RESEARCH REPORT SERIES
(Statistics #2020-05)

Model-Assisted Estimation of Mixed-Effect Model Parameters in Complex Surveys

Eric V. Slud[1, 2]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau
[2]Department of Mathematics, University of Maryland, College Park

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 28, 2020

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

Model-Assisted Estimation of Mixed-Effect Model Parameters in Complex Surveys

Eric V. Slud, US Census Bureau and University of Maryland

August 28, 2020

Abstract. The objective of this paper is to study the feasibility and consistency of complex-survey estimation methods for population models incorporating shared cluster-level random-effect parameters, in contexts where sampling may be informative and where joint inclusion-probabilities are unavailable. There are several well-known papers in this area (Binder 1983, Pfeffermann et al. 1998, Korn and Graubard 2003, Rabe-Hesketh and Skrondal 2006) and several recent papers (Rao et al. 2013, Yi et al. 2016, Kim et al. 2017, Savitsky and Williams 2018, and Williams and Savitsky 2019), but the problem of model- and design-consistent estimation of variance-component parameters in surveys has not yet been solved in a practically effective way.

One contribution of this paper is to propose an EM algorithm applied to the pseudo-loglikelihood estimating an augmented census-loglikelihood incorporating cluster random effects. This algorithm consistently estimates superpopulation variance components under assumed mixed-effect models for survey data under probability sampling designs in which sampling of clusters may be informative but within-cluster sampling is not. A second contribution is to assess the performance of all of the competing proposed methods that use only single-inclusion weights, in the presence of informative sampling under the two-level one-way random-effects Analysis of Variance model. This comparison supports the conclusion that none of these methods is consistent under general informative sampling.

Keywords: Consistency, Informative sampling, Single-inclusion weights, Two-level model.

Disclaimer: This paper is released to inform interested parties of ongoing research and to encourage discussion of it. The views expressed are the author's and not the Census Bureau's.

1 Introduction: Random Effects Models in Complex Surveys

Consider a setting where it is desired to analyze data from a complex survey in which the interesting outcome variables Y_i and explanatory vectors $Z_i \in \mathbb{R}^p$ are dependent within clusters of units i in the underlying frame population $\mathcal{U}$. We restrict attention in this paper to parametric models for data $\mathcal{D} \equiv \{(Y_i, Z_i) : \ i \in \mathcal{U}\}$ with scalar Y_i.

A complex probability survey is a mechanism, partially under the control of the investigator, for designating a random subset $\mathcal{S} \subset \mathcal{U}$ as being **observable**, and part of this mechanism is summarized through the single-inclusion probabilities π_i and weights π_i defined by

$$\pi_i \ = \ P(i \in \mathcal{S} \,|\, Y_i, Z_i) \ \ , \quad w_i \ = \ 1/\pi_i \qquad \text{for} \quad i \in \mathcal{U} \tag{1}$$

In standard sampling theory texts such as Särndal et al. (1992) or Lohr (2009), sampling designs are usually specified with π_i, w_i constant. In real sampling designs, nonresponse usually intervenes, so that the only units i that are truly observable are those within the *respondent set* in the sample $\mathcal{S}$, and the mechanism of response is often modeled with indicators R_i of individual response conditionally independent given $\mathcal{D} \equiv \{(Y_l, Z_l) : \ l \in \mathcal{U}\}$, with probability or *propensity* of response by the i'th unit of the form

$$P(R_i = 1 \,|\, \mathcal{D}) \ = \ g(Y_i, Z_i) \tag{2}$$

depending on the underlying population data $\mathcal{D}$ only through the i'th unit's variables (Y_i, Z_i), through a function g not depending on i. It is also standard in sampling theory and practice to assume that the single-inclusion weights w_i are observable for units $i \in S$. The function g is generally assumed unknown, although sometimes it is modeled parametrically. The observable data from the complex survey then consist of

$$\{ \ (Y_i, \, Z_i, \, w_i) \ : \ \ i \in \mathcal{S} \ \} \tag{3}$$

That is why, in formula (1), the inclusion probabilities and weights are allowed to depend on underlying features of the frame population units, summarized through the random data (Y_i, Z_i) on those respective units. An even more general formulation of a random underlying population and random sampling mechanism has been given by Rubin-Bleuer and Kratina (2006). The level of generality of the special case (1) is common in the sampling and biostatistical literature. We next specify a parametric form (a **superpopulation model**) for the underlying data $\{ \, (Y_i, \, Z_i) : \ i \in \mathcal{U} \, \}$.

If the underlying vector variables (Y_i, Z_i) are modeled as independent identically distributed (*iid*) across $i \in \mathcal{U}$, with parametric joint density $f(y_i, z_i, \theta)$, then the *census loglikelihood* $\sum_{i \in \mathcal{U}} \log f(Y_i, Z_i, \theta)$ is asymptotically approximated by the survey-weighted *pseudo-loglikelihood*

$$\sum_{i \in \mathcal{S}} w_i \, \log f(Y_i, Z_i, \theta) \tag{4}$$

introduced by Binder (1983), and this asymptotic approximation can be used to justify consistent and asymptotically normal inference based on the maximizer of (4) in θ. The same idea works under more general conditions on random or nonrandom (Y_i, Z_i) for (4) or for the conditional pseudo-loglikelihood $\sum_{i \in \mathcal{S}} w_i \, \log f(Y_i \mid Z_i, \theta)$. The applicability and justification of this estimation technique at population level when no model is actually assumed at unit level (and the target estimand is the maximizer of the census-loglikelihood with respect to θ) has given rise to so-called *model-assisted* inferential procedures (Särndal et al. 1992). Also in the setting with *iid* (Y_i, Z_i) for $i \in \mathcal{U}$, the same technique provides design- and model-consistent inferences for θ under general conditions even when sampling is *informative* in the sense that the conditional distributions of (Y_i, Z_i) given $i \in \mathcal{S}$ are different from the distribution of (Y_i, Z_i) for each $i \in \mathcal{U}$.

However, in superpopulation models with dependence across units, such as those in which only clusters are independent, this pseudo-loglikelihood approach to inference must be modified.

1.1 Superpopulation Models with Cluster Structure

We begin by establishing notation for cluster samples, allowing superpopulation data to be dependent within clusters. Then we specialize to *two-level* models with a single random-effect variate shared within each superpopulation cluster, *iid* across clusters. While it is not always true that design inclusion-probabilities are given with separate known factors for sampling of clusters and sampling within the clusters, this hierarchical sampling structure is assumed for single-inclusion probabilities throughout the present paper.

The underlying frame population index set is $\mathcal{U}$, with $|\mathcal{U}| = N$ elements, but its indices i are viewed as standing in one-to-one correspondence with double indices $i = (j, k)$ where $\mathcal{U}$ is partitioned into clusters $\mathcal{U}_k$, $k = 1, \ldots, M$, with respective numbers of elements N_k, and where $k = k(i)$ is the cluster such that $i \in \mathcal{U}_k$, and j indexes units within cluster. In this way, $Y_{j,k}, Z_{j,k}, w_{j,k}$ can be written interchangeably for Y_i, Z_i, w_i. Assume further that the sampling design is a hierarchical cluster design, so that $w_{j,k} = \omega_k \cdot w_{j|k}$ for $j = 1, \ldots, N_k,\ k = 1, \ldots, M,$ where ω_k is a *single-inclusion cluster weight* and $w_{j|k}$ is a *within-cluster single-inclusion (conditional) weight* with

$$\mathcal{S}_C \ = \ \{k = 1, \ldots, M \ : \ k(i) = k \ \text{ for some } \ i \in \mathcal{S}\} \ , \quad \mathcal{S}_k \ = \ \{j = 1, \ldots, N_k : \ (j,k) \in \mathcal{S}\} \tag{5}$$

$$1/\omega_k = P(\, k \in \mathcal{S}_k \,\big|\, \mathcal{D}\,) \quad , \qquad 1/w_{j|k} = P(\,(j,k) \in \mathcal{S} \,\big|\, k \in \mathcal{S}_k \,,\ \mathcal{D}\,) \tag{6}$$

and **within-cluster sampling is assumed to be done independently across clusters**. Denote the numbers of sampled clusters, of sampled units within clusters, and of total sampled units by

$$m \ = \ |\mathcal{S}_C| \ , \quad n_k \ = \ |\mathcal{S}_k| \ , \quad n \ = \ |\mathcal{S}| \ = \ \sum_{k \in \mathcal{S}_C} n_k \tag{7}$$

Then a parametric, clustered superpopulation model could take the form that for *iid* random-effect variables $a_k \sim f_C(a, \eta_2),\ k = 1, \ldots, M$, conditionally given $\underline{Z}_k \equiv (Z_{j,k} : j = 1, \ldots, N_k)$,

$$\underline{Y}_k \ \equiv \ (Y_{j,k} : j = 1, \ldots, N_k) \ \sim \ f_k(\underline{y} \,|\, \underline{Z}_k,\ a_k,\ \eta_1) \ \text{ independently across } \quad k = 1, \ldots, M \tag{8}$$

with $\theta = (\eta_1, \eta_2)$ as unknown parameter. The rest of the paper restricts attention to a more specific, two-level form of such a model and assumes the variables $Y_{j,k}$ are conditionally *iid* across j within cluster k. That is, the model takes the simplified two-level form

$$a_k \ \overset{iid}{\sim} \ f_C(a, \eta_2) \qquad \text{and} \qquad Y_{j,k} \ \overset{iid}{\sim} \ f(y \,|\, Z_{j,k},\ a_k,\ \eta_1) \tag{9}$$

where the functions f_C, f are assumed known and the parameters $\theta = (\eta_1, \eta_2)$ unknown.

In this setting, the sampling is called *informative* if for at least some sets $s_C = \{k_1, \ldots, k_r\}, s_k = \{i_1, \ldots, i_{q_k}\}$, the joint densities of observed variables differ from the corresponding joint densities in the frame population given that these are the selected sets $\mathcal{S}_C$, $\mathcal{S}_k$, i.e.,

$$\mathcal{L}\Big(\, a_k,\ (Y_{j,k} \,:\, j \in s_k),\ k \in s_C) \,\Big|\, \mathcal{S}_C = s_C,\ \mathcal{S}_k = s_k \ \ \forall k \in \mathcal{S}_C,\ (Z_{j,k} \,:\, j \in s_k,\ k \in s_C)\Big) \ \neq$$

$$\mathcal{L}\Big(\, a_k,\ (Y_{j,k} \,:\, j \in s_k),\ k \in s_C) \,\Big|\, (Z_{j,k} \,:\, j \in s_k,\ k \in s_C)\Big)$$

where $\mathcal{L}(\cdot \,|\, \cdot)$ denotes (conditional) probability law. Here $\mathcal{S}_C$ and $\mathcal{S}_k$ respectively denote random sets of selected clusters and of selected units within cluster k, and s_C, s_k respectively denote particular sets of clusters and units that might have been selected. Under our assumptions of independent sampling and independent data across clusters, informativeness means that for some k, s_k,

$$\mathcal{L}\Big(\, a_k,\ (Y_{j,k} \,:\, j \in s_k) \,\Big|\, \mathcal{S}_k = s_k,\ (Z_{j,k} \,:\, j \in s_k),\ k \in S_C \Big) \ \neq \ \mathcal{L}\Big(\, a_k,\ (Y_{j,k} \,:\, j \in s_k) \,\Big|\, Z_{j,k} \,:\, j \in s_k\Big)$$

In what follows, potential informativeness of sampling is considered at each of the two (cluster and within-cluster) stages, where by definition **clusters are sampled noninformatively** if

$$\mathcal{L}\Big(a_k, \, \underline{Y}_k \, \Big| \, \underline{Z}_k \, , \; k \in S_C \Big) \; = \; \mathcal{L}\Big(a_k, \, \underline{Y}_k \, \Big| \, \underline{Z}_k \Big) \tag{10}$$

where recall that $\underline{Y}_k$ denotes the vector of population attributes and $\underline{Z}_k$ denotes the vector of observed explanatory variables, if any, for units within cluster k. Similarly, **within-cluster sampling is noninformative** if for all $k = 1, \ldots, M$ and all within-cluster index sets $s_k \subset \{1, \ldots, N_k\}$,

$$\mathcal{L}\Big(\, (Y_{j,k} : \; j \in s_k) \, \Big| \, a_k, \; (Z_{j,k} : j \in s_k) \, , \; \mathcal{S}_k = s_k, \; k \in S_C \Big) \; = \tag{11}$$

$$\mathcal{L}\Big(\, (Y_{j,k} : \; j \in s_k) \, \Big| \, a_k, \; (Z_{j,k} : j \in s_k) \, , \; k \in S_C \Big)$$

1.2 Model-assisted Estimation and Problem Statement

Model-assisted survey inference (Särndal et al. 1992) is a term ordinarily applied to methods of inference from complex survey data about statistical parameters in a superpopulation model, based on approximating the solution to a population-wide (*census*) estimating equation, such as the maximizer of a population-wide log-likelihood. The parameters estimated are then descriptive parameters of the population, without the superpopulation model necessarily being assumed to hold for population units. *Model-assisted* estimators are required to be consistent in the sense of large-superpopulation and large-sample convergence under the complex-survey probabilistic design, whether or not that design is informative. Without very strong further assumptions, the parameters estimated in this way do not describe model relationships at the level of population subdomains or units. Thus, the estimated parameters **can** be used directly in summarizing population-level data relationships but **cannot** be used in unit-level prediction or domain-level estimation of totals. On the other hand, for many intended applications, such as *small area estimation* (see Molina and Rao 2015), unit-level prediction and domain estimation are precisely the point. In that case, a highly desirable property for an inferential method based on complex survey data is to provide estimators that converge in probability for large samples under the combined sample-design and a complete joint probability model for all population units, whether or not the sampling is informative. This property of an inferential method, which may be proved to hold under some restrictions on the type of informative sampling allowed, is called (joint) design- and model-consistency, which we refer to simply as *consistency* for the rest of this paper. Since this property does rely on the validity of the underlying model, it falls properly within the realm of model-based rather than model-assisted methodology.

In classical sampling theory, the statistical investigator chooses the sampling design and can use joint as well as single inclusion probabilities in analyzing the survey data. However, in the real world of nonresponse, survey methodologists modify single-inclusion weights by methods such as calibration and raking, and while methodological problems and solutions are sometimes described as though the modified weights are reciprocals of inclusion-and-response probabilities, practitioners do not pretend to produce modified joint inclusion probabilities. Therefore, this paper studies the problem of design- and model-consistent estimation of statistical superpopulation parameters for cluster-dependent data, only in terms of observable sampled data and single-inclusion weights.

1.3 First Estimation Methods

The earliest paper attacking the problem of mixed-model consistency under informative sampling seems to be that of Pfeffermann et al. (1998), which considered informatively sampled data from a superpopulation satisfying a normal linear model defined by

$$Y_{j,k} \;=\; \beta' Z_{j,k} \,+\, a_k \,+\, \epsilon_{j,k} \;, \quad a_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_a^2) \;, \quad \epsilon_{j,k} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2) \tag{12}$$

They proposed a complicated iterative Weighted Least Squares procedure involving weight-rescaling. Although they gave no proofs, their method appears from simulations (theirs and also Korn and Graubard's 2003) to provide (model- and design-) consistent estimates of $\theta = (\beta, \sigma_1^2, \sigma_e^2)$ under noninformative sampling and also under sample designs that are noninformative within clusters. Here $\beta \in \mathbb{R}^p$ (with the first element of each vector $Z_{j,k}$ equal to 1, providing an intercept term in the model), and $\eta_1 = (\beta, \sigma_e^2), \eta_2 = \sigma_a^2$, respectively play the roles here of the first and second-level parameters in $\theta = (\eta_1, \eta_2)$.

Korn and Graubard (2003) studied the intercept-only case of (12) where $Z_{j,k} \equiv 1$ (reducing the coefficient p-vector β to the scalar mean parameter μ), under several sorts of highly informative sampling designs. We refer to that model from now on as the **two-level Analysis of Variance (ANOVA) model**. Their simulations showed under that model that the methods of Pfeffermann et al. were *not* consistent under general informative sampling. Korn and Graubard (2003) did also provide consistent weighted method-of-moments estimators expressed in terms of joint inclusion probabilities, estimators that were later generalized by Rao et al. (2013) and Yi et al. (2016).

1.4 Competing Pseudolikelihoods

As mentioned above, cluster-level dependence invalidates simple survey-weighted loglikelihood (4) (or its conditional variant) as a basis for consistent parametric inference under general informative sampling designs, even under designs that sample independently across clusters. The distinction between general sampling designs and those that are noninformative within clusters (but may be informative at cluster level) is important here.

To estimate $\theta = (\eta_1, \eta_2)$, Rabe-Hesketh and Skrondal (2006) maximize

$$\text{logLik}_1 \;=\; \sum_{k \in \mathcal{S}_C} \omega_k \log \int \exp\Big(\sum_{j \in \mathcal{S}_k} w_{j|k} \log f(Y_{jk} \,|\, \mathbf{Z}_{jk}, a_k, \eta_1) \Big)\, f_C(a_k, \eta_2)\, da_k \tag{13}$$

an approximate log-likelihood that Asparouhov (2006) also used iteratively together with weight-rescaling. But the integral expression within (13) is not a likelihood, and consistency of estimation can be justified generally only when the within-cluster sample-sizes go to ∞.

Williams and Savitsky (2018) and Savitsky and Williams (2019, 2020) have developed a Bayesian ‘pseudo-posterior’ methodology applicable when clusters of bounded size are sampled approximately independently. With a few further technical restrictions, they provide Bayesian large-sample theory and supporting simulations, claiming that their method exhibits large-sample posterior concentration (a Bayesian analogue of consistency) under general informative sampling. Although they have not specifically related their results to a frequentist method, they seem to be suggesting that likelihood methods with the approximate loglikelihood

$$\text{logLik}_2 \;=\; \sum_{k \in \mathcal{S}_C} \log \int \exp\Big(\sum_{j \in \mathcal{S}_k} w_{j,k} \log f(Y_{jk} \,|\, \mathbf{Z}_{jk}, a_k, \eta_1) + \omega_k \log f_C(a_k, \eta_2) \Big)\, da_k \tag{14}$$

should have favorable asymptotic properties under informative sampling.

The pseudo-loglikelihood (4) of Binder (1983) is a special case of a more general concept of *composite log-likelihood* (Lindsay 1988), which is essentially a weighted linear combination of terms individually – but not necessarily jointly – justified as log-likelihoods for subsets of the observed data. Expressions (13) and (14) are **not** composite loglikehoods in this sense, although the versions of these expressions without the integration over a_k **are** composite augmented loglikelihoods (*augmented* because the a_k's are not observed), which is exploited in defining an EM method in Section 2 below. The authors proposing (13) and (14) had in mind the approximation of the within-cluster log-likelihood by the survey-weighted pseudo-loglikelihood, but the integral over a_k of the exponentiated pseudo-loglikelihood is no longer a loglikelihood. Using a different approach, Rao et al. (2013) and Yi et al. (2016) revived the composite loglikelihood in the mixed-effects survey estimation context, following preliminary suggestions of Korn and Graubard (2003) and Graubard and Korn (2011), by weighting the log-likelihoods of pairs of observed unit-level data $(Y_i, Y_{i'})$ by the inverse of the joint probability $P(i, i' \in \mathcal{S})$. Under mild conditions, this approach provides a consistent estimator for θ under general informative sampling at the cost of requiring knowledge of correct within-cluster joint inclusion probabilities.

1.5 Scope of This Paper

The objective of this paper is to study the feasibility and consistency of complex-survey estimation methods for population models incorporating shared cluster-level random-effect parameters, in contexts where sampling may be informative and where joint inclusion-probabilities are unavailable. The paper has three parts. In the first, an EM algorithm is proposed in the spirit of model-assisted estimation, to estimate the population-level maximum likelihood parameter values within the census loglikelihood for observable parameters. This algorithm is feasible and is shown to provide consistent estimates when sampling of clusters may be informative but sampling within clusters is noninformative. Second, the major methods of parameter estimation that use only single-inclusion weights are compared theoretically with respect to consistency within the two-way random-effects Analysis of Variance model with normally distributed errors and random effects, where estimators and their limits under noninformative within-cluster sampling are relatively explicit. Third, these methods are compared more broadly within the two-level one-way ANOVA model, under sampling designs ranging from fully noninformative to informative both at cluster-level and within clusters. In these theoretical and simulation-based comparisons, since the goal is to understand consistency of estimation, we ignore methods (primarily, those of Pfeffermann et al. 1998 and Asparouhov 2006) essentially involving iterative weight-rescaling, since such rescaling is directed primarily at variance and interval estimation. Broadly speaking, the thrust of the paper's results is that the new EM method is highly effective and consistent in settings where within-cluster sampling is noninformative, and that **no method** based on single-inclusion probabilities is consistent when within-cluster sampling is informative.

2 The Pseudo-EM Method

The rationale behind (4) was to approximate the census (full-population) loglikelihood, and similarly in the mixed-model setting, the census augmented loglikelihood

$$l_{aug.cens}(\theta) \;=\; \sum_{k=1}^{M} \log f_C(a_k, \eta_2) \;+\; \sum_{(j,k)\in\mathcal{U}} \log f(Y_{j,k} \,|\, Z_{j,k}, a_k, \eta_1)$$

which is the correct loglikelihood for the full-population data $\{Y_{j,k}, Z_{j,k}\}_{(j,k)\in\mathcal{U}}$ along with the intrinsically unobservable random effects $\{a_k\}_{k=1}^{M}$. This augmented loglikelihood is estimated design-consistently (for all parameters $\theta = (\eta_1, \eta_2)$, after normalizing by $1/N$) by:

$$l_w(\theta) = \sum_{k\in\mathcal{S}_C} \omega_k \Big\{ \log f_C(a_k, \eta_2) \;+\; \sum_{j\in\mathcal{S}_k} w_{j|k} \log f(Y_{j,k} \,|\, Z_{j,k}, a_k, \eta_1) \Big\} \;\equiv\; \sum_{k\in\mathcal{S}_C} \omega_k \; l_k^w(\theta)$$

Recall the notation $\mathcal{D}$ for the full-population observable dataset $\{Y_{j,k},\, Z_{j,k} :\; (j,k) \in \mathcal{U}\}$, and note by independence of both the sampling mechanism and superpopulation data across clusters that

$$E_{\theta_0}\Big(l_k^w \,|\, \mathcal{D} \Big) \;=\; E_{\theta_0}(l_k^w(\theta) \,|\, \{Y_{j,k}, Z_{j,k},\; j = 1, \ldots, N_k\})$$

For reference express the usual marginal census observed-data loglikelihood by

$$l_{obs.cens}(\theta, \mathcal{D}) \;=\; \log \Big(\int \exp(l_{aug.cens}(\theta)) \, da_1 \, da_2 \, \cdots \, da_M \Big)$$

The EM algorithm is formulated by taking conditional expectations as though ω_k depends only on $\mathcal{D}_k \;\equiv\; \{Y_{j,k}, Z_{j,k},\; j = 1, \ldots, N_k\}$, not on a_k. The two steps are as follows:

$$\textbf{E-step:}\quad Q_k(\theta, \theta_0) \;\equiv\; \int \Big(\log f_C(x, \eta_2) \;+\; \sum_{j\in\mathcal{S}_k} w_{j|k} \log f(Y_{j,k} \,|\, Z_{j,k}, x, \eta_1) \Big) f_{a_k|\mathcal{D}_k}(x \,|\, \mathcal{D}_k) \, dx \tag{15}$$

$$\textbf{M-step:}\qquad \theta_1 \;\equiv\; \arg\max_\theta \sum_{k\in\mathcal{S}_C} Q_k(\theta, \theta_0) \tag{16}$$

Then, assuming sampling to be noninformative within-cluster, the usual justification of the EM algorithm via the conditional Jensen inequality shows that, starting from an initial guess θ_0 for the true model-parameter θ_*, the EM-step (15)–(16) yields a value such that

$$E_{\theta_0}\Big(l_{aug.cens}(\theta_1) \;-\; l_{obs.cens}(\theta_1, \mathcal{D}) \,\Big|\, \mathcal{D} \Big) \;\le\; E_{\theta_0}\Big(l_{aug.cens}(\theta_0) \;-\; l_{obs.cens}(\theta_0, \mathcal{D}) \,\Big|\, \mathcal{D} \Big) \tag{17}$$

Moreover, an elementary use of bracketing as in the development of empirical-process uniform laws of large numbers, under mild restrictions of continuity of $l_{aug.cens}(\theta)$ with respect to θ, shows for compact neighborhoods $B(\theta_*)$ of the true parameter value θ_*

$$\sup_{\theta \in B(\theta_*)} \frac{1}{N} \left| l_{aug.cens}(\theta) \, - \, l_w(\theta) \right| \;\to\; 0 \qquad \text{in design- and} \quad P_{\theta_*} \quad \text{probability}$$

and as $m \to \infty$,

$$\sup_{\theta \in B(\theta_*)} \frac{1}{N} \left| E_{\theta_0}\Big(l_w(\theta) \, - \, l_{aug.cens}(\theta) \,\Big|\, \mathcal{D}, \mathcal{S}) \Big) \right| \;\to\; 0 \tag{18}$$

(See Boistard et al. 2017 on empirical process theory arising in design-based sampling, and van der Vaart 1998, Ch. 19, as a reference on Glivenko-Cantelli-type limit theorems based on bracketing.)

As a consequence of (17)–(18), for $\theta_0 \in B(\theta_*)$ and under conditions ensuring also $\theta_1 \in B(\theta_*)$,

$$\min\Big(\frac{1}{N} \Big[l_{obs.cens}(\theta_1, \mathcal{D}) \, - \, l_{obs.cens}(\theta_0, \mathcal{D}) \Big], \, 0 \Big) \;\to\; 0 \qquad \text{in design- and} \quad P_{\theta_*} \quad \text{probability.}$$

Under regularity conditions, this kind of argument justifies the fixed-point of the pseudo-EM iteration step (15)–(16) as an approximate stationary point of $l_{obs.cens}(\theta, \mathcal{D})$. Further asymptotic theory could be developed by regarding

$$Q(\theta, \theta_0) \;\equiv\; \sum_{k \in \mathcal{S}_C} \omega_k \, Q_k(\theta, \theta_0)$$

as a bivariate random process indexed by (θ, θ_0). Under assumptions like those made below in the two-level one-way random-effects ANOVA model, with both the sampling and the superpopulation data independent across the large set of primary clusters, the process becomes $Q(\theta, \theta_0) \, = \, E_{\theta_0}\big\{ l_w(\theta) \,\big|\, \mathcal{D} \big\}$. At least in the case of bounded-size clusters, this process (assumed jointly continuously differentiable in θ, θ_0) can be studied by empirical process methods along the lines of Boistard et al. 2017. Under regularity conditions, when sampling is noninformative within clusters (but the sampling of clusters themselves might depend on a_k), this EM approach can be proved consistent. Results along these lines are supplied below in the two-level ANOVA model.

An EM algorithm iterating a step analogous to (15)–(16), but with $l_w(\theta)$ replaced by a differently weighted augmented loglikelihood closer in spirit to (13), was earlier proposed by Kim et al. (2017). Both the pseudo-EM step (15)–(16) proposed here, and that proposed by Kim et al. (2017), integrate conditional expectations for the log-likelihood terms from cluster k using the true conditional law of a_k given $(Y_{j,k}, \; Z_{j,k}, \;\; j \in \mathcal{S}_k)$ and $k \in \mathcal{S}_C$. Since the EM proposal of Kim et al. (2017) works with a cluster pseudo-loglikelihood under an approximate framework making use of within-cluster joint inclusion probabilities, we do not discuss it further here. However, the pseudo-EM method proposed here is different from the other pseudo-loglikelihood maximizers compared below.

If the E-step expectation $E_{\theta_0}(\, l_w(\theta) \,|\, \mathcal{D},\, \mathcal{S})$ can be implemented, then this EM method which approximates the census augmented-data loglikelihood should provide model- and design-consistent estimates of θ, even under informative sampling. Suppose that both the superpopulation and the sampling are independent across clusters, and that sampling is informative with respect to clusters only, and not informative within clusters, in the sense that the conditional density of the k'th-cluster random effect a_k given $k \in \mathcal{S}_C$ and $\{Y_{j,k}, Z_{j,k}\}_{j \in \mathcal{S}_k}$ is proportional to $f_C(a_k, \eta_2)\, f(y_{j,k} \,|\, z_{j,k},\, a_k,\, \eta_1)$. Then the conditional expectation in (15)–(16) can be implemented or approximated analytically.

The pseudo-EM method described here is justified to the extent that the survey data allow accurate approximation to the observable-data census loglikelihood. However, the sketched theoretical argument above does not justify convergence of the successive EM-steps, or say anything about the algorithm when within-cluster sampling is informative. Nevertheless, the pseudo-EM method (15)–(16) is included as one of the competing methods to be applied to survey data satisfying a two-level (intercept-only) analysis-of-variance model.

3 Superpopulation Two-level Analysis of Variance Model

The two-level ANOVA model mentioned in Section 1.3 was defined by

$$Y_{j,k} \,=\, \mu \,+\, a_k \,+\, \epsilon_{j,k}\ , \quad a_k \overset{iid}{\sim} \mathcal{N}(0, \sigma_a^2)\ , \quad \epsilon_{j,k} \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2) \tag{19}$$

Under this model, the various integrals and conditional expectations defined in the competing pseudo-loglikelihoods and EM expressions above can be rendered explicitly, and the performance of the corresponding estimation methods compared in greater detail under different large-sample assumptions and sampling designs.

For uniformity of notation in what follows, note that the standard Horvitz-Thompson estimators in terms of the weights ω_k, $w_{j|k}$, w_{jk} respectively of the numbers M of clusters, N_k of units within the k'th cluster, and N of units within the finite population, are given by

$$\hat{M} \,=\, \sum_{k \in \mathcal{S}_C} \omega_k \ \ , \quad \hat{N}_k \,=\, \sum_{j \in \mathcal{S}_k} w_{j|k} \ \ , \quad \hat{N} \,=\, \sum_{(j,k) \in \mathcal{S}} w_{j,k} \,=\, \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \tag{20}$$

Under this model, the approximate pseudo-loglikelihood and conditional expectations needed for the various estimation methods can be expressed explicitly. Note that under model (19),

$$f_C(a_k, \eta_1) \;=\; \frac{1}{\sqrt{2\pi\,\sigma_a^2}}\, e^{-a_k^2/(2\sigma_a^2)} \;\;, \qquad f(y \,|\, Z_{j,k},\, a_k,\, \eta_2) \;=\; \frac{1}{\sqrt{2\pi\,\sigma_e^2}}\, e^{-(y-a_k-\mu)^2/(2\sigma_e^2)} \tag{21}$$

Substitute (21) respectively into (13) and (14) to obtain (after removing additive constants that do not depend on parameters)

$$\textbf{pslogLik}_1 \;=\; \sum_{k\in\mathcal{S}_C} \omega_k \,\log\Big\{(\sigma_e)^{-\hat{N}_k}\,\sigma_a^{-1} \int \exp\Big(-\sum_{j\in\mathcal{S}_k} w_{j|k}\frac{(Y_{j,k}-\mu-a_k)^2}{2\,\sigma_e^2} \;-\; \frac{a_k^2}{2\,\sigma_a^2}\Big)\, da_k\Big\} \tag{22}$$

$$\textbf{pslogLik}_2 \;=\; \sum_{k\in\mathcal{S}_C} \log\Big\{(\sigma_e)^{-\omega_k\hat{N}_k}\,\sigma_a^{-\omega_k} \int \exp\Big(-\sum_{j\in\mathcal{S}_k} w_{j,k}\frac{(Y_{j,k}-\mu-a_k)^2}{2\,\sigma_e^2} \;-\; \frac{\omega_k\,a_k^2}{2\,\sigma_a^2}\Big)\, da_k\Big\} \tag{23}$$

In simulations below, we also consider estimators based on a variant $\textbf{pslogLik}_0$ of $\textbf{pslogLik}_1$ in which the within-cluster weights $w_{j|k}$ in equation (22) are replaced by 1.

Both expressions (22) and (23) can be unified within a single formula. Let γ_k for $k \in \mathcal{S}_C$ denote positive constants equal to 1 for the case of expression (22) and ω_k for the case of expression (23). Then both expressions have the exact form

$$\sum_{k\in\mathcal{S}_C} \Big\{-\,\frac{\omega_k\,\hat{N}_k}{2}\,\log\sigma_e^2 - \frac{\omega_k}{2}\,\log\sigma_a^2 + \frac{\omega_k}{\gamma_k}\,\log\int \exp\Big(-\,\frac{\gamma_k}{2}\,\Big[\sum_{j\in\mathcal{S}_k} w_{j|k}\frac{(Y_{j,k}-\mu-a_k)^2}{\sigma_e^2} + \frac{a_k^2}{\sigma_a^2}\Big]\Big)\, da_k\Big\}$$

which is written more compactly, using (5) and Lemma 3 in Appendix A, as

$$\text{logL} \;=\; -\,\frac{\hat{N}}{2}\log\sigma_e^2 - \frac{\hat{M}}{2}\log\sigma_a^2 + \frac{1}{2}\sum_{k\in\mathcal{S}_C}\omega_k\Big[\frac{1}{\gamma_k}\log\big(\frac{\sigma_a^2\,\sigma_e^2\,/\,\gamma_k}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\big) - \frac{\text{SSW}_k}{\sigma_e^2} - \frac{\hat{N}_k\,(\bar{Y}^w_{\cdot k}-\mu)^2}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\Big] \tag{24}$$

where

$$\text{SSW}_k \;\equiv\; \sum_{j\in\mathcal{S}_k} w_{j|k}\,(Y_{j,k}-\bar{Y}^w_{\cdot k})^2 \;=\; \sum_{j\in\mathcal{S}_k} w_{j|k}\,(\epsilon_{j,k}-\bar{\epsilon}^w_{\cdot k})^2 \;, \qquad \bar{\epsilon}^w_{\cdot k} = \hat{N}_k^{-1}\sum_{j\in\mathcal{S}_k} w_{j,k}\,\epsilon_{j,k} \tag{25}$$

Note that there is only one term (the first in the square-bracketed summand) in the log-pseudolikelihood expression (24) above where γ_k appears, and this is the only term differing under the two-level ANOVA model (19) between the two log-pseudolikelihood expressions (22) and (23).

The other use made of the densities (21) is in the formula for conditional expectation of cluster log-augmented-density with respect to a_k given $(Y_{j,k} :\; j \in \mathcal{S}_k)$ under the assumption of noninformative sampling within cluster. In this setting, for $s_k \;=\; \{j_1, \ldots, j_{n_k}\}$ a nonrandom set of units within cluster k, (3) implies

$$\text{given}\quad \mathcal{S}_k = s_k \quad\text{and}\quad \{Y_{j,k}\}_{j\in s_k}\;\;, \qquad a_k \;\sim\; \mathcal{N}\Big(\frac{n_k\,\sigma_a^2}{\sigma_e^2+n_k\,\sigma_a^2}\,(\bar{Y}_{\cdot k}-\mu)\,,\; \frac{\sigma_a^2\,\sigma_e^2}{\sigma_e^2+n_k\,\sigma_a^2}\Big) \tag{26}$$

where $\bar{Y}_{\cdot k} = (n_k)^{-1} \sum_{j \in \mathcal{S}_k} Y_{j,k}$ is the unweighted sample average in cluster k and $\bar{\epsilon}_{\cdot k} = \bar{Y}_{\cdot k} - \mu - a_k$ the corresponding average of variables $\epsilon_{j,k}$ sampled in cluster k, and recall that $n_k = |\mathcal{S}_k|$. This standard conditional-density result follows also from Lemma 2 in Appendix A with $\gamma = 1$, $\tau_j \equiv 1$, $V_j = (Y_{j,k} - \mu)$, and $q = n_k$. Therefore

$$m_{rk} \;\equiv\; E_{\theta_0}\Big((a_k)^r \,\Big|\, \mathcal{S}_k = s_k,\, \{Y_{j,k}\}_{j \in s_k} \Big) \qquad \text{for} \quad r = 1, 2$$

is given by

$$q_{k,0} \;\equiv\; \frac{n_k\, \sigma^2_{a,0}}{\sigma^2_{e,0} + n_k\, \sigma^2_{a,0}} \;, \quad \sigma^2_k \;\equiv\; (1 - q_{k,0})\, \sigma^2_{a,0} \;, \quad m_{1k} \;=\; q_{k,0}\, (\bar{Y}_{\cdot k} - \mu_0) \;\;, \quad m_{2k} \;=\; m^2_{1k} + \sigma^2_k \tag{27}$$

where 0 subscripts on parameters reflect the calculation of conditional expectations with respect to $\theta_0 = (\mu_0,\, \sigma^2_{a,0},\, \sigma^2_{e,0})$. Then, starting from (21), a line or two of algebra using (26) leads to

$$E\Big\{ \log \Big(f_C(a_k, \eta_1)^{\omega_k} \prod_{j \in \mathcal{S}_k} f(Y_{j,k} \,|\, a_k, \eta_2)^{w_{j,k}} \Big) \,\Big|\, \mathcal{S}_k,\, \{Y_{j,k}\}_{j \in \mathcal{S}_k} \Big\} = -\,\frac{\omega_k}{2}\, \log((2\pi)^{1+\hat{N}_k}\, \sigma^2_a\, \sigma_e^{2\hat{N}_k})$$
$$-\,\frac{\omega_k}{2} \Big[\frac{m^2_{1k} + \sigma^2_k}{\sigma^2_a} \,+\, \sigma_e^{-2}\, \mathrm{SSW}_k \,+\, \frac{\hat{N}_k}{\sigma^2_e} \left((\bar{Y}^w_{\cdot k} - \mu - m_{1k})^2 \,+\, \sigma^2_k \right) \Big] \tag{28}$$

3.1 Estimation Methods for ANOVA Model

We now proceed to develop estimating formulas under the ANOVA model for each of the (pseudo-) likelihood and EM ideas that have been proposed, with a view to comparing the behavior of the resulting estimators under a variety of assumptions about informative sampling. To begin, ANOVA estimates from a sample-weighted method of moments (replacing $w_{j|k}$ by 1, and assuming all $n_k > 1$ for all sampled clusters within the estimation formulas for variance components) leads to the first set of benchmark estimation formulas, expressed in terms of *residuals* $e_{j,k} \equiv Y_{j,k} - \bar{Y}_{\cdot k}$:

$$\tilde{\mu}^{(M)} = \sum_{(j,k) \in \mathcal{S}} \frac{w_{j,k}}{\hat{N}}\, Y_{j,k} \,, \quad (\tilde{\sigma}^{(M)}_e)^2 = \frac{1}{\hat{M}} \sum_{(j,k) \in \mathcal{S}} \frac{\omega_k\, e^2_{j,k}}{n_k - 1} \,, \quad (\tilde{\sigma}^{(M)}_a)^2 = \sum_{(j,k) \in \mathcal{S}} w_{j,k}\, \frac{(Y_{j,k} - \tilde{\mu})^2}{\hat{N}} - (\tilde{\sigma}^{(M)}_e)^2 \tag{29}$$

Next come estimates expressed as roots of score equations (setting gradients with respect to parameters equal to 0) based on the logL expression (24), respectively with $\gamma_k = 1$ for the Rabe-Hesketh and Skrondal pseudo-loglikelihood (13) and with $\gamma_k = \omega_k$ for the pseudo-loglikelihood (14). The two sets of estimates $\hat{\theta}^{(L)} = (\hat{\mu},\, \hat{\sigma}^2_a,\, \hat{\sigma}^2_e) = (\hat{\mu}^{(L)},\, (\hat{\sigma}^{(L)}_a)^2,\, (\hat{\sigma}^{(L)}_e)^2)$ derived from these equations are written with superscripts $\hat{\theta}^{(L1)}$, $\hat{\theta}^{(L2)}$, and the equations simplified in terms of

$$\hat{\rho}_k \;=\; \hat{N}_k\, (\hat{\sigma}^{(L)}_a)^2 / ((\hat{\sigma}^{(L)}_e)^2 \,+\, \hat{N}_k\, (\hat{\sigma}^{(L)}_a)^2)$$

The score equations obtained by direct differentiation of (24) are:

$$\sum_{k\in\mathcal{S}_C} \frac{\omega_k\, \hat{N}_k\, (\bar{Y}^w_{\cdot k} - \hat{\mu})}{\sigma_e^2 + \hat{N}_k\, \hat{\sigma}_a^2} \;=\; 0 \;\; , \qquad \frac{\hat{M}}{\hat{\sigma}_a^2} \;=\; \sum_{k\in\mathcal{S}_C} \Big[\frac{\omega_k\, \hat{\sigma}_e^2}{\gamma_k\, \hat{\sigma}_a^2\, (\sigma_e^2 + \hat{N}_k\, \hat{\sigma}_a^2)} \;+\; \omega_k\, \frac{\hat{N}_k^2\, (\bar{Y}^w_{\cdot k} - \hat{\mu})^2}{(\sigma_e^2 + \hat{N}_k\, \hat{\sigma}_a^2)^2}\Big]$$

$$\frac{\hat{N}}{\hat{\sigma}_e^2} \;=\; \sum_{k\in\mathcal{S}_C} \Big[\frac{\omega_k\, \hat{N}_k\, \hat{\sigma}_a^2}{\gamma_k\, \hat{\sigma}_e^2\, (\sigma_e^2 + \hat{N}_k\, \hat{\sigma}_a^2)} \;+\; \frac{\omega_k}{\hat{\sigma}_e^4}\, \mathrm{SSW}_k \;+\; \frac{\omega_k\, \hat{N}_k\, (\bar{Y}^w_{\cdot k} - \hat{\mu})^2}{(\sigma_e^2 + \hat{N}_k\, \hat{\sigma}_a^2)^2}\Big]$$

After a further step of simplification with the aid of the notation $\hat{\rho}_k$, the two sets of score equations for the estimates $\hat{\theta}^{(L1)}$ and $\hat{\theta}^{(L2)}$ are respectively given by substituting $\gamma_k = 1$ or $\gamma_k = \omega_k$ into the reduced score equations

$$\hat{\mu}^{(L)} \;=\; \sum_{k\in\mathcal{S}_C} \omega_k\, \hat{\rho}_k\, \bar{Y}^w_{\cdot k} \Big/ \sum_{k\in\mathcal{S}_C} \omega_k\, \hat{\rho}_k \tag{30}$$

$$(\hat{\sigma}_a^{(L)})^2 \;=\; \hat{M}^{-1} \sum_{k\in\mathcal{S}_C} \omega_k \Big[\frac{1}{\gamma_k}\, (\hat{\sigma}_a^{(L)})^2\, (1 - \hat{\rho}_k) \;+\; \hat{\rho}_k^2\, (\bar{Y}^w_{\cdot k} - \hat{\mu}^{(L)})^2)\Big] \tag{31}$$

$$(\hat{\sigma}_e^{(L)})^2 \;=\; \hat{N}^{-1} \sum_{k\in\mathcal{S}_C} \omega_k \Big[\mathrm{SSW}_k \;+\; \frac{(\hat{\sigma}_e^{(L)})^2}{\gamma_k}\, \hat{\rho}_k \;+\; \hat{N}_k\, (1 - \hat{\rho}_k)^2\, (\bar{Y}^w_{\cdot k} - \hat{\mu}^{(L)})^2\Big] \tag{32}$$

The remaining estimation method is the pseudo-EM, obtained by implementing the EM-step equation (15) and then finding its fixed point $\theta_0 = \theta_1 = \hat{\theta}^{(EM)}$. The sum of the expression found in (28) is maximized as in (16) over $\theta = (\mu, \sigma_a^2, \sigma_e^2)$ for fixed θ_0, yielding the EM-equation for $\theta_1 = (\mu_1, \sigma_{a,1}^2, \sigma_{e,1}^2)$ as follows:

$$\mu_1 = \frac{1}{\hat{N}} \sum_{(j,k)\in\mathcal{S}} w_{j,k}(Y_{j,k} - m_{1k}) \; , \quad \sigma_{a,1}^2 = \frac{1}{\hat{M}} \sum_{k\in\mathcal{S}_C} \omega_k(m_{1k}^2 + \sigma_k^2) \tag{33}$$

$$\sigma_{e,1}^2 = \frac{1}{\hat{N}} \sum_{k\in\mathcal{S}_C} \omega_k \Big[\mathrm{SSW}_k \;+\; \hat{N}_k \left((\bar{Y}^w_{\cdot k} - \mu_1 - m_{1k})^2 + \sigma_k^2\right)\Big]$$

Now equation (33) gives only the EM-step for the pseudo-EM algorithm in the ANOVA model. But the fixed-point $\hat{\theta}^{(EM)} = \theta_1 = \theta_0$ of this iterative step is the estimator that we compare with the $\hat{\theta}^{(L1)}$ and $\hat{\theta}^{(L2)}$ estimators defined (implicitly) above. This fixed point is defined, after substituting $\hat{\theta}^{(EM)}$ for both θ_0 and θ_1 and a little algebra, by:

$$\hat{\mu}^{(EM)} \;=\; \frac{\sum_{k\in\mathcal{S}_C} \omega_k\, \hat{N}_k\, (\bar{Y}^w_{\cdot k} - \hat{q}_k \bar{Y}_{\cdot k})}{\sum_{k\in\mathcal{S}_c} \omega_k\, \hat{N}_k\, (1 - \hat{q}_k)} \quad , \quad (\hat{\sigma}_a^{(EM)})^2 \;=\; \frac{\sum_{k\in\mathcal{S}_C} \omega_k\, \hat{q}_k^2\, (\bar{Y}_{\cdot k} - \hat{\mu}^{(EM)})^2}{\sum_{k\in\mathcal{S}_C} \omega_k\, \hat{q}_k} \tag{34}$$

$$(\hat{\sigma}_e^{(EM)})^2 \;=\; \frac{1}{\hat{N}} \sum_{k\in\mathcal{S}_C} \omega_k \Big[\mathrm{SSW}_k + \hat{N}_k \left(\bar{Y}^w_{\cdot k} - \hat{q}_k\, \bar{Y}_{\cdot k} - (1 - \hat{q}_k)\, \hat{\mu}^{(EM)}\right)^2 + (\hat{\sigma}_e^{(EM)})^2\, \frac{\hat{N}_k\, \hat{q}_k}{n_k}\Big] \tag{35}$$

where

$$\hat{q}_k \;\equiv\; n_k\, (\hat{\sigma}_a^{(EM)})^2 \Big/ \Big((\hat{\sigma}_e^{(EM)})^2 + n_k\, (\hat{\sigma}_a^{(EM)})^2\Big)$$

3.2 Uniqueness of Estimating Equation Solutions

The estimators described above are the weighted Method of Moments estimator $\tilde{\theta}^{(M)}$, the maximum weighted-pseudo-loglikelihood estimators $\hat{\theta}^{(L1)}$, $\hat{\theta}^{(L2)}$, and the pseudo-EM estimator $\hat{\theta}^{(EM)}$. Among these, $\tilde{\theta}^{(M)}$ is defined in closed form by (29), but the others must be calculated iteratively.

The maximum weighted-pseudo-loglikelihood estimators $\hat{\theta}^{(L)}$ are found in practice by numerical maximization software, and the global maxima will generally be unique, but the equations (30)–(32) used to reason about them theoretically are only necessary conditions for a stationary point (i.e., a parameter value at which the gradient of pslogLik is 0) and will not in general be known to be unique. Indeed, since the weighted-pseudo-loglikelihoods are not legitimate log-likelihoods and the maximizers not generally consistent, it will not be true as in standard maximum-likelihood theory that with high probability a unique local maximum exists in the neighborhood of the correct parameter value. Consistency and inconsistency for these estimators is discussed in detail below.

The EM iteration defined in (33) is well defined, but is not generally guaranteed to converge. The conjectured large-sample theory alluded to above would apply only in the setting of noninformative within-cluster sampling, which was assumed in deriving the specific formula for the iterative step. But in that setting, the theory leading up to Lemma 12 in Appendix A shows under regularity conditions that with probability approaching 1 in large samples, $\hat{\mu}^{(EM)}$ is consistent (Lemma 10) and the EM iteration with $\hat{\mu}^{(EM)}$ replaced by μ is actually a contraction (Lemma 12) converging uniquely to (σ_a^2, σ_e^2). However, in general informative-sampling settings, the EM iterative equations are somewhat sensitive to initial conditions and may not converge at all.

3.3 Consistency of Methods under the ANOVA Model

In broad outline, the results in the technical appendix imply that the Method-of-Moments and pseudo-EM estimator are consistent under noninformative sampling within clusters and the weighted pseudo-loglikelihood estimators are not. We study differences and similarities among the methods for several different ways in which clusters and cluster weights might be handled. First, we consider the case where the weights within clusters are essentially ignored, which is more or less equivalent to ignoring all units that might have been sampled in a cluster and treating $N_k = n_k$ and $w_{j|k} = 1$ for all j. This is what we *should* do in the interests of efficiency if we are sure that within-cluster sampling is non-informative and observations within cluster k are conditionally i.i.d. given a_k. Second, we consider the case where within-cluster sampling is always simple random sampling, so that $n_k < N_k$ are non-random and $w_{j|k} \equiv N_k/n_k$ for all j. Third, we consider the case where all N_k are large or all n_k are large.

A series of technical lemmas is proved in the Appendix, within the setting of the two-level one-way random effects ANOVA model. All but the first two Lemmas cited below rely on regularity conditions (C1)–(C2) given in the Appendix, the most restrictive aspect of which is that the number of clusters sampled is assumed to be large, that the individual cluster sizes are uniformly bounded, and that the maximum ratio of weights within clusters is also uniformly bounded.

(I) (*Lemma 4:*) When all $n_k \equiv \nu > 1$ have equal size , and all $N_k = N/M$ are of equal size, and all $w_{j|k} = N_k/n_K = N/(M\nu)$, the estimators $\tilde{\theta}^{(M)}$ and $\hat{\theta}^{(EM)}$ are algebraically identical.

(II) (*Lemma 6:*) The estimators $\tilde{\mu}^{(M)}$ for μ and $\tilde{\sigma}_a^{(M)\,2} + \tilde{\sigma}_e^{(M)\,2}$ for $\sigma_a^2 + \sigma_e^2$ are consistent.

(III) (*Lemma 10:*) Under regularity conditions, and the further restriction (C3) that within-cluster sampling is noninformative, $\tilde{\theta}^{(M)}$ is consistent.

(IV) (*Lemma 11:*) Under regularity conditions and noninformative-within-cluster sampling, when $w_{j|k} \equiv 1$, $\hat{\theta}^{(L1)}$ is **inconsistent** when the relative frequency that $N_k - n_k > 0$ is bounded below, and $\hat{\theta}^{(L2)}$ is **inconsistent** either when the fraction of sampled clusters tends to 0 or under conditions like constancy over k of N_k, N_k/n_k that allow certain large-sample formulas to simplify. The pseudo-loglikelihood methods are generally inconsistent even when sampling within clusters is noninformative.

(V) (*Lemma 12:*) Under regularity conditions and noninformative-within-cluster sampling, $\hat{\theta}^{(EM)}$ is consistent; and when also $n_k \equiv N_k$ for all k, $\hat{\theta}^{(L1)}$ is consistent.

4 Simulation Study

The remaining explorations in this paper consist of a simulation study with two objectives. The first is to illustrate empirically the theoretical findings summarized in the previous section for the setting where within-cluster sampling is assumed noninformative. The second goal is to confirm, through a systematic simulation design allowing sample-inclusion of unit (j, k) to depend explicitly on a_k or $\epsilon_{j,k}$ or both, that no available method depending on single-inclusion weights alone is consistent under general informative sampling. The general theoretical justification that this is so will be the subject of another paper.

4.1 Simulation Design

Our simulation design expands slightly on that of Korn and Graubard (2003). Sampling is done hierarchically in two stages: first, clusters k are sampled either SRS or by Poisson sampling, and in informative cases clusters with a_k or $|a_k|$ above a fixed threshold are sub-sampled independently with a fixed probability; second, units within sampled clusters are sampled either SRS or by Poisson sampling and in informative cases, units j with $\epsilon_{j,k}$ or $|\epsilon_{j,k}|$ exceeding a threshold are subsampled with fixed probability. In this way, the possible effects on estimator performance can be examined of unequal sampling weights, of large versus small clusters and cluster-sampling or within-cluster sampling fractions, and of informative sampling either at the whole-cluster or within-cluster level.

The formal simulation steps are as follows. First, the finite frame population is defined: a number M of population clusters is specified, with population clusters of size N_k that may be fixed or randomly selected within some bounded range; then within these clusters, the cluster random-effects a_k and unit 'observable' values $Y_{j,k}$ are specified ($1 \le k \le M,\ 1 \le j \le N_k$)) according to the model (19). In all of the simulations reported here, $M = 20,000$ and $\theta = (\mu, \sigma_a^2, \sigma_e^2) = (1, 2, 3)$. Next, two-stage noninformative cluster sampling is done, with fixed sampling fractions first at whole-cluster and then at within-cluster level. At both levels, single-inclusion weights are either fixed, in which case sampling is SRS, or are made to fall in an arithmetic progression of possible values, in which case sampling is Poisson. Beyond this point, following the approach of Korn and Graubard (2003), sampling of clusters is made informative in some cases by randomly subsampling with selection probability 1/2 those sampled clusters k for which a_k respectively falls outside the range $\pm\, 0.675 \cdot \sigma_a$ (symmetric subsampling criterion) or above $0.675 \cdot \sigma_a$ (asymmetric cluster subsampling criterion). (Here 0.675 is approximately the upper quartile of the standard normal distribution.) Similarly, in those cases where informative within-cluster sampling is chosen, sampled units within each sampled cluster are subsampled independently with probability 1/2 respectively if their unit-level errors $\epsilon_{j,k}$ fall outside the range $\pm\, 0.675 \cdot \sigma_e$ (symmetric subsampling criterion) or above $0.675 \cdot \sigma_e$ (asymmetric unit-within-cluster subsampling criterion). Recall that m and n_k

respectively denote the final numbers, after sampling and possible subsampling, of sampled clusters and of sampled units within cluster k.

While the simulation specification described in the previous paragraph allows a large number of cross-classified factorial parameter-combinations, we describe results for only 20 of them, grouped according to magnitude and uniformity of sizes of clusters, and of noninformative or informative sampling at the two (cluster and within-cluster) levels, by symmetric or asymmetric subsampling criteria. These choices are guided by the theoretical results of Section 3, with the objective of distinguishing the behavior of the different estimation methods studied. Within each parameter setting, a single superpopulation is newly generated along with $R = 5000$ samples drawn. The large number of replications within each generated superpopulation is intended to ensure that large-sample survey-estimators work reliably. Generation of 2 or more superpopulations for each parameter and design combination, not all shown, then confirmed that the large-sample limits are nearly constant across superpopulations. Some of the similar superpopulations displayed among Runs 1-20 in Table 1 differ only in allowing slightly varying random weights in place of constant (SRS) weights [the base-weights in the basic design, before possible informative subsampling]. This is how noninformative-sampling Runs 2, 4 respectively differ from Runs 1, 3, and how informative-sampling Runs 18-20 respectively differ from their constant base-weight counterparts 15-17.

The simulation results consist of means and standard deviations across Monte-Carlo replicated samples of estimates of parameter θ for each of the methods of estimation studied in this paper: Method of Moments (29), maximum-`psLik1` for within-cluster weights replaced by $w_{j|k} = 1$ (a method we denote `psLik0`), maximum-`psLik1` and maximum-`psLik2` as given in (30)–(32) for two different choices of γ_k, and pseudo-EM as defined in (34)–(35).

The various authors who have written on this topic have chosen a number of different, and sometimes idiosyncratic, approaches to simulating superpopulations and informative samples. Some of those informative designs are complicated enough that it is difficult to formulate intuitions about what can go wrong in survey-weighted procedures. We follow the fairly simple idea of Korn and Graubard (2003) for simulation of aggressively informative sampling, modifying that idea to make the designs factorial with respect to choices of numbers, sizes and uniformity of weights of population clusters; sizes and uniformity of weights for sampling within clusters; and a fairly restricted set of types of informative sampling either at cluster or within-cluster levels or both. The particular

Table 1: Characteristics of 20 simulation runs, all based on $M = 20,000$ superpopulation clusters. `CInf=T` denotes subsampling (at rate 0.5) for informative cluster sampling, `CIsym=T` if informative subsampling was symmetrically based on a_k. Within clusters, `WInf=T` indicates informative subsampling (rate 0.5), and `WIsym=T` if symmetrically based on $\epsilon_{j,k}$. Final 3 columns are population method of moments estimators, the targets for parameter estimates for each run.

Run	med(m)	med(N_k)	med(n_k)	CInf	WInf	CIsym	WIsym	μ	σ_a^2	σ_e^2
1	200	100	50	F	F	*	*	0.9787	2.0268	3.0015
2	200	100	50	F	F	*	*	1.0030	2.0118	3.0050
3	200	40	20	F	F	*	*	1.0056	1.9446	3.0018
4	200	40	20	F	F	*	*	1.0086	1.9666	2.9941
5	200	21	4.2	F	F	*	*	1.0022	1.9988	3.0091
6	200	24	4.8	F	F	*	*	0.9947	2.0004	2.9947
7	150	90	36	T	F	T	*	1.0034	1.9706	2.9931
8	150	90	36	T	F	F	*	0.9907	1.9947	3.0030
9	150	12	5	T	F	T	*	0.9987	1.9996	2.9924
10	150	12	5	T	F	F	*	1.0169	1.9869	3.0086
11	200	90	27	F	T	*	T	1.0028	1.9867	2.9989
12	200	90	27	F	T	*	F	0.9901	2.0107	2.9991
13	200	30	11.25	F	T	*	T	1.0035	1.9930	2.9979
14	200	36	13.5	F	T	*	F	1.0063	1.9965	2.9940
15	150	160	60	T	T	T	T	0.9934	1.9983	2.9999
16	150	80	24	T	T	T	T	0.9985	1.9908	2.9989
17	150	40	12	T	T	T	T	1.0107	2.0291	2.9935
18	150	160	60	T	T	T	T	1.0043	2.0075	3.0023
19	150	80	24	T	T	T	T	0.9816	2.0054	2.9954
20	150	46	13.8	T	T	T	T	0.9839	1.9808	3.0041

objectives here are to corroborate the theoretical results of Section 3.3 in the various cases where sampling is noninformative within clusters, and to assess whether any of these methods based only on single-inclusion weights performs adequately when sampling within clusters is informative.

4.2 Simulation Results

Twenty different simulation scenarios were investigated, each based on a newly generated super-population of $M = 20,000$ clusters, with $\theta = (\mu, \sigma_a^2, \sigma_e^2) = (1, 2, 3)$ in model (19), within each of which $R = 5000$ Monte Carlo samples were drawn according to the designs described in Section 4.1. Parameters characterizing the simulations, especially with respect to informative subsampling at or within cluster level, are displayed in Table 1. In addition, since the target parameters for each run reflect the super-population θ estimates for that run rather than the 'true' parameter values, the superpopulation method-of-moments estimates for $(\mu, \sigma_a^2, \sigma_e^2)$ are given in the final columns of Table 1. (The super-populations are large enough that the distinction between moment and ML estimates of θ can be ignored.) Precise description of the runs with non-constant sampling weights at or within cluster level (of which there are a few: Runs 2, 4, and 18-20, corresponding to otherwise identical Runs 1,3, and 15-17 with constant base-weights) is mostly omitted, since the design feature of varying weights turned out to be unimportant for consistency behavior of the estimators studied. (This is confirmed in the Run 1-2 columns of Table 2 and the Run 15-20 rows of Table 4.

Table 2: Monte Carlo averages of estimates over 5000 replicated samples within Simulation Runs 1, 2, 7 and 8 – those with large superpopulation and sampled clusters and noninformative sampling within clusters. Superpopulation target parameters are given in final row, labeled `Popn`. Entries with standardized discrepancies larger than 6 are in boldface.

	Run 1			Run 2			Run 7			Run 8		
Ests	μ	σ_a^2	σ_e^2	μ	σ_a^2	σ_e^2	μ	σ_a^2	σ_e^2	μ	σ_a^2	σ_e^2
`MMom`	0.977	2.015	3.001	1.003	1.994	3.005	1.004	**1.942**	2.992	0.987	1.978	3.004
`psL0`	0.977	2.015	3.001	1.004	1.997	3.006	1.003	1.951	2.992	0.988	1.981	3.003
`psL1`	0.977	**2.045**	**2.971**	1.004	**2.035**	**2.967**	1.003	**2.004**	**2.942**	0.988	**2.034**	**2.952**
`psL2`	0.977	2.016	**2.942**	1.004	2.005	**2.938**	1.003	1.971	**2.910**	0.988	2.000	**2.920**
`psEM`	0.977	2.015	3.001	0.985	**2.065**	3.005	1.004	1.951	2.992	0.988	1.981	3.003
`Popn`	0.978	2.027	3.001	1.003	2.012	3.005	1.003	1.971	2.993	0.991	1.995	3.003

Since finite-sample biases in MLEs do exist, especially for variance estimates, runs with as many as 5000 simulation iterations may appear to signal inconsistent estimates. In our present evaluations of Tables 2 to 4, we apply the rule of thumb that *standardized discrepancies*, defined equal to differences between estimates and population targets of less than 6 measured in units of Monte Carlo standard deviations divided by $\sqrt{5000}$, are acceptable. Thus, entries in the Tables with standardized discrepancies of more than 6 are bolded. Note also that the biases in σ_a^2 and σ_e^2 always have opposite signs.

When clusters are large and sampling is noninformative at all levels, all of the estimation methods are reasonably accurate. This behavior persists when sampling of clusters is informative but sampling within clusters is noninformative. The simulation results can be seen in Table 2, where Monte Carlo averages of estimates are shown for Runs 1, 2, 7 and 8. In this Table, and in all others with noninformative sampling within clusters, Method of Moments `MMom`, the pseudolikelihood estimator `psL0` replacing $w_{j|k}$ by 1, and pseudo-EM (`psEM`) are generally accurate within Monte Carlo sampling error, as theory suggests. The pseudolikelihood estimators `psL1` and `psL2` are also not far off in absolute terms, although their estimates of σ_e^2 are too low by more than 6 standard errors. This behavior becomes much more dramatic when the clusters are much smaller, as can be seen in Table 3. In other simulations related to that Table, numerical experience shows that the smallness of cluster sample sizes matters much more than the within-cluster sampling fractions.

Table 3: Monte Carlo averaged estimates over 5000 samples in Simulation Runs 3, 9, 10, with small clusters in superpopulation and sample, and noninformative within-cluster sampling. Superpopulation targets in `Popn` row. Entries with standardized discrepancies larger than 6 are bolded.

	Run 3			Run 9			Run 10		
Ests	μ	σ_a^2	σ_e^2	μ	σ_a^2	σ_e^2	μ	σ_a^2	σ_e^2
`MMom`	0.996	1.987	2.995	0.997	1.976	2.992	1.015	1.969	3.010
`psL0`	0.995	1.986	2.994	0.997	1.976	2.992	1.015	1.969	3.010
`psL1`	0.994	**2.531**	**2.471**	0.997	**2.356**	**2.611**	1.015	**2.352**	**2.627**
`psL2`	0.994	2.428	**2.373**	0.997	**2.154**	**2.415**	1.015	**2.148**	**2.429**
`psEM`	0.995	1.986	2.994	0.997	1.976	2.992	1.015	1.969	3.010
`Popn`	1.006	1.945	3.002	0.999	2.000	2.992	1.017	1.987	3.009

Table 4: Averaged estimates of σ_a^2 and Standardized Discrepancies (Z) over 5000 samples in Simulation Runs 11–20 informative within-cluster. Superpopulation targets given in Popn column.

Run	MMom	Z	psL0	Z	psL1	Z	psL2	Z	psEM	Z	Popn
11	2.834	21.14	1.975	-0.41	2.103	3.91	2.070	2.84	1.976	-0.38	1.987
12	2.211	6.40	1.998	-0.43	2.098	2.96	2.065	1.85	∞	*	2.011
13	2.837	21.03	1.982	-0.41	2.261	8.38	2.170	5.78	1.923	-2.57	1.993
14	2.190	6.25	1.979	-0.62	2.156	5.24	2.078	2.76	∞	*	1.997
15	2.829	20.76	1.974	-0.88	2.028	1.05	2.010	0.39	2.101	3.46	1.998
16	2.822	20.83	1.968	-0.81	2.113	4.09	2.076	2.91	1.979	-0.42	1.991
17	2.869	20.70	2.014	-0.52	2.301	8.35	2.231	6.40	1.969	-2.15	2.029
18	2.844	20.80	1.987	-0.74	2.042	1.19	2.023	0.55	2.105	3.28	2.008
19	2.836	20.71	1.981	-0.86	2.126	4.00	2.089	2.84	1.991	-0.50	2.005
20	2.815	20.96	1.965	-0.56	2.214	7.44	2.152	5.63	1.928	-1.93	1.981

These results are generally in line with the theoretical predictions of Section 3.3 for designs that are noninformative within clusters. When the simulation parameters allow informative sampling within clusters, the theory did not give a clear guide. Table 4 displays the estimates and standardized discrepancies for the estimates of σ_a^2 in simulation runs 11–20 (parameters of which were given above, in Table 1), together with their standardized discrepancies. Runs 11 and 13 respectively were cases of superpopulations with larger and smaller clusters where sampling within clusters was informative (subsampled according to the symmetric criterion $|\epsilon_{j,k}| > 0.675\,\sigma_e$) while sampling of clusters was noninformative. In these runs, psL0 was surprisingly accurate with psEM only slightly less so, and MMom was heavily biased; while psL1 and psL2 were only slightly upwardly biased in run 11 and more markedly so in run 13. In runs 12 and 14, which were like 11 and 13 except that informative within-cluster subsampling was based on the asymmetric criterion $\epsilon_{j,k} > 0$, method psL0 was quite good, psL1 and psL2 only slightly worse, and MMom awful. Here the psEM method broke down in the sense that its μ and σ_a^2 estimates converged to large (essentially infinite) values. The failure of the EM iterations in runs 12 and 14 was not an artifact based on the choice of starting values, but rather reflects the fact that the EM equations (34)–(35) do in some settings have fixed points with infinite parameter values to which the iterative EMsteps (33) converge. It is not clear what other informative designs might cause similar unpleasant behavior in pseudo-EM.

In the remaining runs covered in Table 4, sampling was informative (by a symmetric subsampling criterion) at both the cluster and within-cluster levels. In these superpopulations and sampling designs, `psL0` holds up as a surprisingly effective method, while `MMom` is again completely inappropriate. For unclear reasons, `psEM` fares clearly worse than `psL1` and `psL2` in the runs (15 and 18) with larger clusters, but better than `psL1` and `psL2` with medium and small clusters.

All of the unequal-weight simulations in Table 1 made use of very symmetric patterns of weights and either SRS or Poisson sampling before *iid* subsampling. This allows `psL0` to shine to undeserved advantage. However, since `psL0` ignores the within-cluster survey weights, completely noninformative sampling designs with strongly unequal within-cluster weights could be used to show why `psL0` is not generally actually an accurate estimation method. Since many classic survey methodology references have discussed the inadmissibility of estimates that ignore strongly unequal survey weights, we do not pursue that issue in the simulations reported here.

5 Conclusions

There are many different and general forms of informative sampling that go well beyond those studied in this paper, but those covered here are already enough to show that no existing method based only on single-inclusion weighting performs adequately in general informative settings. The method `psL0` ignoring within-cluster weights is the single best performer in all of the simulations provided here, since it like `psEM` provides consistent estimators whenever sampling is noninformative within clusters, but for obscure reasons `psL0` continued with near-consistent results in all of the within-cluster-informative simulation runs (11–20) of Table 1. Perhaps realistic informative sampling is in some ways less drastic than the artificial designs simulated here, but it seems unwise to rely on methods like `psL0` guaranteed to work well only when there is noninformative sampling within clusters. The pseudo-EM method advanced in this paper is also not adequate, although it seems to outperform the other methods under the symmetric informative designs studied. On the other hand, when the conditional expectations in the EM are calculated subject to noninformative-within-cluster assumptions, the pseudo-EM iterates can converge to fixed-points with infinite parameter values in the two-level ANOVA model. It seems advisable to develop methods in which the informative-within-cluster selection mechanism is modeled. Even a simplistic and misspecified model may avoid the bad behavior seen here, and may be adequate in many realistic settings with informative sampling. Kim et al. (2017) and Savitsky and Williams (2018) in different ways already considered ways in which informative-missing models might be incorporated into their estimation procedures, and doing the pseudo-EM conditional expectations subject to simple parametric models for selection bias is also a topic for further research.

Unpublished research by the author rules out the possibility that a method of estimation depending only on single-inclusion weights could be model- and design- consistent in general superpopulation two-level models in which the superpopulation data as well as the sampling mechanism are independent across clusters. The evidence from the two-level ANOVA in this paper is that none of the methods simulated can achieve this consistency, since the best-performing method `psL0` can easily be made to fail with unequal-weight within-cluster sample designs.

The variances and efficiency of survey estimates for superpopulation models with mixed-effect cluster-level random effects have not been studied at all in this paper. Other researchers have considered variance behavior of survey estimates in that setting, but the theme of this paper has been that all estimators based on single-inclusion weights are generally biased under informative sampling, so that large-sample variances can be understood usefully only under somewhat restrictive assumptions on the informative-sampling mechanism.

References

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

Binder, D. (1983), On the variance of asymptotically normal estimators from complex surveys, *International Statistical Review* **51**, 279-292.

Boistard, H., Lopuhaä, H. P. and Ruiz-Gazen, A. (2017), Functional central limit theorems for single-stage sampling designs, *Annals of Statistics* **45**, 17281758.

Fuller, W. (2009), *Sampling Statistics*, Wiley.

Graubard, B. and Korn, E. (2011), Conditional logistic regression with survey data, *Statistics in Biopharmaceutical Research* **3**, 398-408.

Kim, J.-K., Park, S. and Lee, Y. (2017), Statistical inference using generalized linear mixed models under informative cluster sampling, *Canadian Journal of Statistics* **45**, 479-497.

Korn, E. and Graubard, B. (2003), Estimating variance components by using survey data, *Journal of the Royal Statistical Society* Ser. B **65**, 175-190.

Lindsay, B (1988) Composite likelihood methods, *Contemporary Math.* **80**, 220-239.

Lohr, S. (2009) *Sampling: Design and Analysis*, Brooks-Cole.

Molina, I. and Rao, JNK (2015), *Small Area Estimation*, 2nd ed., Wiley.

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

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

Rao, JNK, Verret, F. and Hidiroglou, M. (2013), A weighted estimating equations approach to inference for two-level models from survey data. *Survey Methodology* **39**, 263-282.

Rubin-Bleuer, S. and Kratina, I. (2006), On the two-phase framework for joint model and design-based inference, *Annals of Statistics* **33**, 2789-2810.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

Savitsky, T. and Williams, M. (2019), Bayesian mixed models under informative sampling, `arXiv:1904.07680`

Savitsky, T. and Williams, M. (2020), Pseudo-Bayesian Estimation of One-way ANOVA Model in Complex Surveys, preprint.

Van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge Univ. Press.

Williams, M. and Savitsky, T. (2018) Bayesian estimation under informative sampling with unattenuated dependence, *Bayesian Analysis*, advance publication, `doi:10.1214/18-BA1143`

Yi, G., Rao, JNK, and Li, H. (2016) A weighted composite likelihood approach for analysis of survey data under two-level models, *Statistica Sinica* **26**, 569-587.

A Technical Lemmas

A.1 Conditional Distributions and Normal Integral Formulas

Derivations related to the normal two-level ANOVA model (19), and to algebraic manipulations of the pseudo-loglikelihoods and associated estimators, are collected in this Appendix.

Lemma 1 *Let $\gamma > 0$ and q be a positive integer, let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q) \in \mathbb{R}^q$ have all positive entries, and define the entries of a $q \times q$ matrix G for $j, l = 1, \ldots, q$ by*

$$G_{j,l} \;=\; \frac{\sigma_e^2}{\gamma\,\tau_j}\, I_{[j=l]} \,+\, \frac{\sigma_a^2}{\gamma}$$

Then G is positive-definite with inverse G^{-1} defined by the entries

$$(G^{-1})_{j,l} \;=\; \frac{\gamma\,\tau_j}{\sigma_e^2}\, I_{[j=l]} \,-\, \frac{\gamma\,\sigma_a^2\,\tau_j\,\tau_l}{\sigma_e^2\,(\sigma_e^2 + \|\boldsymbol{\tau}\|_1\,\sigma_a^2)} \tag{36}$$

where $\|\boldsymbol{\tau}\|_1 \equiv \sum_{j=1}^q \tau_j$.

Proof. The matrix G is the sum of a positive-definite diagonal matrix and a nonnegative-definite rank-1 matrix, both symmetric. So G is invertible, and an easy calculation shows that G right-multiplied by the matrix defined on the right-hand side of (36) is the $q \times q$ identity matrix. $\square$

Lemma 2 *Let $\gamma > 0$, q be a positive integer, $A \sim \mathcal{N}(0,\, \sigma_a^2/\gamma)$ be a random variable, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q) \in \mathbb{R}^q$ be a vector with all positive entries, and V_j for $j = 1, \ldots, q$ be random variables such that $V_j - A \sim \mathcal{N}(0,\, \sigma_e^2/(\gamma\, \tau_j))$ are jointly independent of one another and of A. Then $\underline{V} \equiv (V_j : j = 1, \ldots, q) \sim \mathcal{N}(\mathbf{0},\, G)$, where G is the same as in Lemma 1. Moreover,*

$$\textit{conditional density of} \quad A \quad \textit{given } \underline{V} \qquad \textit{is} \;\; \mathcal{N}\Big(\sum_{j=1}^{q} \frac{\tau_j\, \sigma_a^2\, V_j}{\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2} \,,\, \frac{\sigma_a^2\, \sigma_e^2}{\gamma\, (\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)} \Big) \qquad (37)$$

Proof. First, $\underline{V}$ is evidently a multivariate-normal random-vector, the sum of independent multivariate-normal vectors $(V_1 - A, \ldots, V_q - A)$ and $A\,\mathbf{1}$, and $\underline{V}$ therefore has mean $\mathbf{0}$ and covariance-matrix G, where $\mathbf{1} \in \mathbb{R}^q$ is the vector with all entries 1. Next, $\underline{V}$ and A are jointly multivariate-normal, so that A is conditionally normal given $\underline{V}$ and that $A^* = A - \sum_{l=1}^{q} \tau_l\, \sigma_a^2\, V_l / (\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)$ is independent of $\underline{V}$ once it is verified that $\text{cov}(A^*, V_j) = 0$ for each $j = 1, \ldots, q$. But this is an easy calculation:

$$\text{cov}(A, V_j) - \sum_{l=1}^{q} \text{cov}\Big(\frac{\tau_l\, \sigma_a^2\, V_l}{\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2}\,, V_j\Big) = \text{var}(A) - \sum_{l=1}^{q} \frac{\tau_l\, \sigma_a^2\, G_{j.l}}{\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2} \;=\; \frac{\sigma_a^2}{\gamma} - \frac{(\sigma_a^2\, (\sigma_e^2 + \|\boldsymbol{\tau}\|_1 \sigma_a^2)}{\gamma\, (\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)} = 0$$

Since A^* is independent of $\underline{V}$, we find $E(A\,|\,\underline{V})$ as given in (37), and the conditional variance is calculated as $\text{var}(A\,|\,\underline{V}) = \text{var}(A^*\,|\,\underline{V}) = \text{var}(A^*)$, with

$$\text{var}(A^*) \;=\; \text{var}(A) - \text{var}\Big(\sum_{j=1}^{q} \frac{\tau_j\, \sigma_a^2\, V_j}{\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2} \Big) \;=\; \frac{\sigma_a^2}{\gamma} \;-\; \frac{\sigma_a^4}{(\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)^2}\, \boldsymbol{\tau}'\, G\, \boldsymbol{\tau}$$

$$=\; \frac{\sigma_a^2}{\gamma} \Big[1 \;-\; \frac{\sigma_a^2}{(\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)^2} \Big(\sigma_e^2\, \|\boldsymbol{\tau}\|_1 \;+\; \sigma_a^2\, \|\boldsymbol{\tau}\|_1^2 \Big) \Big] \;=\; \frac{\sigma_a^2\, \sigma_e^2}{\gamma\, (\sigma_e^2 + \|\boldsymbol{\tau}\|_1\, \sigma_a^2)} \qquad \square$$

The next Lemma applies the previous calculations with fixed $k \in \{1, \ldots, M\}$, $q \equiv n_k$, $\boldsymbol{\tau}^{(k)} \equiv \boldsymbol{\tau} = (w_{j|k},\;\; j = 1, \ldots, n_k)$, where without loss of generality the elements of $\mathcal{U}_k$ are re-ordered so that the first $n_k = |\mathcal{S}_k|$ elements are those of $\mathcal{S}_k$. With those choices, note that $\|\boldsymbol{\tau}^{(k)}\|_1 \;=\; \hat{N}_k$. Next, γ is replaced by γ_k chosen equal either to 1 or ω_k respectively in considering $\textbf{pslogLik}_1$ and

$\mathbf{pslogLik}_2$ from Section 1.4. The matrix G and its inverse in Lemma 1 become

$$G^{(k)} \;=\; \frac{\sigma_e^2}{\gamma_k}\,\mathrm{Diag}(1/\boldsymbol{\tau}^{(k)}) \;+\; \frac{\sigma_a^2}{\gamma_k}\;, \quad G^{(k)\,-1} \;=\; \frac{\gamma_k}{\sigma_e^2}\Big\{\mathrm{Diag}(\boldsymbol{\tau}^{(k)}) \;-\; \frac{\sigma_a^2}{\sigma_e^2 + \hat{N}_k\,\sigma_a^2}\,(\boldsymbol{\tau}^{(k)})\,(\boldsymbol{\tau}^{(k)})'\Big\} \tag{38}$$

where $\mathrm{Diag}(\mathbf{v})$ denotes the diagonal matrix with vector $\mathbf{v}$ along the diagonal, and $1/\mathbf{v}$ denotes the vector of reciprocals of the components of vector $\mathbf{v}$. Note also that in the notations of Lemma 2, $A = a_k$ and $V_j = Y_{j,k} - \mu$.

Lemma 3 *Let $\gamma_k > 0$ for all $k = 1, \ldots, M$, and let n_k and $w_{j|k}$ be as in Section 1.1. Then for each k,*

$$\int_{-\infty}^{\infty} \exp\Big(-\,\frac{\gamma_k}{2\,\sigma_e^2}\sum_{j\in\mathcal{S}_k} w_{j|k}\,(Y_{j,k}-\mu-a_k)^2 \;-\; \frac{\gamma_k\,a_k^2}{2\,\sigma_a^2}\Big)\,da_k \;= \tag{39}$$

$$\sqrt{2\pi}\,\Big(\frac{\sigma_a^2\,\sigma_e^2\,/\,\gamma_k}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\Big)^{1/2}\;\exp\Big(-\,\frac{\gamma_k}{2}\,\Big[\frac{1}{\sigma_e^2}\sum_{j\in\mathcal{S}_k} w_{j|k}\,(Y_{j,k}-\bar{Y}^w_{\cdot k})^2 + \frac{\hat{N}_k}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\,(\bar{Y}^w_{\cdot k}-\mu)^2\Big]\Big)$$

where $\bar{Y}^w_{\cdot k} \;\equiv\; (\hat{N}_k)^{-1}\sum_{j\in\mathcal{S}_k} w_{j|k}Y_{j,k}$ and $\hat{N}_k = \sum_{j\in\mathcal{S}_k} w_{j|k}$ is as defined in (5).

Proof. First, by Lemma 2, the logarithm of the exponent of the joint density of $\underline{V}$ and A (evaluated at $\underline{V}, A$) is equal to

$$-\,\frac{\gamma\,A^2}{2\,\sigma_a^2} \;-\;\sum_{j=1}^{q}\frac{\gamma\,\tau_j}{2\,\sigma_e^2}\,(V_j-A)^2 \;=\; -\,\frac{1}{2}\,\underline{V}'\,(G^{(k)})^{-1}\,\underline{V} \;-\; \frac{\gamma\,(\sigma_e^2+\|\boldsymbol{\tau}\|_1\,\sigma_a^2)}{2\,\sigma_a^2\,\sigma_e^2}\,\Big(A-\sum_{j=1}^{q}\frac{\tau_j\,\sigma_a^2\,V_j}{\sigma_e^2+\|\boldsymbol{\tau}\|_1\,\sigma_a^2}\Big)^2$$

Now substituting $V_j = Y_{j,k} - \mu$ and $a_k = A$, in addition to the other substitutions above (38), we find immediately that the integral in the first line of (39) is equal to

$$\sqrt{2\pi}\,\Big(\frac{\sigma_a^2\,\sigma_e^2\,/\,\gamma_k}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\Big)^{1/2}\;\exp\Big(-\,\frac{1}{2}\,\underline{V}'\,(G^{(k)})^{-1}\,\underline{V}\Big) \tag{40}$$

Moreover, the exponent in the last expression, after substitution of (38), becomes

$$-\,\frac{1}{2}\sum_{j,l\in\mathcal{S}_k}(Y_{j,k}-\mu)\,(Y_{l,k}-\mu)\,\frac{\gamma_k}{\sigma_e^2}\Big\{w_{j|k}\,I_{[j=l]} \;-\; \frac{\sigma_a^2}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\,w_{j|k}\,w_{l,k}\Big\}$$

$$=\; -\,\frac{\gamma_k}{2\,\sigma_e^2}\,\Big[\sum_{j\in\mathcal{S}_k} w_{j|k}(Y_{j,k}-\mu)^2 \;-\; \frac{\sigma_a^2}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\Big(\sum_{j\in\mathcal{S}_k} w_{j|k}\,(\bar{Y}^w_{\cdot k}-\mu)\Big)^2\Big]$$

After expanding the square in the first square-bracketed summation and subtracting and adding $\bar{Y}^w_{\cdot k}$ inside $(Y_{j,k}-\mu)^2$, the last displayed formula becomes equal to

$$-\,\frac{\gamma_k}{2\,\sigma_e^2}\,\Big[\sum_{j\in\mathcal{S}_k} w_{j|k}(Y_{j,k}-\bar{Y}^w_{\cdot k})^2 \;+\; \hat{N}_k\,(\bar{Y}^w_{\cdot k}-\mu)^2 \;-\; \frac{\sigma_a^2\,\hat{N}_k^2}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\,(\bar{Y}^w_{\cdot k}-\mu)^2\Big]$$

Thus in (40), it has been shown that

$$-\frac{1}{2}\,\underline{V}'\,(G^{(k)})^{-1}\,\underline{V} \;=\; -\frac{\gamma_k}{2}\,\Big[\frac{1}{\sigma_e^2}\sum_{j\in\mathcal{S}_k} w_{j|k}(Y_{j,k}-\bar{Y}^w_{\cdot k})^2 \;+\; \frac{\hat{N}_k}{\sigma_e^2+\hat{N}_k\,\sigma_a^2}\,(\bar{Y}^w_{\cdot k}-\mu)^2\Big]$$

and the integral expression in the first line of (39) has been proved equal to the second line. □

Lemma 3 will be applied twice, to simplify expressions in the pseudo-loglikelihoods **pslogLik**$_1$ and **pslogLik**$_2$ that will be used to derive corresponding estimators in the ANOVA model. To help reduce the length of expressions, define (as in (25) in the main text),

$$\mathrm{SSW}_k \;=\; \sum_{j\in\mathcal{S}_k} w_{j|k}\,(Y_{j,k}-\bar{Y}^w_{\cdot k})^2$$

A.2 Equivalence of Moment and EM Estimators

In order to connect the pseudo-EM and Method of Moments estimators, begin with general notations. Recalling the definition of SSW_k from the last line of Section A.1 or (25) in Section 3, now define *within* and *between* weighted variance expressions, as follows:

$$\hat{V}_W = \frac{1}{\hat{N}}\sum_{(j,k)\in\mathcal{S}} w_{j,k}\,(Y_{j,k}-\bar{Y}^w_{\cdot k})^2 = \sum_{k\in\mathcal{S}_C}\frac{\omega_k}{\hat{N}}\,\mathrm{SSW}_k\;, \quad \hat{V}_B = \frac{1}{\hat{N}}\sum_{k\in\mathcal{S}_C}\omega_k\,\hat{N}_k\,(\bar{Y}^w_{\cdot k}-\tilde{\mu}^{(M)})^2 \qquad (41)$$

The equivalence result developed in this section holds exactly only in the case of identical-sized clusters and constant within-cluster weights. In this result, no restriction is placed on the informativeness of the sampling design.

Lemma 4 *In the two-level ANOVA model setting of Section 3.1, assume in addition that*

$$\textit{for all}\quad k=1,\ldots,M:\quad n_k\equiv\nu\;,\quad N_k\equiv\frac{N}{M}\;,\quad \textit{and for all}\;\; j=1,\ldots,N_k,\quad w_{j|k}\equiv\frac{N_k}{n_k}$$

with $\nu>1$. *Then the estimators* $\tilde{\theta}^{(M)}$ *defined in (29) and the pseudo-EM estimators* $\hat{\theta}^{(EM)}$ *defined by (34)–(35) are algebraically identical.*

Proof. First, note under the assumed constancy of cluster-size and within-cluster weights, respectively that $\hat{q}_k\equiv\nu\,(\hat{\sigma}_a^{(EM)})^2/\big[(\sigma_a^{(EM)})^2+\nu\,(\hat{\sigma}_a^{(EM)})^2\big]\equiv\hat{q}$ does not vary with k, and that $\bar{Y}_{\cdot k}\equiv\bar{Y}^w_{\cdot k}$ for all k. Then (34) immediately yields

$$\hat{\mu}^{(EM)} \;=\; \frac{\sum_{k\in\mathcal{S}_C}\omega_k\,\hat{N}_k\,(\bar{Y}^w_{\cdot k}-\hat{q}\bar{Y}_{\cdot k})}{\sum_{k\in\mathcal{S}_C}\omega_k\,\hat{N}_k\,(1-\hat{q})} \;=\; \frac{\sum_{k\in\mathcal{S}_C}\omega_k\,\hat{N}_k\,(1-\hat{q})\,\bar{Y}^w_{\cdot k}}{\sum_{k\in\mathcal{S}_C}\omega_k\,\hat{N}_k\,(1-\hat{q})} \;=\; \frac{\sum_{(j,k)\in\mathcal{S}}w_{j,k}\,Y_{j,k}}{\hat{N}} \;=\; \tilde{\mu}^{(M)}$$

Next, note that $\hat{N}_k = N_k$ does not vary with k, and therefore $\hat{N} = \sum_{k \in \mathcal{S}_C} \omega_k \, N_k = N\hat{M}/M$, and check directly from (29) that

$$\hat{V}_W \;=\; \frac{\nu - 1}{\nu \, \hat{M}} \sum_{k \in \mathcal{S}_C} \omega_k \, \mathrm{var}(\{Y_{j,k} : \; j \in \mathcal{S}_k\}) \;=\; \frac{\nu - 1}{\nu} \, (\tilde{\sigma}_e^{(M)})^2$$

where $\mathrm{var}(\cdot)$ denotes sample variance. Formula (34) then shows that

$$(\hat{\sigma}_a^{(EM)})^2 \;=\; \frac{\hat{q}}{\hat{M}} \sum_{k \in \mathcal{S}_C} \omega_k \, (\bar{Y}_{\cdot k} - \hat{\mu}^{(M)})^2 \;=\; \hat{q} \, \hat{V}_B \tag{42}$$

while formula (35), followed by substitution of (42), gives

$$(\hat{\sigma}_e^{(EM)})^2 \;=\; \hat{V}_W \;+\; (1 - \hat{q})^2 \, \hat{V}_B \;+\; (\hat{\sigma}_a^{(EM)})^2 \, (1 - \hat{q}) \;=\; \hat{V}_W \;+\; (1 - \hat{q}) \, \hat{V}_B \tag{43}$$

Now using (42) again,

$$\hat{q} \;=\; \frac{(\hat{\sigma}_a^{(EM)})^2}{(\hat{\sigma}_a^{(EM)})^2 + (\hat{\sigma}_e^{(EM)})^2/\nu} \quad \Longrightarrow \quad (1 - \hat{q}) \, \hat{V}_B \;=\; (1 - \hat{q}) \Big((\hat{\sigma}_a^{(EM)})^2 + (\hat{\sigma}_e^{(EM)})^2/\nu \Big) \;=\; \frac{(\hat{\sigma}_e^{(EM)})^2}{\nu}$$

and substituting this last expression into (43) shows

$$(\hat{\sigma}_e^{(EM)})^2 \;=\; \hat{V}_W \;+\; (\hat{\sigma}_e^{(EM)})^2/\nu \quad \Longrightarrow \quad (\hat{\sigma}_e^{(EM)})^2 \;=\; \frac{\nu}{\nu - 1} \, \hat{V}_W \;=\; (\tilde{\sigma}_e^{(M)})^2$$

Finally, the equality $(\hat{\sigma}_a^{(EM)})^2 = (\tilde{\sigma}_a^{(M)})^2$ follows from the observation that

$$(\tilde{\sigma}_a^{(M)})^2 + (\tilde{\sigma}_e^{(M)})^2 \;=\; \frac{1}{\hat{N}} \sum_{(j,k) \in \mathcal{S}} w_{j,k} \, (Y_{j,k} - \tilde{\mu}^{(M)})^2 \;=\; \hat{V}_W \;+\; \hat{V}_B$$

together with formulas (42)–(43) showing $(\hat{\sigma}_a^{(EM)})^2 \;=\; \hat{q} \, \hat{V}_B \;=\; \hat{V}_W \;+\; \hat{V}_B \;-\; (\hat{\sigma}_e^{(EM)})^2$. □

A.3 Limits of Estimators in Large ANOVA-model Samples

The results in this section provide large-sample limits for large N of estimators within the very special setting of Sections 1.1 and 3.1. That is, the two-level ANOVA superpopulation model (19) with unknown parameter $\theta = (\mu, \sigma_a^2, \sigma_e^2)$ is assumed, where the *iid* sequence $\{a_k : \; k = 1, \ldots, M\}$ is independent of the *iid* array $\{\epsilon_{j,k} : \; (j,k) \in \mathcal{U}\}$, and conditionally given these superpopulation variables, the sampling design may be informative, but for each k, $(\omega_k, \, I_{[k \in \mathcal{S}_C]}, \, w_{j|k}, \, I_{[j \in \mathcal{S}_k]})$ may depend on a_k and $\{Y_{j,k}\}_{j=1}^{N_k}$. The case where within-cluster sampling is informative will be considered separately from the noninformative case (11).

Throughout the Section, a few more technical assumptions on the sample designs and the range of allowed sampling weights are needed, to simplify the proofs of laws of large numbers. Let m, n, n_k be (possibly random) numbers of sampled clusters and units within cluster (assumed to be defined for all clusters, whether sampled or not) as defined in (7). All limits are taken as population size $N \to \infty$, and the conditions imposed on (possibly random) weights will be seen to ensure that the overall numbers of sampled clusters and units tend to infinity (in design probability).

(C1). The sample design is such that

(i) $(\omega_k,\, I_{[k\in\mathcal{S}_C]}, \mathcal{S}_k = s_k)$ are independent across $k \in \{1,\ldots,M\}$, and may depend on a_k,

(ii) For each $k = 1,\ldots,M$, given $(a_k,\, \omega_k,\, I_{[k\in\mathcal{S}_C]}, \mathcal{S}_k = s_k)$, $Y_{j,k} \sim \mathcal{N}(\mu + a_k,\, \sigma_e^2)$ for $j \in s_k$.

(C2). There are positive constants $m^{(0)}$, $n_k^{(0)}$, K_0, K_1, K_2 such that as $N \to \infty$,

(o) $E(m) \to \infty$,

(i) $\max_{k=1,\ldots,M} N_k \le K_3$,

(ii) for all $k = 1,\ldots,M$, $1/K_2 \le \omega_k\, m^{(0)}/M \le K_2$, with probability 1, and

(iii) for all $(j,k) \in \mathcal{U}$, $1/K_1 \le w_{j|k}\, n_k^{(0)}/N_k \le K_1$, with probability 1.

With these assumptions in place, the distinction between designs informative versus noninformative within cluster is simply that $w_{j|k}$ is allowed to depend on $Y_{j,k}$ (as well as a_k) in the informative case, while $(w_{j|k},\ j = 1,\ldots,N_k)$ is assumed independent of $\{Y_{l,k} :\ l = 1,\ldots,N_k\}$ in the noninformative case.

The next two Lemmas provide basic laws of large numbers under which these Poisson-sampled cluster designs lead to growing samples with design-consistent Horvitz-Thompson estimators.

Lemma 5 *Assume the two-level ANOVA superpopulation model (19) and conditions* **(C1)**–**(C2)** *on the sampling design. Then as* $N \to \infty$,

(a) $m^{(0)} \to \infty$ *and* $n^{(0)} = \sum_{k=1}^{M} n_k^{(0)} \to \infty$, *and* $m, n \to \infty$ *in design probability,*

(b) $\hat{M}/M \to 1$ *and* $\hat{N}/N \to 1$ *in design probability, and*

(c) for $r = 1, 2$, $\hat{N}^{-1}\sum_{(j,k)\in\mathcal{S}} w_{j,k}\, a_k^r \to \sigma_a^2\, I_{[r=2]}$ *and* $\hat{N}^{-1}\sum_{(j,k)\in\mathcal{S}} w_{j,k}\, \epsilon_{j,k}^r \to \sigma_e^2\, I_{[r=2]}$ *in design and model probability.*

Proof. First, in (a),

$$m^{(0)} \;=\; \frac{m^{(0)}}{M}\sum_{k=1}^{M} E(\omega_k\, I_{[k\in\mathcal{S}_C]}) \;=\; E(\sum_{k\in\mathcal{S}_C} \frac{\omega_k\, m^{(0)}}{M})$$

and then (C2)(ii) immediately implies that $E(m)/m^{(0)}$ differs from 1 by at most a factor K_2, implying that $n^{(0)} \ge m^{(0)} \to \infty$ by (C1)(o). Similarly, by (C2)(iii),

$$n^{(0)} \;=\; \sum_{k=1}^{M} \frac{n_k^{(0)}}{N_k} \sum_{j=1}^{N_k} E(w_{j|k} I_{[k\in\mathcal{S}_k]}) \;=\; \sum_{k=1}^{M} E(\sum_{j\in\mathcal{S}_k} \frac{w_{j|k}\, n_k^{(0)}}{N_k})$$

differs from $E(n) \;=\; \sum_{k=1}^{M} E(n_k)$ by a factor at most K_1, as does $n_k^{(0)}$ fron $E(n_k)$. Thus $E(m)$, $E(n) \to \infty$, and both n, m converge to ∞ in design probability.

Assertions (b) concern Horvitz-Thompson estimators. The average $\hat{M}/M \;=\; M^{-1}\sum_{k=1}^{M} \omega_k\, I_{[k\in\mathcal{S}_C]}$ of independent variables has mean 1, since $P(k \in \mathcal{S}_C \,|\, \omega_k) = 1/\omega_k$. Also, the variance of $\hat{M}/M$ is

$$M^{-2}\sum_{k=1}^{M} \text{Var}(\omega_k\, I_{[k\in\mathcal{S}_C]}) \;\le\; M^{-2}\sum_{k=1}^{M} E\big(\omega_k^2\, E(I_{[k\in\mathcal{S}_C]} \,|\, \omega_k)\big) \;=\; M^{-2}\sum_{k=1}^{M} \frac{M}{m^{(0)}}\, E(\frac{\omega_k\, m^{(0)}}{M}) \;\le\; \frac{K_1}{m^{(0)}}$$

which converges to 0 by part (a). Thus the weak law of large numbers $\hat{M}/M \;\to\; 1$ holds by Chebychev's inequality. The proof that $\hat{N}/N \to 1$ is completely analogous and is omitted.

By (b), it suffices to prove (c) with denominator N replacing $\hat{N}$. By assumption, the variables

$$B_k \;=\; \sum_{j=1}^{N_k} I_{[(j,k)\in\mathcal{S}]}\, w_{j,k}\, a_k^r \;=\; I_{[k\in\mathcal{S}_C]}\, \omega_k \sum_{j=1}^{N_k} w_{j|k}\, a_k^r \;=\; I_{[k\in\mathcal{S}_C]}\, \omega_k\, \hat{N}_k\, a_k^r$$

are independent across $k = 1,\ldots,M$. Since $P((j,k)\in\mathcal{S} \,|\, \omega_k,\, a_k\, \underline{Y}_k) \;=\; 1/w_{j,k}$, for $r = 1, 2$,

$$E(B_k) \;=\; N_k\, E(a_k^r) \;=\; N_k\, \sigma_a^2\, I_{[r=2]}\;, \qquad \text{Var}(N^{-1}\sum_{k=1}^{M} B_k) \;\le\; N^{-2}\sum_{k=1}^{M} E(I_{[k\in\mathcal{S}_C]}\, \omega_k^2 \hat{N}_k^2\, a_k^{2r})$$

Therefore, $E(N^{-1}\sum_{k=1}^{M} B_k) \;=\; \sigma_a^2\, I_{[r=2]}$, and by (C.2)

$$\text{Var}(N^{-1}\sum_{k=1}^{M} B_k) \;\le\; N^{-2}\, K_3^2 \sum_{k=1}^{M} E(\omega_k\, a_k^{2r}) \;\le\; N^{-2}\, K_3^2\, K_1, (M/m^{(0)} \sum_{k=1}^{M} E(a_k^{2r})$$

is upper-bounded by a constant divided by $m^{(0)}$, where $m^{(0)} \to \infty$ by part (a). The first convergence in design probability in (c) follows by Chebychev's inequality. The second convergence assertion concerns the sum of the independent variables

$$B_k^* \;=\; \sum_{j=1}^{N_k} I_{[(j,k)\in\mathcal{S}]}\, w_{j,k}\, \epsilon_{j,k}^r \;=\; \omega_k\, I_{[k\in\mathcal{S}_C]} \sum_{j=1}^{N_k} I_{[j\in\mathcal{S}_k]}\, w_{j|k}\, \epsilon_{j,k}^r$$

which for $r = 1, 2$ have expectations $N_k \, I_{[r=2]} \, \sigma_e^2$. By similar reasoning to the first part of (c), $E(N^{-1} \sum_{k=1}^M B_k^*) \; = \; I_{[r=2]} \, \sigma_e^2$, and

$$\mathrm{Var}(N^{-1} \sum_{k=1}^M B_k^*) \; \le \; N^{-2} \sum_{k=1}^M \sum_{j=1}^{N_k} E\big(\omega_k \, \max_j w_{j|k}^2 \, \epsilon^{2r}\big)$$

$$\le \; \frac{M^2}{N^2 \, m^{(0)}} \, K_1^2 \, K_2 \, E\Big(\frac{1}{M} \sum_{k=1}^M \frac{N_k^3}{(n_k^{(0)})^2} \, \epsilon_{j,k}^{2r}\Big) \; \le \; C/m^{(0)}$$

for a constant C, and the Chebychev inequality again implies the result. □

Other Horvitz-Thompson estimators based on the weights $w_{j,k}$ will also be design-consistent under the same assumptions, whether or not sampling is informative. An immediate corollary can be drawn concerning partial consistency of $\tilde{\theta}$.

Lemma 6 *Under the same assumptions as Lemma 5, the estimators* $\tilde{\mu}^{(M)}$ *and* $(\tilde{\sigma}_a^{(M)})^2 + (\tilde{\sigma}_e^{(M)})^2$ *are respectively design and model consistent for* μ *and* $\sigma_a^2 + \sigma_e^2$*.*

Proof. By definition of the estimators $\tilde{\theta}^{(M)}$ in (29), under model (19),

$$\tilde{\mu}^{(M)} - \mu \; = \; \hat{N}^{-1} \sum_{(j,k) \in \mathcal{S}} w_{j,k} \, (a_k + \epsilon_{j,k}) \; \to \; 0$$

according to Lemma 5(c). An analogous argument based on Chebychev's inequality shows that

$$(\tilde{\sigma}_a^{(M)})^2 + (\tilde{\sigma}_e^{(M)})^2 \; \equiv \; \hat{N}^{-1} \sum_{(j,k) \in \mathcal{S}} w_{j,k} \, (a_k + \epsilon_{j,k})^2 \; \to \; \sigma_a^2 \, + \, \sigma_e^2$$

since the right-hand side is the expectation of the left-hand side, and the variance is upper-bounded by a constant divided by $m^{(0)}$. □

A.3.1 Noninformative Within-Cluster Sample Designs

The requirement (11) that within-cluster sampling be noninformative implies conditional independence of $\{Y_{j,k} \, : \;\; j \in \mathcal{S}_k\}$ and $\mathcal{S}_k$ given a_k, $[k \in \mathcal{S}_C]$. The sampling design variables $(\omega_k$, $I_{[k \in \mathcal{S}_C]}$, $\mathcal{S}_k = s_k)$ are independent across clusters $k \in \{1, \ldots, M\}$ (although they may depend on a_k), and given these variables, $\{Y_{j,k} \, : \;\; j \in s_k\}$ is an *iid* $\mathcal{N}(\mu + a_k, \, \sigma_e^2)$ distributed sequence of random variables. Thus the assumptions will be as before, with the addition of

(C3). (i) $(\omega_k, \, \mathcal{S}_k)$ are independent across $k \in \{1, \ldots, M\}$, but may depend on a_k,

(ii) $(w_{j|k},\ j=1,\ldots,N_k)$ do not depend on variables $(a_k,\ \underline{Y}_k)$, and

(iii) given a_k and the variables in (i) above, with $\mathcal{S}_k = s_k$, $\{Y_{j,k} :\ j \in s_k\}$ is an *iid* $\mathcal{N}(\mu+a_k,\ \sigma_e^2)$ distributed sequence of random variables.

Analogous to Lemma 5, the next Lemmas establish law-of-large-numbers limits under assumption (C3) of noninformative sampling within clusters. All convergences $\xrightarrow{P}$ are in model and design probability.

Lemma 7 *Assume the two-level ANOVA superpopulation model (19) and conditions* **(C1)**–**(C3)** *on the sampling design. Let* $g : \mathbb{R}^4 \to \mathbb{R}$ *be a continuous function such that the family of random variables* $g(\hat{N}_k,\ n_k,\ \bar{Y}^w_{\cdot k},\ \bar{Y}_{\cdot k})$ *have second moments uniformly bounded with respect to* $k = 1,\ldots,M$*. By the assumptions, these variables are also independent. Then as* $N \to \infty$,

$$N^{-1} \textstyle\sum_{k\in\mathcal{S}_C} \omega_k \left(g(\hat{N}_k,\ n_k,\ \bar{Y}^w_{\cdot k},\ \bar{Y}_{\cdot k}) \ - \ Eg(\hat{N}_k,\ n_k,\ \bar{Y}^w_{\cdot k},\ \bar{Y}_{\cdot k})\right) \ \xrightarrow{P} \ 0$$

Proof. Let $T_k \ = \ g(\hat{N}_k,\ n_k,\ \bar{Y}^w_{\cdot k},\ \bar{Y}_{\cdot k}) \ - \ Eg(\hat{N}_k,\ n_k,\ \bar{Y}^w_{\cdot k},\ \bar{Y}_{\cdot k})$ and $X_k \ = \ I_{[k\in\mathcal{S}_C]}\, \omega_k\, T_k$. The summand variables X_k are independent, have expectation 0, and by (C2)(ii)

$$\mathrm{Var}(N^{-1}\sum_{k=1}^M X_k) \ \le \ \frac{1}{N^2}\sum_{k=1}^M E\Big[I_{[k\in\mathcal{S}_C]}\, \omega_k^2\, T_k^2\Big] \ = \ \frac{1}{N^2}\sum_{k=1}^M E\Big[\omega_k\, T_k^2\Big] \ \le \ \frac{K_2\, M^2}{m^{(0)}\, N^2} \max_{k=1,\ldots,M} E(T_k^2)$$

which is bounded by a constant over $m^{(0)}$, and therefore $\to 0$ as $N \to \infty$. By Chebychev's inequality, the Lemma is proved. □

Lemma 8 *For* $k = 1,\ldots,M$*, define* $r_k \ \equiv \ \sum_{j\in\mathcal{S}_k} w^2_{j|k}/\hat{N}_k$*, and assume* **(C2)**. *Then for all k,* $\hat{N}_k/n_k \ \le \ r_k \ \le K_1^2\, \hat{N}_k\, n_k/(n_k^{(0)})^2$, *where* $K_1,\ n_k^{(0)}$ *are as in (C2)(iii).*

Proof. The first inequality is Cauchy-Schwarz applied to $(\sum_{j\in\mathcal{S}_k} w_{j|k})^2$, and the second is immediate from (C2)(iii). □

Lemma 9 *Assume the two-level ANOVA superpopulation model (19) and conditions* **(C1)**–**(C3)** *on the sampling design. Then uniformly on* $D \ \equiv \ \{(\xi,\zeta) \in (\mathbb{R}^+)^2 \ : \ \min(|\xi|,\ |\xi|^{-1},\ |\zeta|,\ |\zeta|^{-1}) \ge \delta\}$ *for a fixed number* $\delta \in (0,1)$*, as* $N \to \infty$,

(1°) $N^{-1}\sum_{k\in\mathcal{S}_C} \omega_k\, \hat{N}_k/(\xi + \hat{N}_k\,\zeta)$ *is bounded above and below,*

(2°) $N^{-1}\sum_{k\in\mathcal{S}_C} \omega_k\, (\mathrm{SSW}_k \ + \ \sigma_e^2\, r_k) - \sigma_e^2 \xrightarrow{P} 0,$

(3°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, (\bar{Y}^w_{\cdot k} - \mu)/(\xi + \hat{N}_k \, \zeta) \overset{P}{\to} 0,$

(4°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, \big(\frac{\xi}{\xi + \hat{N}_k \, \zeta}\big)^2 \, \big[(\bar{Y}^w_{\cdot k} - \mu)^2 \, - \, (\sigma_a^2 + \sigma_e^2 \, r_k / \hat{N}_k)\big] \overset{P}{\to} 0,$

(5°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \big(\frac{\hat{N}_k \, \xi}{\xi + \hat{N}_k \, \zeta}\big)^2 \, \big[(\bar{Y}^w_{\cdot k} - \mu)^2 \, - \, (\sigma_a^2 + \sigma_e^2 \, r_k / \hat{N}_k)\big] \overset{P}{\to} 0,$

(6°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, \big(\frac{n_k \, \zeta}{\xi + n_k \, \zeta}\big) \, (\bar{Y}_{\cdot k} - \mu) \overset{P}{\to} 0,$

(7°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, \big(\frac{\xi}{\xi + \hat{N}_k \, \zeta}\big)^2 \, \big[(\bar{Y}_{\cdot k} - \mu)^2 \, - \, (\sigma_a^2 + \sigma_e^2 / n_k)\big] \overset{P}{\to} 0,$

(8°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \big(\frac{n_k \, \zeta}{\xi + n_k \, \zeta}\big)^2 \, \big[(\bar{Y}_{\cdot k} - \mu)^2 \, - \, (\sigma_a^2 + \sigma_e^2 / n_k)\big] \overset{P}{\to} 0,$

(9°) $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, \big(\frac{n_k \, \zeta}{\xi + n_k \, \zeta}\big) \, \big[(\bar{Y}^w_{\cdot k} - \mu)(\bar{Y}_{\cdot k} - \mu) \, - \, (\sigma_a^2 + \sigma_e^2 / n_k)\big] \overset{P}{\to} 0.$

Proof. By (7), (C2)(i), Lemma 5 and because $N_k \geq 1$, assertion (1°) is obvious. For (2°), we calculate using (19)

$$\frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \mathrm{SSW}_k \; = \; \frac{1}{N} \sum_{(j,k) \in \mathcal{S}} w_{j,k} \, (\epsilon_{j,k} - \bar{\epsilon}^w_{\cdot k})^2 \; = \; \frac{1}{N} \sum_{(j,k) \in \mathcal{S}} w_{j,k} \, \epsilon^2_{j,k} \; - \; \frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, (\bar{\epsilon}^w_{\cdot k})^2$$

Of the last two summation terms, the first converges in probability to σ_e^2 by Lemma 5 part (c), and because $\hat{N}_k \, E((\bar{\epsilon}^w_{\cdot k})^2 \, | \, \{w_{j|k}\}) \; = \; r_k \, \sigma_e^2$, a similar proof shows that

$$N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \big[\hat{N}_k \, (\bar{\epsilon}^w_{\cdot k})^2 \, - \, r_k \, \sigma_e^2\big] \overset{P}{\to} 0$$

Putting these facts together, we find $N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, (\mathrm{SSW}_k - \sigma_e^2 \, (\hat{N}_k - r_k)) \overset{P}{\to} 0$, and Lemma 5(b) completes the proof of (2°).

The convergence statements (3°)–(9°) hold by Lemma 7 for each fixed (ξ, ζ) in the compact set $D \subset (0, \infty)^2$. The uniform Lipschitz continuity on D of the left-hand side expressions with respect to (ξ, ζ), together with a standard bracketing argument as in van der Vaart (1998, Sec. 19.2) shows that the convergences in probability are uniform on D. □

The next Lemma applies Lemma 9 to derive equations approximately satisfied in the large-N limit in design and model probability under noninformative within-cluster sampling of the weighted survey estimators $\tilde{\theta}^{(M)}$, $\hat{\theta}^{(L1)}$, $\hat{\theta}^{(L2)}$, and $\hat{\theta}^{(EM)}$. These results are expressed through repeated use of the notation $\approx$ to signify that left- and right-hand expressions differ by a random quantity that converges in probability to 0 as $N \to \infty$. For the estimators $\hat{\theta}^{(L1)}$, $\hat{\theta}^{(L2)}$, and $\hat{\theta}^{(EM)}$, rigorous use of the results of Lemma 9 requires that a separate argument be presented for each to supply a compact subset D of $(0, \infty)^2$ within which the estimated pair (σ_a^2, σ_e^2) must lie. These arguments, and the additional formal hypotheses they entail, are omitted in this paper.

Lemma 10 *Assume the 2-level ANOVA superpopulation model (19) and conditions* **(C1)–(C3)** *on the noninformative-within-cluster sampling design. Then as* $N \to \infty$, $\hat{\theta}^{(L)}$ *and* $\hat{\theta}^{(EM)}$ *respectively satisfy equations (44)–(45) and (46)–(47) below. If also* $n_k \geq 2$ *for all* k*, then* $\tilde{\theta}^{(M)} \to \theta$.

Proof. The weighted moment-based estimators $\tilde{\theta}^{(M)}$ are addressed first. Lemma 6 provides consistency of $\tilde{\mu}^{(M)}$ and $(\tilde{\sigma}_a^{(M)})^2 + (\tilde{\sigma}_e^{(M)})^2$, even without assuming (C3). When all $n_k \geq 2$ and (C3) is assumed, the variables $R_k = \sum_{j \in \mathcal{S}_k} (Y_{j,k} - \bar{Y}_{\cdot k})^2/(n_k - 1) - \sigma_e^2$ are independent with mean 0 and uniformly bounded variances, so as in Lemma 7, $(\tilde{\sigma}_e^{(M)})^2 - \sigma_e^2 \;=\; \hat{M}^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, R_k \;\to\; 0$.

Next examine the estimators $\hat{\theta}^{(L)}$ obtained from (30)–(32). First,

$$\hat{\mu}^{(L)} - \mu \;=\; \frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{\rho}_k \, (\bar{Y}^w_{\cdot k} - \mu) \Big/ \big[\frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{\rho}_k \big]$$

By applying parts (1°) and (3°) of Lemma 9 to the denominator and numerator of this ratio, with $\xi = (\hat{\sigma}_e^{(L)})^2$, $\zeta = (\hat{\sigma}_a^{(L)})^2$, the convergence $\hat{\mu}^{(L)} \to \mu$ is proved. (The same argument applies in both the cases where $\gamma_k = 1$ or ω_k.) Since $\hat{\mu}^{(L)}$ is consistent for μ the large-N in-probability limits of $(\hat{\sigma}_a^{(L)})^2$, $(\hat{\sigma}_e^{(L)})^2$ are respectively the same as when $\hat{\mu}^{(L)}$ in their expressions is replaced by μ. We continue with equation (31), using parts (5°) and (3°) of Lemma 9 and dividing through by a factor $(\hat{\sigma}_a^{(L)})^2$, to find

$$\frac{1}{M} \sum_{k \in \mathcal{S}_C} \big[\omega_k - \frac{\omega_k}{\gamma_k} \, (1 - \hat{\rho}_k) - \omega_k \, \hat{\rho}_k \, \frac{r_k \, \sigma_e^2 + \hat{N}_k \, \sigma_a^2}{(\hat{\sigma}_e^{(L)})^2 + \hat{N}_k \, (\hat{\sigma}_a^{(L)})^2} \big] \;\approx\; 0 \tag{44}$$

Similarly, applying parts (2°) and (4°) of Lemma 9 to (32) yields

$$(\hat{\sigma}_e^{(L)})^2 \approx \frac{1}{N} \Big\{ (N - \sum_{k \in \mathcal{S}_C} \omega_k \, r_k) \, \sigma_e^2 \;+\; (\hat{\sigma}_e^{(L)})^2 \sum_{k \in \mathcal{S}_C} \Big[\frac{\omega_k}{\gamma_k} \, \hat{\rho}_k \;+\; \omega_k \, (1 - \hat{\rho}_k) \, \frac{r_k \, \sigma_e^2 + \hat{N}_k \, \sigma_a^2}{(\hat{\sigma}_e^{(L)})^2 + \hat{N}_k \, (\hat{\sigma}_a^{(L)})^2} \Big] \Big\}$$

which implies using (44) that

$$(\hat{\sigma}_e^{(L)})^2 \approx \frac{1}{N} \Big\{ (N - \sum_{k \in \mathcal{S}_C} \omega_k \, r_k) \, \sigma_e^2 \;+\; (\hat{\sigma}_e^{(L)})^2 \sum_{k \in \mathcal{S}_C} \big[\frac{\omega_k}{\gamma_k} \;-\omega_k + \omega_k \frac{r_k \, \sigma_e^2 + \hat{N}_k \, \sigma_a^2}{(\hat{\sigma}_e^{(L)})^2 + \hat{N}_k \, (\hat{\sigma}_a^{(L)})^2} \big] \Big\} \tag{45}$$

Finally, we consider the estimators $\hat{\theta}^{(EM)}$. According to (34),

$$\hat{\mu}^{(EM)} - \mu \;=\; N^{-1} \sum_{k \in \mathcal{S}_C} \omega \, \hat{N}_k \, (\bar{Y}^w_{\cdot k} - \mu - \hat{q}_k \, (\bar{Y}_{\cdot k} - \mu)) \Big/ \big[N^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, (1 - \hat{q}_k) \big]$$

In this ratio, Lemma 9.(1°) shows the denominator to be bounded above and below, while Lemma 9 parts (3°) and (6°) with $\xi = (\hat{\sigma}_e^{(EM)})^2$, $\zeta = (\hat{\sigma}_a^{(EM)})^2$ show that the numerator converges in probability to 0. Thus, since $\hat{\mu}^{(EM)}$ is consistent for μ the expressions for $(\hat{\sigma}_a^{(EM)})^2$, $(\hat{\sigma}_e^{(EM)})^2$ are consistent if and only if they respectively converge in probability to σ_a^2, σ_e^2 when $\hat{\mu}^{(EM)}$ in those expression are replaced by μ. The expressions to study in (34) and (35) become respectively

$$(\hat{\sigma}_a^{(EM)})^2 \;\approx\; M^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{q}_k^2 \, (\bar{Y}_{\cdot k} - \mu)^2 \Big/ \big[M^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{q}_k\big]$$

$$(\hat{\sigma}_e^{(EM)})^2 \;\approx\; \hat{N}^{-1} \sum_{k \in \mathcal{S}_C} \omega_k \Big[\mathrm{SSW}_k + \hat{N}_k \, \big(\bar{Y}_{\cdot k}^w - \hat{q}_k \, \bar{Y}_{\cdot k} - (1 - \hat{q}_k)\, \mu\big)^2 \,+\, \hat{N}_k \, (\hat{\sigma}_e^{(EM)})^2 \, \frac{\hat{q}_k}{n_k})\Big]$$

where $\hat{q}_k$ was defined immediately following (35). Since for each k,

$$SSW_k \,+\, \hat{N}_k \, \big(\bar{Y}_{\cdot k}^w - \hat{q}_k \, \bar{Y}_{\cdot k} - (1 - \hat{q}_k)\, \mu\big)^2 \;=\; \sum_{j \in S_k} w_{j|k} \, \big(Y_{jk} - \mu - \, \hat{q}_k \, (\bar{Y}_{\cdot\, k} - \mu)\big)^2$$

$$= \sum_{j \in \mathcal{S}_k} w_{j|k} \, (Y_{jk} - \mu)^2 \,+\, \hat{N}_k \, \hat{q}_k^2 \, (\bar{Y}_{\cdot\, k} - \mu)^2 \,-\, 2\hat{N}_k \, \hat{q}_k \, (\bar{Y}_{\cdot\, k}^w - \mu)(\bar{Y}_{\cdot\, k} - \mu)$$

parts $(3°), (6°)$ and $(9°)$ of Lemma 9 imply in the displayed equations for $(\hat{\sigma}_a^{(EM)})^2$ and $(\hat{\sigma}_e^{(EM)})^2$ that asymptotically

$$\frac{1}{M} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{q}_k \, \Big[\frac{n_k \, \sigma_a^2 + \sigma_e^2}{n_k \, (\hat{\sigma}_a^{(EM)})^2 + (\hat{\sigma}_e^{(EM)})^2} - 1\Big] \;\approx\; 0 \tag{46}$$

and

$$(\hat{\sigma}_e^{(EM)})^2 \approx \frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \, \Big[\sigma_a^2 (1 - \hat{q}_k)^2 \,+\, (\hat{\sigma}_e^{(EM)})^2 \, \frac{\hat{q}_k}{n_k} \,+\, \sigma_e^2 \, \big(1 \,-\, \frac{2\hat{q}_k}{n_k} \,+\, \frac{\hat{q}_k^2}{n_k})\Big] \tag{47}$$

□

While Lemma 10 does not directly establish consistency of the θ estimators under noninformative sampling, it *does* provide a direct way of establishing **inconsistency** in some cases. Indeed, since the asymptotic equations provided in the Lemma are continuous functions of the estimators, to show inconsistency it suffices to show that the asymptotic equations are not satisfied when respectively σ_a^2, σ_e^2 are respectively substituted for their estimators. This is enough to draw important conclusions about inconsistency of $\hat{\theta}^{(L1)}$ and $\hat{\theta}^{(L2)}$.

Lemma 11 *Assume the two-level ANOVA superpopulation model (19) and conditions* **(C1)**–**(C3)** *on the sampling design which is noninformative within clusters. Assume also that within-cluster sampling has the same weights as SRS sampling, i.e.,* $w_{j|k} \equiv N_k/n_k$ *for all* k*. Then as* $N \to \infty$*, when* $\hat{\theta}^{(L)}$ *is replaced by the true values* $\theta = (\mu, \, \sigma_a^2, \, \sigma_e^2)$*, if either* *(a)* $\gamma_k \equiv 1$ *and* $\liminf_{N\to\infty} M^{-1}\sum_{k=1}^{M} I_{[N_k>n_k]} > 0$, *or (b)* $\gamma_k \equiv \omega_k$ *and* $m = o(M)$, *or (c)* $N_k \equiv \nu$, $N_k/n_k \equiv r$ *do not vary with* k, *and* m/M *is bounded away from* $1-(r-1)\rho_*$, *where*

$$\rho_* \;=\; \nu\,\sigma_a^2/\,(\nu\sigma_a^2+\sigma_e^2)$$

then as N *gets large, equations (44)–(45) do* **not** *hold approximately. Therefore the estimators* $\hat\theta^{(L1)}$ *and* $\hat\theta^{(L2)}$ *are inconsistent under the respective conditions (a), (b), and* $\hat\theta^{(L2)}$ *is inconsistent under (c) except for the specific limiting first-stage sample fraction* $m/M = 1-(r-1)\rho_*$.

Proof. The formulas simplify somewhat because the SRS within-cluster samples, with $w_{j|k} = N_k/n_k$, imply that $r_k = N_k/n_k$ and $\hat N_k = N_k$. In that setting, replacing $(\hat\sigma_a^{(L)},\,\hat\sigma_e^{(L)})$ by $(\sigma_a,\,\sigma_e)$ has the effect of replacing $\hat\rho_k$ by ρ_k. In the (L1) case (a), the left-hand side of (44) becomes

$$\frac{1}{M}\sum_{k\in\mathcal{S}_C}\omega_k\,\rho_k\,\big[\frac{r_k\sigma_e^2+N_k\sigma_a^2}{\sigma_e^2+N_k\sigma_a^2}-1\big] \;=\; \frac{\sigma_e^2}{M}\sum_{k\in\mathcal{S}_C}\omega_k\,\rho_k\,\frac{N_k/n_k\,-\,1}{\sigma_e^2+N_k\sigma_a^2}\big]$$

which is bounded below away from 0 under the limit condition in (a). In the (L2) case (b), the left-hand side minus the right-hand side of (45) becomes

$$\frac{\sigma_e^2}{N}\sum_{k\in\mathcal{S}_C}\Big[\omega_k\,\frac{N_k}{n_k}-1+\omega_k-\,\omega_k\,N_k\,\frac{\sigma_a^2+\sigma_e^2/n_k}{\sigma_e^2+N_k\sigma_a^2}\Big] \;=\; \frac{\sigma_e^2}{N}\sum_{k\in\mathcal{S}_C}\Big[\frac{\omega_k\,N_k\,\sigma_e^2}{\sigma_e^2+N_k\sigma_a^2}-1\Big]$$

which is again bounded below away from 0 under the limit condition in (b). Finally, in case (c), the ratios $r_k = N_k/n_k$ are all equal to r, and when $(\hat\sigma_a^{(L2)},\,\hat\sigma_e^{(L2)})$ is replaced by $(\sigma_a,\,\sigma_e)$, $\hat\rho_k$ becomes $N_k\sigma_a^2/(N_k\sigma_a^2+\sigma_e^2) = \rho_*$. Then (with $\omega_k = \gamma_k$) the left-hand side of (44) becomes

$$\frac{1}{M}\sum_{k\in\mathcal{S}_C}\Big\{(\omega_k-1)(1-\rho_*)-\omega_k\rho_*\big[\frac{r\sigma_e^2+\nu\sigma_a^2}{\sigma_e^2+\nu\sigma_a^2}-1\big]\Big\} \;=\; \frac{1-\rho_*}{M}\sum_{k\in\mathcal{S}_C}\Big\{\omega_k-1-\omega_k\rho_*(r-1)\Big\}$$

Since $\sum_{k\in\mathcal{S}_C}\omega_k/M \approx 1$ for large M, the last expression $\approx (1-\rho_*)\,\{1-\rho_*(r-1)-m/M\}$.

In all three cases (a)-(c), the fact that (44) or (45) fails to hold when $\hat\theta^{(L)}$ is replaced by the true θ implies that (44)–(45) is incompatible with the convergence $\hat\theta \approx \theta$ as $N\to\infty$. □

It is easy to see that when $\gamma_k \equiv 1$ and $n_k \equiv N_k$, equations (44)–(45) **do** hold as $N\to\infty$. Similarly, as we verify below, (46)–(47) hold approximately without any further conditions when $\hat\theta^{(EM)}$ is replaced by θ. It is also true that when $n_k\to\infty$ for all k and $\hat\theta^{(L)}$ is replaced by θ, both equations (44) and (45) hold approximately. Although the hypothesis in the third of these cases is excluded by our assumptions (C1) and (C2), the first two cases are shown in the next Lemma to imply a consistency result. Some additional arguments are needed to establish this: the asymptotic validity of equations (44)–(45) or (46)–(47) are only *necessary* conditions for consistency of the estimators solving them, not sufficient conditions. The arguments given in the next Lemma are specific to the two-level Analysis of Variance model.

Lemma 12 *Assume the two-level ANOVA superpopulation model (19) and conditions* **(C1)**–**(C3)** *on the sampling design which is noninformative within clusters. Then*

(a) with the true parameter values θ *substituted for* $\hat{\theta}^{(EM)}$*, equations (46)–(47) hold asymptotically as* $N_k \to \infty$*, so that* $\hat{\theta}^{(EM)}$ *is consistent, and*

(b) if $n_k \equiv N_k$ *for all* k*, then* $\hat{\theta}^{(L1)} \approx \theta$ *as* $N \to \infty$*.*

Proof. For (a), begin by checking that (46)–(47) hold when $\hat{\theta}^{(EM)}$ is replaced by θ. This is obvious for (46), and (47) becomes

$$\frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \, \hat{N}_k \Big[- \sigma_e^2 \, + \, \sigma_a^2 (1 - q_k)^2 \, + \, \sigma_e^2 \, \frac{q_k}{n_k} \, + \, \sigma_e^2 \, \big(1 \, - \, \frac{2q_k}{n_k} \, + \, \frac{q_k^2}{n_k}\big)\Big] \; = \; 0$$

The last equation holds because $\sigma_a^2 \, (1 - q_k) = \sigma_e^2 \, q_k / n_k$ by definition of q_k.

Next, using the same limiting results from Lemma 9, we re-write the EM iterative-step equations (33), after replacing μ_1 by the true value μ, as

$$\begin{aligned} \sigma_{a,1}^2 \;\; &\approx \;\; \frac{1}{M} \sum_{k \in \mathcal{S}_C} \omega_k \big[q_{k,0}^2 (\sigma_a^2 + \frac{\sigma_e^2}{n_k}) \, + \, (1 - q_{k,0}) \, \sigma_{a,0}^2 \big] \\ \sigma_{e,1}^2 \;\; &\approx \;\; \frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \hat{N}_k \Big[\sigma_a^2 + \sigma_e^2 + (q_{k,0}^2 - 2 q_{k,0})(\sigma_a^2 + \frac{\sigma_e^2}{n_k}) \, + \, (1 - q_{k,0}) \sigma_{a,0}^2 \Big] \end{aligned}$$

Recognizing that the right-hand sides of these last two displayed equations respectively become $\approx \sigma_{a,0}^2$ and $\approx \sigma_{e,0}^2$ when $(\sigma_a^2, \, \sigma_e^2)$, we further simplify these last two displayed equations algebraically to the form

$$\begin{pmatrix} \sigma_{a,1}^2 - \sigma_a^2 \\ \sigma_{e,1}^2 - \sigma_e^2 \end{pmatrix} \approx \frac{1}{N} \sum_{k \in \mathcal{S}_C} \omega_k \begin{pmatrix} \frac{N}{M} \, (1 - q_{k,0}^2) & -\frac{N}{M} q_{k,0}^2 / n_k \\ -\hat{N}_k \, (1 - q_{k,0})^2 & \frac{\hat{N}_k}{n_k} \, (1 - (1 - q_{k,0})^2) \end{pmatrix} \begin{pmatrix} \sigma_{a,0}^2 - \sigma_a^2 \\ \sigma_{e,0}^2 - \sigma_e^2 \end{pmatrix} \qquad (48)$$

This last equation shows clearly that the true parameter(σ_a^2, σ_e^2) is asymptotically approximately fixed by the EM iterative step (33), but also that the EM step is asymptotically contractive with uniformly high probability. To prove the contraction property, it must be shown that the 2 × 2 matrix on the right-hand side of (48) has real eigenvalues with absolute values less than 1. Denote the diagonal entries of that matrix by α_1, α_4 and the off-diagonal by $-\alpha_2$, $-\alpha_3$. Then by inspecttion all $\alpha_j \in (0,1)$, and $\alpha_1 + \alpha_2 < 1$, $\alpha_3 + \alpha_4 < 1$. Then the roots of the quadratic characteristic polynomial of the matrix are

$$\frac{\alpha_1 + \alpha_4}{2} \;\pm\; \frac{1}{2}\sqrt{(\alpha_1 - \alpha_4)^2 \,+\, 4\alpha_2\,\alpha_3}$$

which are in absolute value less than

$$\frac{\alpha_1 + \alpha_4 + \,|\alpha_1 - \alpha_4|}{2} \;+\; \sqrt{\alpha_2\,\alpha_3} \;<\; \max(\alpha_1, \alpha_4) + (1-\alpha_1)(1-\alpha_4) \;<\; 1 - \min(\alpha_1, \alpha_4) + \alpha_1\alpha_4$$

and the final term is less than 1. Since the asymptotic iteration-step (48) is a contraction, it follows that the unique limiting value for $(\hat{\sigma}_a^{(EM)\,2}$, $(\hat{\sigma}_e^{(EM)\,2}$ is $(\sigma_a^2,\, \sigma_e^2)$, and $\hat{\theta}^{(EM)}$ is consistent.

When $n_k = N_k$, it is easy to see that $r_k = 1$, $w_{j|k} \equiv 1, \hat{N}_k$, and $\hat{\rho}_k = \hat{q}_k$. Equations (44)–(45) take the form

$$\frac{1}{M}\sum_{k \in \mathcal{S}_C} \omega_k\, \hat{q}_k \Big[\frac{\sigma_e^2 + n_k\sigma_a^2}{\hat{\sigma}_e^2 + n_k\hat{\sigma}_a^2} - 1\Big] \;\approx\; 0 \quad , \qquad \frac{\hat{\sigma}_e^2}{\sigma_e^2}\Big[1 - \frac{1}{N}\sum_{k \in \mathcal{S}_C} \omega_k\, \frac{\sigma_e^2 + n_k\sigma_a^2}{\hat{\sigma}_e^2 + n_k\hat{\sigma}_a^2}\Big] \approx 1 - \frac{M}{N}$$

which we re-write in terms of

$$x = \frac{\hat{\sigma}_a^2}{\sigma_a^2} \quad , \quad z = \frac{\hat{\sigma}_e^2}{\sigma_e^2} \quad , \quad t = \frac{\hat{M}}{N} \quad , \quad g = \frac{1}{\hat{M}}\sum_k \hat{q}_k\,\omega_k \quad , \quad h = \frac{1}{\hat{M}}\sum_k \hat{q}_k^2\,\omega_k$$

The rewritten equations (44)–(45) now have the expression

$$g \;=\; (h/x) \,+\, (g-h)/z \quad , \qquad z\big(1 - t\big[\frac{g}{x} \,+\, \frac{1-g}{z}\big]\big) \;=\; t\,g\,(\frac{z}{x} \,-1) \tag{49}$$

or equivalently

$$h\,(z-x) \;=\; gx\,(z-1) \quad , \qquad tg\,(z-x) \;=\; x\,(z-1)$$

Solution of this last pair of equations would be possible for $z \neq 1$ or $z \neq x$ only if $h/t = g^2$. Yet the Cauchy-Schwarz inequality implies, after recalling $\hat{M} = \sum_{k \in \mathcal{S}_C} \omega_k$, that $g^2 \leq h$. Since $t < 1$, this implies $g^2 < h/t$, from which it follows that the only (approximate asymptotic) solution to (49) occurs at $z = x = 1$, i.e. at $\hat{\sigma}_a^2 = \sigma_a^2$ and $\hat{\sigma}_e^2 = \sigma_e^2$. □