Author Query Form

Journal Title	:	JEB
Article Number	:	359793

Dear Author/Editor,

Greetings, and thank you for publishing with SAGE Publications. Your article has been copyedited, and we have a few queries for you. Please address these queries when you send your proof corrections to the production editor. Thank you for your time and effort.
Please assist us by clarifying the following queries:

Query No.	Query	Remarks
1	Please provide e-mail address for the authors Lynne Stokes and Ian R. Harris.	
2	Please provide page range for reference Binder & Roberts, 2003.	
3	Please provide in-text citation for Searle et al., 1992, or delete reference.	
4	Please provide page range for reference of Pfeffermann, D (1989).	
5	Please provide page range for reference of Skinner, C. J. (1989).	

Journal of Educational and Behavioral Statistics
Month 2010, Vol. xxx, No. xx, pp. 1–27
DOI: 10.3102/1076998609359793

Performance of Random Effects Model Estimators Under Complex Sampling Designs

Yue Jia
Educational Testing Service, Southern Methodist University

Lynne Stokes
Ian Harris
Yan Wang
Southern Methodist University

In this article, we consider estimation of parameters of random effects models from samples collected via complex multistage designs. Incorporation of sampling weights is one way to reduce estimation bias due to unequal probabilities of selection. Several weighting methods have been proposed in the literature for estimating the parameters of hierarchical models, of which random effects models are a special case. We evaluate the bias of the weighted analysis of variance (ANOVA) estimators of the variance components for a two-level, one-way random effects model. For these estimators, analytic bias expressions are developed and the accuracy of the expressions is evaluated through Monte Carlo simulation. The expressions are used to examine the impact of sample size, the size of the intraclass correlation coefficient (ICC), and the sampling design on the estimators' performance. The sampling designs considered are two-stage, with a general probability design at Level 2 and simple random sampling without replacement (SRS) at Level 1. The study shows that variance component estimators using only first-order weights perform well when both cluster size and ICC are moderate. However, this weighting method should be used with caution for small cluster sizes (less than 20), particularly when ICC is small (less than 0.2). In such scenarios, scaled first-order weighted (SFW) estimators provide an alternative to the difficult-to-use second-order weighted estimators for designs in which SRS is used at the ultimate sampling unit level (Level 1). This article is discussed in the context of large educational survey assessments.

The authors acknowledge American Educational Research Association (AERA) for partially supporting the research through its AERA Grants Program (NSF Grant #REC-0310268). Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies or any of the affiliated institutions.

Keywords: *random effects model; variance components; estimation bias; ANOVA estimators; complex sampling designs; selection probability; sampling weights; ICC; NAEP*

1. Introduction

Large-scale survey assessments, such as the National Assessment of Educational Progress (NAEP), typically collect cognitive data from a complex multistage sample of schools and students, along with a rich amount of background information. Researchers often fit models designed to understand the relationships or interdependencies between students' performance and student or school characteristics.

It is cost-efficient to use a multistage sampling design to test groups of students from the same school (cluster). However, the selection probabilities for different schools and different students within a school may be unequal. And if they are, sampling weights are needed in the estimation procedure when the design is informative, that is, when units at any level of the hierarchy are selected in ways that are not accounted for by the model (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). See Sugden and Smith (1984) and Binder and Roberts (2001) for more detailed discussion on the informativeness of a sampling design.

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. Weighting in single-level regression models uses methods appropriate for pseudomaximum likelihood (PML) estimation (Binder, 1983; Skinner, 1989). Yet weighting in hierarchical models is not a trivial extension of PML (Pfeffermann et al., 1998). Several methods have been proposed in recent years for incorporating first-order weights into hierarchical models (Asparouhov, 2006; Graubard & Korn, 1996; Grilli & Pratesi, 2004; Kovacevic & Rai, 2003; Pfeffermann et al., 1998). A number of commercially available software packages also implement the method by Pfeffermann et al. to obtain estimates for parameters of hierarchical models (e.g., hierarchical linear model [HLM] 6.0, MLWIN, LISREL, and Stata GLLAMM. See Chantala & Suchindran, 2006, for detailed discussions). In hierarchical models, however, for estimators that are nonlinear in the data (such as estimators of model variance components), the property of asymptotic unbiasedness requires the sample sizes at both levels to increase (Pfeffermann et al., 1998), while in practice, the cluster size (e.g., number of students within school) is often small. 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 through an analytic approach. The article is organized as follows. Section 2 reviews the background of sampling weights and hierarchical models. Section 3 presents analytical expressions for the bias of the first-order weighted analysis of variance (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 (Pfeffermann et al., 1998) are evaluated in Section 6. Finally, a summary and recommendations for users of NAEP data follows in Section 7.

2. Sampling Weights and Hierarchical Models

When the purpose of a survey assessment is to make valid inferences from a sample to a finite 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. The estimation procedure needs to take into account the unequal selection probabilities by weighting 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 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 $\pi_{s|i} = P(\text{student } s \text{ in sample} | \text{school } i \text{ in sample})$. The first-order weights are defined as $w_i = 1/\pi_i$ and $w_{s|i} = 1/\pi_{s|i}$.

Frequently scientific research questions require inference based on stochastic models, rather than finite population descriptive statistics. For example, researchers might be interested in examining relationships between student assessment scores and the background questionnaires about schools, teachers, and the students themselves. A simple two-level HLM (Raudenbush & Bryk, 2002) that could describe such relationships can be written as

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

$$\text{Level2}: \; \beta_{0i} = \gamma_{00} + \gamma_{01} z_i + a_{0i},$$

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

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, k is the number of schools (clusters), m_i is the school size, $\underline{\beta} = [\beta_{0i}, \beta_{1i}]^T$ 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, $\mathrm{Var}(\underline{a_i}) =$ and $\mathrm{Var}(\varepsilon_{is}) = \sigma_e^2$.

One special case of Model 1 is the two-level, one-way random effects model, in which $\beta_{0i} = \mu$ is the grand mean and $\beta_{1i} = 0$, that 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. As a common practice, data analysts often start with Model 2 with no predictors to establish a baseline for the decomposition of the total variance into variance components associated with each level of the model. Another quantity of interest here is 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.

As stated before, asymptotically unbiased estimators of the model variances with first-order weights require the sample sizes at both levels to increase. This is because the full-population functions of the data being estimated are nonlinear, specifically involving squares of sums of the individual scores. Even though Pfeffermann and LaVange (1989) and Korn and Graubard (2003) proposed various second-order weighted estimators that are asymptotically unbiased when the number of clusters increases, no commercial software package is currently available to incorporate those methods. Furthermore, 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.

Because the cluster size is often small in practice, in the absence of the second-order weights, it is important to study the asymptotic bias of the first-order weighted estimators of variance components, under the conditions of an increasing number of clusters while leaving the cluster size as a fixed small finite value. Results from simulations studies (Asparouhov, 2006; Grilli & Pratesi, 2004; Pfeffermann et al., 1998) seem to suggest that the biases of the first-order weighted variance component estimators of a hierarchical model are related to cluster size, design informativeness, and interclass correlation. Theoretical evaluation of the weighted estimators becomes rapidly intractable when the estimation procedure involves iterative methods, so the focus of this article is on the weighted ANOVA estimators for one-way random effects model that

were proposed by Graubard and Korn (1996). These are easier to examine analytically and are identical to the restricted maximum likelihood estimators when the estimator of the between cluster variance is greater than zero. This focus allows systematic and insightful examination of the estimation bias for a larger range of sampling designs and population scenarios than can easily be handled by simulation.

3. Bias of First-Order Weighted ANOVA Estimators

3.1. First-Order Weighted ANOVA Estimators

To study the properties of an estimator from a survey sample, the randomization from both complex survey designs and stochastic models are considered through the "superpopulation" approach. This framework was first introduced by Hartley and Sielken (1975), in which a superpopulation model was assumed with the finite population as a realization, and the sample is selected from the finite population using certain sampling designs. The large sample properties of the sample estimators are then evaluated with regard to the joint distribution induced by the model and the sampling scheme. This view has been adapted by many researchers, including Fuller (1975), Binder and Roberts (2003), Rubin-Bleuer and Kratina (2005), among others.

If all students from all schools in a finite population are observed, the population mean and within- and between-school variances can be written as

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

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

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

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 ith school average, $\bar{Y}$ is the overall average, and

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

In this article, it is assumed that the finite population has arisen from a superpopulation described by Model 2, and we are interested in estimating the model parameters μ, σ_e^2, and σ_a^2. The population quantities in Equations 4 to 6 are consistent for these model parameters. Of course, access to data from all students in the population is usually not available; instead, the model parameters must be estimated

from a sample by replacing the sums over all population units with the analogous sums over all sample units in Equations 4–6. 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 the estimators from the sample are design consistent for the finite population quantities and are consistent for the model parameters as well. But the sample estimators can be design-biased, even asymptotically, if either the students or the schools have unequal selection probabilities (see Jia, 2007, for detailed discussion).

To reduce design bias, Graubard and Korn (1996) suggested the first-order weighted ANOVA estimators:

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

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

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

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 4–7 with weighted sample sums. It is straightforward to show that the weighted estimator $\bar{y}_{..FW}$ is consistent for μ with respect to both design and model randomizations. However, large sample sizes at both levels are required for s^2_{eFW} and s^2_{aFW} to be unbiased. The number of students within each 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 estimation bias and relative 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. A design that is approximately SRS at the lower level is common in educational surveys, including NAEP. Under

such designs and Model 2, 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 a random variable with respect to the jointed model-design distribution. The student within school conditional selection probability $\pi_{s|i} = m_i/M_i$ is constant within each school. In addition, a random indictor variableI_i is denoted with value 1 if the *i*th school is included in the sample and 0 otherwise. Similarly, $I_{s|i}$ is defined as the inclusion indicator for the *s*th student in the *i*th school. As an example, the quantity $\bar{y}_{..FW}$ that appears in Expressions 8 and 10 can be rewritten as

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

The expectations of I_i and $I_{s|i}$ with respect to sampling designs are denoted by π_i and $\pi_{s|i}$, respectively. Let *pI*and *pII* denote the randomization due to sampling designs at the school and the student level, ξI and ξII the randomization due to the model at both levels, the expectation of a given statistic is defined as

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

and its corresponding bias and relative bias are defined as

$$\text{Bias}_{\xi p}\left(\hat{\theta}\right) = E_{\xi p}\left(\hat{\theta}\right) - \theta,$$

$$RB_{\xi p}\left(\hat{\theta}\right) = \frac{E_{\xi p}\left(\hat{\theta}\right) - \theta}{\theta}.$$

The expectation of s^2_{eFW} is approximated by taking the expectation of the first term of the Taylor expansion (see the Appendix A).This yields an approximate bias and approximate relative bias for s^2_{eFW} of

$$\text{Bias}_{\xi p}(s^2_{eFW}) \approx -\left(\frac{\sum_{i=1}^{K} (M_i/m_i) - K}{N - K}\right)\sigma^2_e = -\frac{avg(M/m) - 1}{\bar{M} - 1}\sigma^2_e,$$

$$RB_{\xi p}(s^2_{eFW}) \approx -\frac{avg(M/m) - 1}{\bar{M} - 1}, \tag{11}$$

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} is 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 does not affect the approximate relative bias.

The bias and relative bias of s^2_{aFW} can be approximated using similar methods (see Appendix A). The resulting bias Expression A20 is too complicated to be

helpful for drawing general conclusions, so we consider a simpler balanced case in which $M_i = M$ and $m_i = m$ for all i. Then

$$RB_{\xi p}\left(s^2_{aFW}\right) \approx \frac{1}{m}\frac{1-ICC}{ICC}\left(\frac{K-E_{\xi I}(w_i)}{K-1}-\frac{m-1}{M-1}\right)-\frac{E_{\xi I}(w_i)-1}{K-1}$$
$$-\rho_{\xi I}\left(\pi_{ij}w_iw_j, a_ia_j\right)sd_{\xi I}\left(\pi_{ij}w_iw_j\right)-\frac{\rho_{\xi I}\left(w_i,a_i^2\right)sd_{\xi I}(w_i)}{K-1},$$

where $E_{\xi I}(), sd_{\xi I}(w_i)$ and $\rho_{\xi I}()$ are defined as the expectation, standard deviation, and correlation of the random variables with respect to the school-level model randomization.

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

$$-\frac{E_{\xi I}(w_i)-1}{K-1},$$

is 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 in Equation 12 are related to the informativeness of the sample. The third term is usually small unless the design has π_{ij} is considerably different from $\pi_i\pi_j$, for example, if a small school-level sampling rate is present. Otherwise, $\pi_{ij} \approx \pi_i\pi_j = 1/w_iw_j$. If extreme schools (those with either high or low scores) are oversampled, then the last term in Equation 12,

$$-\frac{\rho_{\xi I}\left(w_i,a_i^2\right)sd_{\xi I}(w_i)}{K-1},$$

will contribute a positive component to the relative bias.

Because the bias expressions reported in this section are approximations based on Taylor expansion, a simulation study was conducted to check their accuracy. In the simulation, we assumed a population of $K = 1{,}500$ schools, each of size $M = 56$ students (which was the estimated average school population size in the NAEP 2003 fourth-grade reading national 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 ($k_1 = K_1; k_2 = K_2/2$) and Level B_2 denoted selecting 90 schools from Stratum 1 and 9 schools from

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.

Stratum 2 ($k_1 = 90; k_2 = 9$). 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 $\sigma^2_e = 1$[1] and ICC $= 0.23$.[2] Then 5,000 samples were simulated from the data for each of the $2 \times 2 \times 2 = 8$ conditions just described. To obtain the estimation bias from simulation, the first-order weighted estimators s^2_{eFW} and s^2_{aFW} 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. Expressions for relative bias were then computed from Equations 11 and 12 for each of the eight designs. The results are reported in Table 1. 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 Weighted ANOVA Estimators

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 is done by systematically varying features of the model parameters and sampling design and using the analytic expressions of bias for evaluation. Equations 11 and 12 show that the relative bias of the first-order weighted estimators of the

variance components is affected by sample size, ICC, sampling rates, and the informativeness of the design, which is consistent with results from many simulation studies in the literature. In this section, we use the derived expressions to examine how much these factors affect the bias.

4.1. Effect of Cluster Size Under Balanced Noninformative Designs

In the first example, we examine the simple case of a single-stage sample from a population of equal-sized schools. That is, all schools are selected and a simple random sample of m students within each school are selected. From Equations 11 and 12,

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

$$RB_{\xi p}(s^2_{aFW}) = \frac{M-m}{(M-1)m}\frac{1-ICC}{ICC}. \tag{14}$$

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 is considered unacceptably large, then s^2_{eFW} has too large of a bias if $m < 10$ for M ranging from about 40 to 140. The estimator s^2_{aFW} requires even larger values of m to have an acceptably small bias. For example, m needs to be at least 20 when $M = 40$ and at least 30 when $M = 100$.

4.2. Effect of Variable Cluster Population and Sample Sizes Under an Unbalanced Noninformative Design

The second example is designed to examine whether variable school population sizes or school sample sizes affects the bias of the first-order weighted variance component estimators. It is assumed that the school population size M_i follows a specified distribution and that all schools and a simple random sample of m_i students per school are selected. Equation A20 (see Appendix A) simplifies to

$$RB_{\xi p}(s^2_{aFW}) = \frac{\sum_{i=1}^K \frac{M_i}{m_i}\sum_{i=1}^K M_i - \sum_{i=1}^K \frac{M_i^2}{m_i}}{\sum_{i\neq j=1}^k M_i M_j}\frac{1-ICC}{ICC}$$
$$-\frac{(K-1)\sum_{i=1}^K \left(\frac{M_i(m_i-1)}{m_i}\right)\sum_{i=1}^K M_i}{\sum_{i\neq j=1}^k M_i M_j \sum_{i=1}^K (M_i-1)}\frac{1-ICC}{ICC}.$$

As in the first example, we set ICC = 0.2. To examine a realistic range of distributions of school population size, we first fitted a γ distribution to the empirical distribution of estimated school population sizes from the NAEP 2003 fourth-grade reading assessment by matching the first two moments ($\bar{M}_{\text{weighted}} = 56, \text{S}_{\text{weighted}}(M) = 44$). The corresponding coefficient of variation (CV) is 0.78. Figure 2 plots the histogram of the estimated school population size

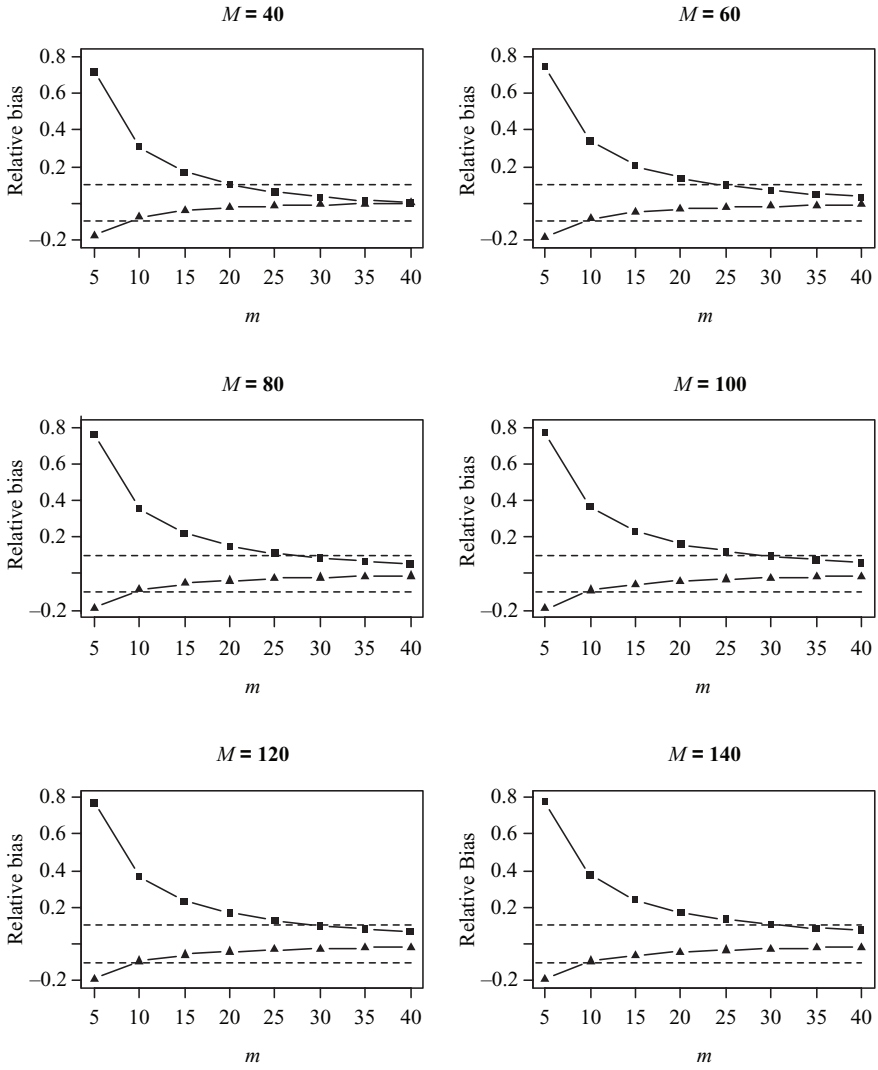

FIGURE 1. *Relative bias of first-order weighted variance estimators as a function of school population and sample sizes for a noninformative design in which all schools are sampled and a simple random sample of* m *students are selected within each school. The dashed lines are the bench marks for −10% 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.*

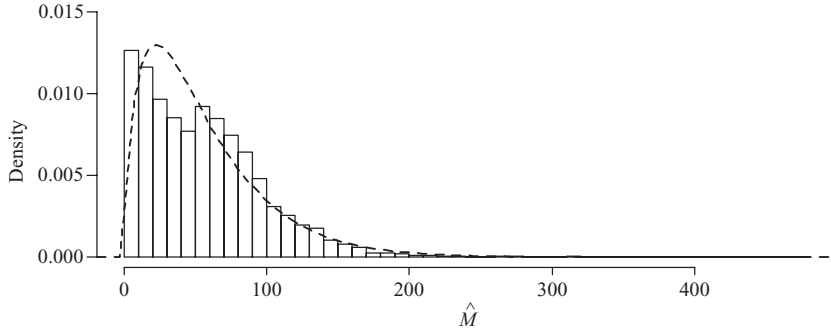

FIGURE 2. *Histogram of the estimated school population size for National Assessment of Educational Progress (NAEP) 2003 fourth-grade national assessment.* $\hat{M}$ = *estimated school population size*.

along with the γ density approximation. Then K (=1,500) units were generated from that γ distribution. To have varying school sample sizes, $m_i = M_i/2$ was used. In addition, cases were considered for which the school population sizes were generated from three other γ 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.

Table 2 shows the relative biases computed from Equations 12 and 15. Note that γ (1.70, 0.030) in the third row is the γ distribution that most accurately approximates the NAEP school population size distribution. It can be seen that 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 ($RB_{\xi p}(s^2_{eFW}) = -1.8\%$ and $RB_{\xi p}(s^2_{aFW}) = 7.3\%$). The results suggest that the relative biases of s^2_{eFW} ands^2_{aFW} are mainly driven by the average school size and do not seem to be substantial affected by varying school population sizes and sample sizes. Thus, an extremely accurate modeling of the school size distributions may not be particularly necessary in practice.

4.3. Joint Effect of Cluster Sample Size and ICC

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 reported in the simulation study conducted by Asparouhov (2006). The analytic bias expression and Table 1 show that the effect of ICC on $RB_{\xi p}(s^2_{aFW})$ is mitigated by large cluster sample size (m). The third example looks systematically at the joint effect of these factors for both informative and noninformative designs. The analysis is restricted to equal cluster population size and equal cluster sample size for simplicity.

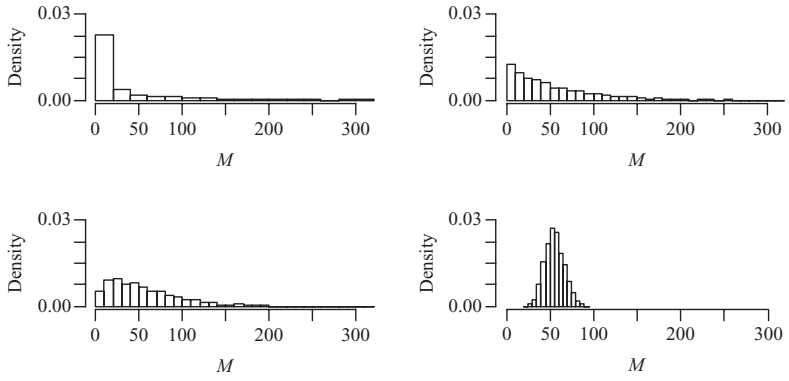

FIGURE 3. *Histogram of the simulated school population size. The distributions from which the finite population of school was generated from top left to the bottom right:* γ *(0.25, 0.004),* γ *(1,0.018),* γ *(1.70, 0.030), and* γ *(25, 0.448).* M = *school population size.*

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)	$\mathrm{RB}_{I,a,e}\left(s^2_{eFW}\right)$	$\mathrm{RB}_{I,a,e}\left(s^2_{aFW}\right)$
γ (0.25, 0.004)	2	−1.9%	7.6%
γ (1.00, 0.018)	1	−1.8%	7.1%
γ (1.70, 0.030)	0.78	−1.8%	7.2%
γ (25, 0.448)	0.2	−1.8%	7.3%

Note: The RBs 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.

In this example, the number of schools in the population is fixed at 1,500, and the population is assumed to follow the model in Equation 2. Four different school-level designs are considered. The first three are informative designs, and were all stratified, with strata defined by varying cut points on the school random effect. Design 1 oversamples high-performing schools (that is, a school belonged to Stratum 1 if $a_i \geq \sigma_a$ and to Stratum 2 otherwise); Design 2 oversamples above-average schools (strata defined by $a_i \geq 0$ and a_i); and Design 3 oversamples extreme-performing schools (strata defined by $|a_i| \geq 0.6745 \cdot \sigma_a$ and $|a_i|$). In a real application, the stratification design would likely be less informative than these, so in some sense, this example represents a worst case. Design 4 selects schools by SRS and so is not informative. For the first three designs, 90 schools were sampled from the oversampled stratum and 9 from the other one; 99 schools were

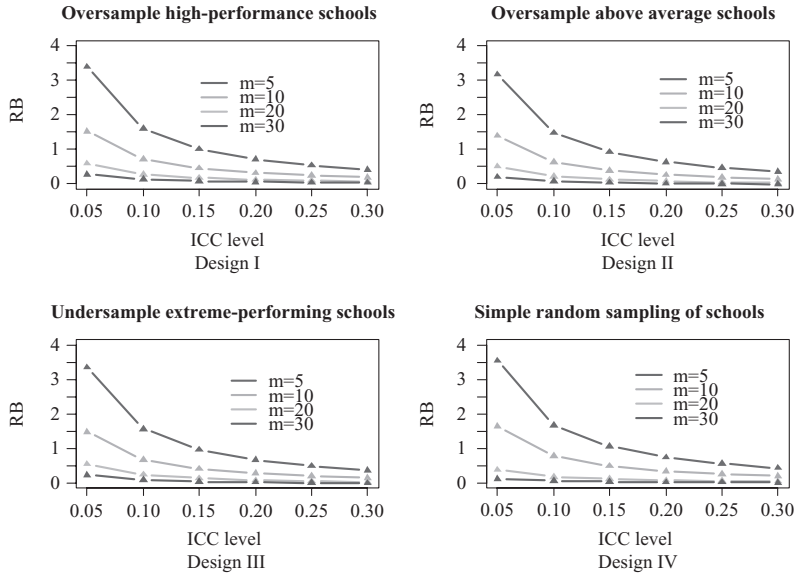

FIGURE 4. *Effect of intraclass correlation coefficient (ICC), school sample size (*m*), and sampling design on the magnitude of the relative bias of the first-order weighted estimator of the between-school variance component.*

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. The ICC ranges from 0.05 to 0.30.

The relative bias of s^2_{aFW} is calculated using Equation 12, where w_i and π_{ij} are all functions of the normally distributed random variable a_i. Figure 4 plots $RB_{\xi p}(s^2_{aFW})$ as a function of ICC and m under the four given designs. The trends are similar for the four designs, showing that the relative bias increases as ICC decreases and as school sample size decreases. A design having small school sample sizes could make the relative bias unacceptable. The informative designs show similar magnitudes of bias to the noninformative design. It appears that the relative bias of the first-order weighted estimators of the between-school variance component is mainly due to the school sample size and ICC effect and is insensitive to design informativeness.

4.4. Summary

The purpose of this section was to examine whether the first-order weighted estimators have an acceptably small bias for estimation of variance components

in the random effects model. Our examples show that the first-order weighted variance components estimators are biased under both informative and noninformative designs. However, the degree of informativeness of the school sampling design is not the main factor contributing to the bias. The first-order weights remove most of the bias due to this source. Rather, the relative bias was large when the ICC and the within school sampling rate were both small. In any particular case, when a data analyst has an idea about the size of ICC, m, and M, he or she can investigate the magnitude of the relative bias by using the simplified expressions in Equations 13 and 14.

5. Application—NAEP 2003 Fourth-Grade Reading Assessment

The 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 standard one-level 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.

In this section, we calculate first-order and second-order weighted estimates 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 are not known, the second-order weighted estimators are approximately unbiased (Korn & Graubard, 2003). Thus, we will judge the first-order weighted estimators comparing them to the estimators based on second-order weights.

More than 187,000 students from over 7,700 schools in 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 random sample of students drawn from each school, so that students within sampled schools were selected with equal probability. The average school sample size for the national sample was 23; the estimated average school population size was 56. The NAEP restricted use database contains both school and student overall weights (w_i and w_{is}), from which the

student conditional weights $w_{s|i}$are calculated ($w_{s|i} = w_{is}/w_i$). The estimation procedure was carried out entirely in the R language environment.

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. 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. 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, $\pi_{ij} = \pi_i\pi_j$. See Brewer and Donadio (2003) for alternatives to estimate π_{ij} for high entropy sampling designs. Based on this analysis, the ICC was estimated by the second-order weighted estimators to be approximately 0.24. Both Figure 4 and Equation 11 suggest 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 about 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 suggests that bias of estimators of the within-school variance component is not likely to be a problem for either jurisdiction. Figure 4 suggests that the first-order weighted estimator of the between-school variance for Jurisdiction 2 is 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-

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	1,026.5 (−2.3%)	355.9 (7.2%)	0.26 (8.3%)
Second-order weights	1,050.6	331.9	0.24
	NAEP Jurisdiction 1 data		
First-order weights	1,616.3 (−1.7%)	175.1 (19.6%)	0.10 (25%)
Second-order weights	1,644.8	146.4	0.08
	NAEP Jurisdiction 2 data		
First-order weights	1,111.8 (−2.8%)	573.9 (4.7%)	0.34 (3.0%)
Second-order weights	1,144.4	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.

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.

6. Weight Scaling

In Section 4, we saw 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 is to scale the weights. Recent statistical literature provides several scaling methods (Asparouhov, 2006; Korn & Graubard, 2003; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002). Pfeffermann et al. (1998) propose 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 is

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

and the sum of $w_{s|i}^{(1)}$ over s is 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 is

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

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

For designs that are SRS at the student level, Pfeffermann's Scaling Method 2 is more appropriate to produce an approximately unbiased estimator of the within-school variance. For such designs, the scaled student conditional weight in Equation 17 is 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 ($s_{e\mathrm{SFW}}^2$) reduces to the unweighted one (with weight of 1), which is approximately unbiased, so that

$$RB_{\xi p}\left(s_{eSFW}^2\right) \approx 0. \tag{18}$$

However, the SFW estimator $\left(s_{a\mathrm{SFW}}^2\right)$ of the between-school variance is still biased. For the same sampling design assumed before with constant M and m,

$$RB_{\xi p}\left(s_{aSFW}^2\right) \approx \left(\frac{1 - E_{\xi I}(w_i)}{(K-1)m}\right) \frac{1 - ICC}{ICC} - \rho_{\xi I}\left(\pi_{ij} a_i a_j, z_i z_j\right) sd_{\xi I}\left(\pi_{ij} w_i w_j\right) + \frac{1 - E_{\xi I}(w_i)}{K-1} - \frac{\rho_{\xi I}\left(w_i, a_i^2\right) sd_{\xi I}(w_i)}{(K-1)}.$$

Note that Equation 19 was approximately zero for large K while the first two moments of w_i are finite or if a large fraction of schools is 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. The SFW estimators of within-school variance were approximately unbiased and those of between-

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%
			C2 ($m = 5$)		
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.

school variance were negatively biased. The relative bias of $s^2_{a\mathrm{SFW}}$ 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) eliminates most of the bias from estimators of the variance components for designs that are SRS at the student level, along with a large number of schools in the population or a large fraction of schools being selected. In the most current version of the HLM software (HLM v 6.0), the weighting option automatically rescales the student level weights for the users using Pfeffermann's Scaling Method 2 and also rescales the school weights to the total number of schools in the sample.

7. Summary and Discussion

This article covers the possible bias in variance component estimators that can arise when fitting a one-way random effects model for the data obtained from complex sampling designs. The primary purpose is to examine when the first-order weighted estimators are adequate. The results suggest that the first-order weighted estimators take care of much of the bias due to the informativeness of the design; but, they can still suffer from a large relative bias when both school sample size and ICC are small. That is, when school sample size is less than 20,

and particularly when ICC is less than 0.2. These problems occur even for non-informative designs. Incorporating sampling weights in the regression coefficient estimators has been widely discussed (Fuller, 2002; Pfeffermann & Holmes, 1985; Pfeffermann & Lavange, 1989; Skinner, 1989). However, we emphasize the variance component estimators in this article and use an analytic approach to successfully identify the factors that affect the estimation bias. The analytic bias expressions derived 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 baseline information about the outcome variability at each of level of the model (Raudenbush & Bryk, 2002).

Under the superpopulation framework that was adopted in this study, an underlying model is assumed to generate the finite population, and the sampling weights are incorporated to adjust the effect of sampling designs. However, as discussed in Pfeffermann (1993), weighting can also protect against misspecification of the model in producing design consistent estimators, particularly for descriptive statistics, such as the variance components. This issue is not explored here, but it would be an interesting direction to pursue.

One limitation of the analytic expressions presented in this article is that the obtained bias expressions for the first-order weighted estimators only apply to the specific model and sampling designs they were derived for. The results can shed light on the bias of the variance component estimators of more general hierarchical models and sampling designs, but the more direct derivation of the bias expressions would be much more sophisticated.

One suggestion for data users who estimate variance components with small school sample sizes and with multilevel models for which they expect small ICC is to first examine the weights. If the weights are relatively constant at both student and school levels (as for Jurisdiction 1 in the NAEP example), then unweighted estimators of variance components will be less biased than the first-order weighted estimator. If the weights vary at either level, then the second-order weighted estimators are needed for estimating variance components. This difference presents a problem for the typical data user, not only because of the unavailability of commercial software to compute these estimators but also because constructing second-order weights accurately requires a level of knowledge about the design that is not likely to be available. Some examples can be found in which inaccurate assessment of second-order weights used actually could cause more bias than using the first-order weighted estimators. SFW estimators 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.

Appendix A

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

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

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

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

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

and

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

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

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

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

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

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

$$\sum_{s=1}^{M_i} I_{s|i} = m_i, \quad E\left(I_{s|i}\right) = E\left(I^2_{s|i}\right) = \pi_{s|i} = \frac{m_i}{M_i}. \tag{A4}$$

Given the designs, Expression A3 can be further simplified as

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

Therefore,

(continued)

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

and

$$E_{\xi p}\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \bar{y}_{i.FW}^2\right) = E_{\xi I} E_{pI|\xi I} E_{\xi II} E_{pII}\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i} \bar{y}_{i.FW}^2\right)$$
$$= E_{\xi I}\left[\sum_{i=1}^{K} \pi_i w_i \left(\mu^2 M_i + a_i^2 M_i + \frac{\sum_{s=1}^{M_i} \pi_{s|i} w_{s|i}^2}{M_i} \sigma_e^2 + 2\mu a_i M_i\right)\right]$$
$$= E_{\xi I}\left[\sum_{i=1}^{K} \left(\mu^2 + a_i^2 + \frac{1}{m_i} \sigma_e^2 + 2\mu a_i\right) M_i\right].$$

As a result,

$$E_{\xi p}(\mathrm{sse}_{FW}) = E_{\xi I}\left[\sum_{i=1}^{K} (\mu^2 + a_i^2 + \sigma_e^2 + 2\mu a_i) M_i - \sum_{i=1}^{K} \left(\mu^2 + a_i^2 + \frac{1}{m_i} \sigma_e^2 + 2\mu a_i\right) M_i\right]$$
$$= E_{\xi I}\left[\sigma_e^2 \sum_{i=1}^{K} \left(\frac{M_i (m_i - 1)}{m_i}\right)\right]$$
$$= \sigma_e^2 \sum_{i=1}^{K} \left(\frac{M_i (m_i - 1)}{m_i}\right).$$

Meanwhile,

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

The right-hand side of Expression A7 can be written as

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

Equations A6 and A8 together yield

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

and

(continued)

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

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

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

with

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

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

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

Given A4, we have

$$m_{0FW} = \frac{1}{\left(\sum_{i=1}^{K} I_i w_i - 1\right)} \left(\sum_{i=1}^{K} I_i w_i M_i - \frac{\sum_{i=1}^{K} I_i w_i M_i^2}{\sum_{i=1}^{K} I_i w_i M_i} \right). \tag{A16}$$

Note that

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

(continued)

$$+2\mu\sum_{i=1}^{K} I_i w_i a_i M_i + 2\mu\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i}\varepsilon_{is} + 2\frac{\left(\sum_{i=1}^{K} I_i w_i a_i M_i\right)\left(\sum_{i=1}^{K} I_i w_i \sum_{s=1}^{M_i} I_{s|i} w_{s|i}\varepsilon_{is}\right)}{\sum_{i=1}^{K} I_i w_i M_i}.$$

Because

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

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

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

we have

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

However, the expectation of Equation A16 is

$$E_{\xi p}(m_{0FW}) = E_{\xi I}E_{pI|\xi I}\left[\frac{1}{\left(\sum_{i=1}^{K} I_i w_i - 1\right)}\left(\sum_{i=1}^{K} I_i w_i M_i - \frac{\sum_{i=1}^{K} I_i w_i (M_i)^2}{\sum_{i=1}^{K} I_i w_i M_i}\right)\right] \approx E_{\xi I}E_{\xi II}\left[\frac{1}{\left(\sum_{i=1}^{K} \pi_i w_i - 1\right)}\left(\sum_{i=1}^{K} \pi_i w_i M_i - \frac{\sum_{i=1}^{K} \pi_i w_i (M_i)^2}{\sum_{i=1}^{K} \pi_i w_i M_i}\right)\right] = \frac{1}{K-1}\left(\sum_{i=1}^{K} M_i - \frac{\sum_{i=1}^{K} M_i^2}{\sum_{i=1}^{K} M_i}\right) = \frac{1}{K-1}\frac{\sum_{i\neq j}^{K} M_i M_j}{\sum_{i=1}^{K} M_i}.$$

Combining Equations A6, A17, and A18, the delta method gives

(continued)

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

and

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

Notes

1. As indicated in Expressions 11 and 12, the relative biases of s^2_{eFW} and s^2_{aFW} do not depend on the actual value of σ^2_e. So the analytic results in Table 1 can be generalized to cases where ICC = 0.23 with σ^2_e different from 1.

2. The ICC value estimated from NAEP Grade 4 reading 2003 assessment.

References

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

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

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

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers & C. J. Skinner (Eds), *Analysis of survey data* (chap. 3). New York, NY: John Wiley.

2

Brewer, K., & Donadio, M. E. (2003). The high entropy variance of the Hovitz-Thompson estimator. *Survey Methodology*, *29*, 189–196.

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.

Fuller, W. A. (1975). Regression analysis for sample surveys. *Sankhyā Ser. C*, *37*, 117–132.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Ethodology*, *28*, 5–23.

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.

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, *30*, 4–14.

Hartley, H. O., & Sielken, R. L., Jr. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics*, *31*, 411–422.

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*, 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.

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.

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

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

Rubin-Bleuer, S., & Kratina, I. S. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, *33*, 2789–2810.

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

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of complex surveys.* Chichester, England: John Wiley.

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

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, *71*, 495–506.

Authors

YUE JIA is a psychometrician at Educational Testing Service, MS 02T, Rosedale Road, Princeton, NJ 08541; yjia@ets.org. Her primary research interests include hierarchical

linear modeling, sampling and weighting in large educational surveys, and item response theory.

LYNNE STOKES is a professor in the Department of Statistical Science at Southern Methodist University, Dallas, TX 75275. Her primary research interests are in sample design and modeling and estimation of nonsampling errors.

1

IAN R. HARRIS is an associate professor in the Department of Statistical Science, Southern Methodist University, Dallas, TX 75275. His primary research interests include random effects and hierarchical models, sampling issues, and robust estimation procedures.

YAN WANG is currently an economist at the Fannie Mae, 3900 Wisconsin Ave., DC 20016; yan_wang@fanniemae.com. This work was done while she was a student at Southern Methodist University. Her primary research interests include dependence on data with hierarchical structure and item response theory.

Manuscript received August 8, 2009
Accepted October 21, 2009