Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses

Tianji Cai

Sociological Methodology 2013 43: 178
DOI: 10.1177/0081175012460221

Sociological Methodology
43(1) 178–219

DOI: 10.1177/0081175012460221
http://sm.sagepub.com

5

INVESTIGATION OF WAYS TO HANDLE SAMPLING WEIGHTS FOR MULTILEVEL MODEL ANALYSES

Tianji Cai*

Abstract

When analysts estimate a multilevel model using survey data, they often use weighted procedures using multilevel sampling weights to correct the effect of unequal probabilities of selection. This study addresses the impacts of including sampling weights and the consequences of ignoring them by assessing the performance of four approaches: the multilevel pseudo–maximum likelihood (MPML), the probability-weighted iterative generalized least squares (PWIGLS), the naive (ignoring sampling weights), and the sample distribution methods for a linear random-intercept model under a two-stage clustering sampling design. When inclusion probabilities are correlated with the values of outcome variable conditioning on the model covariates, the sampling design becomes informative. The results show that whether a sampling design is informative and at which stage of the sampling design it is informative have substantial impacts on the estimation. The results also show that the level of variation of sampling weights is correlated with the bias of estimates. A higher level of variation of sampling weights is associated with a higher level of bias when a sampling design is informative; however, under a noninformative design, the level of variation of sampling weights may not necessarily associate with biased results. Ignoring an informative sampling design at the first stage will result in biased estimates on the intercept and variance of

*University of Macau, Taipa, Macau

Corresponding Author:
Tianji Cai, University of Macau, Department of Sociology, Av. Padre Tomás Pereira, Taipa, Macau
Email: tjcai@umac.mo

random effect, whereas ignoring an informative sampling design at the second stage will lead to slightly underestimated fixed effects and residual variance, in addition to the biased estimates on the intercept and variance of random effect. Including the sampling weights as the hybrid methods (MPML and PWIGLS) may still produce biased estimates on the intercept and variance of random effect and slightly underestimated fixed effects and residual variance. The sample distribution method may give unbiased estimates, but it depends on the correct specification of the sampling process.

Keywords

sampling weights, multilevel model, sample distribution method

1. INTRODUCTION

Classical statistical models assume that any data being analyzed are a simple random sample; however, using simple random sampling in large-scale surveys is rare. Most data available for social science researchers are from large-scale surveys, in which complex sampling techniques such as unequal probabilities of selection, stratification, and/or cluster sampling are implemented to save money and time. For example, under a two-stage cluster with unequal probabilities design, the population elements are grouped into clusters according to characteristics such as city blocks, schools, and hospitals. Before the elements are selected, a subset of clusters called primary sampling units (PSUs) is selected with unequal probabilities. Elements are then drawn with unequal probabilities from each of the selected PSUs. The key features of such a design are that individual elements within a PSU are usually positively correlated, and the sample is disproportional with respect to the population from which the sample is drawn (Kish 1965; Mickey, Goodwin, and Costanza 1991). Furthermore, if inclusion probabilities are correlated with the values of outcome variable conditioning on the model covariates, the sampling design then becomes informative, because even if the proposed model is true in the population under study, the corresponding model holding in the sample is different from the true population model. Therefore, ignoring the sampling process may lead to biased estimates (Pfeffermann, Da Moura, and Do Nascimento Silva 2006).

Questions have been asked (e.g. Skinner, Holt, and Smith 1989) about how survey data should be modeled when individuals are correlated

within clusters and the sampling design generates a nonrepresentative sample. On one hand, to correct for the disproportionality, usually after the data are collected in a survey, sampling weights are generated to reflect unequal probabilities of inclusion and to compensate for the nonignorable nonresponse and the frame undercoverage. On the other hand, a multilevel model has been widely adopted to handle the correlated data, because it not only accounts for clustering of responses in estimating standard errors of regression coefficients but also allows modeling different effects for different clusters.

Unfortunately, using multilevel model and sampling weights together is not straightforward. It has been widely accepted that weighting of sample data is necessary for the inference on descriptive finite population quantities such as mean, sum, ratio, and so on. However, whether sampling weights need to be used and how they should be used in analytical models have been debated extensively in the literature (e.g. Little 1993; Pfeffermann 1993). The debate is largely between modelers and survey statisticians (see, e.g., Lehtonen, and Pahkinen 2004), who have different theoretical perspectives on the target of inference for survey data.

To a modeler, the target of inference is the parameters of a specified model that generates outcome variable, for example, the regression coefficients, expected values, variances, and so on. A modeler would consider sampling weights irrelevant and thereby to be excluded in the analytical model as long as the probabilities of inclusion were not correlated with the response variable, while other design features, such as clustering or stratification, could be incorporated as an inherent part of the proposed model. If the proposed model between the response variable and covariates is correct, excluding sampling weights does not lead to biased estimates. This type of analysis is referred to as the model-based approach, which relies on the data-generating process. The actual finite population from which the sample is drawn is considered a realization of the infinite possible ones from the specified superpopulation model. The model-based inference proceeds with respect to the sampling distribution of the parameters of interest over repeated realizations generated by the superpopulation model while the selected sample is held fixed (Skinner et al. 1989).

In contrast, a survey statistician usually focuses on a fixed finite population, and the parameters of interest are some descriptive finite population quantities, such as mean, sum, or ratio. Because the population is fixed, to draw a sample, one needs to define a sampling design in which

a random indicator is introduced to each of the population elements. Given the sampling design, the realization of the random indictor is the only source of randomness. If a census were taken, there would be no variability for the parameters of interest. The inference proceeds with respect to the sampling distribution of the parameters of interest of repeated samples (Skinner et al. 1989). Including sampling weights in the estimation and inference is necessary; ignoring them leads to biased estimates, because they reflect unequal sample inclusion probabilities and compensate for differential nonresponse and frame undercoverage. This type of analysis is usually referred to as a design-based or randomization approach.

It seems as if there are irreconcilable differences between design-based and model-based approaches; however, it has been shown that a hybrid approach does exist by linking the two approaches through a so-called census parameter, which is a finite population parameter and is close to the superpopulation parameter (Pfeffermann 1993, 1996). Pfeffermann (1993) argued that under the mixed $p\xi$ distribution, the inference of superpopulation parameters can be made in two steps: (1) inference from the sample to the finite population and (2) inference from the finite population to the superpopulation model. For example, assume that a linear superpopulation model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ generates a finite population with size N in which $\boldsymbol{Y}$ takes value $y_1, \ldots, y_N$ and that a sample with size n ($n \leq N$) is drawn from the finite population with observed $y_1, \ldots, y_n$ and the corresponding covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. If the whole population is observed, one could estimate the regression parameters $\boldsymbol{\beta}$ by the least squares estimator

$$\boldsymbol{B} = \left(\sum\nolimits_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^{'}\right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i y_i.$$

In this case, $\boldsymbol{B}$ are census parameters and are model consistent (ξ consistent) to $\boldsymbol{\beta}$ as the size of finite population N increases to infinite. If data are available only for a subset of the finite population, one must estimate $\boldsymbol{B}$ by $\boldsymbol{b}$ using the sample data,

$$\boldsymbol{b} = \left(\sum\nolimits_{i=1}^{n} \boldsymbol{x}_i w_i \boldsymbol{x}_i^{'}\right)^{-1} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i w_i y_i.$$

The estimator $\boldsymbol{b}$ is design consistent, which means it approaches to $\boldsymbol{B}$ as both n, the sample size, and N, the finite population size, tend to be

infinite (e.g., Binder and Roberts 2003). The reason one should restrict to the design consistent estimator $\boldsymbol{b}$ is the robustness. As Box (1979) stated, ''all models are wrong; some models are useful.'' If the superpopulation model is misspecified, the $\boldsymbol{B}$ still has a meaningful substantive interpretation, which is the best linear approximation to the Y value in the finite population, whereas the $\boldsymbol{\beta}$ may not (Pfeffermann 1993).

As mentioned above, the design-based and model-based approaches are also different in how they define the variability of the estimated parameters: the model-based approach estimates the variance under the model, whereas the design-based approach estimates the variance under the randomization distribution. Theoretically, a hybrid approach should include variability under both the model and the randomization distribution. However, empirically, in most cases, the population size is much larger than the sample size; the variance of the design consistent estimator $\boldsymbol{b}$ under the $p\xi$ distribution is approximately the same as the design-based variance (Pfeffermann 1993).

In addition to the hybrid approach, model-based approaches have also been proposed to incorporate the probabilities of inclusion when they are correlated with the response variable. Although the design variables and their interactions can be included as an inherent part of the proposed model, doing this may lead to unstable estimates (Cook and Gelman 2006). Krieger and Pfeffermann (1992, 1997) proposed the so-called sample distribution method, which extracted the parametric distribution of the sample data as a function of the superpopulation model and the sampling design. For example, denote the superpopulation distribution of response variable Y as $f_p(Y_i|X_i)$, and let I_i be the indicator of whether or not a population member is selected into the sample. Then the sample distribution of Y $f_s(y_i|x_i)$ can be written as

$$f_s(y_i|x_i) = f_p(Y_i|X_i, I_i = 1) = \frac{Prob(I_i = 1|Y_i, X_i) \times f_p(Y_i|X_i)}{Prob(I_i = 1|X_i)},$$

where the subscripts p and s refer to the population and sample, respectively. When an unequal probability sample design is used, the selected sample may not be representative of the population. Although the theoretical model $f_p(Y_i|X_i)$ might not be exactly the same as the model $f_s(y_i|x_i)$ that holds in the sample, $f_s(y_i|x_i)$ is a function of $f_p(Y_i|X_i)$ and the distribution of I_i conditional on Y_i and X_i. In general, the probability of inclusion $\pi_i = Prob(I_i = 1|Y_i, Z_i)$ is not the same as $Prob(I_i = 1|Y_i, X_i)$,

where Z_i denotes the other design variables, and X_i is the set of independent variables in the population model $f_p(Y_i|X_i)$. Pfeffermann, Skinner, et al. (1998) showed that

$$Prob(I_i = 1|Y_i, X_i) = \int Prob(I_i = 1|Y_i, X_i, \pi_i) f(\pi_i|Y_i, X_i) d\pi_i = E_p(\pi_i|Y_i, X_i).$$

Thus,

$$f_s(y_i|x_i) = = \frac{E_p(\pi_i|Y_i, X_i) \times f_p(Y_i|X_i)}{E_p(\pi_i|X_i)} = \frac{E_p(\pi_i|Y_i, X_i) \times f_p(Y_i|X_i)}{\int E_p(\pi_i|Y_i, X_i) \times f_p(Y_i|X_i) dY_i}.$$

If one can specify $E_p(\pi_i|Y_i, X_i)$, then the model held in the sample $f_s(Y_i|X_i)$ can be derived. Although the exact form of $E_p(\pi_i|Y_i, X_i)$is usually unknown, under some regularity conditions, it can be approximated by low-order polynomials in terms of Y_i and X_i or by exponentials via the Taylor series approximation. For example, under an exponential model,

$$E_p(\pi_i|Y_i, X_i) \approx exp\left(\sum_{j=0}^{J} A_j Y_i^j + h(X_i)\right).$$

The corresponding sample distribution can be written as

$$f_s(y_i|x_i) = \frac{exp\left(\sum_{j=0}^{J} A_j Y_i^j\right) \times f_p(Y_i|X_i)}{\int exp\left(\sum_{j=0}^{J} A_j Y_i^j\right) \times f_p(Y_i|X_i) dY_i}.$$

The unknown parameters A_js can be estimated using the sample data. Pfeffermann and Sverchkov (1999) showed that those population expectations can be identified and estimated from the sample data using the following relationships (see Pfeffermann and Sverchkov 1999, for detailed proofs):

$$E_p(\pi_i|Y_i, X_i) = \frac{1}{E_s(w_i|y_i, x_i)},$$

$$E_p(\pi_i|X_i) = \frac{1}{E_s(w_i|x_i)}.$$

Although the hybrid and the sample distribution approaches have been extended to incorporate multilevel sampling design (e.g., Korn and Graubard 1995, 2003; Rabe-Hesketh and Skrondal 2006; Pfeffermann et

al. 2006; Pfeffermann, Skinner, et al. 1998; Eideh and Nathan 2009), it is still unclear which of those methods is more appropriate, because the performance of each approach may depend on the actual survey design and data features (Asparouhov 2006). Published studies have focused on comparing weighting schemes for the hybrid approach. The literature is sparse in terms of comparing the sample distribution and hybrid approaches. The goal of this article is to fill these gaps in the literature by focusing on various ways to handle sampling weights for a linear random-intercept model under a two-stage cluster with unequal probabilities design. In this study, the sampling weights are defined as the reciprocal of the PSUs, or individual probabilities of inclusion. Issues such as using weights to compensate for nonresponse or undercoverage are not discussed.

The article is organized as follows. The details of the hybrid and sample distribution methods that are evaluated are presented in section 2. Section 2.1 gives details of how two widely used hybrid procedures incorporate sampling weights, section 2.2 describes how a sample distribution approach includes sampling information, and section 2.3 discusses the scaling issue of sampling weights for the hybrid approach. Closer inspection of these methods, mainly through simulation studies, is provided in section 3. The results of simulation studies are presented and discussed in section 4. Section 5 includes conclusions and a real data example with more discussion. Technical details and codes used in the study are provided in Appendices A and B, respectively.

2. MULTILEVEL MODELS FOR SURVEY DATA

Laird and Ware (1982) outlined a general form of the two-level linear model,

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\beta + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{e_i}.$$

In the above equation, i indexes the cluster, with $i = 1$ to m, where m is the number of clusters. For the ith cluster with size n_i, $\boldsymbol{Y_i}$ is an $n_i \times 1$ vector of observed response, $\boldsymbol{X_i}$ is an $n_i \times p$ observed matrix for fixed effects, β is a $p \times 1$ vector of unknown coefficients, $\boldsymbol{Z_i}$ denotes an $n_i \times q$ random-effect design matrix, $\boldsymbol{b_i}$ is a $q \times 1$ vector of cluster-specified random effects, and $\boldsymbol{e_i}$ is an $n_i \times 1$ vector of random residual errors, where p is the number of unknown coefficients including

the intercept and q is the number of random effects. Because the focus is on the random-intercept model, q equals 1. Introducing a cluster-specified random effect not only controls the correlation within clusters, but also corrects the denominator degrees of freedom for the number of clusters. Searle, Casella, and McCulloch (1992) provided a detailed derivation of the maximum likelihood estimator. For example, the likelihood function for the linear mixed model above is defined as

$$L\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\beta,\boldsymbol{D},\sigma_e^2\right)=\frac{\exp\left(-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta)\right)}{(2\pi)^{\frac{N}{2}}|\boldsymbol{V}|^{\frac{1}{2}}},$$

where $\boldsymbol{V}$ is the covariance matrix of vector $\boldsymbol{Y}$, $\boldsymbol{V}=\boldsymbol{ZDZ}'+\sigma_e^2\boldsymbol{I}$, $\boldsymbol{D}$ denotes the covariance matrix for the random effect vector $\boldsymbol{b_i}$, and in our case, it is a scalar $\boldsymbol{\sigma}_u^2$, and $\boldsymbol{\sigma}_e^2$ is the variance of the error term. Then the log likelihood function can be written as

$$l=\log L\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{D},\sigma_e^2\right)=-\frac{1}{2}N\log(2\pi)-s\frac{1}{2}\log|\boldsymbol{V}|-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta),$$

where N is the total number of observations, $N=\sum_{i=1}^{m}n_i$. The unknown parameters (fixed coefficients and variance components) can be solved by either the full maximum likelihood or the restricted maximum likelihood method.

2.1. The Hybrid Approaches

One difficulty of using sampling weights in multilevel models is the proper incorporation of sampling weights into the estimation. Unlike a single-level regression, in which sampling weights are inserted into sums of squares and cross-products, direct insertion of final-level weights, which are the product of multilevel weights, might lead to biased estimates in multilevel models (Christ, Biemer, and Wiesen 2007). Also, single final-level weights may not carry adequate information to correct for higher level unequal probabilities of inclusion (Pfeffermann, Skinner, et al. 1998). For example, under a two-stage cluster sampling design, both clusters and individuals can be chosen with unequal probabilities.

If the whole finite population were included in the data, a finite population likelihood function could be constructed. However, although

only sample data and the sampling weights are available, the unknown parameters have to be solved by maximizing the weighted sample likelihood, which is design consistent to the finite population likelihood. There are two widely used estimation methods to solve the weighted sample likelihood. Rabe-Hesketh and Skrondal (2006) and Asparouhov (2004, 2006) proposed the multilevel pseudo–maximum likelihood (MPML) method, which directly estimated the population likelihood function by weighting the sample likelihood function,

$$L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \prod_{i=1}^{m} \left(\int \left(\prod_{j=1}^{n_i} f\left(y_{ij} | \boldsymbol{x}_{ij}, u_i, \boldsymbol{\theta}_1\right)^{\lambda_{2i} w_{j|i}} \right) \phi(u_i | \boldsymbol{z}_i, \boldsymbol{\theta}_2) du_{\mathrm{i}} \right)^{\lambda_{1i} w_{\mathrm{i}}},$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are census parameters for the fixed effects for the individual level and the cluster level, respectively; μ_i is the cluster-specific random effect; and λ_{1i} and λ_{2i} are the scaling factors for the cluster-level and individual-level sampling weights, respectively. Numerical techniques are needed to integrate out the unobserved random effect μ_i or to approximate the weighted likelihood. Mplus (Muthén and Muthén 1998–2011) has implemented this method with the variance estimated by the robust variance estimator, which takes the following form:

$$\left(\frac{\partial^2 \log L}{\partial \theta^2} \right)^{-1} \left(\sum_{i=1}^{m} (\lambda_{1i} w_i) \frac{\partial \log L}{\partial \theta} \left(\frac{\partial \log L}{\partial \theta} \right)' \right) \left(\frac{\partial^2 \log L}{\partial \theta^2} \right)^{-1}.$$

Goldstein (1986) developed an iterative generalized least squares (IGLS) algorithm involving iterations between estimation of the fixed effects and variance components for a linear multilevel model. Pfeffermann, Skinner, et al. (1998) proposed the probability-weighted iterative generalized least squares (PWIGLS) approach, in which the population quantities in IGLS solutions of the fixed effects and variance components for a linear multilevel model were replaced by their corresponding weighted sample statistics. For instance, for a linear random-intercept model, the IGLS solution for the fixed effects is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}\right) = \left(\sum\nolimits_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i' \right)^{-1} \left(\sum\nolimits_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{Y}_i' \right),$$

The PWIGLS replaced the population quantities, for example, $\sum_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i'$, by the weighted sample statistics

$$\sum_i \boldsymbol{w}_i \left(\sum_j \boldsymbol{w}_{j|i} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}' - \frac{\left(\sum_j \boldsymbol{w}_{j|i} \boldsymbol{x}_{ij} \right)^2}{\sum_j \boldsymbol{w}_{j|i} + \frac{\sigma_e^2}{\sigma_u^2}} \right).$$

This algorithm has been implemented in commercial packages such as LISREL (Mels 2006). Given the sufficiently small sampling fractions at both levels, the variance of PWIGLS estimates is close to the randomization variance and can be estimated using the delta method (Pfeffermann, Skinner, et al. 1998).

2.2. Scaling Sampling Weights for Multilevel Models

For the weighted multilevel procedures, it is well known that the estimated variance components are biased for small cluster sizes such as 20 (e.g., Rabe-Hesketh and Skrondal 2006). Many studies have been done to correct this by rescaling sampling weights. For example, Pfeffermann, Skinner, et al. (1998) suggested that if the cluster-level sampling weights were noninformative, scaling the individual-level sampling weights for the *i*th cluster by the factor

$$\lambda_{2i} = \frac{\sum_{j=1}^{n_i} w_{j|i}}{\sum_{j=1}^{n_i} w_{j|i}^2},$$

produced approximately unbiased estimators for both variance components. If both levels of sampling design were informative, in their simulation study, scaling the individual-level sampling weights for the *i*th cluster by the factor λ_{2i}, defined as

$$\lambda_{2i} = \frac{n_i}{\sum_{j=1}^{n_i} w_{j|i}},$$

worked better, where n_i is the number of sample units in *i*th cluster.

These two scaling methods, referred to as ECluster and Cluster in the Mplus documentation, have been implemented in Mplus Asparouhov (2008). LISREL automatically applies Cluster scaling for both level of sampling weights during the estimation (Scientific Software International 2005–2012).

2.3. The Sample Distribution Approach

The sample distribution method proposed by Krieger and Pfeffermann (1992, 1997), Pfeffermann, Krieger, and Rinott (1998), and Pfeffermann

and Sverchkov (1999) has been extended to multilevel cases. Pfeffermann, Moura, and Silva (2006) and Eideh and Nathan (2009) showed that the conditional sample distribution of the cluster-level random effect u_i is

$$f_s(u_i|\boldsymbol{z}_i) = \frac{E_p(\pi_i|u_i, \boldsymbol{z}_i) \times f_p(u_i|\boldsymbol{z}_i)}{E_p(\pi_i|\boldsymbol{z}_i)},$$

$$E_p(\pi_i|\boldsymbol{z}_i) = \int E_p(\pi_i|u_i, \boldsymbol{z}_i) \times f_p(u_i|\boldsymbol{z}_i) du_i,$$

where $\boldsymbol{z}_i$ is an $n_i \times q$ matrix of cluster-level auxiliary predictors for the random effect u_i, and $f_p(u_i|\boldsymbol{z}_i)$ is the population distribution of the random effect u_i conditional on covariates $\boldsymbol{z}_i$. Similarly, the conditional sample distribution of y_{ij} given $u_i, \boldsymbol{x}_{ij}$ is

$$f_s\left(y_{ij}|\boldsymbol{x}_{ij}, u_i\right) = \frac{E_p\left(\pi_{j|i}|u_i, \boldsymbol{x}_{ij}, y_{ij}\right) \times f_p\left(Y_{ij}|\boldsymbol{X}_{ij}, u_i\right)}{E_p\left(\pi_{j|i}|u_i, \boldsymbol{x}_{ij}\right)},$$

$$E_p\left(\pi_{j|i}|u_i, \boldsymbol{x}_{ij}\right) = \int E_p\left(\pi_{j|i}|u_i, \boldsymbol{x}_{ij}, y_{ij}\right) \times f_p\left(y_{ij}|\boldsymbol{x}_{ij}, u_i\right) dy_{ij},$$

where $\boldsymbol{x}_{ij}$ is a $1 \times p$ vector of individual level predictors. Thus, the joint-sample likelihood function can be written as

$$f_s(\boldsymbol{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_s\left(y_{ij}|\boldsymbol{x}_{ij}, u_i\right) f_s(u_i|\boldsymbol{z}_i) du_i,$$

which can be maximized by any standard procedures.

Eideh and Nathan (2009) approximated the conditional expectations on both levels by the first-order exponential model, and then the joint-sample likelihood function was simplified to a multivariate normal distribution with a shifted intercept, while the variance remained the same as in the population model. The detailed derivation for a linear multilevel model with two-stage first-order exponential model adopted from Eideh and Nathan (2009) can be found in Appendix A1.

For simplicity's sake, the exponential model is appealing. But usually, researchers may not have enough information to determine whether the exponential approximation is appropriate. Besides the exponential approximation, the conditional expectations $E_p(\pi_i|u_i, \boldsymbol{z}_i)$ and $E_p\left(\pi_{j|i}|u_i, \boldsymbol{x}_{ij}, y_{ij}\right)$ can also be modeled by the logistic model, and

then $E_p(\pi_i|\boldsymbol{z}_i)$ and $E_p\left(\pi_{j|i}|u_i,\boldsymbol{x}_{ij}\right)$ can be approximated by the Laplace approximation. For example, assume that the cluster-level sample indicator I_i follows a logistic model,

$$E_p(\pi_i|u_i) = \frac{1}{1+exp(-\alpha_0 - \alpha_1 u_i)},$$

$$E_p(\pi_i) = \int \frac{1}{1+exp(-\alpha_0 - \alpha_1 u_i)} du_i.$$

Because u_i follows the normal distribution, the expectation can be approximated by the Laplace approximation. For example, the marginal conditional expectation can be approximated by

$$E_p(\pi_i) \approx \frac{exp\left(\alpha_0 + \alpha_1\sigma_u u_{(1)} - \frac{u_{(1)}^2}{2}\right)}{\sqrt{\alpha_1^2\sigma_u^2 exp\left(\alpha_0 + \alpha_1\sigma_u u_{(1)}\right) + \left[1+exp\left(\alpha_0 + \alpha_1\sigma_u u_{(1)}\right)\right]^2}},$$

where

$$u_{(1)} = \frac{\alpha_1\sigma_u[1+exp(\alpha_0)]}{\alpha_1^2\sigma_u^2 exp(\alpha_0) + [1+exp(\alpha_0)]^2}.$$

Similarly, assume that $I_{j|i}$ follows a logistic model

$$E_p\left(\pi_{j|i}|y_{ij}\right) = \frac{1}{1+exp\left(-b_0 - b_1 y_{ij}\right)},$$

$$E_p\left(\pi_{j|i}\right) = \int \frac{1}{1+exp\left(-b_0 - b_1 y_{ij}\right)} dy_{ij}.$$

Thus, the marginal conditional expectation can be approximated by

$$E_p\left(\pi_{j|i}\right) \approx \frac{exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right) - \frac{s_{(1)}^2}{2}\right)}{\sqrt{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1+exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2}},$$

$$s_{(1)} = \frac{b_1\sigma_e\left(1+exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right)}{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1+exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2}.$$

Substituting the approximated $E_p(\pi_i)$ and $E_p\left(\pi_{j|i}\right)$, the joint-sample likelihood $f_s(\boldsymbol{y})$ can then be maximized by numerical techniques to

integrate out the unobserved random effect u_i. The detailed derivation and proof can be found in Appendix A2.

The sample distribution method relies on both model and randomization distributions. The variability of the parameter estimates is from both the model that generates the data and the differences across all possible samples. Eideh and Nathan (2009) estimated the variance of the sample distribution estimator by bootstrapping. It follows from Pfeffermann (1993) that if the sampling fraction is small, the variance under the $p\xi$ distribution can be estimated by the randomization variance, which again can be implemented using the delta method.

3. SIMULATION STUDY

To evaluate the performance of the hybrid and sample distribution approaches for linear multilevel model under informative sampling designs, I conduct a simulation study on the following model:

$$Y_{ij} = 2 + .5X_{1ij} + .8X_{2ij} - .5X_{3ij} + u_i + e_{ij},$$

where u_i is the normally distributed cluster-level random effect with mean 0 and variance 3, and e_{ij} is the normally distributed individual-level error term with mean 0 and variance 6. Explanatory variable X_1 follows a Bernoulli distribution with mean 5 (e.g., gender). X_2 is from a normal distribution $N(12, 9)$ (e.g., years of schooling). X_3 follows a uniform distribution $U(0, 20)$ (e.g., centered age). I adopt the infinite target population approach (Asparouhov 2005) to generate samples. For example, let *Y, X*, and *I* be the model outcome, independent variables, and the inclusion indicator, respectively. A selection model for *I*, $p(I = 1) = f(Y, X)$, is specified. Individual elements (*Y, X, I*) are included in a sample only if $I = 1$. The informative selection model at each level is defined by

$$p(I_i = 1) = \frac{1}{1 + exp(-\alpha_0 - \alpha_1 \times u_i)},$$

$$p\left(I_{j|i} = 1\right) = \frac{1}{1 + exp\left(-b_0 - b_1 \times y_{ij}\right)}.$$

When the selection at cluster or individual level is noninformative, u_i or y_{ij} is replaced by another random variable that is not part of the population model. Because there are many factors that have substantial influence on the quality of the estimation (e.g., the sample size of

cluster, the intraclass correlation [Asparouhov and Muthén 2006]), to simplify our study, I focus only on the effect of the informative sampling design and the unequal probabilities inclusion. Other parameters are set to be fixed as in a typical social science scenario, for instance, a relatively large sample size (100 clusters of size 50) and a moderate intraclass correlation such as .33.

By varying α_0, α_1, b_0, and b_1, samples with different levels of the informativeness of selection and variation of sampling weights can be produced. In reality, the variation of sampling weights can vary greatly across levels and studies. For example, for the data from the 2005–2006 National Survey of Children with Special Health Care Needs, the relative variances of the sampling weights (defined as variance over the squared mean) are 1.80 and 1.92 at household and individual levels, respectively (Carle 2009), while for the first-wave core sample from the National Longitudinal Study of Adolescent Health (Add Health; Harris et al. 2009), they are 3.00 and 0.89 at the school and individual levels, respectively. In this study, I fix α_0 and b_0 at negative and vary α_1 and b_1 (both take positive values) to obtain the relative variances of the sampling weights for two levels, for example, approximately 6.0 and 4.0 as high, 3.0 and 2.0 as moderate, and 2.0 and 1.0 as low. Therefore, for each level of variation of sampling weights, I conduct four simulations according to the level of informativeness. In design 1, both levels of selection are informative; design 2 and design 3 are informative at the second level and the first level of selection, respectively; and in design 4, neither level is informative.

Each simulation is replicated 500 times, with two hybrid estimators—MPML using Mplus and PWIGLS using the LISREL MULTILEV module—and two approximation models for the sample distribution estimator. For the MPML estimator, I consider three scaling methods: method A (the Cluster method in Mplus), method B (the ECluster method in Mplus), and method U (unscaled). SAS PROC NLMIXED is used to estimate the sample distribution estimator with the two-stage first-order exponential model, and the variance is estimated by the model-based approach (see Appendix A1 for more discussion of this), where the one using the logistic model is implemented in SAS IML and its variance is estimated using the delta method. The nonlinear optimization of the likelihood function is carried out by the IML function NLPNRA, and the corresponding Hessian matrix is obtained by using function NLPFDD.

The informativeness of sampling design is evaluated by the χ^2 test constructed by Pfeffermann (1993):

$$I = \left(\hat{\theta}_w - \hat{\theta}_0\right)' \left[\hat{V}\left(\hat{\theta}_w\right) - \hat{V}\left(\hat{\theta}_0\right)\right]^{-1} \left(\hat{\theta}_w - \hat{\theta}_0\right) \tilde{\chi}_p^2,$$

where $\hat{\theta}_w$ and $\hat{\theta}_0$ are the estimates of weighted and unweighted analyses, respectively, and $\hat{V}\left(\hat{\theta}_w\right)$ and $\hat{V}\left(\hat{\theta}_0\right)$ are their variance estimates. The statistic approximately follows a χ_p^2 distribution with $p = dim(\theta)$ degrees of freedom. The quality of estimates is evaluated by using the empirical relative bias, the empirical mean square error (MSE), and the coverage rate. The relative bias is defined as

$$RBias\left(\hat{\theta}\right) = \frac{1}{\theta}\left(\frac{1}{500}\sum_1^{500}\left(\hat{\theta}_i - \theta\right)\right).$$

The root of MSE is calculated using the following formula

$$RMSE\left(\hat{\theta}\right) = \sqrt{\left(\frac{1}{500}\sum_1^{500}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2\right)},$$

where $\bar{\hat{\theta}} = \frac{1}{500}\sum_1^{500}\hat{\theta}_i$. The coverage rate is calculated as the percentage of true parameter that falls within the *t* test–based 95 percent confidence region of estimates for each replicated sample.

4. RESULTS

4.1. The Effect of Informativeness at a High Level of Variation

Figure 1 shows the relative bias, root of MSE, and coverage rate at high-level variation of sampling weights. About 99.6 percent, 78.8 percent, 81.8 percent, and 49.2 percent of the tests of the informativeness are significant at the .05 level in designs 1 to 4, respectively. When both of the sampling stages are informative in our simulation setup, the clusters with higher value of u_i and the individuals with higher response value are more likely to be included in the sample. Hence, if the sampling design is ignored, the variance of u_i, σ_u^2, would be underestimated, and the intercept would be overestimated. Because the explanatory variables in a linear model are treated as fixed, ignoring the sampling design would not produce biased estimates for other fixed effects. If σ_u^2 is

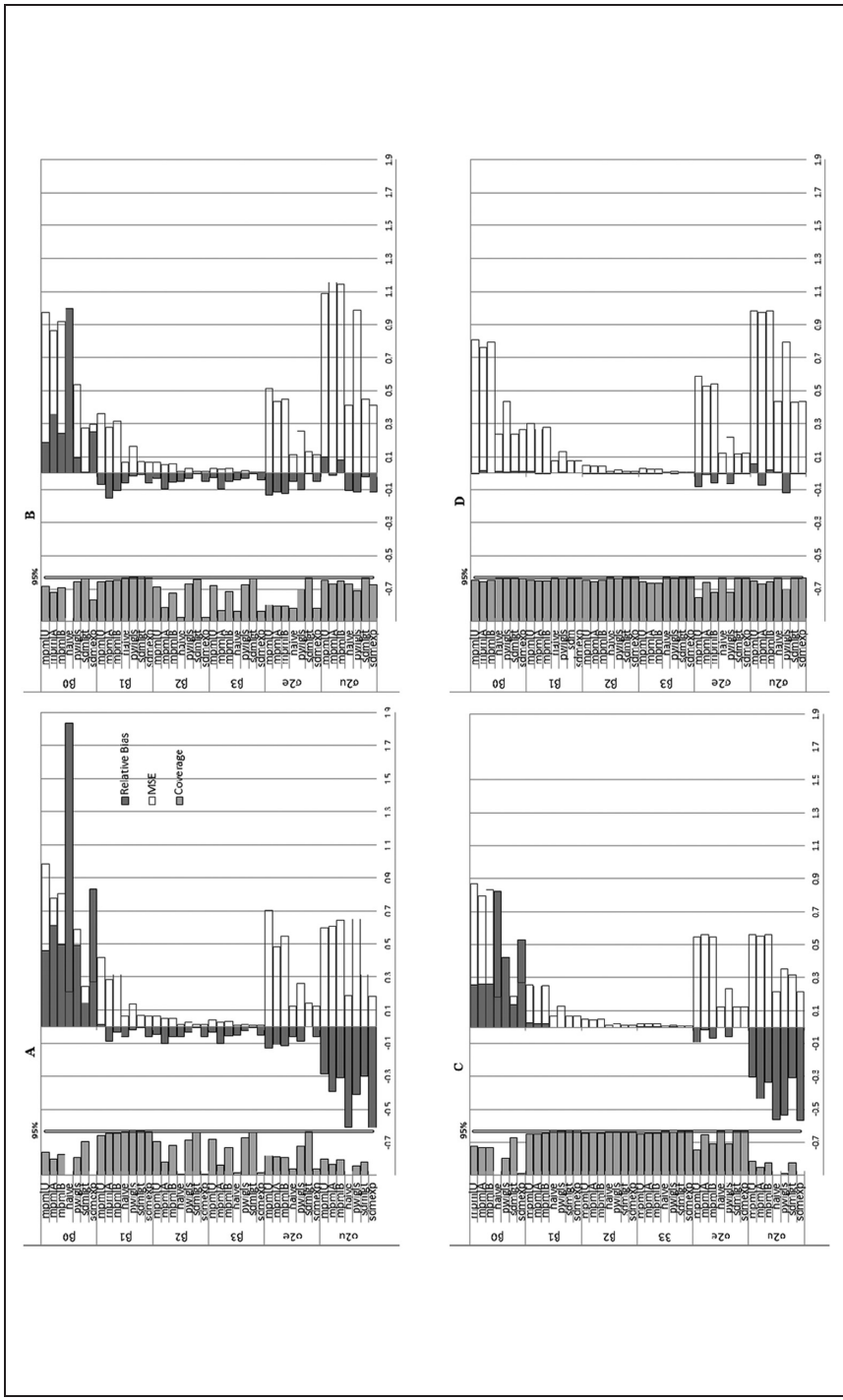

Figure 1. The relative bias, mean square error (MSE), and coverage rate at a high level of variation of the sampling weights.
Note: The left panel of each chart represents the coverage rate. The relative bias is indicated by the blue bar, and the blank bar represents the MSE. (A) Design 1: both stages informative. (B) Design 2: informative at the second stage. (C) Design 3: informative at the first stage. (D) Design 4: both stages noninformative.

underestimated, then the confidence interval would be too narrow to cover the true parameter.

The results from design 1 are presented in Figure 1A. All estimates for β_0 and σ_u^2 are biased: β_0 is overestimated, and σ_u^2 is underestimated. All other fixed effects and σ_e^2 are also underestimated except for those from the sample distribution estimator using the logistic model, which produces unbiased results.

Because σ_u^2 is substantially underestimated, the coverage rates of the naive method are very poor, for example, 37.6 percent, 43 percent, and 68 percent for β_2, β_3, and σ_e^2, respectively, along with 0 percent coverage for β_0 and σ_u^2. The MPML estimates are slightly biased for β_0 and σ_u^2, where the biases are within 15 percent. Compared with other estimators, the MPML estimates have the highest MSE. The coverage rates of three MPML estimates for fixed effects and variance components vary from 69 percent to 95 percent and from 54.4 percent to 76.4 percent, respectively. Scaling method B slightly outperforms method A, which confirms Pfeffermann, Skinner, et al.'s (1998) finding. However, the unscaled method works slightly better than methods A and B. The PWIGLS estimates are relatively better than MPML estimates for the fixed effects and σ_e^2 but are a bit worse for σ_u^2. The sample distribution estimator using logistic model produces unbiased estimates and very good coverage rates on all parameters except β_0 and σ_u^2. Although the estimates for β_0 and σ_u^2 are biased, the β_0 estimates are less biased, and coverage rates for β_0 and σ_u^2 are better than any other estimates. The sample distribution estimator using the first-order exponential model does not work well. It suffers all the problems the naive method has, except for having a less biased β_0.

When the sampling design is informative at the second stage, the individuals with higher value of y_{ij} are more likely to be selected. Therefore, the estimates of β_0 will be positively biased if the sampling design is ignored. Figure 1B shows the results from this design. All estimators tend to overestimate β_0 and underestimate other fixed effects and σ_e^2, except for the sample distribution estimator using logistic model, which has almost unbiased estimates for all parameters.

As expected, the naive estimates are positively biased on β_0. The naive method also underestimated the variance components, which leads to poor coverage rates for some parameters, such as β_2, β_3, and σ_e^2. Negative biases appear for most of the MPML estimates, except β_0 and σ_u^2 under methods B and U. Similar to the naive estimator, all the MPML

estimates provide poor coverage rates for β_2, β_3, and σ_e^2. It is hard to say which scaling method works better. Still, method A works slightly better than method B only for the variance components, and method U works slightly better than methods A and B in estimating the fixed effects. The PWIGLS estimates are better than all MPML estimates in terms of bias and coverage rate, except for σ_u^2, which is underestimated by 11.4 percent with a 65.4 percent coverage rate. Again, the sample distribution method with the logistic model produces the best estimates for all parameters.

When the sampling design is informative at the first stage, ignoring the sampling design will result in overestimated β_0 and underestimated σ_u^2. Figure 1C presents the results from design 3. Similar to design 1, all estimates for β_0 and σ_u^2 are biased. The hybrid methods also tend to underestimate σ_e^2. All estimates on the other fixed effects are unbiased, with almost 95 percent coverage rates.

If the sampling design is noninformative at both stages, all parameters should be correctly estimated, and the naive estimator should be most efficient. As shown in Figure 1D, all estimates for the fixed effects are unbiased, but the hybrid methods tend to slightly underestimate σ_e^2, and the PWIGLS also tend to underestimate σ_u^2. Among all estimators, the naive estimator produces the lowest MSE, while the MPML estimator gives the highest MSE.

To summarize the results presented in Figure 1, whether a sampling design is informative and at which stage of the sampling design is informative have substantial impacts on the estimation. For the naive method, an informative sampling design at the first stage will result in biased estimates on β_0 and σ_u^2, whereas an informative sampling design at the second stage will lead to slightly underestimated fixed effects and σ_e^2, besides the biased estimates on β_0 and σ_u^2. Although the hybrid methods produce biased estimates when a sampling design is informative, only estimates on β_0 and σ_u^2 are severely biased, and the biases on the other fixed effects and σ_e^2 usually are within 15 percent. Nevertheless, at a high level of variation of the sampling weights, the sample distribution estimator using logistic model, which is the correct sampling model, works reasonably well and outperforms most of other estimators. However, the sample distribution method with exponential model does not work well. It works better than the naive method only on estimation of β_0. Hence, the sample distribution method is not robust to the misspecification of the sampling process in our setup.

4.2. The Effect of Variation

To further explore the effect of variation of sampling weights on estimation, I conduct a simulation using the same setup for the informativeness but vary the level of variation. The parameters α_1, and b_1 in the selection model are kept the same for each design, while α_0 and b_0 are changed to obtain different level of variation. Figure 2 reports the results of designs 1 to 3 under moderate and low levels of variation. It is clear that all biases reduce as the level of variation reduces, which confirms Asparouhov's (2006) finding. The naive estimates are still substantially biased for β_0 and σ_u^2. Although the estimates for the other fixed effects and σ_e^2 are close to unbiased, the coverage rates are not satisfactory when the second stage is informative: designs 1 and 2. The hybrid methods estimate all fixed effects and σ_e^2 reasonably well, but the estimated σ_u^2 is still biased about 30 percent, even under the low level of variation. Again, the sample distribution estimator using the logistic model produces unbiased estimates for almost all parameters, with 95 percent of coverage rates for all informative designs. With the exponential model, the sample distribution method works better than the naive method but worse than the hybrid methods.

Since the estimates of fixed effects are close to unbiased for design 4, only the results for β_0, σ_u^2, and σ_e^2 are reported in Table 1. In brief, with a large number of clusters and a sufficient cluster sample size, all methods perform well when the sample design is noninformative. The increased level of variation does not associate with the increase of bias, except for the PWIGLS method: The negative bias increases as the level of variation at the second level decreases.

4.3. The Effect of Trimming Sampling Weights

Large variation of sampling weights is associated with high variability and bias of parameter estimates. In practice, variation reduction techniques such as weight truncation or trimming can be imposed to reduce biases. To investigate the performance of weight trimming on hybrid methods, I use eight trimming levels in the simulation conducted in section 4.1.[1] For example, the weight beyond or below the following quantiles is trimmed at the left and right tails: 1 percent, 2.5 percent, 5 percent, 10 percent, 15 percent, 20 percent, 25 percent, and 30 percent. The relative variation of sampling weights for each level of trimming is

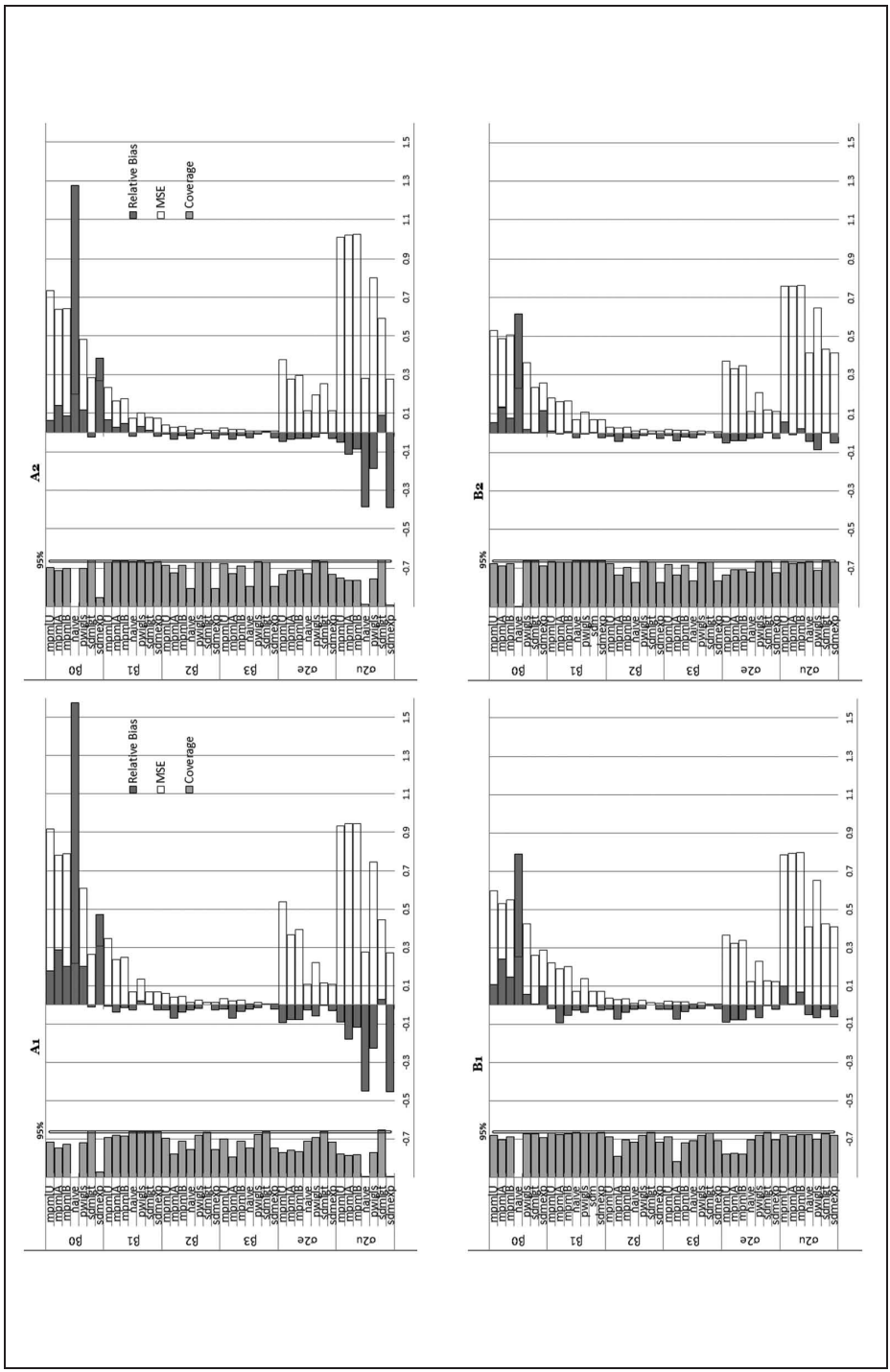

Figure 2. *(continued)*

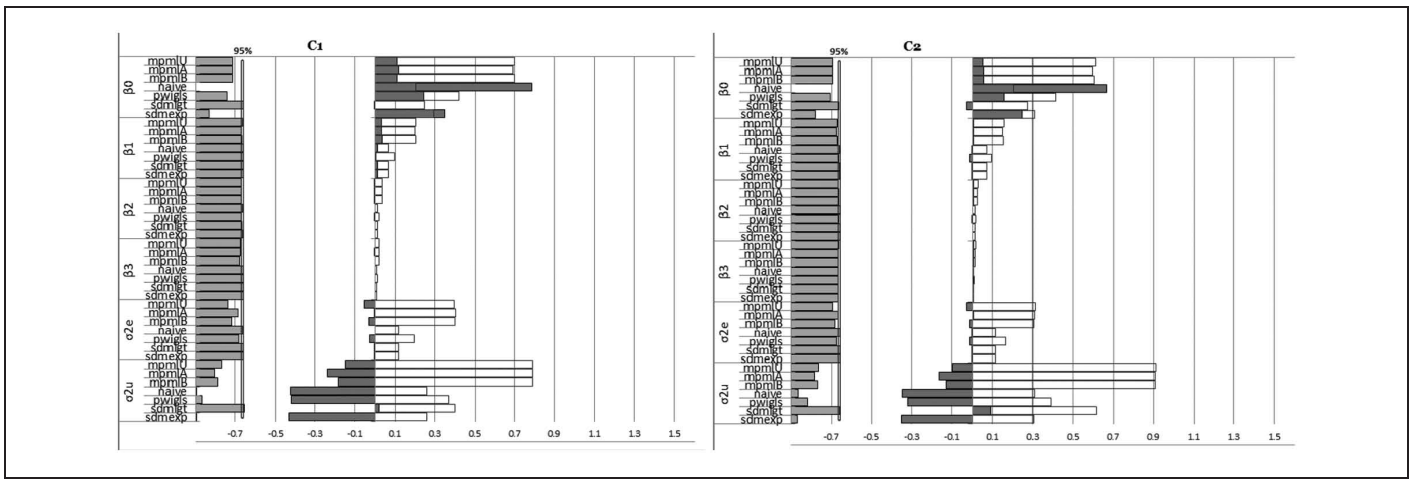

Figure 2. The relative bias, mean square error (MSE), and coverage rate at different levels of variation of the sampling weights.

Note: The left panel of each chart represents the coverage rate. The relative bias is indicated by the blue bar, and the blank bar represents the MSE. (A1) Design 1 at moderate-level variation. (A2) Design 1 at low-level variation. (B1) Design 2 at moderate-level variation. (B2). Design 2 at low-level variation. (C1) Design 3 at moderate-level variation. (C2) Design 3 at low-level variation.

Table 1. Relative Bias and Coverage Rates under Different Levels of Variation of Sampling Weights

First-level Variation Parameter	Second-level Variation	High: 6			Moderate: 3			Low: 2		
		High: 4	Moderate: 2	Low: 1	High: 4	Moderate: 2	Low: 1	High: 4	Moderate: 2	Low: 1
β_0	MPML U	–0.014 (88.8)	0.001 (88.8)	–0.020 (86.2)	0.024 (92.8)	0.007 (91.2)	–0.011 (91.4)	0.023 (93.2)	0.007 (92.8)	0.018 (90.6)
	MPML A	–0.015 (88.2)	0.012 (90.2)	–0.014 (88.8)	0.026 (90.2)	0.009 (91.8)	–0.005 (92.0)	0.028 (93.2)	0.015 (92.4)	0.016 (90.8)
	MPML B	–0.015 (88.4)	0.005 (88.6)	–0.018 (87.6)	0.026 (91.4)	0.008 (91.2)	–0.010 (91.0)	0.027 (93.6)	0.011 (92.4)	0.017 (91.0)
	Naive	–0.001 (94.0)	–0.007 (91.0)	–0.018 (94.8)	0.010 (94.2)	0.008 (96.6)	–0.003 (94.8)	–0.003 (94.2)	0.014 (94.4)	0.009 (93.2)
	PWIGLS	–0.010 (92.2)	–0.005 (94.2)	–0.029 (94.2)	0.017 (95.4)	0.018 (93.0)	–0.008 (93.4)	0.000 (93.8)	0.011 (95.8)	0.005 (93.4)
	SDMLGT	–0.001 (94.0)	–0.007 (91.0)	–0.018 (94.8)	0.010 (94.2)	0.008 (96.6)	–0.003 (94.8)	–0.003 (94.0)	0.014 (93.4)	0.009 (93.2)
	SDMEXP	–0.008 (90.6)	–0.002 (87.2)	–0.019 (93.2)	0.014 (90.8)	0.018 (91.6)	–0.005 (91.0)	–0.002 (90.0)	0.019 (91.0)	0.004 (88.4)
σ_e^2	MPML U	–0.080 (55.6)	–0.052 (70.8)	–0.039 (77.6)	–0.081 (59.0)	–0.049 (74.0)	–0.032 (83.2)	–0.079 (55.2)	–0.049 (73.8)	–0.033 (77.6)
	MPML A	–0.010 (89.6)	–0.007 (92.6)	–0.005 (86.8)	–0.011 (88.6)	–0.004 (88.8)	0.000 (91.8)	–0.008 (90.8)	–0.006 (93.2)	0.000 (93.2)
	MPML B	–0.053 (68.2)	–0.034 (80.6)	–0.023 (84.0)	–0.053 (69.4)	–0.031 (80.6)	–0.017 (86.6)	–0.050 (72.4)	–0.032 (85.4)	–0.018 (87.2)
	Naive	0.002 (95.0)	0.002 (98.2)	0.001 (95.0)	0.002 (94.0)	0.003 (95.4)	0.000 (96.0)	0.000 (93.2)	–0.001 (93.0)	0.002 (95.8)
	PWIGLS	–0.059 (63.8)	–0.03 (87.6)	–0.018 (90.4)	–0.059 (63.6)	–0.028 (82.8)	–0.019 (91.6)	–0.058 (64.4)	–0.034 (80.8)	–0.017 (93.2)
	SDMLGT	0.001 (95.0)	0.001 (98.2)	0.000 (94.8)	0.001 (94.0)	0.002 (95.6)	0.000 (97.0)	0.000 (94.0)	–0.001 (93.0)	0.002 (95.8)
	SDMEXP	0.001 (95.0)	0.001 (98.2)	0.000 (94.8)	0.001 (94.0)	0.002 (95.6)	0.000 (97.0)	0.000 (94.0)	–0.001 (93.0)	0.002 (95.8)
σ_u^2	MPML U	0.041 (90.4)	0.003 (86.0)	0.009 (86.6)	0.074 (87.8)	0.027 (87.0)	–0.008 (86.0)	0.080 (93.8)	0.054 (93.4)	0.041 (91.0)
	MPML A	–0.065 (84.6)	–0.072 (81.96)	–0.055 (80.4)	–0.031 (84.2)	–0.053 (84.6)	–0.071 (82.2)	–0.030 (89.6)	–0.023 (85.4)	–0.022 (87.4)
	MPML B	0.017 (89.0)	–0.023 (84.0)	–0.018 (84.2)	0.050 (87.4)	0.000 (87.0)	–0.035 (84.2)	0.055 (94.4)	0.028 (89.8)	0.014 (90.0)
	Naive	0.018 (95.8)	–0.007 (96.8)	0.018 (95.6)	0.019 (94.4)	0.007 (96.2)	0.001 (94.8)	0.015 (95.8)	0.000 (94.4)	0.000 (93.6)
	PWIGLS	–0.115 (74.0)	–0.153 (71.0)	–0.171 (59.8)	–0.094 (77.6)	–0.125 (72.4)	–0.152 (68.4)	–0.033 (86.6)	–0.062 (81.6)	–0.085 (76.8)
	SDMLGT	0.008 (95.0)	–0.017 (96.4)	0.008 (95.6)	0.008 (94.4)	–0.004 (95.8)	–0.010 (93.6)	0.004 (95.2)	–0.011 (94.2)	–0.011 (93.2)
	SDMEXP	0.008 (96.2)	–0.017 (96.4)	0.008 (95.6)	0.009 (94.0)	–0.004 (95.8)	–0.010 (92.2)	0.005 (95.6)	–0.010 (94.2)	–0.011 (93.4)

Note: MPML = multilevel pseudo–maximum likelihood; PWIGLS = probability-weighted iterative generalized least squares; SDMEXP = Sample Distribution Method using Exponential model; SDMLGT = Sample Distribution Method using Logistic model.

reported in Table 2. Only the results for β_0, σ_u^2, and σ_e^2 using scaling method B are reported in Figure 3.

In general, imposing trimming reduces the relative variation but does not change the informativeness for all designs. It also reduces the MSE for all parameters. For β_0, trimming the sampling weights does not reduce any bias under an informative sampling design: As the level of trimming increases, the bias gets larger and the coverage rate drops to zero. When the sampling design is noninformative at both stages, trimming does slightly increase the coverage rate.

For σ_e^2, trimming weights lead to lower biases and MSE. When a sampling design is informative at the second stage or both stages, the coverage rate drops to zero as the level of trimming increases; when a sampling design is noninformative at the second level or both (designs 2 and 1), higher level of trimming is associated with higher level of coverage rate.

When the first stage is informative (designs 1 and 3), a higher level of trimming leads to higher biases and lower coverage rates for σ_u^2. When the first stage is noninformative (designs 2 and 4), trimming improves the estimates from MPML but does not do the same for PWIGLS.

The results clearly indicate that the effect of trimming depends on the informativeness at certain stage. It works well for the noninformative situation but does not reduce the bias for the parameter of a variable that is involved in the selection process.

5. DISCUSSION AND AN ILLUSTRATED EXAMPLE

5.1. Discussion

In many cases, data available for social scientists contain sampling weights that are assigned to each observation at one or multiple levels to adjust for the unequal probability of selection. Hybrid approaches have been widely recommended when one analyzes survey data. However, in some cases, the hierarchical structure of the multilevel model that one wants to estimate does not fit the structure of the sampling design. A decision must be made to either change the model or to ignore the sampling weights. One may want to investigate whether there are other options besides a hybrid approach and the consequences if the informative sampling design is ignored. To aid future investigations, in this study, I conduct simple simulations to evaluate the hybrid and sample distribution methods in a linear random-intercept model.

Table 2. Relative Variation (RV) and Proportion of Rejection (PR), Test for Informativeness at the $\alpha = .05$ Level

	Design 1			Design 2			Design 3			Design 4		
Trimming	RV 1	RV 2	PR	RV 1	RV 2	PR	RV 1	RV 2	PR	RV 1	RV 2	PR
[1%, 99%]	3.833	2.124	0.986	4.468	2.514	0.876	4.067	1.793	0.838	4.069	1.795	0.494
[2.5%, 97.5%]	1.377	1.580	0.988	1.371	1.833	0.944	1.349	1.316	0.826	1.483	1.318	0.398
[5%, 95%]	0.711	1.164	0.998	0.681	1.345	0.990	0.677	0.961	0.808	0.701	0.964	0.370
[10%, 90%]	0.308	0.774	1.000	0.290	0.896	0.998	0.304	0.630	0.762	0.298	0.633	0.346
[15%, 85%]	0.172	0.562	0.996	0.159	0.650	1.000	0.165	0.451	0.708	0.167	0.451	0.384
[20%, 80%]	0.101	0.413	0.990	0.096	0.482	0.992	0.100	0.327	0.698	0.098	0.328	0.446
[25%, 75%]	0.059	0.298	0.978	0.058	0.350	0.992	0.061	0.233	0.674	0.060	0.234	0.500
[30%, 70%]	0.035	0.204	0.966	0.034	0.241	0.964	0.036	0.157	0.670	0.036	0.157	0.418

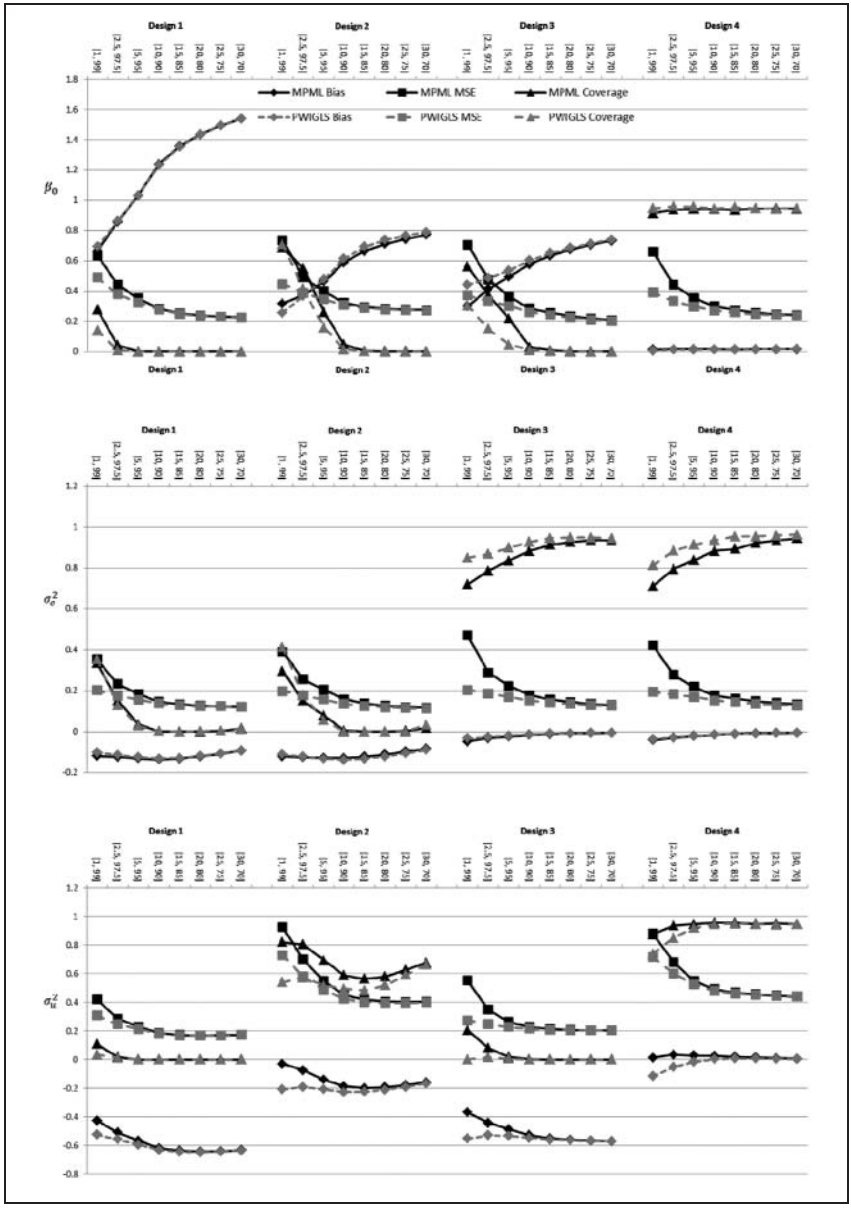

Figure 3. The relative bias, mean square error (MSE), and coverage rate at different levels of weight trimming at high variation of the sampling weights.

Note: The solid line and dashed line represent the result from the multilevel pseudo–maximum likelihood (MPML) and probability-weighted iterative generalized least squares (PWIGLS) approaches, respectively. The relative bias is indicated by the diamond marker, the squared marker represents the MSE, and the triangle marker represents the coverage rate.

I found substantial differences among these methods. The naive method does not work well in estimating β_0 and σ_u^2, though the estimates for other fixed effects and σ_e^2 are nearly unbiased or slightly biased within 10 percent of the true value. Although in many cases, biased estimates for the β_0 and σ_u^2 may not cause serious concern, ignoring the uncertainties of parameters that are due to the randomization leads to incorrect inference for all parameters, because the confidence interval is too narrow to cover the true parameter.

Generally, for point estimates, the hybrid methods have the same problems as the naive method, although the randomization variance gives a better coverage rate than the naive method. For example, the estimated β_0 and σ_u^2 are severely biased when a sampling design is informative. Furthermore, including the sampling weights substantially increases the MSE.

The sampling distribution method reduces the biases on β_0 dramatically, but it is sensitive to the specification of the sampling model. Without the correct form of sampling selection, the bias reduction is very limited, only on β_0 for instance. If the sampling model is correctly specified, the sample distribution method outperforms any other methods considered in this study.

The test of informativeness might fail to detect the informativeness of a sampling design, because it depends on the naive and hybrid estimates of model parameters and their variances, both which can be biased. For example, in design 2, only 78.8 percent of the tests are significant at the .05 level, and in design 4, 49.2 percent of the tests are significant at the .05 level. In addition, if the selection process at the first level is negatively associated with u_i and positively correlated with y_{ij}, the test might even be more unreliable.

To sum up, in this study, I compare the hybrid and sample distribution methods on estimation of linear random-intercept model under a two-stage sampling design. Although there are many factors that could affect the quality of the estimation, to reduce the complexity, I focus on the effect of informativeness, and the level of variation of sampling weights. I find the following:

1. Whether a sampling design is informative and at which stage of the sampling design is informative have substantial impacts on the estimation. When a sampling design is informative, the level of variation of sampling weights is also correlated with the bias of estimates.

2. A higher level of variation of sampling weights is associated with a higher level of bias when a sampling design is informative; this may not be true under a noninformative design.
3. Ignoring an informative sampling design at the first stage will result in biased estimates on β_0 and σ_u^2, whereas ignoring an informative sampling design at the second stage will lead to slightly underestimated fixed effects and σ_e^2, besides the biased estimates on β_0 and σ_u^2.
4. Including the sampling weights as in the hybrid methods may still produce biased estimates on β_0 and σ_u^2 and slightly underestimated fixed effects and σ_e^2.
5. The sample distribution method may give unbiased estimates, but it depends on the correct specification of the sampling process.
6. The effect of sampling weight trimming depends on whether a sampling stage is informative. Imposing weight trimming does not reduce the bias for the parameter of a variable that involves in sampling selection.

The design of this simulation study captures the general features of data sets available in social science that have large numbers of clusters, large cluster sizes, and moderate intraclass correlation coefficients. Further analyses may be required to generalize the conclusions drawn here to other settings. Some of the results obtained from our study are different from those obtained by Asparouhov and Muthén (2007); for example, they reported that the MPML estimator outperforms substantially the PWIGLS estimator. The difference might be due to the different settings of simulation, in particular the relative variation of sampling weights.[2] Therefore, it is highly possible that our results might not be replicated in a different setting. Some of the simulation settings in our study might be very rare in reality, such as having the relative variation of weights at 6. Such an extreme case only serves to evaluate the performance of different estimators.

5.2. An Illustrated Example

To illustrate the multiple ways of handing sampling weights, I fit a linear random-intercept model using the wave I data from Add Health, a longitudinal study of a nationally representative sample of adolescents in grades 7 to 12 in the United States during the 1994–1995 school year (Harris et al. 2009). Add Health provides a rich set of information on respondents' social, economic, psychological, and physical well-being,

with contextual data on families, neighborhoods, communities, schools, friendships, peer groups, and romantic relationships. Add Health was designed as a stratified two-stage cluster sample with unequal probabilities of inclusion for individuals. The dependent variable is a delinquency index, which summarizes responses from 12 survey questions that ask how often the respondent engaged in activities such as stealing amounts larger or smaller than $50, breaking and entering, selling drugs, serious physical fighting, the use of weapons to get something from someone, involvement in physical fighting between groups, shooting or stabbing someone, deliberately damaging property, and pulling a knife or gun on someone in the past 12 months. The responses to those survey items are coded from 1 (''never'') to 4 (''more than 3 times''). The predictors are age, gender, and Picture Vocabulary Test score.

The results are presented in Table 3. The data include 17,681 completed cases from 130 schools with relative variances of the sampling weights of 3.00 and 0.89 at the school and student levels, respectively. The calculated χ^2 statistic for the informativeness test is 31.68, with 6 degrees of freedom, which yields a p value $< .0001$. Pfeffermann and Sverchkov (1999) suggested that the t test for the coefficient of sampling weights against residuals in a linear regression can also be used to detect whether the sampling weights is informative. For the school level, the t-test result is significant at the .05 level, which suggests that the sampling design is informative, whereas the same test result is not significant at individual level. I also tested the effect of sampling weights within each school and found that only 6 percent of coefficients were significant at .05 level. Therefore, according to the model I estimate, the sampling design might fit scenario C1 in Figure 2. All the estimated parameters are in same direction and have similar sizes, no matter which estimator is used, except for the effect of Picture Vocabulary Test score estimated by the unscaled MPML. The effect size of estimates from MPML is smaller than that from other methods. The MPML estimates also have higher standard error than other estimates. The sample distribution estimates are very close to the naive estimates. Choosing the best method is not an easy task. Some traditional index may not work well. For example, the Akaike information criterion and the Bayesian information criterion are almost identical among the naive and sample distribution methods and not comparable with those from MPML.

Table 3. Coefficients (Standard Errors) of Random-intercept Model of Delinquency among Adolescents (Add Health Wave I)

	Estimator							
		MPML[a]						
Parameter	Naive	U	A	B	PWIGLS	SDMEXP	SDMEXP[b]	SDMLGT
Intercept	3.743 (0.348)***	2.258 (0.749)**	2.625 (0.442)***	2.777 (0.403)***	3.383 (0.399)***	3.559 (0.317)***	3.532 (0.372)***	3.559 (0.32)***
Age	–0.048 (0.019)*	–0.015 (0.043)	–0.001 (0.023)	–0.014 (0.022)	–0.027 (0.020)	–0.048 (0.016)**	–0.052 (0.019)**	–0.048 (0.017)**
Female	–1.321 (0.069)***	–1.142 (0.080)***	–1.171 (0.064)***	–1.183 (0.061)***	–1.250 (0.064)***	–1.321 (0.047)***	–1.326 (0.101)***	–1.321 (0.047)***
PVT score	–0.062 (0.020)**	0.011 (0.089)	–0.042 (0.051)	–0.037 (0.044)	–0.069 (0.027)*	–0.062 (0.017)***	–0.059 (0.044)	–0.062 (0.017)***
σ_e^2	9.780 (0.104)***	8.785 (1.452)***	8.655 (0.918)***	9.040 (0.812)***	9.410 (0.887)***	9.779 (0.104)***	9.690 (0.420)***	9.779 (0.104)***
σ_u^2	0.173 (0.034)***	0.427 (0.111)***	0.270 (0.082)**	0.292 (0.081)***	0.162 (0.034)***	0.170 (0.033)***	0.256 (0.039)***	0.170 (0.033)***
a_0						–4.844***	—	–4.834***
a_1						0.576***	0.576	0.581***
b_0						–6.406***	—	–6.401***
b_1						0.009**	0.009	0.009**
–2 LL	90,636.46				101,743.19	90,636.46		90,636.47
AIC	90,648.46		11,127,530.64	15,384,886.25	101,755.19	90,648.46		90,648.47
BIC	90,649.15		11,127,606.32	15,384,963.82	101,755.87	90,649.15		90,649.15

Note: AIC = Akaike information criterion; BIC = Bayesian information criterion; LL = log likelihood; MPML = multilevel pseudo–maximum likelihood; PWIGLS = probability-weighted iterative generalized least squares; PVT = Picture Vocabulary Test; SDMEXP = Sample Distribution Method using Exponential model; SDMLGT = Sample Distribution Method using Logistic model.

[a]The AIC and BIC were calculated on the basis of pseudo-likelihood.

[b]The estimates are based on 500 bootstrapping replicates.

*$p < .05$. **$p < .01$. ***$p < .001$.

Nevertheless, some conclusions are clear. Although the estimates are different, their confidence intervals actually overlap. Those estimates are not substantially different from each other. It does not matter which method one chooses in this example.

I also further investigate some possible reasons why the hybrid estimates on intercept are lower than those from other methods. I find that the missing pattern on delinquency is correlated with the individual-level sampling weights. Because the sampling weights were adjusted according to survey nonresponse (Tourangeau and Shin 1999), but not for the missing on delinquency, the sampling weights at the individual level may not carry enough information to recover the possible nonrandom missing on delinquency. Cautiously using the same set of sampling weights for all analyses should be advised.

APPENDIX A1

All 1st Order Exponential Approximation From Eideh And Nathan 2009

Under the first-order exponential approximation, the cluster level probabilities of inclusion can be written as

$$E_p(\pi_i|u_i, \boldsymbol{z}_i) \approx g(\boldsymbol{z}_i) exp(a_1 u_i).$$

Thus, the sample distribution of random effect u_i can be derived as

$$f_s(u_i|\boldsymbol{z}_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left(- \frac{\left(u_i - \left(a_1\sigma_u^2 + z_i'\gamma\right)\right)^2}{2\sigma_u^2} \right).$$

If the individual level probabilities of inclusion can also be approximated by the first-order exponential, according to Eideh and Nathan (2009), the conditional sample distribution of individual elements can be written as

$$E_p\left(\pi_{j|i}|y_{ij}, \boldsymbol{x}_{ij}, u_i\right) \approx k\left(\boldsymbol{x}_{ij}, u_i\right) exp\left(b_1 y_{ij}\right),$$

$$f_s\left(y_{ij}|x_{ij}, u_i\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}} exp\left(- \frac{\left(y_{ij} - \left(u_i + x_{ij}'\beta + b_1\sigma_e^2\right)\right)^2}{2\sigma_e^2} \right),$$

and then the marginal sample distribution for the *i*th cluster and the full sample distribution function are given as follows:

$$f_s(\boldsymbol{y}_i) = 2\pi^{-\frac{n_i}{2}}\left(\sigma_e^2\right)^{-\frac{n_i-1}{2}}\left(m_i\sigma_u^2+\sigma_e^2\right)^{-\frac{1}{2}}\times$$
$$exp\left(-\frac{1}{2\sigma_e^2}\sum\nolimits_{j=1}^{n_i}\left(y_{ij}-\left(a_1\sigma_u^2+z_i'\gamma+x_{ij}'\boldsymbol{\beta}+b_1\sigma_e^2\right)^2\right)\right)$$
$$\times exp\left(\frac{\sigma_u^2}{2\sigma_e^2\left(n_i\sigma_u^2+\sigma_e^2\right)}\sum\nolimits_{j=1}^{n_i}\left(y_{ij}-\left(a_1\sigma_u^2+z_i'\gamma+x_{ij}'\beta+b_1\sigma_e^2\right)\right)^2\right),$$

$$f_s(\boldsymbol{y}) = \prod_{i=1}^{m} f_s(\boldsymbol{y}_i),$$

where γ and $\boldsymbol{\beta}$ are unknown fixed effects, and σ_u^2, and σ_e^2 are variance components. Because the sample likelihood function contains some unknown informativeness parameters a_1 and b_1, they must be estimated first. A two-stage estimation procedure was proposed by Eideh and Nathan (2009). The first step is to estimate α_1 and b_1, by regressing $-log(w_i)$ and $-log\left(w_{j|i}\right)$ against u_i and y_{ij}, respectively. One problem is that the random effect u_i is not observed, though it can be measured by the cluster mean, for example, $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ (Eideh and Nathan 2009). However, substituting u_i with $\bar{y}_i$ is not an ideal solution, because for u_i, $\bar{y}_i$ is a measure with error. A better solution is to use $\hat{u}_i$, which is the predicted u_i from the unweighted method, because $\hat{u}_i$ is the BLUP (e.g., Robinson 1991).

The second step is to maximize the full sample likelihood function, with $\hat{a}_1$ and $\hat{b}_1$. Because the two-stage procedure treats $\hat{a}_1$ and $\hat{b}_1$ as fixed, the model-based variance estimator underestimates the true variability. Although the variance can be estimated by bootstrapping (Eideh and Nathan 2009), for computational conveniences, we ignore the variation due to the randomization distribution and use the model variance, which is directly obtained by PROC NLMIXED. Another reason we do so is that the first-order exponential model is an oversimplified approximation and does not reduce much bias on β_0 and σ_u^2. It produces identical estimates as the naive method for other parameters. Although variance can be better estimated, and the variability of a_1 and b_1 can be implemented using methods other than the two-stage estimation, such as EM, we still adopt the model-based variance and acknowledge that by doing so, the coverage rates can be very poor.

APPENDIX A2

When the cluster-level sample indicator I_i follows a logistic model conditional on the cluster-specified random effect u_i, the conditional expectation and its marginal can be written as

$$E_p(\pi_i|u_i) = \frac{1}{1+exp(-\alpha_0 - \alpha_1 u_i)},$$

$$E_p(\pi_i) = \int \frac{1}{1+exp(-\alpha_0 - \alpha_1 u_i)} du_i = \int \frac{1}{1+exp(-\alpha_0 - \alpha_1 u_i)} \times$$

$$\frac{1}{\sqrt{2\pi\sigma_u^2}} exp\left(-\frac{u_i^2}{2\sigma_u^2}\right) du_i.$$

Define $h(u_i)$ as

$$h(u_i) = -log(1+exp(-\alpha_0 - \alpha_1 u_i)) - \frac{u_i^2}{2\sigma_u^2}.$$

The first and second derivatives for $h(u_i)$ are

$$\frac{dh(u_i)}{du_i} = \frac{\alpha_1}{1+exp(\alpha_0+\alpha_1 u_i)} - \frac{u_i}{\sigma_u^2},$$

$$\frac{dh^2(u_i)}{du_i^2} = -\frac{\alpha_1^2 exp(\alpha_0+\alpha_1 u_i)}{[1+exp(\alpha_0+\alpha_1 u_i)]^2} - \frac{1}{\sigma_u^2}.$$

Because $\frac{dh^2(u_i)}{du_i^2} < 0$, the function $h(u_i)$ is concave. Set $\frac{dh(u_i)}{du_i} = 0$, and use the Newton algorithm to find the solution:

$$u_{(t+1)} = u_{(t)} - \frac{\frac{dh(u_i)}{du_i}}{\frac{dh^2(u_i)}{du_i^2}}\Big|_{u_i = u_{(t)}}.$$

Practice shows that the algorithm quickly converges at iteration 1 to the maximum point from $u_{(0)} = 0$:

$$u_{(1)} = \frac{\alpha_1 \sigma_u [1+exp(\alpha_0)]}{\alpha_1^2 \sigma_u^2 exp(\alpha_0) + [1+exp(\alpha_0)]^2}.$$

Thus, the Laplace approximation of $E_p(\pi_i)$ can be written as

$$E_p(\pi_i) \approx exp\left(h\left(u_{(1)}\right)\right)\left(\frac{dh^2(u_i)}{du_i^2}\right)^{-\frac{1}{2}}\Big|_{u_i = u_{(1)}}$$

$$= \frac{exp\left(\alpha_0 + \alpha_1 \sigma_u u_{(1)} - \frac{u_{(1)}^2}{2}\right)}{\sqrt{\alpha_1^2 \sigma_u^2 exp\left(\alpha_0 + \alpha_1 \sigma_u u_{(1)}\right) + \left[1+exp\left(\alpha_0 + \alpha_1 \sigma_u u_{(1)}\right)\right]^2}}.$$

Therefore, the approximated sample distribution of u_i is

$$f_s(u_i) \approx \frac{exp\left(\alpha_1\left(u_i - \sigma_u u_{(1)}\right) + \frac{u_{(1)}^2}{2} - \frac{u_i^2}{2\sigma_u^2}\right)}{\sqrt{2\pi\sigma_u^2}\left[1 + exp(\alpha_0 + \alpha_1 u_i)\right]}$$
$$\times \sqrt{\alpha_1^2\sigma_u^2 exp\left(\alpha_0 + \alpha_1\sigma_u u_{(1)}\right) + \left[1 + exp\left(\alpha_0 + \alpha_1\sigma_u u_{(1)}\right)\right]^2}.$$

Smilarly, if the individual-level sample indicator $I_{j|i}$ follows a logistic model conditional on y_{ij}, the conditional expectation, and its marginal, can be written as

$$E_p\left(\pi_{j|i}|y_{ij}\right) = \frac{1}{1 + exp\left(-b_0 - b_1 y_{ij}\right)},$$

$$E_p\left(\pi_{j|i}\right) = \int \frac{1}{1 + exp\left(-b_0 - b_1 y_{ij}\right)} dy_{ij} = \int \frac{1}{1 + exp\left(-b_0 - b_1 y_{ij}\right)}$$
$$\times \frac{1}{\sqrt{2\pi\sigma_e^2}} exp\left(-\frac{\left(y_{ij} - x_{ij}\beta - u_i\right)^2}{2\sigma_e^2}\right) dy_{ij}.$$

Define s_{ij} as

$$s_{ij} = \frac{y_{ij} - x_{ij}\beta - u_i}{\sigma_e}.$$

Thus,

$$E_p\left(\pi_{j|i}\right) = \int \frac{1}{1 + exp\left(-b_0 - b_1\left(x_{ij}\beta + u_i + \sigma_e s_{ij}\right)\right)} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{{s_{ij}}^2}{2}\right) ds_{ij}.$$

Define $g\left(s_{ij}\right)$ as

$$g\left(s_{ij}\right) = -log\left(1 + exp\left(-b_0 - b_1\left(x_{ij}\beta + u_i + \sigma_e s_{ij}\right)\right)\right) - \frac{s_{ij}^2}{2}.$$

The first and second derivatives for $g\left(s_{ij}\right)$ are

$$\frac{dg\left(s_{ij}\right)}{ds_{ij}} = \frac{b_1\sigma_e}{1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{ij}\right)\right)} - s_{ij},$$

$$\frac{dg^2\left(s_{ij}\right)}{ds_{ij}^2} = -\frac{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{ij}\right)\right)}{\left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{ij}\right)\right)\right]^2} - 1.$$

Because $\frac{dg^2(s_{ij})}{ds_{ij}^2} < 0$, the function $g(s_{ij})$ is concave. Again, we set $\frac{dg(s_{ij})}{ds_{ij}} = 0$ and use the Newton algorithm to find the solution:

$$s_{(t+1)} = s_{(t)} - \frac{\frac{dg(s_{ij})}{ds_{ij}}}{\frac{dg^2(s_{ij})}{ds_{ij}^2}}\Big|_{s_{ij}=s_{(t)}}.$$

The algorithm quickly converges at iteration 1 to the maximum point from $s_{(0)} = 0$:

$$s_{(1)} = \frac{b_1\sigma_e\left(1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right)}{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2}.$$

Thus, the marginal conditional expectation can be approximated by

$$E_p\left(\pi_{j|i}\right) \approx \frac{exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right) - \frac{s_{(1)}^2}{2}\right)}{\sqrt{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2}}.$$

Therefore, the approximated sample distribution of y_{ij} can be written as

$$f_s\left(y_{ij}\right) \approx \frac{exp\left(b_1\left(y_{ij} - x_{ij}\beta - u_i - \sigma_e s_{(1)}\right) - \frac{\left(y_{ij} - x_{ij}\beta - u_i\right)^2}{2\sigma_e^2} + \frac{s_{(1)}^2}{2}\right)}{\left[1 + exp\left(b_0 + b_1 y_{ij}\right)\right] \times \sqrt{2\pi\sigma_e^2}}$$

$$\times \sqrt{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2}.$$

The approximated marginal sample distribution of $\boldsymbol{y}_i$ is

$$f_s(\boldsymbol{y}_i) = \int \prod_{j=1}^{n_i} f_s\left(y_{ij}|\boldsymbol{x}_{ij}, u_i\right) du_i$$

$$\approx c_2 \times \int \frac{exp\left(-\frac{1}{2\sigma_e^2}\sum_{j=1}^{n_i}\left(y_{ij} - x_{ij}\beta - u_i - b_1\sigma_e^2\right)^2 - \frac{\left(u_i - \alpha_1\sigma_u^2\right)^2}{2\sigma_u^2}\right)}{1 + exp(\alpha_0 + \alpha_1 u_i)}$$

$$\times \prod_{j=1}^{n_i} \sqrt{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2} du_i,$$

where

$$c_2 = \frac{c_1 \times exp\left(\frac{1}{2}\sum_{j=1}^{n_i}\left[s_{(1)} - b_1\sigma_e\right]^2\right)}{\prod_{j=1}^{n_i} exp\left(b_0 + b_1 y_{ij}\right)} \times \left(2\pi\sigma_e^2\right)^{-\frac{n_i}{2}},$$

$$c_1 = \frac{exp\left(\frac{1}{2}(u_{(1)} - \alpha_1\sigma_u)^2\right)}{\sqrt{2\pi\sigma_u^2}} \times \sqrt{\alpha_1^2\sigma_u^2 exp\left(\alpha_0 + \alpha_1 u_{(1)}\right) + \left[1 + exp\left(\alpha_0 + \alpha_1 u_{(1)}\right)\right]^2}.$$

The approximated marginal sample distribution of $\boldsymbol{y}_i$ still needs to be approximated according to the random effect u_i.

Define Δ as

$$\Delta = b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right).$$

Then $\frac{d\Delta}{du_i} = b_1$.

Define $k(u_i)$ as

$$k(u_i) = -\frac{\sum_{j=1}^{n_i}\left(y_{ij} - x_{ij}\beta - u_i - b_1\sigma_e^2\right)^2}{2\sigma_e^2} - \frac{\left(u_i - \alpha_1\sigma_u^2\right)^2}{2\sigma_u^2} - log(1 + exp(\alpha_0 + \alpha_1 u_i)) + \frac{\sum_{j=1}^{n_i} log\left(b_1^2\sigma_e^2 exp(\Delta) + [1 + exp(\Delta)]^2\right)}{2}.$$

Thus, the first and second derivatives of $k(u_i)$ are

$$\frac{dk(u_i)}{du_i} = \frac{\sum_{j=1}^{n_i}\left(y_{ij} - x_{ij}\beta - u_i - b_1\sigma_e^2\right)}{\sigma_e^2} - \frac{\left(u_i - \alpha_1\sigma_u^2\right)}{\sigma_u^2} - \frac{\alpha_1}{1 + exp(-\alpha_0 - \alpha_1 u_i)} + \frac{1}{2}\sum_{j=1}^{n_i}\frac{2b_1 exp(2\Delta) + \left(b_1^3\sigma_e^2 + 2b_1\right)exp(\Delta)}{b_1^2\sigma_e^2 exp(\Delta) + [1 + exp(\Delta)]^2},$$

$$\frac{dk^2(u_i)}{du_i^2} = -\frac{n_i}{\sigma_e^2} - \frac{1}{\sigma_u^2} - \frac{\alpha_1^2 exp(\alpha_0 + \alpha_1 u_i)}{[1 + exp(\alpha_0 + \alpha_1 u_i)]^2} + \sum_{j=1}^{n_i}\frac{b_1^2\left(b_1^2\sigma_e^2 + 2\right)[exp(3\Delta) + exp(\Delta)] + 4b_1^2 exp(2\Delta)}{2\left[b_1^2\sigma_e^2 exp(\Delta) + [1 + exp(\Delta)]^2\right]^2}.$$

$$\frac{b_1^2\left(b_1^2\sigma_e^2 + 2\right)[exp(3\Delta) + exp(\Delta)] + 4b_1^2 exp(2\Delta)}{2\left[b_1^2\sigma_e^2 exp(\Delta) + [1 + exp(\Delta)]^2\right]^2} = \frac{b_1^2}{exp(\Delta) + exp(-\Delta) + b_1^2\sigma_e^2 + 2} \times \left(\frac{b_1^2\sigma_e^2 + 2}{2} - \frac{\left[b_1^2\sigma_e^2 + 2\right]^2 - 4}{2\left(exp(\Delta) + exp(-\Delta) + b_1^2\sigma_e^2 + 2\right)}\right).$$

Because function $exp(\Delta) + exp(-\Delta) \geq 2$, and it takes value 2 at $\Delta = 0$.

Therefore,

$$\frac{b_1^2\left(b_1^2\sigma_e^2+2\right)[exp(3\Delta)+exp(\Delta)]+4b_1^2exp(2\Delta)}{2\left[b_1^2\sigma_e^2exp(\Delta)+[1+exp(\Delta)]^2\right]^2} \leq \frac{b_1^2}{b_1^2\sigma_e^2+4}\times$$
$$\left(\frac{b_1^2\sigma_e^2+2}{2}-\frac{1}{2}\times\frac{\left[b_1^2\sigma_e^2+2\right]^2-4}{b_1^2\sigma_e^2+4}\right)=\frac{b_1^2}{b_1^2\sigma_e^2+4}<\frac{1}{\sigma_e^2}.$$

Thus,

$$\sum_{j=1}^{n_i}\frac{b_1^2\left(b_1^2\sigma_e^2+2\right)[exp(3\Delta)+exp(\Delta)]+4b_1^2exp(2\Delta)}{2\left[b_1^2\sigma_e^2exp(\Delta)+[1+exp(\Delta)]^2\right]^2}<\frac{n_i}{\sigma_e^2},$$

and $\frac{dk^2(u_i)}{du_i^2}<0$ and $k(u_i)$ is concave. Set $\frac{dk(u_i)}{du_i}=0$, and use the Newton algorithm to find the solution.

Update $u_{(t+1)}=u_{(t)}-\frac{\frac{dk(u_i)}{du_i}}{\frac{dk^2(u_i)}{du_i^2}}|_{u_i=u_{(t)}}$, till convergence.

The Laplace approximation for $\int exp(k(u_i))du_i$ is

$$\int exp(k(u_i))du_i \approx exp\left(k\left(u_{(2)}\right)\right)\times\sqrt{2\pi\hat{\sigma}^2},$$

where $u_{(2)}$ is the limit point of the Newton iterations, and

$$\hat{\sigma}^2=\left(-\frac{dk^2(u_i)}{du_i^2}\right)^{-1}|_{u_i=u_{(2)}}.$$

Put all together, the log likelihood of the approximated $f_s(\boldsymbol{y}_i)$ is

$$\log(f_s(\boldsymbol{y}_i))\approx log(c_2)+k\left(u_{(2)}\right)+\frac{1}{2}log\left(2\pi\hat{\sigma}^2\right),$$

where

$$\log(c_2)=log(c_1)+\frac{\sum_{j=1}^{n_i}\left[s_{(2)}-b_1\sigma_e\right]^2}{2}-\sum_{j=1}^{n_i}log\left(1+exp\left(b_0+b_1y_{ij}\right)\right)-\frac{n_i}{2}log\left(2\pi\sigma_e^2\right),$$

$$\log(c_1)=\frac{\left[u_{(1)}-\alpha_1\sigma_u\right]^2}{2}-\frac{log\left(2\pi\sigma_u^2\right)}{2}+$$
$$\frac{log\left(\alpha_1^2\sigma_u^2exp\left(\alpha_0+\alpha_1\sigma_u u_{(1)}\right)+\left[1+exp\left(\alpha_0+\alpha_1\sigma_u u_{(1)}\right)\right]^2\right)}{2},$$

$$s_{(2)} = s_{(1)}\big|_{u_i = u_{(2)}}.$$

The sample information parameters α_0, α_1, b_0 and b_1 can be estimated from the sample data.

If the cluster-level sample indicator I_i does not depend on the cluster-specified random effect u_i, then $\alpha_0 = \alpha_1 = 0$, and $f_s(\boldsymbol{y}_i)$ becomes

$$f_s(\boldsymbol{y}_i) \approx c_0 \times \int exp\left(-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} \left(y_{ij} - x_{ij}\beta - u_i - b_1\sigma_e^2 \right)^2 - \frac{(u_i)^2}{2\sigma_u^2} \right)$$

$$\times \prod_{j=1}^{n_i} \sqrt{b_1^2\sigma_e^2 exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right) + \left[1 + exp\left(b_0 + b_1\left(x_{ij}\beta + u_i + \sigma_e s_{(1)}\right)\right)\right]^2} du_i$$

where

$$c_0 = \sqrt{2\pi\sigma_u^2} \times \frac{exp\left(\frac{1}{2} \sum_{j=1}^{n_i} \left[s_{(1)} - b_1\sigma_e \right]^2 \right)}{\prod_{j=1}^{n_i} exp\left(b_0 + b_1 y_{ij}\right)} \times \left(2\pi\sigma_e^2\right)^{-\frac{n_i}{2}}.$$

If the individual-level sample indicator $I_{j|i}$ does depend on the response variable y_{ij}, then $b_0 = b_1 = 0$, and $f_s(\boldsymbol{y}_i)$ becomes

$$f_s(\boldsymbol{y}_i) \approx \left(2\pi\sigma_e^2\right)^{-\frac{n_i}{2}} exp\left(\frac{(u_{(1)} - \alpha_1\sigma_u)^2}{2} \right) \times \int \frac{exp\left(-\frac{1}{2\sigma_e^2} \sum_{j=1}^{n_i} \left(y_{ij} - x_{ij}\beta - u_i \right)^2 - \frac{\left(u_i - \alpha_1\sigma_u^2\right)^2}{2\sigma_u^2} \right)}{1 + exp(\alpha_0 + \alpha_1 u_i)}$$

$$\times \sqrt{\alpha_1^2\sigma_u^2 exp(\alpha_0 + \alpha_1 u_i) + \left[1 + exp(\alpha_0 + \alpha_1 u_i)\right]^2} du_i.$$

The value of α_0, α_1, b_0 and b_1 can be estimated by given

$$w_i = 1 + exp(-\alpha_0 - \alpha_1 u_i),$$

$$w_{j|i} = 1 + exp\left(-b_0 - b_1 y_{ij}\right).$$

where u_i can be replaced by $\hat{u}_i$.

APPENDIX B

SAS code for the naive method:

```
proc mixed data=data covtest method=REML empirical;
  model y=x1 x2 x3/s;
  random intercept/sub=clu;
run;
```

SAS code for the sample distribution method: the exponential approximation under the two-stage informative design:

```
proc nlmixed data=_sample QPOINTS=100 NOAD NOADSCALE;
  parms beta0-beta3 1 s2u s2e 2;
  mu=a1*s2u+b1*s2e+beta0+beta1*x1+beta2*x2+beta3*x3+u;
  model y~normal(mu,s2e);
  random u~normal(0,s2u) sub=cluster;
run;
```

SAS IML code for the sample distribution method: the Laplace approximation under the two-stage informative design:

```
proc iml;
start lfs(initial) global(cluster,x,y,a0,a1,b0,b1);
  beta=initial[1:4];
  s2u=initial[5];
  s2e=initial[6];
  n=ncol(UNIQUE(cluster));
  lfs=0;
  /*u(1)*/
  u1=a1*sqrt(s2u)*(1+exp(a0))/(a1**2*s2u*exp(a0)+(1+
  exp(a0))**2);
  /*log C1 */
  lc1=.5*(u1-a1*sqrt(s2u))**2+.5*log(a1**2*s2u*exp(a0+a1
  *sqrt(s2u)*u1)+
      (1+exp(a0+a1*sqrt(s2u)*u1))**2)-.5*log(2*CONSTANT ('PI'))-
      .5*log(s2u);
  do i=1 to n;
    yi=y[loc(cluster=i)];
    xi=x[loc(cluster=i),];
    mui=xi*beta;
    ei=yi-mui;
    mi=nrow(yi);
    c=b1**2*s2e+2;
    t=0;
    u=-1;
    /*u(2) and s(2) for ijth individual*/
    do until (abs(u2-u0)<.0001);
      u0=u;
      s2=(b1*sqrt(s2e)*(1+exp(b0+b1*(mui+u0))))
        /(b1**2*s2e*exp(b0+b1*(mui+u0))+(1+exp(b0+b1*
        (mui+u0)))##2);
      delta=b0+b1*(mui+u0+sqrt(s2e)*s1);
```

```
        d1=-u0/s2u+a1+1/s2e*sum(ei-u0-b1*s2e)-a1/(1+exp
        (-a0-a1*u0))+
            .5*sum((b1*c*exp(-delta)+2*b1)/(c*exp(-delta) +1+exp
            (-2*delta)));
        d2=-1/s2u-mi/s2e-a1**2/(exp(-.5*a0-.5*a1*u0)+exp
        (.5*a0+.5*a1*u0))**2
            +.5*sum((b1**2/(c+exp(-delta)+exp(delta)))#(c-(c**2-4)/
            (c+exp(delta)
            +exp(-delta))));
        u=sum(u0,-d1/d2);
        u2=u;
        t=t+1;
      end;
      /*log C2 for ith cluster*/
      lc2i=lc1+.5*sum((s1-b1*sqrt(s2e))##2)-sum(log(1+exp
      (b0+b1*yi)))
            -.5*mi*(log(s2e)+log(2*CONSTANT('PI')));
      ki=-(u2-a1*s2u)**2/(2*s2u)-sum((ei-u2-b1*s2e)##2/(2*s2e))-
      log(1+exp(a0+a1*u2))
            +.5*sum(log(b1**2*s2e*exp(delta)+(1+exp(delta))##2));
      lfsi=lc2i+ki-.5*sum(log(-d2))+.5*log(2*CONSTANT('PI'));
      lfs=lfs+lfsi;
    end;
    return(-lfs);
  finish lfs;
```

Mplus code:

```
VARIABLE:
Names are clu wt1 x1 x2 x3 y wt2 ind;
Usevariables are clu y x1 x2 x3;
Cluster is clu;
WITHIN =x1 x2 x3;
Weight is wt2;
Bweight = wt1;
Wtscale =CLUSTER;
Bwtscale =SAMPLE;
MODEL:
%WITHIN%
y on x1 x2 x3;
%BETWEEN%
ANALYSIS:
ALGORITHM = INTEGRATION;
```

```
ITERATIONS =2000;
Type =twolevel;
OUTPUT:
sampstat tech1;
```

LISREL code

#Make a PSF file

```
Raw Data from File c:\temp\lis.dat
DA NI=8 NO=0
LA
id x1 x2 x3 y wt_ind wt_clu clu
RA FI=C:\temp\lis.dat
OU RA=lis.PSF
```

#MULTILEV module

```
OPTIONS OLS=YES CONVERGE=0.0001 MAXITER=2000 OUTPUT=STANDARD;
SY=lis.psf;
ID2=clu;
ID1=id;
WEIGHT2=wt_clu;
WEIGHT1=wt_ind;
RESPONSE=y;
FIXED=intcept x1 x2 x3;
RANDOM1=intcept;
RANDOM2=intcept;
```

Notes

1. I also evaluated the effect of weight trimming for the simulations in section 4.2. The results were very close to those reported here and are available on request.
2. I replicated their results and found that in our replication, the average relative variations of sampling weights were 0.184, and 0.452, respectively. The results and code are available on request.

References

Asparouhov, Tihomir. 2004. ‘‘Weighting for Unequal Probability of Selection in Multilevel Modeling.’’ Mplus Web Notes: No. 8. Retrieved August 21, 2012 (http://www.statmodel.com/download/webnotes/MplusNote81.pdf).

Asparouhov, Tihomir. 2005. ‘‘Sampling Weights in Latent Variable Modeling.’’ *Structural Equation Modeling* 12:411–34.

Asparouhov, Tihomir. 2006. ‘‘General Multi-level Modeling with Sampling Weights.’’ *Communications in Statistics: Theory and Methods* 35:439–60.

Asparouhov, Tihomir. 2008. ''Scaling of Sampling Weights for Two Level Models in Mplus 4.2.'' Retrieved August 21, 2012 (http://www.statmodel.com/download/Scaling3.pdf).

Asparouhov, Tihomir and Bengt Muthén. 2006. ''Multilevel Modeling of Complex Survey Data.'' Retrieved August 21, 2012 (http://www.statmodel.com/download/Survey
JSM1.pdf).

Asparouhov, Tihomir and Bengt Muthén. 2007. ''Testing for Informative Weights and Weights Trimming in Multivariate Modeling with Survey Data.'' Retrieved August 21, 2012 (http://www.statmodel.com/download/JSM2007000745.pdf).

Binder, David A. and Georgia R. Roberts. 2003. ''Design-based and Model-based Methods for Estimating Model Parameters.'' Pp. 29–48 in *Analysis of Survey Data*, edited by R. L. Chambers and Chris Skinner. Chichester, UK: Wiley.

Box, George E. P. 1979. ''Some Problems of Statistics and Everyday Life.'' *Journal of the American Statistical Association, 74*:1–4.

Carle, Adam C. 2009. ''Fitting Multilevel Models in Complex Survey Data with Design Weights: Recommendations.'' *BMC Medical Research Methodology 9*(49):1–13.

Christ, Sharon, Paul Biemer, and Christopher Wiesen. 2007. *Guidelines for Applying Multilevel Modeling to the NSCAW Data*. Ithaca, NY: National Data Archive on Child Abuse and Neglect.

Cook, Samantha R. and Andrew Gelman. 2006. ''Survey Weighting and Regression.'' Technical Report. New York: Columbia University, Department of Statistics.

Eideh, Abdulhakeem and Gad Nathan. 2009. ''Two-stage Informative Cluster Sampling with Application in Small Area Estimation.'' *Journal of Statistical Planning and Inference* 139:3088–3101.

Goldstein, Harvey. 1986. ''Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares.'' *Biometrika* 73:43–56.

Harris, K. M., C. T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. R. Udry. 2009. ''The National Longitudinal Study of Adolescent Health: Research Design.'' Retrieved August 21, 2012 (http://www.cpc.unc.edu/projects/addhealth/design).

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley.

Korn, Edward L. and Barry I. Graubard. 1995. ''Examples of Differing Weighted and Unweighted Estimates from a Sample Survey.'' *American Statistician* 49:291–95.

Korn, Edward L. and Barry I. Graubard. 2003. ''Estimating Variance Components by Using Survey Data.'' *Journal of the Royal Statistical Society, Series B* 65:175–90.

Krieger, Abba. M. and Danny Pfeffermann. 1992. ''Maximum Likelihood from Complex Sample Surveys.'' *Survey Methodology* 18:225–39.

Krieger, Abba M. and Danny Pfeffermann. 1997. ''Testing of Distribution Functions from Complex Sample Surveys.'' *Journal of Official Statistics* 13:123–42.

Laird, Nan M. and James H. Ware. 1982. ''Random-effects Models for Longitudinal Data.'' *Biometrics* 38:963–74.

Lehtonen, Risto and Erkki J. Pahkinen. 2004. *Practical Methods for Design and Analysis of Complex Surveys*. Hoboken, NJ: John Wiley.

Little, R. J. A. 1993. ''Post-stratification: A Modeler's Perspective.'' *Journal of the American Statistical Association* 88:1001–12.

Mels, Gerhard. 2006. *LISREL for Windows: Getting Started Guide*. Lincolnwood, IL: Scientific Software International.

Mickey, Ruth M., Gregory D. Goodwin, and Michael C. Costanza. 1991. ''Estimation of the Design Effect in Community Intervention Studies.'' *Statistics in Medicine* 10:53–64.

Muthén, Linda K. and Bengt O. Muthén. 1998–2011. *Mplus User's Guide*. 6th ed. Los Angeles: Muthén & Muthén.

Pfeffermann, Danny. 1993. ''The Role of Sampling Weights When Modeling Survey Data.'' *International Statistical Review* 61(2):317–37.

Pfeffermann, Danny. 1996. ''The Use of Weights in Survey Analysis.'' *Statistical Methods in Medical Research* 5:239–61.

Pfeffermann, Danny, Fernando Antonio Da Moura, and Pedro Luis Do Nascimento Silva. 2006. ''Multilevel Modeling under Informative Sampling.'' *Biometrika* 93:943–59.

Pfeffermann, Danny, Abba M. Krieger, and Yosef Rinott. 1998. ''Parametric Distributions of Complex Survey Data under Informative Probability Sampling.'' *Statistica Sinica* 8:1087–1114.

Pfeffermann, Danny, C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. ''Weighting for Unequal Selection Probabilities in Multilevel Models (with Discussion).'' *Journal of the Royal Statistical Society, Series B* 60:23–56.

Pfeffermann, Danny and Michail Sverchkov. 1999. ''Parametric and Semi-parametric Estimation of Regression Models Fitted to Survey Data.'' *Sankhya Series B* 61: 166–86.

Rabe-Hesketh, Sophia and Anders Skrondal. 2006. ''Multilevel Modeling of Complex Survey Data.'' *Journal of the Royal Statistical Society, Series A* 169:805–27.

Robinson, G. K. 1991. ''That BLUP Is a Good Thing: The Estimation of Random Effects.'' *Statistical Science* 6:15–51.

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

Searle, Shayle R., George Casella, and Charles E. McCulloch. 1992. *Variance Components*. New York: John Wiley.

Skinner, C. J., D. Holt, and T. M. F. Smith, eds. 1989. *Analysis of Complex Surveys*. Chichester, UK: Wiley.

Tourangeau, Roger and Hee-Choon Shin. 1999. ''Grand Sample Weight.'' Unpublished paper. Chapel Hill: Carolina Population Center, University of North Carolina at Chapel Hill.

Author Biography

Tianji Cai is an assistant professor of sociology at the University of Macau. His research focuses on quantitative research methods, especially the issues of sampling weights in multilevel and longitudinal models. In addition, he is also interested in integrating genetics and sociology in the studies of social and health behaviors.