_{i=1}^{K} M_i^2 E_{\xi I}\left(w_i a_i^2\right) + \sum_{i\neq j}^{K} M_i M_j E_{\xi I}\left(\pi_{ij} w_i w_j a_i a_j\right) \quad ,$$

we have

$$E_{\xi p}\left(\bar{y}_{..FW}^2 \sum_{i=1}^{K} I_i w_i M_i\right) \approx \mu^2 \sum_{i=1}^{K} M_i + \frac{\sum_{i=1}^{K} \frac{M_i^2}{m_i} E_{\xi I}\left(w_i\right)}{\sum_{i=1}^{K} M_i} \sigma_e^2 + \frac{\sum_{i=1}^{K} M_i^2 E_{\xi I}\left(w_i a_i^2\right)}{\sum_{i=1}^{K} M_i} + \frac{\sum_{i\neq j}^{K} M_i M_j E_{\xi I}\left(\pi_{ij} w_i w_j a_i a_j\right)}{\sum_{i=1}^{K} M_i} \quad . \tag{A17}$$

On the other hand, the expectation of Equation A15 is

$$\begin{aligned} E_{\xi p}\left(m_{0FW}\right) &= E_{\xi I} E_{pI|\xi I} E_{\xi II} E_{pII}\left[\frac{\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} - \frac{\sum_{i=1}^{K} I_i w_i \left(\sum_{s=1}^{M_i} I_{s|i} w_{s|i}\right)^2}{\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i}}}{\left(\sum_{i=1}^{K} I_i w_i - 1\right)}\right] \\ &\approx E_{\xi I} E_{pI|\xi I}\left[\frac{1}{\left(\sum_{i=1}^{K} I_i w_i - 1\right)}\left(\sum_{i=1}^{K} I_i w_i M_i - \frac{\sum_{i=1}^{K} I_i w_i M_i^2}{\sum_{i=1}^{K} I_i w_i M_i}\right)\right] \\ &\approx E_{\xi I}\left[\frac{1}{\left(\sum_{i=1}^{K} \pi_i w_i - 1\right)}\left(\sum_{i=1}^{K} \pi_i w_i M_i - \frac{\sum_{i=1}^{K} \pi_i w_i M_i^2}{\sum_{i=1}^{K} \pi_i w_i M_i}\right)\right] \\ &= \frac{1}{K-1}\left(\sum_{i=1}^{K} M_i - \frac{\sum_{i=1}^{K} M_i^2}{\sum_{i=1}^{K} M_i}\right) \\ &= \frac{1}{K-1}\frac{\sum_{i\neq j}^{K} M_i M_j}{\sum_{i=1}^{K} M_i} \end{aligned} \tag{A18}$$

Combining Equations A10, A17, and A18, the delta method gives

$$E_{\xi p}\left(s^2_{aFW}\right) \approx \sigma_a^2 \frac{\left(\sum_{i=1}^{K} M_i\right)^2}{\sum_{i\neq j}^{K} M_i M_j} - \frac{\sum_{i=1}^{K} M_i^2 E_{\xi I}\left(w_i a_i^2\right)}{\sum_{i\neq j}^{K} M_i M_j} - \frac{\sum_{i\neq j}^{K} M_i M_j E_{\xi I}\left(\pi_{ij} w_i w_j a_i a_j\right)}{\sum_{i\neq j}^{K} M_i M_j}$$
$$+\sigma_e^2 \left(\frac{\sum_{i=1}^{K} \frac{M_i}{m_i} \sum_{i=1}^{K} M_i - \sum_{i=1}^{K} \frac{M_i^2}{m_i} E_{\xi I}\left(w_i\right)}{\sum_{i\neq j}^{K} M_i M_j} - \frac{\left(K-1\right) \sum_{i=1}^{K} M_i \sum_{i=1}^{K} \left(\frac{M_i\left(m_i-1\right)}{m_i} \right)}{\sum_{i\neq j}^{K} M_i M_j \sum_{i=1}^{K} \left(M_i-1\right)} \right) \quad \text{(A19)}$$

and

$$RB_{\xi p}\left(s^2_{aFW}\right) = \frac{\left(\sum_{i=1}^{K} M_i\right)^2}{\sum_{i\neq j}^{K} M_i M_j} - \frac{\sum_{i=1}^{K} M_i^2 E_{\xi I}\left(w_i a_i^2\right)}{\sigma_a^2 \sum_{i\neq j}^{K} M_i M_j} - \frac{\sum_{i\neq j}^{K} M_i M_j E_{\xi I}\left(\pi_{ij} w_i w_j a_i a_j\right)}{\sigma_a^2 \sum_{i\neq j}^{K} M_i M_j}$$
$$+\frac{1-ICC}{ICC} \left(\frac{\sum_{i=1}^{K} \frac{M_i}{m_i} \sum_{i=1}^{K} M_i - \sum_{i=1}^{K} \frac{M_i^2}{m_i} E_{\xi I}\left(w_i\right)}{\sum_{i\neq j}^{K} M_i M_j} - \frac{\left(K-1\right) \sum_{i=1}^{K} M_i \sum_{i=1}^{K} \left(\frac{M_i\left(m_i-1\right)}{m_i} \right)}{\sum_{i\neq j}^{K} M_i M_j \sum_{i=1}^{K} \left(M_i-1\right)} \right) \quad \text{(A20)}$$