Received: 10 February 2017 | Revised: 23 November 2018 | Accepted: 27 November 2018

DOI: 10.1002/sim.8070

RESEARCH ARTICLE

WILEY Statistics in Medicine

Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative

Francesco Innocenti[1] | Math J.J.M. Candel[1] | Frans E.S. Tan[1] | Gerard J.P. van Breukelen[1,2]

[1] Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, The Netherlands

[2] Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, The Netherlands

Correspondence
Francesco Innocenti, Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, P.O. Box 616, 6200 MD, Maastricht, The Netherlands.
Email: francesco.innocenti@maastrichtuniversity.nl

In multilevel populations, there are two types of population means of an outcome variable ie, the average of all individual outcomes ignoring cluster membership and the average of cluster-specific means. To estimate the first mean, individuals can be sampled directly with simple random sampling or with two-stage sampling (TSS), that is, sampling clusters first, and then individuals within the sampled clusters. When cluster size varies in the population, three TSS schemes can be considered, ie, sampling clusters with probability proportional to cluster size and then sampling the same number of individuals per cluster; sampling clusters with equal probability and then sampling the same percentage of individuals per cluster; and sampling clusters with equal probability and then sampling the same number of individuals per cluster. Unbiased estimation of the average of all individual outcomes is discussed under each sampling scheme assuming cluster size to be informative. Furthermore, the three TSS schemes are compared in terms of efficiency with each other and with simple random sampling under the constraint of a fixed total sample size. The relative efficiency of the sampling schemes is shown to vary across different cluster size distributions. However, sampling clusters with probability proportional to size is the most efficient TSS scheme for many cluster size distributions. Model-based and design-based inference are compared and are shown to give similar results. The results are applied to the distribution of high school size in Italy and the distribution of patient list size for general practices in England.

KEYWORDS

design-based inference, hierarchical population, informative cluster size, model-based inference, two-stage sampling

1 | INTRODUCTION

Hierarchical or multilevel populations arise when individuals or micro-units are nested within clusters or macro-units.[1,2] Considering, for the sake of simplicity, only populations with two levels of nesting, examples include patients clustered in general practices, elderly people nested in nursing homes, and students grouped in schools. In these populations, the overall mean of an outcome variable (eg, cholesterol level, blood pressure, body mass index) can be defined in two ways, ie, as the mean of all individuals in the population ignoring cluster membership (ie, first, pooling all patients from all clusters in the population, and then computing the average cholesterol level); or as the mean of all cluster-specific means (ie, first, computing the mean cholesterol level within each cluster, and then averaging all the cluster-specific means). These two definitions coincide only under special conditions, as will be seen later, but this paper focuses on the first definition only. Related to these two definitions is the concept of informative cluster size.

When clusters vary in size in the population (eg, small versus large general practices), cluster sizes can be seen as realizations of a random variable,[3] and the outcome variable of interest may be related to cluster size (eg, surgeons operating on many patients might have better performances than those operating on fewer patients[4]). If this is the case, then cluster size is said to be informative.[5] Nevalainen et al[6] describe and give practical examples of three data-generating mechanisms that can lead to informative cluster size. Briefly, a latent variable (eg, the competence of the surgeon) influences cluster size (eg, the number of patients) and the outcome variable (eg, success of the operation) at the same time; or cluster size affects the outcome variable (eg, surgeons become better by practice); or vice versa, the outcome variable affects cluster size (eg, better surgeons get more referrals). In relation, Seaman et al[5] point out that the standard methods to analyze clustered data, namely, generalized linear mixed models (GLMMs) and generalized estimating equations (GEEs), implicitly assume that cluster size is unrelated to the outcome variable, and discuss different methods to handle informative cluster size for cluster-specific inference with GLMM and population-average inference with GEE.

The topic of this paper is the unbiased and efficient estimation of the population mean in the presence of informative cluster size. To estimate the population mean, individuals can be sampled either with simple random sampling (SRS), that is, directly from the population, or with two-stage sampling (TSS), that is, sampling first clusters and then individuals within the sampled clusters.[7-9] Given cluster size variation in the population, at least three alternative TSS schemes can be considered.

1. Sampling clusters with probability proportional to cluster size and then sampling the same number of individuals from each sampled cluster.
2. Sampling clusters with equal probability and then sampling per sampled cluster a number of individuals proportional to cluster size.
3. Sampling clusters with equal probability and then sampling the same number of individuals per cluster.

In order to evaluate each sampling scheme in terms of unbiasedness and efficiency of mean estimation, it is useful to distinguish two approaches to inference in survey sampling literature[10]: the design-based paradigm[7-9] and the model-based approach.[11-13] In the design-based approach, the outcome value for each unit (eg, patient) in the population is assumed to be a fixed unknown quantity. The random variable is then the *inclusion indicator*, that is, the variable that states whether or not a unit is included into the sample. Thus, inference is based on the distribution of the inclusion indicator over repeated samples with a probability sampling design. In contrast, the model-based approach assumes that the outcome value in the real finite population is a realization of a stochastic model, representing a hypothetical infinite population. Inference is then based on the probabilistic model. As long as the assumptions of the model are met, model-based inference can then ignore the sampling scheme and condition on the observed sample.[8,10,12,13] However, if the model residuals (ie, the stochastic part) are correlated with the variables which determine the sampling probabilities (and then with the sampling probabilities themselves), the sampling design is said to be informative.[2(p222),10,13-16] When this is the case, model-based analysis is biased, unless the sampling design is taken into account.[2(p237)] In the multilevel modeling literature, many authors have investigated unbiased estimation when TSS with unequal sampling probabilities is informative, but they assumed noninformative cluster size.[16-20] In this paper, this sampling scheme is informative due to the cluster size being informative.

In this paper, cluster size is treated as a random variable and assumed to be informative, but the special case of noninformative cluster size will also be covered briefly. Furthermore, a simple hierarchical linear model,[1,2] for the outcome variable in the population, is assumed and used to define the parameter of interest (ie, the population mean). We thus adopt a model-based approach but will also make a comparison with design-based inference. It will be shown that the type of analysis (ie, unweighted versus weighted analysis) needed for unbiased estimation of the population mean depends on the chosen sampling scheme. Furthermore, the three aforementioned TSS schemes will be compared with each other and with SRS in terms of their efficiency under the constraint of a fixed total sample size. It will also be shown that their relative efficiencies depend on the cluster size distribution.

The rest of this paper is organized as follows. In Section 2, the assumptions on which our findings are based and the considered sampling schemes are presented in more detail. In Section 3, the population mean is derived under a linear mixed model for a two-level hierarchical population with varying and informative cluster size. Furthermore, Section 3 deals with the estimation of the population mean under different sampling schemes, presenting both the expectation and sampling variance of the estimator under each scheme. In Section 4, the three TSS schemes are compared with each other and with SRS in terms of efficiency for a given total sample size (number of individuals). In Section 5, the relative efficiencies of the three TSS schemes are derived under the design-based approach and compared with those obtained under the model-based framework. The results of this paper are applied in Section 6 to two real populations, ie, high schools in Italy and general practices in England. Some final remarks are offered in Section 7. The online Supplementary Material contains part of the derivations of the equations given in this paper as well as additional tables and figures.

2 | ASSUMPTIONS AND SAMPLING SCHEMES

The structure of the data is hierarchical with two levels of nesting (eg, pupils are nested within schools, patients within general practitioners (GPs)). The results of this paper are based on the following assumptions (the notation is summarized in Table A1 in Appendix A).

Assumption 1. The population is composed of K clusters (eg, schools, GPs) and each cluster j contains N_j individuals (eg, students, patients), that is, clusters are allowed to have different sizes. The total number of individuals in the population (ie, the population size) is $N_{\text{pop}} = \sum_{j=1}^{K} N_j$.

Assumption 2. Sampling is either SRS of individuals in one stage, or else TSS. In TSS, we first sample k clusters, and then sample n or n_j individuals from each sampled cluster j. In case of TSS, the population is very large relative to the sample size at each design level, that is, $\frac{k}{K} \to 0$ and $\frac{\overline{n}}{\theta_N} \to 0$, where $\overline{n} = \frac{\sum_{j=1}^{k} n_j}{k}$ is the average number of individuals sampled per sampled cluster and $\theta_N = \frac{N_{\text{pop}}}{K}$ is the mean cluster size in the population. In case of SRS, N_{pop} is very large relative to m, the number of individuals sampled (ie, $\frac{m}{N_{\text{pop}}} \to 0$).

Assumption 3. The outcome variable Y_{ij} is quantitative (eg, cholesterol level) and measured at the individual (eg, patient) level. Furthermore, Y_{ij} shows variation at the cluster level as well as at the individual level. Therefore, sampling error occurs at each design level. This is taken into account by assuming the following two-level random intercept model for the outcome of the ith individual from the jth cluster:

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}, \tag{1}$$

where $u_j|N_j \sim N(0, \sigma_v^2)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $u_j \perp \varepsilon_{ij}$, and σ_v^2 will be defined in the next assumption. Note that multilevel models, such as Equation (1), are not only a standard procedure for modeling hierarchical populations[1,2] but also a natural way for taking into account the clustering induced by TSS in a model-based approach.*

Assumption 4. The cluster effect u_j is allowed to be linearly related to the size of the cluster in the population N_j, that is, $u_j = \alpha + \gamma N_j + v_j = \gamma(N_j - \theta_N) + v_j$, where $\alpha = -\gamma\theta_N$ for model identifiability, $v_j \sim N(0, \sigma_v^2)$, and $v_j \perp N_j$.

In order to deal with cluster size variation and informative cluster size in estimating the population mean (ie, the average of all individual outcomes), three competing TSS schemes are considered, which will be compared with SRS of individuals and with each other, under the constraint that all sampling schemes have the same total sample size.

Two-Stage Sampling 1 (TSS1):

Stage 1: Sample k clusters with probability proportional to cluster size N_j, that is, $\frac{N_j}{\sum_{j=1}^{K} N_j}$, is the probability of cluster j being sampled if one cluster is randomly sampled, and so the inclusion probability for the jth cluster, that is, the probability that cluster j is sampled given a total of k sampled clusters, is $\pi_j = 1 - \left(1 - \frac{N_j}{\sum_{j=1}^{K} N_j}\right)^k$.[9(p51)] If $\frac{N_j}{\sum_{j=1}^{K} N_j} \to 0, \quad \forall j = 1, \dots, K$, then $\pi_j \approx \frac{kN_j}{\sum_{j=1}^{K} N_j}$; this approximation will be used.

Stage 2: Sample the same number of individuals n per cluster, so that $\pi_{i|j} = \frac{n}{N_j}$, where $\pi_{i|j}$ denotes the probability of including the ith individual from cluster j in the sample, given that, at the first stage, the jth cluster is sampled.

*See other works.[1(pp212,213),2(pp218,223),8(pp200,262-264),10,11(p256),12(p65),13,21,22]

Note that, under this sampling scheme, all individuals have the same unconditional probability of selection, that is, $\pi_{ij} = \pi_j \pi_{i|j} \approx \frac{kN_j}{\sum_{j=1}^{K} N_j} \frac{n}{N_j} = \frac{nk}{N_{\text{pop}}}$. A potential drawback of TSS1 is that we must know the sizes of all clusters in the population to draw the k clusters for the sample.

Two-Stage Sampling 2 (TSS2):

Stage 1: Sample k clusters with SRS, that is, $\pi_j = \frac{k}{K}, \quad \forall j = 1, \dots, K$.

Stage 2: Sample the same percentage of individuals per cluster p, that is, the number of individuals sampled per cluster (ie, n_j) is proportional to the cluster size in the population (ie, N_j), and so $\pi_{i|j} = \frac{n_j}{N_j} = p \ \forall i = 1, \dots, N_j$ and $\forall j = 1, \dots, K$. Under this sampling scheme, the unconditional probability of being included into the sample is the same for all individuals, that is, $\pi_{ij} = \pi_j \pi_{i|j} = \frac{k}{K}\frac{n_j}{N_j} = \frac{k}{K}p$. In contrast to what was the case for TSS1, we now need to know only the cluster sizes for the sampled clusters before sampling individuals from those sampled clusters.

Two-Stage Sampling 3 (TSS3):

Stage 1: Sample k clusters with SRS, that is, $\pi_j = \frac{k}{K}, \quad \forall j = 1, \dots, K$.

Stage 2: Sample the same number of individuals n per cluster, then $\pi_{i|j} = \frac{n}{N_j}$.

The unconditional sample inclusion probability of the ith individual in the jth cluster is $\pi_{ij} = \pi_j \pi_{i|j} = \frac{k}{K}\frac{n}{N_j}$. Thus, individuals from different clusters have a different probability to be drawn from their cluster (the larger N_j, the smaller this probability). This has consequences for the data analysis as will be seen in the next section.

As a final remark on this section, note that the three TSS schemes considered here can be seen as three particular cases of a larger family of alternative TSS schemes. At the first stage, a more general expression for π_j is $\pi_j = \frac{kX_j}{\sum_{j=1}^{K} X_j}$, where X_j is an arbitrary auxiliary variable available before sampling. At the second stage, a general form for $\pi_{i|j}$ is $\pi_{i|j} = \frac{n_j Z_{ij}}{\sum_{i=1}^{N_j} Z_{ij}}$, where Z_{ij} is an auxiliary variable for individuals prior of sampling. Thus, TSS1 follows by imposing $X_j = N_j$, $Z_{ij} = 1$, and $n_j = n$. Instead, TSS2 results from $X_j = 1$, $Z_{ij} = 1$, and $n_j = pN_j$, whereas TSS3 is obtained with $X_j = 1$, $Z_{ij} = 1$, and $n_j = n$.

3 | DEFINITION AND ESTIMATION OF THE POPULATION MEAN μ

To find the population mean $E(Y_{ij})$ and variance $V(Y_{ij})$, defined from model (1) as the marginal expectation and variance of Y_{ij} over cluster effect u_j and individual effect ε_{ij}, the marginal expectation and variance of cluster effect u_j (ie, $E(u_j)$ and $V(u_j)$, respectively) are needed. If cluster size is noninformative (ie, $\gamma = 0$ in Assumption 4), then $E(u_j) = 0$ and $V(u_j) = \sigma_v^2$ leading to $E(Y_{ij}) = \beta_0$ and $V(Y_{ij}) = \sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2$. In contrast, if cluster size is informative (ie, $\gamma \neq 0$ in Assumption 4), $E(u_j) = 0$ or $E(u_j) \neq 0$ depending on the sampling scheme. To prevent misunderstanding, note that the cluster effect u_j in the population does not depend on the sampling design, and its marginal distribution in the population is $f(u_j) = \int f(u_j|N_j)f(N_j)dN_j$ (where $f(.)$ indicates a probability density function). Nevertheless, the sampling design determines the cluster effect sampling distribution, which is, for a sample of size one, equal to $\int f(u_j|N_j)f(N_j)dN_j$ if clusters are sampled with equal probabilities, and equal to $\int \left(\frac{N_j}{\theta_N}\right) f(u_j|N_j)f(N_j)dN_j$, if clusters are sampled with probabilities proportional to their size.

Under TSS2 or TSS3, the k clusters are sampled with equal probabilities from the population of K clusters, and then (for proofs, see Appendix A)

$$\text{(a)} \quad E_{\text{TSS2/TSS3}}(u_j) = 0, \quad \text{and} \quad \text{(b)} \quad V_{\text{TSS2/TSS3}}(u_j) = \sigma_v^2 + \gamma^2\sigma_N^2 = \sigma_u^2. \tag{2}$$

Note that $\gamma^2\sigma_N^2$ is the component of $V_{\text{TSS2/TSS3}}(u_j)$ explained by N_j, and σ_v^2 is the unexplained variance of u_j. Hence, the following expression for $E(Y_{ij})$ comes from model (1) and Equation (2a):

$$E_{\text{TSS2/TSS3}}(Y_{ij}) = \beta_0, \tag{3}$$

which can be interpreted as the expected outcome for an arbitrary individual (ie, $E(\varepsilon_{ij}) = 0$) from an arbitrary cluster (ie, $E(u_j) = 0$). To estimate β_0 unbiasedly, large and small clusters should be weighted equally, both in the sampling scheme and in the estimator (see Appendix B). However, β_0 is not the parameter of interest in this paper.

Under SRS m individuals are sampled directly from the population of $N_{\text{pop}} = \sum_{j=1}^{K} N_j$ individuals and with equal probabilities (ie, $\pi_i = \frac{m}{N_{\text{pop}}}, \forall i = 1, \dots, N_{\text{pop}}$). Now, the probability that a selected individual belongs to a cluster of size N_j is proportional to cluster size, meaning that large clusters have higher chance of being represented in the SRS sample. Hence, under SRS, k_{SRS} clusters are indirectly sampled from the population with sampling probability proportional to size, and k_{SRS} can run from 1 to m. Likewise, under TSS1 k clusters are sampled with probabilities proportional to their size, and so large clusters are more likely to be drawn. Therefore, under SRS and TSS1, the marginal expectation and variance of cluster effect u_j are (for proofs, see Appendix A)

$$\text{(a)} \quad E_{\text{SRS/TSS1}}(u_j) = \gamma \theta_N \tau_N^2, \quad \text{and} \quad \text{(b)} \quad V_{\text{SRS/TSS1}}(u_j) = \sigma_v^2 + \gamma^2 \sigma_N^2 \left[\tau_N \left(\zeta_N - \tau_N \right) + 1 \right], \tag{4}$$

where $\tau_N = \frac{\sigma_N}{\theta_N}$ and $\zeta_N = \frac{E[(N_j-\theta_N)^3]}{\sigma_N^3}$ are the coefficient of variation and the skewness of cluster size distribution in the population, respectively. Note that $V_{\text{SRS/TSS1}}(u_j) = V_{\text{TSS2/TSS3}}(u_j)$ if one of the following conditions holds: (i) $\tau_N = 0$ (ie, no cluster size variation), (ii) $\gamma = 0$ (ie, cluster size is noninformative), or (iii) $\zeta_N = \tau_N$ (eg, N_j is Poisson distributed, see Table S.M.1 in the Supplementary Material). Likewise, $E_{\text{SRS/TSS1}}(u_j) = E_{\text{TSS2/TSS3}}(u_j)$ if either condition (i) or (ii) holds. Thus, from model (1) and Equation (4a), the population mean that we here want to estimate as follows:

$$E_{\text{SRS/TSS1}}(Y_{ij}) = \beta_0 + \gamma \theta_N \tau_N^2 = \mu. \tag{5}$$

This mean can be interpreted as the expected outcome for an individual randomly sampled from the population ignoring cluster membership by SRS. Note that the two definitions of $E(Y_{ij})$ in Equations (3) and (5) coincide if either clusters have the same size in the population (ie, $\tau_N = 0$) or cluster size is not related to the outcome (ie, $\gamma = 0$). Given the focus of this paper on μ, model (1) can be rewritten from Equation (5) as follows:

$$y_{ij} = \mu + b_j + \varepsilon_{ij}, \tag{6}$$

where $b_j = u_j - \gamma \theta_N \tau_N^2 = u_j - E_{\text{SRS/TSS1}}(u_j)$ (see Equation (4a)) with $E_{\text{SRS/TSS1}}(b_j) = 0$ and $V_{\text{SRS/TSS1}}(b_j) = V_{\text{SRS/TSS1}}(u_j)$ (see Equation (4b)).

To estimate μ unbiasedly, the weight of a cluster should be proportional to its size, either in the sampling scheme or in the estimator (for details, see Appendices A and B). For each sampling scheme, the first row of Table 1 presents the unbiased or approximately unbiased (ie, for k sufficiently large) estimator of μ under model (6), the second and third row present the conditional expectation and variance of $\hat{\mu}$, the fourth row gives the marginal expectation of $\hat{\mu}$, and the last two rows show the two components of the marginal variance of $\hat{\mu}$ (ie, $\text{Var}(\hat{\mu}) = E(V(\hat{\mu}|\mathbf{N}_*)) + V(E(\hat{\mu}|\mathbf{N}_*))$, where $\mathbf{N}_* = \mathbf{N} = (N_1, \dots, N_k)^T$ under TSS and $\mathbf{N}_* = \mathbf{N}_{\text{SRS}} = (N_1, \dots, N_{k_{\text{SRS}}})^T$ under SRS) (for proofs, see Appendix B). As the first row of Table 1 shows, the estimator of μ is a weighted sum of cluster means in each sampling scheme, but the weights differ between schemes. Under SRS k_{SRS} clusters are indirectly sampled from the population and large clusters have higher chance of being sampled, thus the unweighted estimator is unbiased for μ (recall that from Assumption 2, $\frac{m}{N_{\text{pop}}} \to 0$, which implies that $k_{\text{SRS}} \to m$). Under TSS1 clusters are sampled with probabilities proportional to their size, and so μ is estimated unbiasedly by the unweighted average of cluster means. Under TSS3 and TSS2 cluster means must be weighted by cluster size (ie, N_j in TSS3, and also in TSS2 because $n_j = pN_j$) in the analysis, because clusters are weighted equally by these sampling designs, that is, all clusters have equal sampling probability (for details, see Appendix B). An exception to this is the special case of noninformative cluster size (ie, $\gamma = 0$), in which the two definitions of population means coincide (ie, $\mu = \beta_0$). It then follows that $E(u_j) = 0$ for any sampling scheme (see Appendix A), and from model (1), then results that $E(\bar{y}_j) = \beta_0$. Thus, any estimator of $\mu = \beta_0$ of the form $\hat{\mu} = \frac{\sum_{j=1}^k w_j \bar{y}_j}{\sum_{j=1}^k w_j}$ is unbiased then, although some weights w_j are more efficient than others.[23,24]

4 | RELATIVE EFFICIENCIES OF TSS SCHEMES VERSUS SRS AND EACH OTHER

Under the constraint of a fixed total sample size (ie, $m = \overline{n}k$), the efficiency of the three TSS schemes can be investigated by computing their *relative efficiencies*, defined as the ratio of the sampling variances of $\hat{\mu}$ under two competing sampling schemes (ie, the variances obtained as the sum of the last two rows of Table 1). For instance, the relative efficiency of TSS1 versus SRS is defined as the ratio of $V(\hat{\mu})$ for SRS to $V(\hat{\mu})$ for TSS1 (ie, RE (TSS1 vs SRS) $= V(\hat{\mu}_{\text{SRS}}) / V(\hat{\mu}_{\text{TSS1}})$). The relative efficiencies are given in Table 2 (for proof, see section 2 of the Supplementary Material), whereas the relative efficiency of TSS2 versus TSS1 is plotted in Figure 1. As shown by Table 2, the numerator and denominator of the relative efficiency are both a weighted sum of two components, respectively $E(V(\hat{\mu}|\mathbf{N}))$ and $V(E(\hat{\mu}|\mathbf{N}))$ from last two rows of Table 1, with weights determined by the correlation between cluster effect and cluster size $\text{corr}(u_j, N_j)$. The component $E(V(\hat{\mu}|\mathbf{N}))$ with weight $(1 - \text{corr}(u_j, N_j)^2)$ depends on the intraclass correlation $\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2}$, the coefficient of variation of cluster size τ_N, and the average number of individuals sampled per cluster $\overline{n}$. The other component, ie, $V(E(\hat{\mu}|\mathbf{N}))$, weighted by $\text{corr}(u_j, N_j)^2$,

TABLE 1 Estimators of the population mean $\mu = \beta_0 + \gamma \theta_N \tau_N^2$: conditional and marginal expectations and variances[a]

	SRS	TSS1	TSS2	TSS3
$\hat{\mu}$	$\sum_{i=1}^{m} \frac{y_i}{m}$	$\sum_{j=1}^{k} \frac{\overline{y}_j}{k}$	$\frac{\sum_{j=1}^{k} pN_j \overline{y}_j}{\sum_{j=1}^{k} pN_j}$	$\frac{\sum_{j=1}^{k} N_j \overline{y}_j}{\sum_{j=1}^{k} N_j}$
$E(\hat{\mu}\|\mathbf{N}_*)$	$\beta_0 + \gamma \left(\overline{N}_{\text{SRS}} - \theta_N\right)$	$\beta_0 + \gamma \left(\overline{N} - \theta_N\right)$	$\beta_0 + \gamma \left(\overline{N}\left(CV_N^2 + 1\right) - \theta_N\right)$	$\beta_0 + \gamma \left(\overline{N}\left(CV_N^2 + 1\right) - \theta_N\right)$
$V(\hat{\mu}\|\mathbf{N}_*)$	$\frac{\sigma_v^2 + \sigma_\varepsilon^2}{m}$	$\frac{n\sigma_v^2 + \sigma_\varepsilon^2}{nk}$	$\frac{\overline{n}(CV_N^2+1)\sigma_v^2 + \sigma_\varepsilon^2}{\overline{n}k}$	$\frac{n\sigma_v^2 + \sigma_\varepsilon^2}{nk} \times \left(CV_N^2 + 1\right)$
$E(\hat{\mu})$	$\beta_0 + \gamma \theta_N \tau_N^2$	$\beta_0 + \gamma \theta_N \tau_N^2$	$\beta_0 + \gamma \theta_N \left(\frac{k-1}{k}\right) \tau_N^2$	$\beta_0 + \gamma \theta_N \left(\frac{k-1}{k}\right) \tau_N^2$
$E(V(\hat{\mu}\|\mathbf{N}_*))$	$\frac{\sigma_v^2 + \sigma_\varepsilon^2}{m}$	$\frac{n\sigma_v^2 + \sigma_\varepsilon^2}{nk}$	$\frac{p\theta_N \left(\frac{k\left(\tau_N^2+1\right)}{\tau_N^2+k}\right)\sigma_v^2 + \sigma_\varepsilon^2}{p\theta_N k}$	$\frac{\left(n\sigma_v^2 + \sigma_\varepsilon^2\right)\left(\frac{k\left(\tau_N^2+1\right)}{\tau_N^2+k}\right)}{nk}$
$V(E(\hat{\mu}\|\mathbf{N}_*))$	$\gamma^2 \frac{\sigma_N^2\left[\tau_N(\zeta_N - \tau_N)+1\right]}{m}$	$\gamma^2 \frac{\sigma_N^2\left[\tau_N(\zeta_N - \tau_N)+1\right]}{k}$	$\gamma^2 \frac{\sigma_N^2}{k} \left[\left(\frac{k-1}{k}\right)^2 \tau_N^2 \left(\eta_N - \frac{k-3}{k-1} + \tau_N\left(\tau_N - 2\zeta_N\right)\right) + \right.$ $\left. + 2\left(\frac{k-1}{k}\right) \tau_N \left(\zeta_N - \tau_N\right) + 1\right]$	$\gamma^2 \frac{\sigma_N^2}{k} \left[\left(\frac{k-1}{k}\right)^2 \tau_N^2 \left(\eta_N - \frac{k-3}{k-1} + \tau_N\left(\tau_N - 2\zeta_N\right)\right) + \right.$ $\left. + 2\left(\frac{k-1}{k}\right) \tau_N \left(\zeta_N - \tau_N\right) + 1\right]$

[a] Derivations are given in Appendix B. Note that $m = \overline{n}k$ where k is the number of clusters sampled with any TSS scheme; under SRS $\mathbf{N}_* = \mathbf{N}_{\text{SRS}} = (N_1, \dots, N_{k_{\text{SRS}}})^T$ and $\overline{N}_{\text{SRS}} = \frac{\sum_{j=1}^{k_{\text{SRS}}} N_j}{k_{\text{SRS}}}$, where k_{SRS} is the number of clusters indirectly sampled with SRS; under any TSS scheme $\mathbf{N}_* = \mathbf{N} = (N_1, \dots, N_k)^T$; $CV_N = \frac{S_N}{\overline{N}}$ is the sample coefficient of variation of cluster size, where $\overline{N} = \frac{\sum_{j=1}^{k} N_j}{k}$ and $S_N = \sqrt{\frac{\sum_{j=1}^{k}(N_j - \overline{N})^2}{k}}$; $\tau_N = \frac{\sigma_N}{\theta_N}$ is the population coefficient of variation of cluster size; $\zeta_N = E\left[\left(\frac{N_j - \theta_N}{\sigma_N}\right)^3\right]$ is the skewness and $\eta_N = E\left[\left(\frac{N_j - \theta_N}{\sigma_N}\right)^4\right]$ is the kurtosis of cluster size distribution. The fourth row shows whether $\hat{\mu}$ is unbiased or approximately unbiased (ie, for k sufficiently large). SRS, simple random sampling; TSS1, two-stage sampling 1; TSS2, two-stage sampling 2; TSS3, two-stage sampling 3.

TABLE 2 Relative efficiencies of two-stage sampling (TSS) schemes versus simple random sampling (SRS) and each other[a]

RE (TSS1 vs SRS)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)+\mathrm{corr}(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho[\tau_N(\zeta_N-\tau_N)+1]}$
RE (TSS2 vs SRS)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)+\mathrm{corr}(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left[1+\left(\overline{n}\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)-1\right)\rho\right]+\mathrm{corr}(u_j,N_j)^2\overline{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$
RE (TSS3 vs SRS)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)+\mathrm{corr}(u_j,N_j)^2\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$
RE (TSS2 vs TSS1)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left[1+\left(\overline{n}\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)-1\right)\rho\right]+\mathrm{corr}(u_j,N_j)^2\overline{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$
RE (TSS3 vs TSS1)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho[\tau_N(\zeta_N-\tau_N)+1]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$
RE (TSS3 vs TSS2)	$\dfrac{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left[1+\left(\overline{n}\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)-1\right)\rho\right]+\mathrm{corr}(u_j,N_j)^2\overline{n}\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}{\left(1-\mathrm{corr}(u_j,N_j)^2\right)\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)[1+(n-1)\rho]+\mathrm{corr}(u_j,N_j)^2 n\rho\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N-\frac{k-3}{k-1}+\tau_N(\tau_N-2\zeta_N)\right)+2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N-\tau_N)+1\right]}$

[a] Derivations are given in section 2 of the Supplementary Material. Recall that ρ is the intraclass correlation, defined as $\frac{\sigma_v^2}{\sigma_y^2} \in (0,1)$, where $\sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2$ is the total unexplained outcome variance.

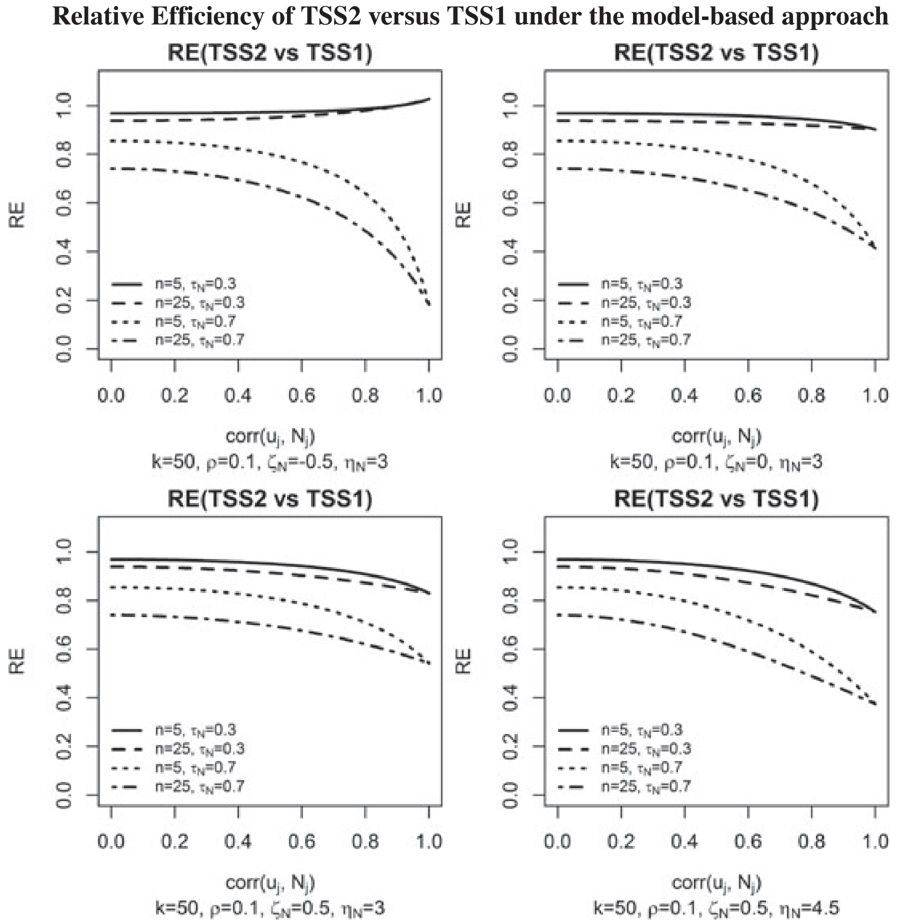

FIGURE 1 Model-based Relative Efficiency of TSS2 versus TSS1 for a given total sample size $\overline{n}k$, as a function of the (absolute value of the) correlation between cluster effect and cluster size (ie, $\mathrm{corr}(u_j, N_j)$), for different values of the average number of individuals sampled per cluster (ie, $\overline{n}$) and of the coefficient of variation of cluster size (ie, τ_N) (curves), and different cluster size distributions (panels). The values of the relative efficiency at $\mathrm{corr}(u_j, N_j) = 0$ and $\mathrm{corr}(u_j, N_j) = 1$ refer to ω and λ, respectively

is a function of the coefficient of variation τ_N, the skewness ζ_N, and (for TSS2 and TSS3 only) the kurtosis η_N of cluster size distribution. Denote by ω the relative efficiency under noninformative cluster size (ie, $RE = \omega$ if $\text{corr}(u_j, N_j) = 0$), and by λ the relative efficiency under a perfect linear relation between u_j and N_j (ie, $RE = \lambda$ if $\text{corr}(u_j, N_j)^2 = 1$). These two extremes can be derived directly from Table 2 and Figure 1, which plots the RE against $\text{corr}(u_j, N_j)$. Therefore, the RE moves from ω to λ as $\text{corr}(u_j, N_j)$ moves from zero to one. For small to moderate correlations (say, $|\text{corr}(u_j, N_j)| < 0.7$), ω receives more weight in the relative efficiency. If ω and λ are both smaller than or equal to one, the relative efficiency is also smaller than or equal to one. Now, the ω's shown in Table 2 are all smaller than one, which entails the following ordering of the sampling schemes in terms of efficiency based on ω (from most to least efficient): SRS, TSS1, TSS2, and TSS3. Under a perfect linear relation between cluster effect and cluster size (ie, $\text{corr}(u_j, N_j)^2 = 1$), $RE = \lambda$, and SRS is more efficient than TSS1, whereas TSS2 and TSS3 are equally efficient. Furthermore, TSS1 is more efficient than TSS2 and TSS3 (ie, $\lambda \leq 1$) if one of the following conditions is met (for proofs, see section 2 of the Supplementary Material): the cluster size distribution is positively skewed (ie, $\zeta_N > 0$) with $\tau_N \in [0, \zeta_N]$, or is symmetric (ie, $\zeta_N = 0$) with $\tau_N \in [0, 1]$ and $k \in \left[1, \frac{(2-\tau_N^2)+\sqrt{2-\tau_N^2}}{(1-\tau_N^2)}\right]$, or is Normal. Thus, for any value of $\text{corr}(u_j, N_j)$, the ordering of the sampling schemes in terms of efficiency based on $V(\hat{\mu})$ is (from most to least efficient) as follows: SRS, TSS1, TSS2, and TSS3. However, if none of the aforementioned conditions is met, λ might be bigger than one and then, to see whether TSS1 is more efficient than TSS2 and TSS3, the relative efficiency must be computed for the specific cluster size distribution.

Given that $RE = \omega$ if $\text{corr}(u_j, N_j) = 0$ and ω has more weight than λ in the RE for $|\text{corr}(u_j, N_j)| < 0.7$, it is useful to have a closer look at the patterns of the ω's shown in Table 2. First, the ω of any TSS scheme versus SRS is a decreasing function of the intraclass correlation ρ, the average number of individuals sampled per cluster $\bar{n}$, and (only for TSS2 and TSS3) of the coefficient of variation of cluster size τ_N. Second, ω(TSS2 vs TSS1), ω(TSS3 vs TSS1), and ω(TSS3 vs TSS2) are decreasing functions of the coefficient of variation of cluster size τ_N. Third, as the intraclass correlation ρ and/or the average number of individuals sampled per cluster $\bar{n}$ increase, TSS2 moves away from TSS1 and toward TSS3 in terms of efficiency as expressed by ω (see Figure 2).

When the outcome variable is unrelated to the cluster size (ie, $\gamma = 0$ and so also $\text{corr}(u_j, N_j) = 0$), the population mean μ is equal to β_0, as shown in Section 3. In this special case, any estimator of μ of the form $\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{y}_j}{\sum_{j=1}^{k} w_j}$ is unbiased. However, some weights are more efficient than others. For TSS2, weighting cluster means by their inverse variance (ie, $w_j = \text{Var}(\bar{y}_j)^{-1} = \left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_j}\right)^{-1}$, where $\sigma_u^2 = \sigma_v^2$ because $\gamma = 0$) is optimal, and unweighted analysis (ie, $w_j = 1$) is more or less efficient than cluster size weighting (ie, $w_j = pN_j$), depending on the intraclass correlation ρ and the average cluster size in the sample.[3,23] The conditional variance of the optimal estimator is $\text{Var}\left(\frac{\sum_{j=1}^{k} \bar{y}_j \text{Var}(\bar{y}_j)^{-1}}{\sum_{j=1}^{k} \text{Var}(\bar{y}_j)^{-1}} \middle| \mathbf{N}\right) = \left(\sum_{j=1}^{k} \frac{pN_j}{pN_j\sigma_u^2+\sigma_\varepsilon^2}\right)^{-1}$.[3, (eq.(6))] Under TSS1 and TSS3, the same number of individuals is sampled per cluster (ie, $n_j = n, \forall j = 1, \dots, k$), so the estimator with $w_j = \text{Var}(\bar{y}_j)^{-1}$ reduces to $\frac{\sum_{j=1}^{k} \bar{y}_j}{k}$. Thus, for TSS1 and TSS3, $w_j = 1$ is optimal and its sampling variance is given in the fifth row of the TSS1 column in Table 1 (for proof, see Appendix B or section 2.3 of the Supplementary Material),

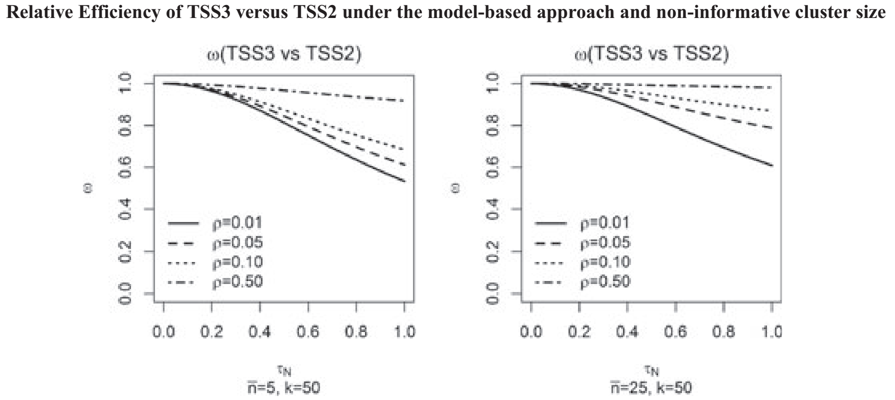

FIGURE 2 Model-based Relative Efficiencies of TSS3 versus TSS2, for a given total sample size $\bar{n}k$ and noninformative cluster size (ie, $\gamma = 0$), as a function of the coefficient of variation of cluster size (ie, τ_N), for different values of the intraclass correlation (ie, ρ) (curves) and for different average numbers of individuals sampled per cluster (ie, $\bar{n}$) (panels)

so TSS1 and TSS3 are equally efficient then, given equal weighting of cluster means, but TSS3 is more practical because, unlike TSS1, it does not require the knowledge of all cluster sizes in the population. The optimal estimator of TSS2 is less efficient than that of TSS3 and TSS1 (ie, RE $\left(\frac{\sum_{j=1}^{k} \overline{y}_j \mathrm{Var}(\overline{y}_j)^{-1}}{\sum_{j=1}^{k} \mathrm{Var}(\overline{y}_j)^{-1}} \text{vs} \frac{\sum_{j=1}^{k} \overline{y}_j}{k} \right) \leq 1$, for proof see section 2.3 of the Supplementary Material). Therefore, TSS3 combined with $\frac{\sum_{j=1}^{k} \overline{y}_j}{k}$ is the best strategy to estimate μ if cluster size is not informative. To prevent misunderstanding, note that the ordering of sampling schemes in this last paragraph only holds if noninformative cluster size is combined with optimal weighting of cluster means. Those weights differ from the ones in Table 1 first row, on which Table 2 and Figures 1 and 2 are based, and which are needed for unbiased estimation of the population mean if cluster size is informative.

5 | DESIGN-BASED INFERENCE FOR TSS WHEN CLUSTER SIZE IS INFORMATIVE

The aim of this section is to study the relative efficiencies of the three TSS schemes compared with SRS and with each other under the design-based approach. It is important to emphasize that the inferential framework of this section is different from the model-based approach adopted in the rest of this paper. So far, the outcome variable Y_{ij} and cluster size N_j were both seen as random variables, and inference was based on the probability distribution of Y_{ij} given in model (1). In contrast, in the design-based approach (ie, this section), the outcome variable Y_{ij} and cluster size N_j are fixed quantities, the inclusion indicator is the only random variable (eg, for cluster j, it is defined as $I_j = 1$ if cluster j is included into the sample, which occurs with probability π_j, and $I_j = 0$ otherwise), and inference is based on the probability distribution induced by the sampling scheme.

The notation of this section remains the same as before with the important distinction that all population quantities here must be interpreted as relating to the finite population. Thus, the two types of population means can be expressed as $\mu = \frac{\sum_{j=1}^{K} N_j \overline{Y}_j}{\sum_{j=1}^{K} N_j}$ and $\beta_0 = \frac{\sum_{j=1}^{K} \overline{Y}_j}{K}$, respectively, where $\overline{Y}_j$ is the mean of all N_j individuals within cluster j. Furthermore, in the population the outcome variable for the ith individual within the jth cluster can be decomposed (combining model (1) with Assumption 4) as follows:

$$Y_{ij} = \beta_0 + \gamma(N_j - \theta_N) + \nu_j + \varepsilon_{ij}, \tag{7}$$

where ν_j is the cluster effect with $E(\nu_j) = \frac{\sum_{j=1}^{K} \nu_j}{K} = 0$ and $V(\nu_j) = \frac{\sum_{j=1}^{K} \nu_j^2}{K} = \sigma_\nu^2$, and $\nu_j \perp N_j$, whereas ε_{ij} is the individual effect with $E(\varepsilon_{ij}) = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}}{N_j} = 0$, $V(\varepsilon_{ij}) = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}^2}{N_j} = \sigma_\varepsilon^2$, and $\nu_j \perp \varepsilon_{ij}$, which entails that $\overline{Y}_j$ here represents $\beta_0 + u_j$ in model (1). Note that, in this section, no distributional assumptions are made for Equation (7), all quantities (ie, Y_{ij}, N_j, ν_j, and ε_{ij}) are just fixed constants, the only random variable is the inclusion indicator and its probability distribution is the foundation of inference. From Equation (7), it follows that $\mu = \beta_0 + \gamma\theta_N\tau_N^2$, an expression that is similar to Equation (5) but refers to the finite population (for proof, see section 3 of the Supplementary Material). Hence, under both inferential paradigms, the two population means coincide (ie, $\mu = \beta_0$) only if either there is no cluster size variation in the population (ie, $\tau_N = 0$), or cluster size is noninformative (ie, $\gamma = 0$).

For each sampling scheme, Table 3 shows in the first row the estimator of the population mean μ, in the second row the sampling variance of $\hat{\mu}$ as available in the design-based literature,[7-9,25] and in the third row again the sampling variance of $\hat{\mu}$ but under the assumption that Equation (7) describes the outcome variable Y_{ij} in the population (for proofs, see section 3 of the Supplementary Material). For large enough k (say, $k \geq 30$), the model-based variances given in Table 1 are equal to the design-based variances given in the third row of Table 3. Furthermore, the estimators of Table 3 are the same as those of the model-based approach (ie, Table 1, first row). The estimators under SRS and TSS1 are unbiased,[7(p308),8(p236)] whereas the estimator under TSS2 and TSS3, the so-called ratio estimator, is only approximately unbiased,[8(p186),25(pp323,324)] and then the number of sampled clusters k is assumed to be large enough to neglect this bias. It is important to emphasize that, under the design-based paradigm, the properties of an estimator (ie, approximate unbiasedness, variance as given in the second row of Table 3) are based only on the sampling scheme.[8(p147),9(p239)] The assumption that the outcome variable is described by Equation (7) (ie, Table 3, third row) is needed to allow a fair comparison with the results obtained under the model-based approach. However, the assumption of a model, like Equation (7), to evaluate competing sampling schemes is appropriate under the design-based framework, provided that inference is then based on the sampling scheme only.[7(p256),8(p205),26,27]

TABLE 3 Population mean μ estimator and sampling variance per sampling scheme under the design-based approach[a]

	SRS	TSS1	TSS2	TSS3
$\hat{\mu}$	$\sum_{i=1}^{m} \frac{y_i}{m}$ 7, (p22),8, (eq.(2.8),p35), 25, (eq.(5),p21),	$\sum_{j=1}^{k} \frac{\overline{y}_j}{k}$ 7, (eq.(11.39),p308),8, (p236), 25, (eq.(2),p359),	$\frac{\sum_{j=1}^{k} N_j \overline{y}_j}{\sum_{j=1}^{k} N_j}$ 7, (eq.(11.25),p303), 8, (eq.(5.26),p186), 25, (eq.(76),p317),	$\frac{\sum_{j=1}^{k} N_j \overline{y}_j}{\sum_{j=1}^{k} N_j}$ 7(eq.(11.25),p303),8(eq.(5.26),p186), 25, (eq.(76),p317),
$V(\hat{\mu})$	$\frac{\sum_{i=1}^{N_{\text{pop}}} (Y_i-\mu)^2}{m(N_{\text{pop}}-1)}$ 7, (eq.(2.8),p23), 8, (eq.(2.9),p36), 25, (eq.(39),p29),	$\frac{1}{k} \sum_{j=1}^{K} \frac{N_j}{N_{\text{pop}}} (\overline{Y}_j - \mu)^2$ $+\frac{1}{k} \sum_{j=1}^{K} \frac{N_j}{N_{\text{pop}}} \sum_{i=1}^{N_j} \frac{(Y_{ij}-\overline{Y}_j)^2}{n(N_j-1)}$ 7, (eq.(11.33),p307), 25, (eq.(14),p362),	$\frac{1}{k} \sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \frac{(\overline{Y}_j-\mu)^2}{(K-1)}$ $+\frac{1}{kK} \sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \sum_{i=1}^{N_j} \frac{(Y_{ij}-\overline{Y}_j)^2}{n_j(N_j-1)}$ 7, (eq.(11.27),p304), 25, (eq.(96),p325),	$\frac{1}{k} \sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \frac{(\overline{Y}_j-\mu)^2}{(K-1)}$ $+\frac{1}{kK} \sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \sum_{i=1}^{N_j} \frac{(Y_{ij}-\overline{Y}_j)^2}{n(N_j-1)}$ 7, (eq.(11.27),p304), 25, (eq.(96),p325),
$V(\hat{\mu})$ Under Equation (7)	$\frac{\sigma_\nu^2+\sigma_\varepsilon^2+\gamma^2\sigma_N^2\left[\tau_N\left(\zeta_N-\tau_N\right)+1\right]}{m}$	$\frac{n\sigma_\nu^2+\sigma_\varepsilon^2}{nk} + \frac{\gamma^2\sigma_N^2\left[\tau_N\left(\zeta_N-\tau_N\right)+1\right]}{k}$	$\frac{p\theta_N(\tau_N^2+1)\sigma_\nu^2+\sigma_\varepsilon^2}{p\theta_N k}$ $+\frac{\gamma^2\sigma_N^2\left(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1\right)}{k}$	$\frac{(\tau_N^2+1)\left(n\sigma_\nu^2+\sigma_\varepsilon^2\right)}{nk}$ $+\frac{\gamma^2\sigma_N^2\left(\tau_N^4+\tau_N^2(\eta_N-3)+2\zeta_N\tau_N(1-\tau_N^2)+1\right)}{k}$

[a] Note that $m = \overline{n}k$ is the number of individuals sampled with SRS, k is the number of clusters sampled with a TSS scheme, and $\overline{n} = \frac{\sum_{j=1}^{k} n_j}{k}$. For any TSS scheme, we assume $\frac{k}{K} \rightarrow 0$ and $\frac{\overline{n}}{\theta_N} \rightarrow 0$ or sampling with replacement at each stage, and for SRS $\frac{m}{N_{\text{pop}}} \rightarrow 0$ or sampling with replacement. In the third row, the outcome variable is assumed to be described by Equation (7). For large enough k, the variances in the third row are equal to those in the last two rows of Table 1. Note that $\zeta_N = \left(\frac{1}{\sigma_N^3}\right)\left(\frac{\sum_{j=1}^{K} (N_j-\theta_N)^3}{K}\right)$ is the skewness, and $\eta_N = \left(\frac{1}{\sigma_N^4}\right)\left(\frac{\sum_{j=1}^{K} (N_j-\theta_N)^4}{K}\right)$ is the kurtosis of cluster size distribution in the population. Derivations are given in section 3 of the Supplementary Material. SRS, simple random sampling; TSS1, two-stage sampling 1; TSS2, two-stage sampling 2; TSS3, two-stage sampling 3.

Similarly to Section 4, the relative efficiency of two competing sampling schemes is defined as the ratio of their variances (as given in the third row of Table 3). For large enough k (say, $k \geq 30$), it turns out that these relative efficiencies (given in Table S.M.2 and shown in Figures S.M.1-2 of the Supplementary Material) are approximately equal to those shown in Table 2 because the variances in Table 1 and those in the third row of Table 3 are approximately equal. The only distinction to be made is that $\text{corr}(u_j, N_j)$ is replaced with the correlation between cluster mean and cluster size $\text{corr}(\overline{Y}_j, N_j)$. Like in Section 4, numerator and denominator of the relative efficiency are both made up of two components, weighted by $\text{corr}(\overline{Y}_j, N_j)^2$ and $(1 - \text{corr}(\overline{Y}_j, N_j)^2)$, respectively, and only the component weighted by $\text{corr}(\overline{Y}_j, N_j)^2$ depends on the skewness and kurtosis of the cluster size distribution. The extreme cases of the relative efficiency, namely, under noninformative cluster size and a perfect relation between cluster mean and cluster size, are denoted by ω and λ, respectively. The patterns and the ordering of the relative efficiencies are then those of Section 4. Specifically, for any value of $\text{corr}(\overline{Y}_j, N_j)$, SRS is the most efficient sampling scheme, followed by TSS1 (under the conditions given in Section 4), TSS2, and finally TSS3.

To conclude, even though the mathematical foundations of the two inferential approaches are different, in the considered setting, they yield almost the same results, ie, the population mean estimators are the same, as well as the relative efficiencies, provided that k is large enough and Equation (7) holds in the population. An advantage of the design-based approach is robustness because the unbiasedness and the variance of a design-based estimator do not depend on the assumptions of a model. Nevertheless, the model-based approach has a practical advantage when designing a survey, more specifically for choosing a sampling scheme and computing the sample size. The sampling variances in Table 1 (last two rows) and Table 3 (last row), and the relative efficiencies in Table 2, all based on Equation (7), can be obtained by specifying the intraclass correlation ρ, the correlation $\text{corr}(u_j, N_j)$, and four parameters of cluster size distribution (ie, θ_N, τ_N, ζ_N, and η_N). In contrast, the sampling variances in Table 3 (second row) from the design-based approach require the knowledge of cluster size N_j and cluster mean $\overline{Y}_j$ for all the K clusters in the population. If that information were available, then the population mean μ would also be known, making the survey superfluous.

6 | APPLICATION TO TWO REAL CLUSTER SIZE DISTRIBUTIONS

With the aim of planning a survey to estimate the population mean μ of a quantitative outcome variable Y_{ij} in a two-level population, we want to establish whether TSS1 is more efficient than TSS2 for the population under study and assess its efficiency gain relative to TSS2. The outcome variable Y_{ij} is assumed to be decomposed, as shown in Equation (7), but the analysis is carried out for both the model-based and the design-based approach. Two real cluster size distributions are considered, ie, the distribution of public high school size in Italy and the distribution of patient list size for general practices in England.

School size and alcohol consumption. In adolescent health literature, it has been shown that greater connection between students and school (eg, positive relations with teachers and peers, participation in school activities) is associated with less emotional distress, substance consumption (eg, alcohol, cigarettes, marijuana), violence, and suicidal intentions.[28] Furthermore, it has been found that *school connectedness* and school size are inversely related,[29,30] which suggests that school size can be informative for health risk behaviors in adolescents. Suppose that we want to estimate the average weekly alcohol consumption (in liters) among high school students in Italy. According to the Italian Ministry of Education,[31] in the school year 2016/2017 in Italy, there were $6,235 = K$ public high schools with a total of $2,515,060 = N_{\text{pop}}$ students enrolled. The distribution of public high school size in Italy (with parameters $\theta_N = 403$, $\tau_N = 0.912$, $\zeta_N = 1.256$, and $\eta_N = 4.315$) is plotted in Figure 3 (first column, first row). The first row of Figure 3 also shows the relative efficiency of TSS2 versus TSS1, for a sample of $50 = k$ schools and $20 = \overline{n}$ students per school, as a function of the (absolute value of the) correlation between school size and school specific-mean, for different values of the intraclass correlation, under the model-based (second column) and the design-based approach (third column). As can be seen from Figure 3, under both inferential approaches TSS1 is more efficient than TSS2 and allows a sizeable efficiency gain (about 15%) even for noninformative school size and a small intraclass correlation ($\rho = 0.01$).

Patient list size for general practices and government expenditure on health. According to Eurostat,[32] in 2016, health was the second largest area of government expenditure in the United Kingdom with a share of 7.6% of the Gross Domestic Product (GDP). Spending for hospital services represented the largest component of the government expenditure on health, with a share of 5.7% of the GDP.[32] In reducing such costs, general practices can play a role by effectively

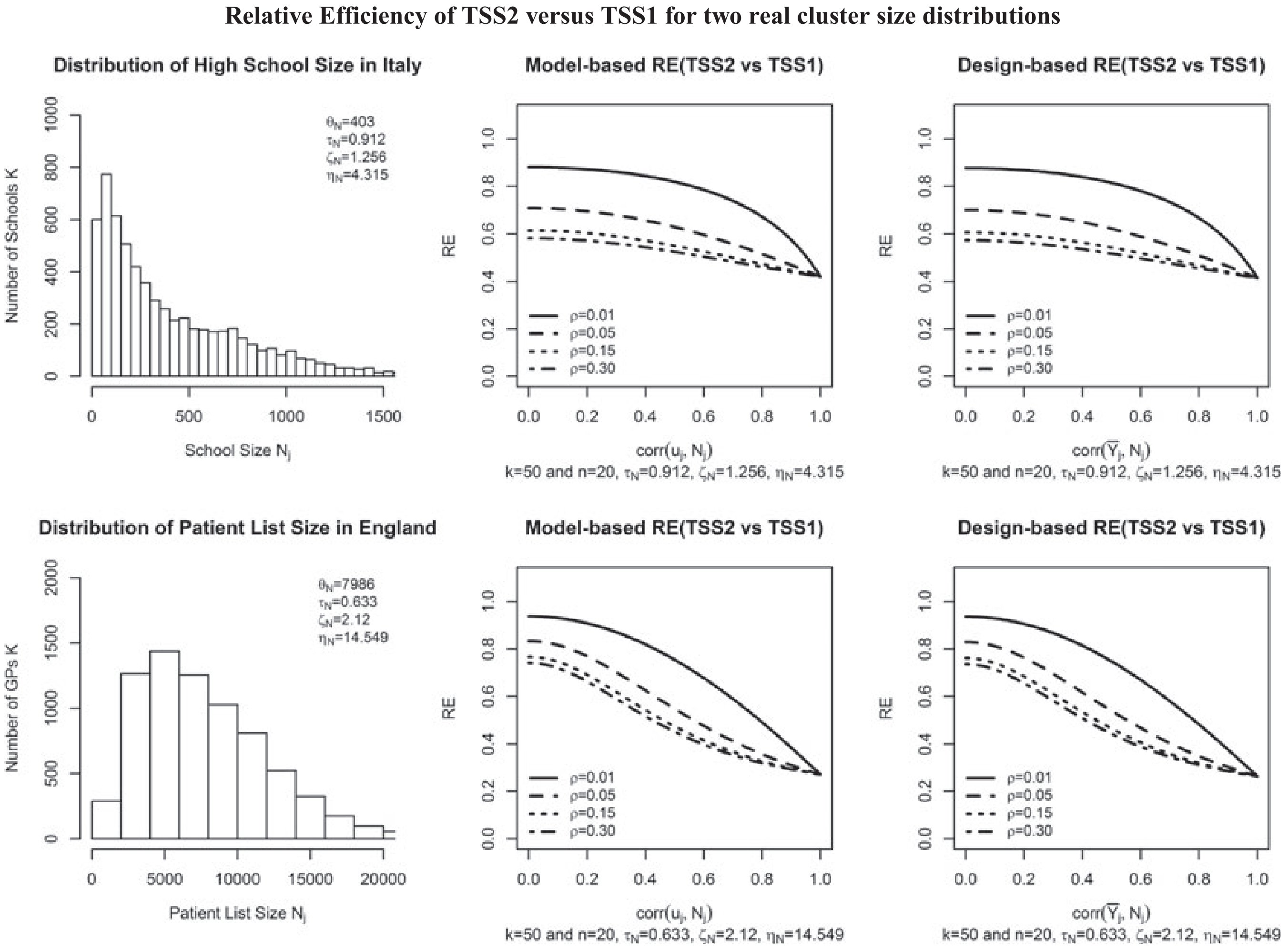

FIGURE 3 First column: Distribution of public high school size in Italy (first row), distribution of patient list size for general practices in England (second row). Second column: Model-based Relative Efficiency of TSS2 versus TSS1, as a function of the (absolute value of the) correlation between cluster effect and cluster size (ie, corr(u_j, N_j)), for different values of the intraclass correlation coefficient ρ (curves). Third column: Design-based Relative Efficiency of TSS2 versus TSS1, as a function of the (absolute value of the) correlation between cluster mean and cluster size (ie, corr($\overline{Y}, N_j$)), for different values of the intraclass correlation coefficient ρ (curves). TSS1, two-stage sampling 1; TSS2, two-stage sampling 2

treating those conditions, which can lead to avoidable hospitalisations (eg, influenza, diabetic complications). Kelly and Stoye[33] have found that small practices (defined as those with three or fewer full-time equivalent (FTE) practitioners) had higher rates of hospitalizations for such preventable conditions in 2010/2011 in England. This suggests that patient list size can be informative for government expenditure on health, given that patients per general practice were proportional to the number of FTE practitioners (see figure 2.6 and table 2.3 in the work of Kelly and Stoye[33]). Suppose we want to estimate the average per capita government expenditure on health in England. According to the Health and Social Care Information Centre,[34] in October 2017, $58,719,921 = N_{\text{pop}}$ patients were registered at $7,353 = K$ general practices in England. The distribution of patient list size for general practices in England (with parameters $\theta_N = 7,986$, $\tau_N = 0.633$, $\zeta_N = 2.12$, and $\eta_N = 14.549$) is plotted in Figure 3 (first column, second row). The second row of Figure 3 shows the relative efficiency of TSS2 versus TSS1, for a sample of $50 = k$ practices and $20 = \overline{n}$ patients per practice, as a function of the (absolute value of the) correlation between patient list size and general practice specific-mean, for different values of the intraclass correlation, under the model-based (second column) and the design-based approach (third column). As shown in the second row of Figure 3, TSS1 is more efficient than TSS2 under both inferential paradigms and its efficiency gain increases as the intraclass correlation and/or the correlation between patient list size and the general practice specific-mean increase.

To conclude, the two examples show that TSS2 leads to important efficiency losses relative to TSS1, and that, in planning a survey, it is more practical to use variances based on a model, like those given in Table 1 or third row of Table 3, than the design-based variances in the second row of Table 3, which require the prior knowledge of all cluster sizes N_j as well as all cluster means $\overline{Y}_j$ in the population.

7 | DISCUSSION

In multilevel populations, two types of overall means can be defined, ie, the mean of all individual outcomes in the population ignoring cluster membership and the mean of all cluster-specific means. For unbiased estimation of the first population mean, individuals can be sampled not only by SRS but also with three alternative TSS schemes, ie, sampling clusters with probability proportional to cluster size and then taking a SRS of the same number of individuals within sampled clusters (ie, TSS1); drawing a SRS of clusters and then sampling the same percentage of individuals per cluster (ie, TSS2); and taking a SRS of clusters and then of individuals within the sampled clusters (ie, TSS3).

The results of this paper are the following. First, it was shown that the first population mean gives equal weight to all individuals and thus more weight to large clusters than to small clusters, the second mean gives equal weight to all clusters irrespective their size, and these two means coincide only if cluster size does not vary or is unrelated (ie, noninformative) to the outcome variable of interest. Second, for estimation of the first population mean (ie, the average of all individual outcomes), the unweighted average of cluster means is unbiased under TSS1, and weighting cluster means by cluster size is asymptotically unbiased under TSS2 or TSS3. Third, it was shown that the relative efficiency of any TSS scheme versus SRS is a decreasing function of the intraclass correlation, the average number of individuals sampled per cluster, and (only for TSS2 and TSS3) of the coefficient of variation of cluster size. Furthermore, the relative efficiencies of TSS2 and TSS3 versus TSS1 and of TSS3 versus TSS2 are decreasing functions of the coefficient of variation of cluster size, but the efficiency loss of TSS3 compared with TSS2 improves with an increase of the intraclass correlation and/or the average number of individuals sampled per cluster. All relative efficiencies also depend on other features of the cluster size distribution, in particular, on its skewness and (only for those involving TSS2 and TSS3) kurtosis. Nevertheless, SRS is always the most efficient sampling scheme, followed (for many cluster size distributions) by TSS1, and then by TSS2, which, in turn, is always more efficient than TSS3. With respect to choosing between the three TSS schemes, we do not expect TSS1 to be less efficient than TSS2 in practice, and thus we recommend TSS1 provided all cluster sizes are known before sampling. Fourth, it was shown that model-based and design-based inference in survey sampling yield almost the same results, at least if the model assumptions are met.

Although design-based inference has the advantage of being robust against violations of the model assumptions, comparing the four sampling schemes in terms of their relative efficiencies, as well as sample size planning, can only be done taking a model-based approach. Sample size planning within the design-based approach would require knowledge of the size and outcome mean of all clusters in the population (see Table 3, second row), which, in turn, would imply that the population mean is already known. Furthermore, models are also needed to deal with missing data and measurement error.[9]

The results of this paper could be extended by (i) deriving the optimal design of these three TSS schemes under a cost constraint and comparing their efficiencies under that constraint instead of the present constraint of a fixed total sample size, (ii) investigating different variance estimation methods, (iii) considering binary outcome variables, and (iv) deriving the optimal design for a scheme, which samples different numbers and percentages of individuals at the second stage, that is, a sampling scheme in-between TSS2 and TSS3.

ORCID

Francesco Innocenti https://orcid.org/0000-0001-6113-8992
Math J.J.M. Candel https://orcid.org/0000-0002-2229-1131
Gerard J.P. van Breukelen https://orcid.org/0000-0003-0949-0272

REFERENCES

1. Goldstein H. *Multilevel Statistical Models*. 4th ed. Chichester, UK: John Wiley & Sons; 2011.
2. Snijders TAB, Bosker RJ. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. 2nd ed. London, UK: SAGE Publishers; 2012.
3. van Breukelen GJP, Candel MJJM, Berger MPF. Relative efficiency of unequal *versus* equal cluster sizes in cluster randomized and multicentre trials. *Statist Med*. 2007;26(13):2589-2603.
4. Panageas KS, Schrag D, Localio AR, Venkatraman ES, Begg CB. Property of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statist Med*. 2007;26(9):2017-2035.
5. Seaman S, Pavlou M, Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statist Med*. 2014;33(30):5371-5387.
6. Nevalainen J, Datta S, Oja H. Inference on the marginal distribution of clustered data with informative cluster size. *Stat Pap*. 2014;55(1):71-92.
7. Cochran WG. *Sampling Techniques*. 3rd ed. New York, NY: John Wiley & Sons; 1977.
8. Lohr SL. *Sampling: Design and Analysis*. 2nd ed. Boston, MA: Brooks/Cole; 2010.
9. Särndal C-E, Swensson B, Wretman J. *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag New York; 1992.
10. Skinner C, Wakefield J. Introduction to the design and analysis of complex survey data. *Stat Sci*. 2017;32(2):165-175.
11. Valliant R, Dorfman AH, Royall RM. *Finite Population Sampling and Inference: A Prediction Approach*. New York, NY: John Wiley & Sons; 2000.
12. Chambers RL, Clark RG. *An Introduction to Model-Based Survey Sampling With Applications*. Oxford, UK: Oxford University Press; 2012.
13. Little RJ. To model or not to model? Competing modes of inference for finite population sampling. *J Am Stat Assoc*. 2004;99(466): 546-556.
14. Sudgen RA, Smith TMF. Ignorable and informative designs in survey sampling inference. *Biometrika*. 1984;71(3):495-506.
15. Pfeffermann D. The role of sampling weights when modeling survey data. *Int Stat Rev*. 1993;61(2):317-337.
16. Pfeffermann D, Skinner CJ, Holmes DJ, Goldstein H, Rasbash J. Weighting for unequal selection probabilities in multilevel models. *J R Stat Soc Series B Stat Methodol*. 1998;60(1):23-40.
17. Grilli L, Pratesi M. Weighted estimation in multilevel ordinal and binary models in the presence of informative designs. *Surv Methodol*. 2004;30(1):93-103.
18. Asparouhov T. General multi-level modeling with sampling weights. *Commun Stat Theory Methods*. 2006;35(3):439-460.
19. Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. *J R Stat Soc Ser A Stat Soc*. 2006;169(4):805-827.
20. Koziol NA, Bovaird JA, Suarez S. A comparison of population-averaged and cluster-specific approaches in the context of unequal probabilities of selection. *Multivariate Behav Res*. 2017;52(3):325-349.
21. Makela S, Si Y, Gelman A. Bayesian inference under cluster sampling with probability proportional to size. *Statist Med*. 2018;37(26):3849-3868.
22. Zheng H, Little RJA. Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Surv Methodol*. 2004;2(30):209-218.
23. Searle SR, Pukelsheim F. Effect of intraclass correlation on weighted averages. *Am Stat*. 1986;40(2):103-105.
24. van Breukelen GJP, Candel MJJM. Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests. *Statist Med*. 2012;31(4):397-400.
25. Sukhatme PV. *Sampling Theory of Surveys With Applications*. Ames, IA: Iowa State College Press; 1954.
26. Hansen MH, Madow WG, Tepping BJ. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *J Am Stat Assoc*. 1983;78(384):776-793.
27. Smith TMF. Sample surveys 1975-1990; an age of reconciliation? *Int Stat Rev*. 1994;62(1):5-19.
28. Resnick MD, Bearman PS, Blum RW, et al. Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health. *JAMA*. 1997;278(10):823-832.
29. McNeely CA, Nonnemaker JM, Blum RW. Promoting school connectedness: evidence from the National Longitudinal Study on Adolescent Health. *J Sch Health*. 2002;72(4):138-146.
30. Thompson DR, Iachan R, Overpeck M, Ross JG, Gross LA. School connectedness in the health behavior in school-aged children study: the role of student, school, and school neighborhood connectedness. *J Sch Health*. 2006;76(7):379-386.
31. Direzione generale per i contratti gli acquisti e per i sistemi informativi e la statistica. Studenti per anno di corso e fascia di etá: scuola statale. DGCASIS; 2018. http://dati.istruzione.it/opendata/opendata/catalogo/elements1/?area=Studenti. Accessed May 1, 2018.
32. Eurostat. Government expenditure by function. 2018. http://ec.europa.eu/eurostat/statistics-explained/index.php/Government_expenditure_by_function. Accessed July 1, 2018.
33. Kelly E, Stoye G. Does GP practice size matter? GP practice size and the quality of primary care. IFS report R101. London, UK: Institute for Fiscal Studies; 2014. https://www.ifs.org.uk/uploads/publications/comms/R101.pdf. Accessed July 1, 2018.
34. Salt K. Patients registered at a GP practice, October 2017; special topic-practice list size comparison. Health and Social Care Information Centre. 2017. https://files.digital.nhs.uk/publication/b/k/gp-reg-pat-prac-topic-int-oct-17.pdf. Accessed November 8, 2017.
35. Casella G, Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury; 2002.

36. Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics*. 3rd ed. New York, NY: McGraw-Hill; 1974.
37. Zhang L. Sample mean and sample variance: their covariance and their (in)dependence. *Am Stat*. 2007;61(2):159-160.
38. Pearson K. Editorial note to 'Inequalities for moments of frequency functions and for various statistical constants'. *Biometrika*. 1929;21:361-375.
39. Gurland J. The teacher's corner: an inequality satisfied by the expectation of the reciprocal of a random variable. *Am Stat*. 1967;21(2):24-25.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Innocenti F, Candel MJJM, Tan FES, van Breukelen GJP. Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine*. 2019;38:1817–1834. https://doi.org/10.1002/sim.8070

APPENDIX A

DERIVATION OF THE POPULATION MEAN μ

Assuming that $u_j = \gamma(N_j - \theta_N) + v_j$ (ie, Assumption 4) and then plugging this equation into model (1) give

$$y_{ij} = \beta_0 + \gamma(N_j - \theta_N) + v_j + \varepsilon_{ij}. \tag{A1}$$

Now, before deriving the population mean μ under model (1) with informative cluster size as in (A1), we will first show how the sampling scheme affects the sampling distribution of cluster size, which, in turn, influences the cluster effect sampling distribution if cluster size is informative.

Denote by $f(N_j|\theta_N, \sigma_N^2)$ the probability density function of cluster size N_j, where θ_N and σ_N^2 are its mean and variance, respectively, and by $\mathbf{N} = (N_1, \dots, N_k)^T$ the vector of the cluster sizes of the k sampled clusters. Under TSS2 or TSS3, the k clusters are sampled with equal probabilities from the population of K clusters, then

$$f_{\text{TSS2/TSS3}}(N_j) = f(N_j), \quad \text{and} \quad f_{\text{TSS2/TSS3}}(N_1, \dots, N_k) = \prod_{j=1}^{k} f(N_j),$$

where the subscript of $f_*(.)$ (here, TSS2/TSS3) indicates how the k clusters are drawn at the first stage (here, with equal probabilities, ie, under TSS2 or TSS3). Thus, under TSS2 or TSS3, clusters are weighted equally in the cluster size sampling distribution, and then integrating over that distribution gives

$$\text{(a)} \quad E_{\text{TSS2/TSS3}}(N_j) = \theta_N, \quad \text{and} \quad \text{(b)} \quad V_{\text{TSS2/TSS3}}(N_j) = \sigma_N^2. \tag{A2}$$

In contrast, under SRS m individuals are sampled directly from the population of N_{pop} individuals. The probability that a selected individual belongs to a cluster of size N_j is $\frac{N_j f(N_j) dN_j}{\int N_j f(N_j) dN_j}$ and then, under SRS, k_{SRS} clusters are indirectly sampled from the population, where k_{SRS} can run from 1 to m. Thus, large clusters have higher chance of being represented in an SRS sample, as well as, in a TSS1 sample, because under both sampling schemes clusters are sampled (directly or indirectly) with probabilities proportional to their size. Denote by k_* the number of clusters sampled with an arbitrary sampling scheme, then $k_* = k_{\text{SRS}}$ under SRS and $k_* = k$ under any TSS scheme. Thus, under SRS or TSS1, we have that

$$f_{\text{SRS/TSS1}}(N_j) = \frac{N_j f(N_j) dN_j}{\int N_j f(N_j) dN_j}, \quad \text{and} \quad f_{\text{SRS/TSS1}}(N_1, \dots, N_{k_*}) = \prod_{j=1}^{k_*} \left(\frac{N_j f(N_j) dN_j}{\int N_j f(N_j) dN_j} \right),$$

and so each cluster of size N_j is weighted by the factor $\frac{N_j}{\theta_N}$ in the cluster size sampling distribution, which gives (for proofs, see section 1 in the Supplementary Material)

$$\text{(a)} \quad E_{\text{SRS/TSS1}}(N_j) = E_{\text{TSS2/TSS3}}(N_j)\left(\tau_N^2 + 1\right), \quad \text{and} \quad \text{(b)} \quad V_{\text{SRS/TSS1}}(N_j) = V_{\text{TSS2/TSS3}}(N_j)\left[\tau_N(\zeta_N - \tau_N) + 1\right], \tag{A3}$$

TABLE A1 Notation

	Population	Sample
Number of clusters	K	k
Number of individuals within cluster j	N_j	n_j or n
Number of individuals	$N_{\text{pop}} = \sum_{j=1}^{K} N_j$	$m = \overline{n}k = \sum_{j=1}^{k} n_j$
Average cluster size	θ_N	$\overline{N} = \frac{\sum_{j=1}^{k} N_j}{k}$
Cluster size variance	σ_N^2	$S_N^2 = \frac{\sum_{j=1}^{k}\left(N_j - \overline{N}\right)^2}{k}$
Coefficient of variation of cluster size	$\tau_N = \frac{\sigma_N}{\theta_N}$	$CV_N = \frac{S_N}{\overline{N}}$
Skewness of cluster size distribution	$\zeta_N = \frac{E[(N_j-\theta_N)^3]}{\sigma_N^3}$	-
Kurtosis of cluster size distribution	$\eta_N = \frac{E[(N_j-\theta_N)^4]}{\sigma_N^4}$	-
Correlation between cluster effect and cluster size	$\text{corr}(u_j, N_j)$	-
Unexplained between-cluster variance	σ_ν^2	-
Within-cluster variance	σ_ε^2	-
Total unexplained outcome variance	$\sigma_y^2 = \sigma_\nu^2 + \sigma_\varepsilon^2$	-
Intraclass correlation coefficient	$\rho = \frac{\sigma_\nu^2}{\sigma_y^2}$	-

where $\tau_N = \frac{\sigma_N}{\theta_N}$ and $\zeta_N = \frac{E\left[(N_j-\theta_N)^3\right]}{\sigma_N^3}$ are the coefficient of variation and the skewness of cluster size distribution in the population, respectively.

Now, let us consider how the sampling distribution of the cluster effect u_j is affected by the sampling distribution of cluster size N_j. For all sampling schemes, the expectation and the variance of cluster effect u_j conditional on N_j are

$$E(u_j|N_j) = E\left(\gamma(N_j - \theta_N) + \nu_j|N_j\right) = \gamma(N_j - \theta_N) \tag{A4a}$$

$$V(u_j|N_j) = V\left(\gamma(N_j - \theta_N) + \nu_j|N_j\right) = \sigma_\nu^2. \tag{A4b}$$

In contrast, the marginal expectation (ie, $E(u_j) = E(E(u_j|N_j))$) and the marginal variance (ie, $V(u_j) = E(V(u_j|N_j)) + V(E(u_j|N_j))$) of u_j are affected by the sampling scheme because they are obtained by integrating $E(u_j|N_j)$ and $V(u_j|N_j)$ over the cluster size sampling distribution. Thus, if clusters are weighted equally in the cluster size sampling distribution (ie, under TSS2 or TSS3), it follows from (A2) that

$$E_{\text{TSS2/TSS3}}(u_j) = E_{\text{TSS2/TSS3}}\left(\gamma(N_j - \theta_N)\right) = 0$$

$$V_{\text{TSS2/TSS3}}(u_j) = E_{\text{TSS2/TSS3}}\left(\sigma_\nu^2\right) + V_{\text{TSS2/TSS3}}\left(\gamma(N_j - \theta_N)\right) = \sigma_\nu^2 + \gamma^2\sigma_N^2 = \sigma_u^2,$$

that is, Equations (2a) and (2b), respectively. In contrast, if each cluster of size N_j is weighted by the factor $\frac{N_j}{\theta_N}$ in the cluster size sampling distribution (ie, under SRS or TSS1), it follows from (A3) that

$$E_{\text{SRS/TSS1}}(u_j) = E_{\text{SRS/TSS1}}\left(\gamma(N_j - \theta_N)\right) = \gamma\theta_N\tau_N^2$$

$$V_{\text{SRS/TSS1}}(u_j) = E_{\text{SRS/TSS1}}\left(\sigma_\nu^2\right) + V_{\text{SRS/TSS1}}\left(\gamma(N_j - \theta_N)\right) = \sigma_\nu^2 + \gamma^2\sigma_N^2\left[\tau_N(\zeta_N - \tau_N) + 1\right],$$

that is, Equations (4a) and (4b), respectively. The two definitions of population means (ie, Equations (3) and (5)) now follow from $E_{\text{TSS2/TSS3}}(u_j)$ and $E_{\text{SRS/TSS1}}(u_j)$, respectively, given model (1).

APPENDIX B

RESULTS OF TABLE 1

The following facts will be used in this appendix. First, $\mathbf{N} = (N_1, \ldots, N_k)^T$ denotes the vector of the cluster sizes of the k clusters drawn with TSS, $\mathbf{N}_{\text{SRS}} = (N_1, \ldots, N_{k_{\text{SRS}}})^T$ is the vector of the cluster sizes of the k_{SRS} clusters indirectly sampled with SRS, whereas $\mathbf{N}_*$ is used when the sampling scheme is not specified (ie, $\mathbf{N}_* = \mathbf{N}$ for TSS and $\mathbf{N}_* = \mathbf{N}_{\text{SRS}}$ for SRS). Second, note that the four estimators in the first row of Table 1 are all of the form $\hat{\mu} = \frac{\sum_{j=1}^{k_*} w_j \bar{y}_j}{\sum_{j=1}^{k_*} w_j}$, where $k_* = k$ for the three TSS schemes and $k_* = k_{\text{SRS}}$ for SRS (recall that $\frac{m}{N_{\text{pop}}} \rightarrow 0$ (ie, Assumption 2), which entails that $k_{\text{SRS}} \rightarrow m$). Third, from Equation (A1), we have that $\overline{Y}_j = \beta_0 + \gamma(N_j - \theta_N) + \nu_j + \bar{\varepsilon}_j$.

Conditional expectations and unbiasedness.

The conditional expectation of any estimator in Table 1 has the form $E(\hat{\mu}|\mathbf{N}_*) = \frac{\sum_{j=1}^{k_*} w_j E(\bar{y}_j|\mathbf{N}_*)}{\sum_{j=1}^{k_*} w_j}$, where $E(\bar{y}_j|\mathbf{N}_*) = \beta_0 + \gamma(N_j - \theta_N)$. Thus, the second row of Table 1 follows $E(\hat{\mu}_{\text{TSS1}}|\mathbf{N}) = \beta_0 + \gamma(\overline{N} - \theta_N)$ because $w_j = 1$, where $\overline{N} = \frac{\sum_{j=1}^{k} N_j}{k}$; $E(\hat{\mu}_{\text{TSS3}}|\mathbf{N}) = \beta_0 + \gamma(\overline{N}(CV_N^2 + 1) - \theta_N)$ because $w_j = N_j$ and $\frac{\sum_{j=1}^{k} N_j^2}{\sum_{j=1}^{k} N_j} = \frac{S_N^2 + \overline{N}^2}{\overline{N}} = \overline{N}(CV_N^2 + 1)$, where $S_N^2 = \frac{\sum_{j=1}^{k} (N_j - \overline{N})^2}{k}$; $E(\hat{\mu}_{\text{TSS2}}|\mathbf{N}) = E(\hat{\mu}_{\text{TSS3}}|\mathbf{N})$ because $w_j = n_j = pN_j$; and $E(\hat{\mu}_{\text{SRS}}|\mathbf{N}_{\text{SRS}}) = \beta_0 + \gamma(\overline{N}_{\text{SRS}} - \theta_N)$ because $w_j = 1$, where $\overline{N}_{\text{SRS}} = \frac{\sum_{j=1}^{k_{\text{SRS}}} N_j}{k_{\text{SRS}}}$. To prove the unbiasedness of the four estimators (fourth row of Table 1), we need to derive their marginal expectation, that is, integrating the conditional expectation over the cluster size sampling distribution. Thus, Equation (A3a) implies that $\hat{\mu}_{\text{TSS1}}$ and $\hat{\mu}_{\text{SRS}}$ are unbiased because $E(\hat{\mu}_{\text{TSS1}}) = E_{\text{TSS1}}(E(\hat{\mu}_{\text{TSS1}}|\mathbf{N})) = \beta_0 + \gamma(E_{\text{TSS1}}(\overline{N}) - \theta_N) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu$ and $E(\hat{\mu}_{\text{SRS}}) = E_{\text{SRS}}(E(\hat{\mu}_{\text{SRS}}|\mathbf{N}_{\text{SRS}})) = \beta_0 + \gamma(E_{\text{SRS}}(\overline{N}_{\text{SRS}}) - \theta_N) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu$. In contrast, $\hat{\mu}_{\text{TSS3}}$ and $\hat{\mu}_{\text{TSS2}}$ are asymptotically unbiased because $E(\hat{\mu}_{\text{TSS3}}) = E(\hat{\mu}_{\text{TSS2}}) = E_{\text{TSS3}}(E(\hat{\mu}_{\text{TSS3}}|\mathbf{N})) = E_{\text{TSS2}}(E(\hat{\mu}_{\text{TSS2}}|\mathbf{N})) = \beta_0 + \gamma\left(E_{\text{TSS2/TSS3}}\left(\frac{S_N^2 + \overline{N}^2}{\overline{N}}\right) - \theta_N\right) \approx \beta_0 + \gamma\theta_N\left(\frac{k-1}{k}\right)\tau_N^2 \approx \mu$, where $E_{\text{TSS2/TSS3}}\left(\frac{S_N^2 + \overline{N}^2}{\overline{N}}\right) = E\left(\frac{S_N^2 + \overline{N}^2}{\overline{N}}\right) = E(\overline{N}(CV_N^2 + 1)) \approx \theta_N\left(\left(\frac{k-1}{k}\right)\tau_N^2 + 1\right)$ comes from (i) $E(S_N^2) = \left(\frac{k-1}{k}\right)\sigma_N^2$ and (ii) the multivariate version of the Delta Method.[35(pp241-242)] To better understand why the unweighted average of cluster means is an unbiased estimator of μ under SRS and TSS1 but biased under TSS2 and TSS3, note that (i) $E\left(\sum_{j=1}^{k_*} \frac{\bar{y}_j}{k_*} \middle| \mathbf{N}_*\right) = \beta_0 + \gamma(\overline{N}_* - \theta_N)$ for any sampling scheme, and (ii) $E(N_j)$ depends on the sampling scheme (see Equations (A2a) and (A3a)), and so

$$E_{\text{TSS2/TSS3}}\left(E\left(\sum_{j=1}^{k} \frac{\bar{y}_j}{k} \middle| \mathbf{N}\right)\right) = \beta_0 + \gamma\left(E_{\text{TSS2/TSS3}}(\overline{N}) - \theta_N\right) = \beta_0 \neq \mu.$$

This also points out that the unweighted average of cluster means is a biased estimator for β_0 under SRS and TSS1 because clusters are weighted proportionally to their size by the latter two sampling schemes.

Conditional variances.

The conditional variance of any estimator in Table 1 has the form $V(\hat{\mu}|\mathbf{N}_*) = \frac{\sum_{j=1}^{k_*} w_j^2 V(\bar{y}_j|\mathbf{N}_*)}{(\sum_{j=1}^{k_*} w_j)^2}$. Furthermore, note that under TSS1 and TSS3 n individuals are sampled per cluster and so $V(\bar{y}_j|\mathbf{N}) = \sigma_\nu^2 + \frac{\sigma_\varepsilon^2}{n}$, whereas under TSS2 n_j individuals are sampled per cluster, then $V(\bar{y}_j|\mathbf{N}) = \sigma_\nu^2 + \frac{\sigma_\varepsilon^2}{n_j}$. Under SRS k_{SRS} clusters are sampled indirectly from the population, but given that $k_{\text{SRS}} \rightarrow m$ (which follows from $\frac{m}{N_{\text{pop}}} \rightarrow 0$ in Assumption 2), we have that $V(\bar{y}_j|\mathbf{N}_{\text{SRS}}) = \sigma_\nu^2 + \sigma_\varepsilon^2$. Thus, the third row of Table 1 follows $V(\hat{\mu}_{\text{TSS1}}|\mathbf{N}) = \frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk}$ because $w_j = 1$; $V(\hat{\mu}_{\text{TSS3}}|\mathbf{N}) = \frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk} \times (CV_N^2 + 1)$ because $w_j = N_j$ and $\frac{\sum_{j=1}^{k} N_j^2}{(\sum_{j=1}^{k} N_j)^2} = \frac{S_N^2 + \overline{N}^2}{k\overline{N}^2} = \frac{(CV_N^2 + 1)}{k}$; $V(\hat{\mu}_{\text{TSS2}}|\mathbf{N}) = \frac{p\overline{N}(CV_N^2 + 1)\sigma_\nu^2 + \sigma_\varepsilon^2}{p\overline{N}k}$ because $w_j = n_j = pN_j$ and $\frac{\sum_{j=1}^{k} n_j^2}{(\sum_{j=1}^{k} n_j)^2} = \frac{\sum_{j=1}^{k} N_j^2}{(\sum_{j=1}^{k} N_j)^2}$; and $V(\hat{\mu}_{\text{SRS}}|\mathbf{N}_{\text{SRS}}) = \frac{\sigma_\nu^2 + \sigma_\varepsilon^2}{m}$ because $w_j = 1$.

Marginal variances.

Recall that the marginal variance is defined as $V(\hat{\mu}) = E(V(\hat{\mu}|\mathbf{N}_*)) + V(E(\hat{\mu}|\mathbf{N}_*))$. From Equation (A3b) follows that $V(\hat{\mu}_{\text{TSS1}}) = E_{\text{TSS1}}\left(\frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk}\right) + V_{\text{TSS1}}(\beta_0 + \gamma(\overline{N} - \theta_N)) = \frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk} + \gamma^2\frac{V_{\text{TSS1}}(N)}{k} = \frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk} + \gamma^2\frac{\sigma_N^2[\tau_N(\zeta_N - \tau_N) + 1]}{k}$ and that $V(\hat{\mu}_{\text{SRS}}) = E_{\text{SRS}}\left(\frac{\sigma_\nu^2 + \sigma_\varepsilon^2}{m}\right) + V_{\text{SRS}}(\beta_0 + \gamma(\overline{N}_{\text{SRS}} - \theta_N)) = \frac{\sigma_\nu^2 + \sigma_\varepsilon^2}{m} + \gamma^2\frac{V_{\text{SRS}}(N)}{m} = \frac{\sigma_\varepsilon^2 + \sigma_\nu^2 + \gamma^2\sigma_N^2[\tau_N(\zeta_N - \tau_N) + 1]}{m}$. The derivation of the marginal variances of TSS3 and TSS2 requires more steps. The first component of $V(\hat{\mu})$ (fifth row of Table 1) for TSS3 and TSS2 are, respectively, $E(V(\hat{\mu}_{\text{TSS3}}|\mathbf{N})) = \frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk} \times (E(CV_N^2) + 1) \approx \left(\frac{n\sigma_\nu^2 + \sigma_\varepsilon^2}{nk}\right) \times \left(\frac{k(\tau_N^2 + 1)}{\tau_N^2 + k}\right)$ and $E(V(\hat{\mu}_{\text{TSS2}}|\mathbf{N})) = \frac{(E(CV_N^2) + 1)\sigma_\nu^2}{k} + \frac{\sigma_\varepsilon^2}{pk}E\left(\frac{1}{\overline{N}}\right) \approx \frac{p\theta_N\left(\frac{k(\tau_N^2 + 1)}{\tau_N^2 + k}\right)\sigma_\nu^2 + \sigma_\varepsilon^2}{p\theta_N k}$, where both $E(CV_N^2) + 1 = E\left(\frac{S_N^2}{\overline{N}^2}\right) + 1 \approx \frac{E(S_N^2)}{E(\overline{N}^2)} + 1 = \frac{\left(\frac{k-1}{k}\right)\sigma_N^2}{\frac{\sigma_N^2}{k} + \theta_N^2} + 1 = \frac{k(\tau_N^2 + 1)}{\tau_N^2 + k}$

and $E\left(\frac{1}{\overline{N}}\right) \approx \frac{1}{\theta_N}$ follow from the Delta Method.[35(pp241-242)] The second component of $V(\hat{\mu})$ (sixth row of Table 1) is the same under TSS2 and TSS3 because $E(\hat{\mu}_{\mathrm{TSS2}}|\mathbf{N}) = E(\hat{\mu}_{\mathrm{TSS3}}|\mathbf{N})$ (see Table 1, second row), then $V(E(\hat{\mu}_{\mathrm{TSS3}}|\mathbf{N})) = V(E(\hat{\mu}_{\mathrm{TSS2}}|\mathbf{N})) = \gamma^2 V_{\mathrm{TSS2/TSS3}}\left(\frac{S_N^2+\overline{N}^2}{\overline{N}}\right) \approx \gamma^2 \frac{\sigma_N^2}{k}\left[\left(\frac{k-1}{k}\right)^2 \tau_N^2\left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N)\right) + 2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N - \tau_N) + 1\right]$, which is derived as follows. To apply the Delta Method, compute the first derivatives of $\mathrm{g}(S_N^2, \overline{N}) = \frac{S_N^2+\overline{N}^2}{\overline{N}}$ at $(E(S_N^2), E(\overline{N}))^T$: $\left.\frac{\partial \mathrm{g}(S_N^2,\overline{N})}{\partial S_N^2}\right|_{S_N^2=\left(\frac{k-1}{k}\right)\sigma_N^2, \overline{N}=\theta_N} = \frac{1}{\theta_N}$, $\left.\frac{\partial \mathrm{g}(S_N^2,\overline{N})}{\partial \overline{N}}\right|_{S_N^2=\left(\frac{k-1}{k}\right)\sigma_N^2, \overline{N}=\theta_N} = 1 - \left(\frac{k-1}{k}\right)\tau_N^2$. Then, plug these derivatives into equation (5.5.9) in the work of Casella and Berger,[35(p242)]: $\mathrm{Var}(\mathrm{g}(S_N^2, \overline{N})) \approx \frac{1}{\theta_N^2}\mathrm{Var}(S_N^2) + \left(1 - \left(\frac{k-1}{k}\right)\tau_N^2\right)^2 \mathrm{Var}(\overline{N}) + 2\frac{1}{\theta_N}\left(1 - \left(\frac{k-1}{k}\right)\tau_N^2\right)\mathrm{Cov}(S_N^2, \overline{N})$. Finally, in the previous expression replace $\mathrm{Var}(S_N^2)$, $\mathrm{Var}(\overline{N})$, and $\mathrm{Cov}(S_N^2, \overline{N})$ with $\mathrm{Var}(S_N^2) = \left(\frac{k-1}{k}\right)^2 \mathrm{Var}\left(\frac{\sum_{j=1}^{k}(N_j - \overline{N})^2}{k-1}\right) = \left(\frac{k-1}{k}\right)^2 \frac{\sigma_N^4}{k}\left(\eta_N - \frac{k-3}{k-1}\right)$ (Theorem 2, p229, in the work of Mood et al[36]), where $\eta_N = E\left[\left(\frac{N_j - \theta_N}{\sigma_N}\right)^4\right]$ is the kurtosis of cluster size distribution, $\mathrm{Var}(\overline{N}) = \frac{\sigma_N^2}{k}$ and $\mathrm{Cov}(S_N^2, \overline{N}) = \left(\frac{k-1}{k}\right)\frac{\sigma_N^3 \zeta_N}{k}$.[37]