

Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative

Statistical Methods in Medical Research

2021, Vol. 30(2) 357–375

© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0962280220952833

journals.sagepub.com/home/smm

Francesco Innocenti¹ , Math JJM Candel¹ , Frans ES Tan¹ and Gerard JP van Breukelen^{1,2} 

Abstract

To estimate the mean of a quantitative variable in a hierarchical population, it is logistically convenient to sample in two stages (two-stage sampling), i.e. selecting first clusters, and then individuals from the sampled clusters. Allowing cluster size to vary in the population and to be related to the mean of the outcome variable of interest (informative cluster size), the following competing sampling designs are considered: sampling clusters with probability proportional to cluster size, and then the same number of individuals per cluster; drawing clusters with equal probability, and then the same percentage of individuals per cluster; and selecting clusters with equal probability, and then the same number of individuals per cluster. For each design, optimal sample sizes are derived under a budget constraint. The three optimal two-stage sampling designs are compared, in terms of efficiency, with each other and with simple random sampling of individuals. Sampling clusters with probability proportional to size is recommended. To overcome the dependency of the optimal design on unknown nuisance parameters, maximin designs are derived. The results are illustrated, assuming probability proportional to size sampling of clusters, with the planning of a hypothetical survey to compare adolescent alcohol consumption between France and Italy.

Keywords

Cross-national comparisons, informative cluster size, maximin design, optimal design, sample size calculation, two-stage sampling

1 Introduction

For the purpose of estimating the mean or prevalence of an outcome variable (e.g. alcohol consumption or smoking) in a hierarchical population (e.g. students within schools, patients within general practices), or of comparing subpopulations with respect to such a mean or prevalence, it is often convenient, for economic or logistic reasons, to sample in two stages: first, clusters (e.g. schools, general practices) are sampled and then individuals (e.g. students, patients) are drawn from the sampled clusters.^{1–3} Examples of these multi-stage sampling designs include school-based surveys for monitoring substance use among adolescents,^{4–6} and national surveys for estimating the average length of stay for discharges from hospitals,⁷ or nursing homes.⁸ The topic of this paper is the efficient design of two-stage sampling (TSS) schemes for estimating the mean of a quantitative outcome variable in a two-level population.

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

²Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands

Corresponding author:

Francesco Innocenti, Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, Peter Debyeplein 1, 6229 HA, Maastricht, Netherlands.

Email: francesco.innocenti@maastrichtuniversity.nl

In practice, clusters usually vary in size (e.g. small versus large schools) and then, to estimate the population mean, a sample can be drawn with at least three alternative TSS schemes: sampling clusters with probability proportional to cluster size, and then sampling the same number of individuals from each selected cluster (TSS1); sampling clusters with equal probability, and then sampling the same percentage of individuals from each sampled cluster (TSS2); sampling clusters with equal probability, and then sampling the same number of individuals per cluster (TSS3). These three TSS schemes will be considered in this paper and compared with Simple Random Sampling (SRS) of individuals.

Additionally to cluster size variation, further complications arise with informative cluster sizes, that is, when cluster size is related to the outcome of interest.^{9,10} For instance, cluster size is informative when the amount of alcohol consumed by an adolescent is related to the number of students enrolled in the school, as small schools might provide a more supportive environment,^{11–13} or when the number of patients registered to a general practice affects its efficacy in preventing expensive hospitalisations,¹⁴ thus impacting on public expenditure on health per patient. Informative cluster sizes not only can have direct policy implications, such as introducing a limit to school or general practice size, they also have consequences for statistical data analysis and sample size planning. In informative cluster size literature (see the review by Seaman et al.,⁹ and references therein), the main focus has been on how to handle informative cluster size when the target of inference is the association between the outcome variable and some covariates (e.g. a risk factor). For instance, Seaman et al.⁹ have discussed several methods to make cluster-specific inferences with Generalized Linear Mixed Models and population-average inferences with Generalized Estimating Equations when cluster size is informative. Innocenti et al.,¹⁵ instead, have investigated a different topic: the implications of informative cluster size for unbiased and efficient estimation of a population mean in surveys conducted with the three aforementioned TSS schemes. The present paper is also about mean estimation for these three TSS schemes when cluster size is informative, but focuses instead on sample size planning, and the consequences of informative cluster size for the required sample sizes and budget.

Innocenti et al.'s results¹⁵ are the starting point of this paper and therefore summarized here. First, there are two definitions of overall mean in a two-level population, namely the average of all individual outcomes and the average of all cluster-specific means. These two definitions coincide only if cluster sizes are either equal or non-informative. Second, when cluster size is informative, estimation of the mean of all individual outcomes (i.e. the definition used in this paper) is unbiased under TSS1 with the unweighted average of cluster means, and asymptotically unbiased under TSS2 and TSS3 with the average of cluster means weighted by cluster size. In contrast, when cluster size is non-informative, the unweighted average of cluster means is unbiased for all sampling schemes, but optimally efficient for TSS1 and TSS3 only. Third, under the constraint of a fixed total sample size, SRS is more efficient than any TSS scheme, TSS3 is the least efficient TSS scheme, and TSS1 is the most efficient for many cluster size distributions. Indeed, when cluster size is informative, the relative efficiency of these sampling schemes depends on some features of the cluster size distribution in the population, such as the coefficient of variation, the skewness, and the kurtosis. However, when cluster size is non-informative, TSS1 and TSS3 are equally efficient and outperform TSS2. Fourth, the two inferential paradigms in survey sampling, namely the model-based³ and the design-based approach,^{1,2} give similar results in terms of unbiased and efficient estimation of the average of all individual outcomes with the three aforementioned TSS schemes, at least if the model assumptions are met. Furthermore, sample size planning and sampling schemes comparisons, which are the topics of this paper, are much more feasible with the assumption of a model for the outcome variable of interest.¹⁵ For these two reasons, the model-based approach is adopted here.

This work extends the results of Innocenti et al.¹⁵ in the following ways. First, for each of the three aforementioned TSS schemes, the optimal design is derived. Here, the optimal design is defined as that design (i.e. number of clusters and number of individuals per cluster) that minimizes the sampling variance of the population mean estimator subject to a cost constraint. Second, the three optimal TSS schemes are compared with SRS and with each other under the constraint of a fixed budget. Third, to take care of uncertainty with respect to model parameters and distributional features of cluster size, as a practical alternative, maximin designs are derived. Fourth, sample size calculations for making comparisons between populations are derived and illustrated.

This paper is structured as follows. In section 2, the assumptions of this paper are presented, as well as the sampling schemes and the corresponding mean estimators. Furthermore, the findings of a simulation study to assess the accuracy of some results in Innocenti et al.¹⁵ that are relevant to the present paper are summarized. In section 3, the optimal design for each TSS scheme is derived, and these optimal TSS designs are compared with each other and with SRS for a fixed budget. Furthermore, the consequences of ignoring informative cluster size at the design phase of a study are investigated. Section 4 deals with the maximin approach, that is, a strategy to solve the dependency of the optimal design on unknown nuisance parameters. Section 5 provides a procedure for

computing sample sizes for surveys aimed to make cross-population comparisons, and the procedure is illustrated in planning a survey for comparing the average alcohol consumption among adolescents in France and Italy. Section 6 offers some final remarks. The mathematical derivations of the results, the description of the simulation study discussed in section 2, and additional figures and tables can be found in the Supplementary Material 1 (S.M.1). The Supplementary Material 2 (S.M.2) provides the R¹⁶ code of the simulation study and other R codes to apply some of the mathematical results of this paper.

2 Assumptions, sampling schemes and mean estimators

The results of Innocenti et al.¹⁵ and this paper are based on the following assumptions (the notation used in the main text is summarized in the Appendix).

Assumption 1: The population is composed of K clusters and each cluster j contains N_j individuals, that is, in the population clusters vary in size (N_j). The population size is $N_{pop} = \sum_{j=1}^K N_j$.

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 per selected cluster j . In case of TSS, the population is very large relative to the sample size at each design level (i.e. $\frac{k}{K} \rightarrow 0$ and $\frac{\bar{n}}{\theta_N} \rightarrow 0$, where $\bar{n} = \frac{\sum_{j=1}^k n_j}{k}$ is the average sample size per sampled cluster, and $\theta_N = \frac{N_{pop}}{K}$ is the population mean of cluster size). In case of SRS, N_{pop} is very large relative to m , the number of individuals sampled (i.e. $\frac{m}{N_{pop}} \rightarrow 0$).

Assumption 3: The outcome variable Y_{ij} is quantitative (e.g. alcohol consumption) and measured at the individual level. Further, 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 i -th individual from the j -th cluster^{3,17}

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij} \quad (1)$$

where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, and cluster effect u_j and individual effect ε_{ij} are unrelated (i.e. $u_j \perp \varepsilon_{ij}$). The distribution of u_j will be defined in the next assumption.

Assumption 4: Cluster effect u_j is linearly related to cluster size N_j , that is, $u_j = \alpha_0 + \alpha_1 N_j + \nu_j = \alpha_1(N_j - \theta_N) + \nu_j$, where $\alpha_0 = -\alpha_1 \theta_N$ for model identifiability, $\nu_j \sim N(0, \sigma_\nu^2)$, and ν_j is the component of cluster effect u_j that does not depend on cluster size (i.e. $\nu_j \perp N_j$). Thus, the conditional distribution of u_j given N_j is $u_j | N_j \sim N(\alpha_1(N_j - \theta_N), \sigma_\nu^2)$.

Innocenti et al.¹⁵ show that β_0 in model (1) is the average of all cluster-specific means in the population, and differs from the average of all individual outcomes in the population μ , unless cluster size is non-informative ($\alpha_1 = 0$) or constant across clusters, as can be seen from the following expression

$$\mu = \beta_0 + \alpha_1 \theta_N \tau_N^2 \quad (2)$$

where θ_N , $\tau_N = \frac{\sigma_N}{\theta_N}$, and σ_N^2 are, respectively, the population mean, the coefficient of variation, and the variance of cluster size. The distinction between β_0 and μ comes from considering the distribution of cluster effect u_j over either the population of clusters (which yields β_0) or the population of individuals (which yields μ).¹⁵ This paper focuses on μ .

With the aim of estimating μ , the three aforementioned TSS schemes are studied in this paper. For each of these TSS schemes and SRS, Table 1 summarizes the sampling procedure (i.e. sample size and inclusion probability per design stage) and the required knowledge before sampling. Furthermore, Table 1 shows the population mean estimator $\hat{\mu}$ and the sampling variance $V(\hat{\mu})$ for each sampling scheme. Denote by $\rho_{uN} = \frac{E[(u_j - E(u_j))(N_j - E(N_j))]}{\sigma_u \sigma_N} = \frac{E[u_j(N_j - \theta_N)]}{\sigma_u \sigma_N}$ the correlation between u_j and N_j , where $E(u_j) = 0$ and $V(u_j) = \sigma_u^2 = \sigma_\nu^2 + \alpha_1^2 \sigma_N^2$, and by $\psi = \left(\frac{\rho_{uN}^2}{1 - \rho_{uN}^2} \right)$

the degree of informativeness of cluster size. From Table 1, note that $\bar{n} = \frac{\sum_{j=1}^k n_j}{k} = n$ for TSS1 and TSS3, while $\bar{n} = \frac{\sum_{j=1}^k n_j}{k} = p \frac{\sum_{j=1}^k N_j}{k} = p \bar{N}$ for TSS2, where \bar{N} is the average population size of the k sampled clusters (not to be

Table 1. Sampling schemes, required prior knowledge, population mean estimators, and sampling variances.

TSS1	Stage 1	k clusters with probability $\pi_j \approx \frac{kN}{\sum_{j=1}^k N_j}$
	Stage 2	n individuals per sampled cluster with probability $\pi_{tij} = \frac{n}{N_j}$ List of all K clusters in the population and their sizes N_j . List of all individuals within the k sampled clusters.
	$\hat{\mu}$	$\sum_{j=1}^k \frac{y_j}{k}$
	$V(\hat{\mu})$	$\frac{\sigma_y^2}{nk} \{1 + \rho [(n-1) + n\psi(\tau_N(\zeta_N - \tau_N) + 1)]\}$
TSS2	Stage 1	k clusters with probability $\pi_j = \frac{k}{N_j}$
	Stage 2	$n_j = pN_j$ individuals per sampled cluster with probability $\pi_{tij} = \frac{n_j}{N_j} = p$ List of all K clusters in the population. List of all individuals within the k sampled clusters.
	$\hat{\mu}$	$\frac{\sum_{j=1}^k n_j \bar{y}_j}{\sum_{j=1}^k n_j} = \frac{\sum_{j=1}^k pN_j \bar{y}_j}{\sum_{j=1}^k pN_j} = \frac{\sum_{j=1}^k N_j \bar{y}_j}{\sum_{j=1}^k N_j}$
	$V(\hat{\mu})$	$\frac{\sigma_y^2}{nk} \left\{ 1 + \rho \left[\left(\frac{\tau_N^2 + 1}{\tau_N^2 + 1} \right) n - 1 + n\psi \left(\left(\frac{k-1}{k} \right)^2 \tau_N^2 \left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N) \right) + 2 \left(\frac{k-1}{k} \right) \tau_N(\zeta_N - \tau_N) + 1 \right) \right] \right\}$
TSS3	Stage 1	k clusters with probability $\pi_j = \frac{k}{N_j}$
	Stage 2	n individuals per sampled cluster with probability $\pi_{tij} = \frac{n}{N_j}$ List of all K clusters in the population. List of all individuals within the k sampled clusters.
	$\hat{\mu}$	$\frac{\sum_{j=1}^k N_j \bar{y}_j}{\sum_{j=1}^k N_j}$
	$V(\hat{\mu})$	$\frac{\sigma_y^2}{nk} \left\{ \frac{\tau_N^2 + 1}{\tau_N^2 + 1} + \rho \left[\left(\frac{\tau_N^2 + 1}{\tau_N^2 + 1} \right) (n-1) + n\psi \left(\left(\frac{k-1}{k} \right)^2 \tau_N^2 \left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N) \right) + 2 \left(\frac{k-1}{k} \right) \tau_N(\zeta_N - \tau_N) + 1 \right) \right] \right\}$
SRS	Stage 1	m individuals with probability $\pi_i = \frac{m}{N_{pop}}$
	Stage 2	List of all N_{pop} individuals in the population.
	Required prior knowledge	$\sum_{i=1}^m \frac{y_i}{m}$
	$\hat{\mu}$	$\sum_{i=1}^m \frac{y_i}{m}$
	$V(\hat{\mu})$	$\frac{\sigma_y^2}{m} \{1 + \rho\psi[\tau_N(\zeta_N - \tau_N) + 1]\}$

Note: For TSS1, $\pi_j \approx \frac{kN}{\sum_{j=1}^k N_j}$ follows from $\pi_j = 1 - \left(1 - \frac{N_j}{\sum_{j=1}^k N_j} \right)^k$ if $\frac{N_j}{\sum_{j=1}^k N_j} \rightarrow 0 \forall j = 1, \dots, K$. For TSS2, $n = E(\bar{n}) = p\theta_N$, where $\bar{n} = \frac{\sum_{j=1}^k n_j}{k} = p \frac{\sum_{j=1}^k N_j}{k} = p\bar{N}$, and $E(N) = \theta_N$.

confused with \bar{n} , that is, the average sample size of the sampled clusters). Furthermore, for TSS2 $n = E(\bar{n}) = pE(\bar{N}) = p\theta_N$. The sampling variances in Table 1 are functions of the total unexplained outcome variance $\sigma_y^2 = \sigma_\nu^2 + \sigma_\varepsilon^2$, the intraclass correlation coefficient $\rho = \frac{\sigma_\nu^2}{\sigma_y^2} \in [0,1]$, the sample sizes (k, n) , the parameter ψ , and some features of the cluster size distribution in the population: the coefficient of variation τ_N , the skewness ζ_N , and (for TSS2 and TSS3 only) the kurtosis η_N . When cluster size is non-informative ($\psi = 0$), $V(\hat{\mu})$ depends only on σ_y^2, ρ, k, n , and (for TSS2 and TSS3 only) τ_N . The estimators $\hat{\mu}$ associated with SRS and TSS1 are unbiased, and their sampling variances $V(\hat{\mu})$ are exact expressions.¹⁵

The estimators associated with TSS2 and TSS3 are only asymptotically unbiased, and the corresponding sampling variances are based on first-order Taylor series approximations.¹⁵ The accuracy of these approximations was evaluated through a simulation study discussed in supplementary material S.M.1 (section 1), but the main findings are summarized here. Sampling $k = 20$ clusters guarantees nearly unbiased estimates of μ under TSS2 and TSS3 independently of the cluster size distribution, and fair accuracy (i.e. bias $\leq 5\%$) of the variances in Table 1 (TSS2 and TSS3 row) when $|\rho_{uN}| \leq 0.75$, $\rho \leq 0.3$, and ζ_N and η_N are relatively close (say, ± 1.5) to those of the Normal distribution (i.e. $\zeta_N = 0$ and $\eta_N = 3$). However, for cluster size distributions with extreme skewness and kurtosis (e.g. $\zeta_N \geq 2$ and $\eta_N \geq 9$) at least $k = 100$ clusters must be sampled to achieve a reasonable accuracy (i.e. bias $\leq 6\%$) of the sampling variances in Table 1, for $\rho_{uN} \leq 0.5$ and $\rho \leq 0.3$. Furthermore, the simulations showed that the two lower-bounds for k (i.e. 20, and 100) guarantee the corresponding accuracy level across different values for n (at least for $2 \leq n \leq 100$). To contextualize these two lower-bounds for k , in a school-based survey for studying substance use among adolescents in 21 European countries, Shackleton et al.¹⁸ have reported that, across countries, $k \in [36,531]$ (Median = 123) and $\bar{n} \in [5.92,119.62]$ (Median = 20.74).

3 Optimal design and relative efficiencies for a given budget

3.1 Optimal design

For any sampling scheme, the precision of the estimator $\hat{\mu}$, and thus also the width of a confidence interval for μ and the statistical power for testing a hypothesis on μ , depends on the number of clusters and on the sample size per cluster (Table 1). This raises the question of the best combination of sample sizes at each design stage (i.e. sampling many clusters versus sampling many individuals per cluster). Define the optimal design as that design (i.e. number of clusters and number of individuals per cluster), which minimizes $V(\hat{\mu})$ subject to a cost constraint, given that time and budget are limited in practice. For TSS, the cost constraint is assumed to be $C = k(c_2 + c_1n)$, where C is the budget for sampling and measuring (excluding costs for constructing the sampling frame and other costs not related to sample size). From now on C is called the *research budget*. Furthermore, c_2 is the average cost for sampling a cluster, c_1 is the average cost for sampling an individual from a sampled cluster, and $(c_2 + c_1n)$ is the cost per cluster including the costs for sampling n individuals from that cluster (recall that for TSS2 $n = p\theta_N$). For SRS, the cost constraint is $C = c_0 + c_{srs}m$, where m is the number of individuals to sample, c_{srs} is the average cost for sampling an individual directly from the population, and c_0 represents the extra-cost due to constructing the sampling frame for a SRS compared with the sampling frame for a TSS.

For each TSS scheme, the optimal design (i.e. the optimal sample sizes k^* and n^*) for estimating μ and the optimal variance $V(\hat{\mu})^*$ (i.e. $V(\hat{\mu})$ under the optimal design) are given in Table 2 (for proofs, see section 2.2 of S.M.1). For TSS2, one can obtain the optimal proportion of individuals to sample per cluster p^* from the optimal n^* , by dividing n^* as given in Table 2 (TSS2 row) by θ_N . The optimal TSS2 and TSS3 designs depend on two approximations of $V(\hat{\mu})$: the first-order Taylor approximation mentioned in section 2 and evaluated in S.M.1 (section 1), which underlies the equations in Table 1, and an approximation based on large k (i.e. k such that $\frac{\tau_N^2}{k} \approx 0$, $\frac{k-1}{k} \approx 1$, and $\frac{k-3}{k-1} \approx 1$) to simplify the expressions in Table 1. These two approximations give the following equations (for details, see section 2.1 of S.M.1)

$$V_{TSS2}(\hat{\mu}) \approx \frac{\sigma_y^2}{nk} \{ 1 + \rho [n((\tau_N^2 + 1) + \psi(\tau_N^4 + \tau_N^2(\eta_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1)) - 1] \} \tag{3}$$

Table 2. Optimal design and optimal variance $V(\hat{\mu})^*$ for each sampling scheme.

SRS	$V(\hat{\mu})^*$	$\frac{c_{SRS} \sigma_y^2 (1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1])}{C - c_0}$
TSS1	Optimal design	$n^* = \sqrt{c_r \left(\frac{1 - \rho}{\rho} \right) \left(\frac{1}{1 + \psi [\tau_N (\zeta_N - \tau_N) + 1]} \right)}, k^* = \frac{C}{c_1 (c_r + n^*)}$
	$V(\hat{\mu})^*$	$\frac{c_1 \sigma_y^2 \left(\sqrt{c_r \rho (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1])} + \sqrt{1 - \rho} \right)^2}{C}$
TSS2	Optimal design	$n^* = \sqrt{c_r \left(\frac{1 - \rho}{\rho} \right) \frac{1}{(\tau_N^2 + 1) + \psi [\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1]}}, k^* = \frac{C}{c_1 (c_r + n^*)}$
	$V(\hat{\mu})^*$	$\frac{c_1 \sigma_y^2 \left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)]} + \sqrt{1 - \rho} \right)^2}{C}$
TSS3	Optimal design	$n^* = \sqrt{c_r \left(\frac{1 - \rho}{\rho} \right) \frac{(\tau_N^2 + 1)}{(\tau_N^2 + 1) + \psi [\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1]}}, k^* = \frac{C}{c_1 (c_r + n^*)}$
	$V(\hat{\mu})^*$	$\frac{c_1 \sigma_y^2 \left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)]} + \sqrt{(1 - \rho) (\tau_N^2 + 1)} \right)^2}{C}$

Note: Derivations are given in section 2.2 in supplementary material S.M.1. Note that $c_r = \frac{c_2}{c_1} > 1$, $\zeta_N \geq \tau_N - \frac{1}{\tau_N}$ implies that $[\tau_N (\zeta_N - \tau_N) + 1] \geq 0$ that, in turn, entails that $1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] > 0$ since $\psi \geq 0$. Note that $[\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1] \geq 0$ for any distribution (for proof, see section 2.1, S.M.1). Recall that for TSS2 $n^* = p^* \theta_N$.

and

$$V_{TSS3}(\hat{\mu}) \approx \frac{\sigma_y^2}{nk} \left\{ \tau_N^2 + 1 + \rho [(\tau_N^2 + 1)(n - 1) + m \psi (\tau_N^4 + \tau_N^2 (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] \right\} \tag{4}$$

where for TSS2, $n = p \theta_N$. Recall from section 2 that, for TSS2 and TSS3, k must be large anyway, because the estimators $\hat{\mu}_{TSS2}$ and $\hat{\mu}_{TSS3}$ given in Table 1 are only asymptotically unbiased. As a special case, $\psi = 0$ gives the optimal design and optimal variance for non-informative cluster size (for which case $\beta_0 = \mu$), which under TSS1 coincide with the equations available for cluster randomized trials (for instance, see Moerbeek et al. ¹⁹). There is no such equivalence under TSS2 due to sample size variation between clusters, and under TSS3 due to weighting cluster means by cluster size if informative cluster size is assumed in the design phase. Indeed, under non-informative cluster size, no weighting is needed under TSS3,¹⁵ and then the optimal design equations for TSS1 apply to TSS3 as well.

Note from Table 2 that the optimal number of clusters k^* and the optimal number of individuals per cluster n^* are inversely related, and that n^* is an increasing function of the cluster-to-individual cost ratio $c_r = \frac{c_2}{c_1} > 1$ and a decreasing function of ρ and ψ . These relations between the optimal design and c_r , ρ , and ψ hold, under TSS1, for $\zeta_N > \tau_N - \frac{1}{\tau_N}$, and always under TSS2 and TSS3 (for proof, see section 2.1 of S.M.1). The condition $\zeta_N > \tau_N - \frac{1}{\tau_N}$ is met by all the distributions in Tables S.2 and S.7 (S.M.1). Hence, this condition is assumed to be satisfied when considering results for TSS1 in the sequel.

3.2 Effect of cluster size informativeness on the optimal design and study budget needed

The optimal number of individuals per cluster n^* for TSS1 and TSS3 is plotted in Figure 1, for two real-life cluster size distributions: the general practice list size distribution in England, and the public high school size distribution in Italy (both distributions are shown in Figure S.1, S.M.1). The behaviour of n^* for other cluster size distributions is shown in Figures S.2 and S.4 (S.M.1) for TSS1 and TSS3, respectively, and in Figure S.3 (S.M.1) for TSS2. In most scenarios in Figure 1 and Figures S.2–S.4 (S.M.1), the difference between n^* for $\psi = 0.35$ (i.e. $\rho_{uN} = \pm 0.51$) and n^* for $\psi = 0$ (i.e. $\rho_{uN} = 0$) is small, which means that the ratio of $V(\hat{\mu})$ under the design assuming $\psi = 0.35$ to $V(\hat{\mu})$ under the design assuming $\psi = 0$, when the true $\psi = 0.35$, is close to 1. So, the optimal designs in Table 2 are quite robust against misspecification of ψ , in the sense of being efficient relative to the optimal design for the true

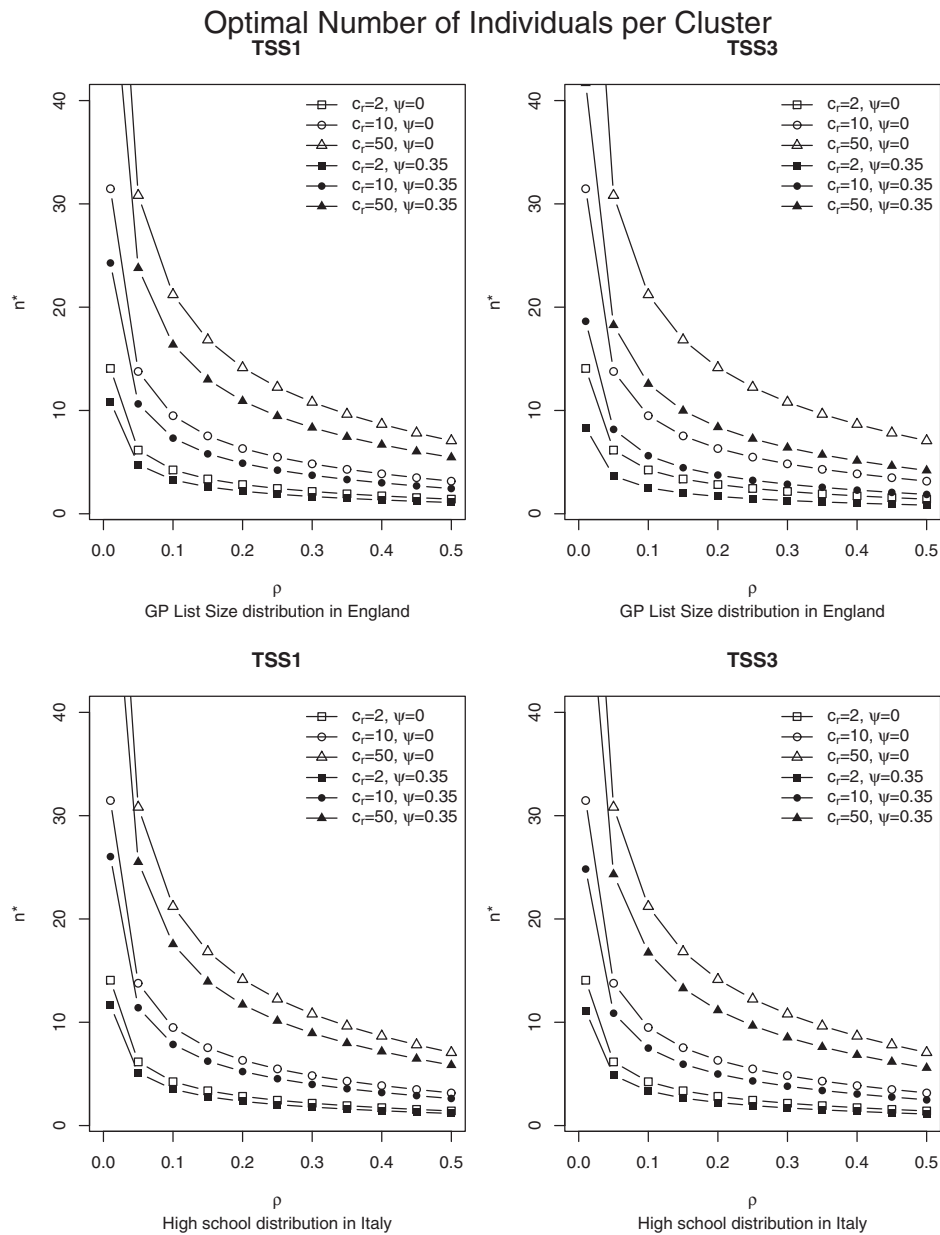


Figure 1. Optimal number of individuals per cluster n^* under TSS1 (left column) and TSS3 (right column), as a function of ρ , for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

ψ and given a fixed research budget C . However, ignoring informativeness can lead to serious underestimation of the sampling variance of the mean estimator, and thereby also of the budget needed, as will be seen below. Further, the optimal design depends not only on ψ , but also on ρ and the cluster size distribution $(\tau_N, \zeta_N, \eta_N)$. That dependence will be addressed in section 4.

An example will now show that (a) given a study budget, the optimal design is robust against misspecification of cluster size informativeness, but (b) the budget needed is very sensitive to misspecification. Suppose we plan a survey to estimate μ in the population of all patients of all general practices in England. The parameters of the general practice patient list size distribution are $\tau_N = 0.633$, $\zeta_N = 2.12$, and $\eta_N = 14.549$ (Table S.2, S.M.1). Furthermore, suppose that $\rho = 0.05$, $c_r = 10$, and $C/c_1 = 1000$. The optimal TSS1 samples $n^* = 10.74$ individuals and $k^* = 48.22$ clusters assuming $\psi = 1/3$, and $n^* = 13.78$ and $k^* = 42.04$ assuming $\psi = 0$ (see Table 2, TSS1 row). If the true $\psi = 1/3$, $V(\hat{\mu}) = \sigma_y^2 \times 0.00354$ for the design correctly assuming $\psi = 1/3$,

and $V(\hat{\mu}) = \sigma_y^2 \times 0.00360$ for the design incorrectly assuming $\psi = 0$ (see variance equation in Table 1, TSS1 row), giving a variance ratio $0.00354/0.00360 = 0.983$. Additional results for TSS1, TSS2, and TSS3 are given in Table S.8 (S.M.1), which shows that even in some more extreme cases (e.g. $\psi = 1$, i.e. $\rho_{uN} = \pm 0.707$) the variance ratio still exceeds 0.8. The example given here and those in Table S.8 (S.M.1) show that the optimal designs in Table 2 are quite robust against misspecification of ψ , in the sense of being efficient relative to the optimal design for the true ψ and given a fixed research budget C .

However, ignoring informativeness can lead to serious underestimation of the budget needed. Suppose one wants to test the null hypothesis H_0 that $\mu = \mu_0$ against the alternative hypothesis H_1 that $\mu \neq \mu_0$. The budget that guarantees the desired power level $1 - \gamma$ for the chosen type I error rate α , is then obtained by equating $V(\hat{\mu})^*$ in Table 2 with $\left(\frac{\mu - \mu_0}{z_{1-\gamma} + z_{1-\frac{\alpha}{2}}}\right)^2$, where z_q is the q th percentile of the standard normal distribution. This gives $C = \frac{g(\rho, \psi)(z_{1-\gamma} + z_{1-\frac{\alpha}{2}})^2}{d_0^2}$, where $g(\rho, \psi)$ is the numerator of $V(\hat{\mu})^*$ in Table 2 excluding σ_y^2 , and $d_0 = \frac{\mu - \mu_0}{\sigma_y}$ is the standardized difference between true mean and mean according to H_0 . Since $g(\rho, \psi)$ is an increasing function of c_1 , c_2 , and ψ , the required budget C for the desired power level also increases with c_1 , c_2 , and ψ . Likewise, C increases with ρ , at least up to $\rho = 0.5$ (for proofs, see section 2.2 in S.M.1). The required budget C to detect a standardized difference of medium size ($d_0 = 0.5$), with 90% power and two-tailed $\alpha = 0.05$, is plotted in Figure 2 for TSS1 and TSS3, as function of ψ , for the general practice list size distribution in England and the public high school size distribution in Italy, and assuming $c_1 = 10$. As can be seen in Figure 2, the research budget C is not robust against misspecification of ψ . For example, the required budget C for the optimal TSS1, assuming the English general practice list size distribution, $c_r = 30$, $c_1 = 10$, and $\rho = 0.10$ (Figure 2, left column, first row), is underestimated by 29% if one incorrectly assumes $\psi = 0$ when the true $\psi = 0.35$. The required budget C is also shown, for other cluster size distributions, in Figures S.5 and S.7 (S.M.1) for TSS1 and TSS3, respectively, and in Figure S.6 (S.M.1) for TSS2. These figures show that C increases with ρ , c_2 , and ψ , and that the impact of the cluster size distribution on C becomes more relevant as ψ increases. Hence, ignoring informative cluster size at the design phase of the survey can lead to underestimating the required budget for the chosen effect size and desired power level. Finally, for the desired power level, the required budget is smallest with the optimal TSS1, and largest with the optimal TSS3.

3.3 Relative efficiencies for a given budget

We now compare the efficiency of the optimal designs in Table 2 with each other and with SRS, under the constraint of a fixed research budget. The relative efficiency (RE) of the optimal designs for two sampling schemes is defined as the ratio of their optimal variances $V(\hat{\mu})^*$ in Table 2, more specifically, $RE(D1 \text{ vs } D2) = \frac{V_{D2}(\hat{\mu})^*}{V_{D1}(\hat{\mu})^*}$. These RE s are shown in Table 3 (for proofs, see section 2.3, S.M.1), which also gives the sufficient (but not necessary) conditions under which each RE is smaller than one.

The RE of a TSS scheme compared with SRS (Table 3, first three rows) is composed of three ratios. The first ratio is a function of ρ , ψ , τ_N , ζ_N , η_N , and c_r , and is always smaller than one for $\psi = 0$, and also for $\psi \neq 0$ at least under the conditions for ζ_N given in the rightmost column of Table 3 (for proofs, see section 2.3, S.M.1). The other two components of $RE(\text{TSS vs SRS})$ are the ratio $\frac{c_{srs}}{c_1}$, for the costs per individual in SRS relative to TSS, and the budget ratio $\frac{C}{C - c_0}$. Since sampling an individual directly from the population will be more expensive than sampling an individual after having sampled the cluster to which he/she belongs (i.e. $c_{srs} > c_1$), and constructing the sampling frame for a SRS has extra-costs compared with constructing the sampling frame for a TSS (i.e. $c_0 > 0$), the ratios $\frac{c_{srs}}{c_1}$ and $\frac{C}{C - c_0}$ will always be at least one and often larger than one. As a result, the RE can become larger than one, implying that SRS can be less efficient than TSS under the constraint of a fixed budget.

The RE s of the optimal TSS1 and TSS3 versus SRS are shown in Figure 3, for the general practice list size distribution in England and the public high school size distribution in Italy, and assuming $\left(\frac{c_{srs}}{c_1}\right) = \left(\frac{C}{C - c_0}\right) = 1$ (note that values greater than 1 give a higher RE of TSS versus SRS). Further, Figures S.8–S.10 (S.M.1) show the RE s of the optimal TSS1, TSS2, and TSS3 versus SRS for other cluster size distributions. For $\psi = 0$, the RE of any optimal TSS versus SRS is a decreasing function of (i) c_r (Table 3), (ii) ρ (at least for $\rho \leq 0.5$, see Figure 3, and Figures S.8–S.10 in S.M.1), and (iii), only for TSS2 and TSS3, τ_N (Table 3). For $\psi \neq 0$, the patterns remain almost the same as before and the RE s also do not seem to vary much across cluster size distributions (Figure 3, and Figures S.8–S.10 in S.M.1).

The RE s of the three TSS schemes compared with each other (Table 3, last three rows) are functions of ρ , c_r , ψ , τ_N , ζ_N , and η_N . The optimal TSS2 is more efficient than the optimal TSS3 since $RE(\text{TSS3 vs TSS2}) < 1$ (unless

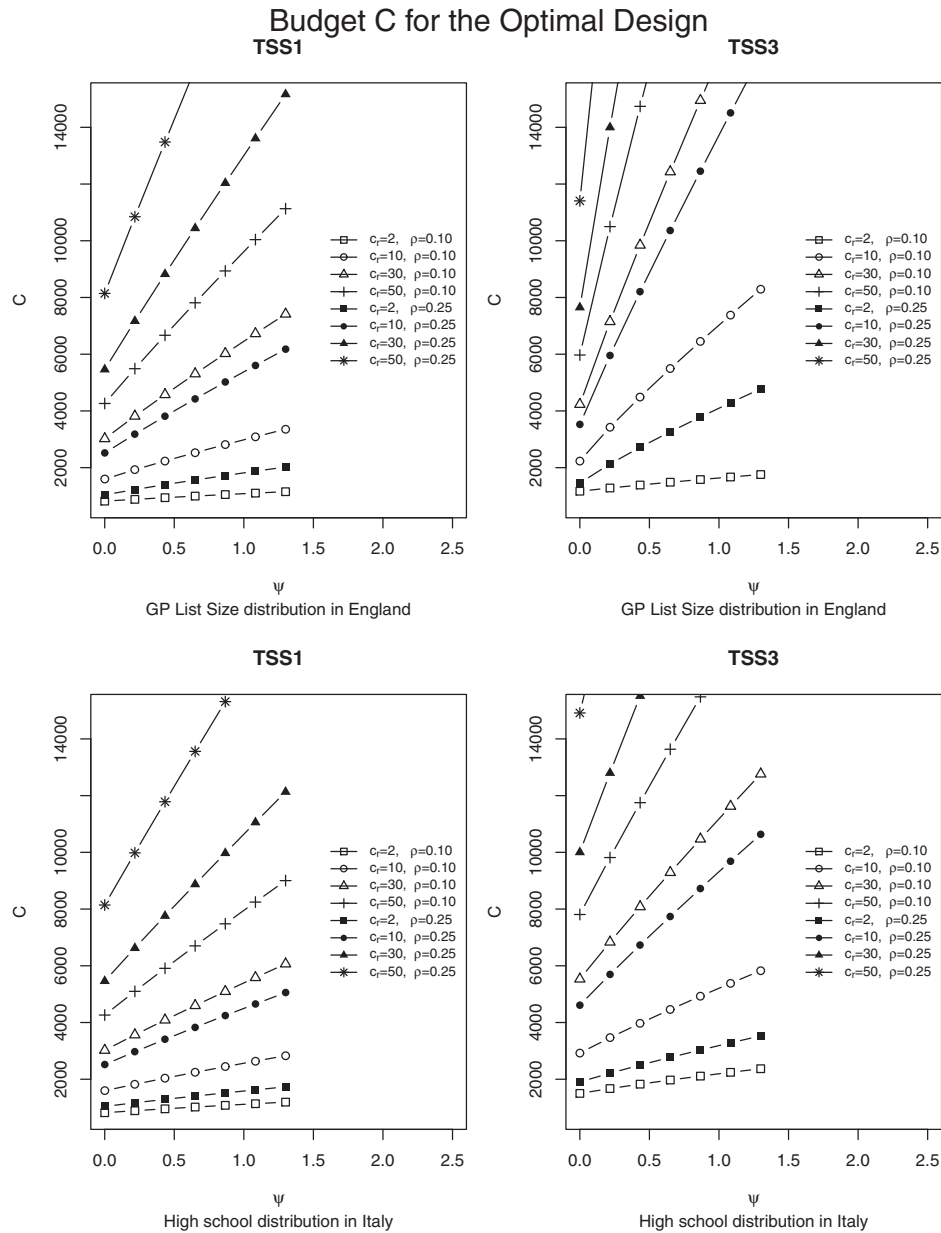


Figure 2. Budget C needed for the optimal design to detect a standardized difference between hypothesized and true population mean of medium size ($d_0 = 0.5$), with 90% power using a two-tailed test with $\alpha = 0.05$, as a function of ψ , for different values of ρ and c_r (curves) with $c_1 = 10$, different sampling schemes (columns), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi \in [0, 1.3]$ corresponds to $\rho_{uN} \in [-0.75, +0.75]$.

$\tau_N = 0$, Table 3, or $\rho_{uN} \approx \pm 1$,¹⁵ since in both cases $RE(TSS3 \text{ vs } TSS2) = 1$). The REs of TSS2 and TSS3 versus TSS1 are smaller than one, and so the optimal TSS1 is the most efficient TSS scheme, at least for cluster size distributions satisfying the conditions in Table 3 (rightmost column), such as all distributions in Table S.7 (S.M.1). For other cluster size distributions, one must compute the RE for that particular distribution to see whether $RE < 1$. However, for $\psi = 0$, the REs in the last three rows of Table 3 are all smaller than one for any cluster size distribution, making TSS1 the most efficient TSS scheme, followed by TSS2. Note that this only holds if informative cluster size ($\psi \neq 0$) is assumed at the design stage, such that in TSS3 cluster means are weighted by cluster size to estimate μ (Table 1). If non-informative cluster size ($\psi = 0$) is assumed already in the design stage, then no weighting is needed for TSS3,¹⁵ and TSS3 then is as efficient as TSS1.

The RE of the optimal TSS2 and TSS3 versus the optimal TSS1 are shown in Figure 4, for the general practice list size distribution in England and the public high school size distribution in Italy, and in Figures S.11–S.12

Table 3. Relative efficiencies of TSS schemes versus SRS and each other for a given budget.

D1 vs D2	$RE(D1 \text{ vs } D2) = \frac{V_{D2}(\hat{\mu})^*}{V_{D1}(\hat{\mu})^*}$	Sufficient (but not necessary) conditions such that $RE \leq 1$
TSS1 vs SRS	$\frac{1 + \rho\psi[\tau_N(\zeta_N - \tau_N) + 1]}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{1 - \rho})^2} \times \left(\frac{c_1}{c - c_0}\right) \times \left(\frac{c}{c - c_0}\right)$	$\zeta_N \geq \tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi} \text{ and } \left(\frac{c_1}{c - c_0}\right) \times \left(\frac{c}{c - c_0}\right) = 1$
TSS2 vs SRS	$\frac{1 + \rho\psi[\tau_N(\zeta_N - \tau_N) + 1]}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{1 - \rho})^2} \times \left(\frac{c_1}{c - c_0}\right) \times \left(\frac{c}{c - c_0}\right)$	$\zeta_N \leq \tau_N - \frac{1}{\tau_N} \text{ or } \zeta_N \geq \tau_N + \frac{1}{\tau_N c} - \frac{1}{\tau_N} \text{ or } N_j \sim N(\theta_N, \sigma_N^2), \text{ and } \left(\frac{c_1}{c - c_0}\right) = 1$
TSS3 vs SRS	$\frac{1 + \rho\psi[\tau_N(\zeta_N - \tau_N) + 1]}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{(1 - \rho)(\tau_N^2 + 1)})^2} \times \left(\frac{c_1}{c - c_0}\right) \times \left(\frac{c}{c - c_0}\right)$	$\zeta_N \leq \tau_N - \frac{1}{\tau_N} \text{ or } \zeta_N \geq \tau_N + \frac{1}{\tau_N c} - \frac{1}{\tau_N} \text{ or } N_j \sim N(\theta_N, \sigma_N^2), \text{ and } \left(\frac{c_1}{c - c_0}\right) = 1$
TSS2 vs TSS1	$\frac{(\sqrt{c_1\rho[1 + \psi(\tau_N(\zeta_N - \tau_N) + 1)] + \sqrt{1 - \rho}})^2}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{1 - \rho})^2}$	$\tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi} \leq \zeta_N \leq \tau_N - \frac{1}{\tau_N} \text{ or } \zeta_N \geq \tau_N \text{ or } N_j \sim N(\theta_N, \sigma_N^2)$
TSS3 vs TSS1	$\frac{(\sqrt{c_1\rho[1 + \psi(\tau_N(\zeta_N - \tau_N) + 1)] + \sqrt{1 - \rho}})^2}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{(1 - \rho)(\tau_N^2 + 1)})^2}$	$\tau_N - \frac{1}{\tau_N} - \frac{1}{\tau_N\psi} \leq \zeta_N \leq \tau_N - \frac{1}{\tau_N} \text{ or } \zeta_N \geq \tau_N \text{ or } N_j \sim N(\theta_N, \sigma_N^2)$
TSS3 vs TSS2	$\frac{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{1 - \rho})^2}{(\sqrt{c_1\rho[\tau_N^2 + 1 + \psi(\tau_N^4 + \tau_N^2)(n_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1]} + \sqrt{(1 - \rho)(\tau_N^2 + 1)})^2}$	

Note: Derivations are given in section 2.3 in supplementary material S.M.1. Recall that $V(\hat{\mu})^*$ is the optimal variance in Table 2, $c_1 = \frac{\sigma}{c} > 1$, $\psi = \left(\frac{\rho_{\theta_N}}{1 - \rho_{\theta_N}}\right)$, $\left(\frac{c_1}{c - c_0}\right) \geq 1$ and $\left(\frac{c}{c - c_0}\right) \geq 1$. The conditions for ζ_N in the rightmost column are valid for $\psi \neq 0$ and are satisfied by all distributions in Table S.7 (S.M.1).

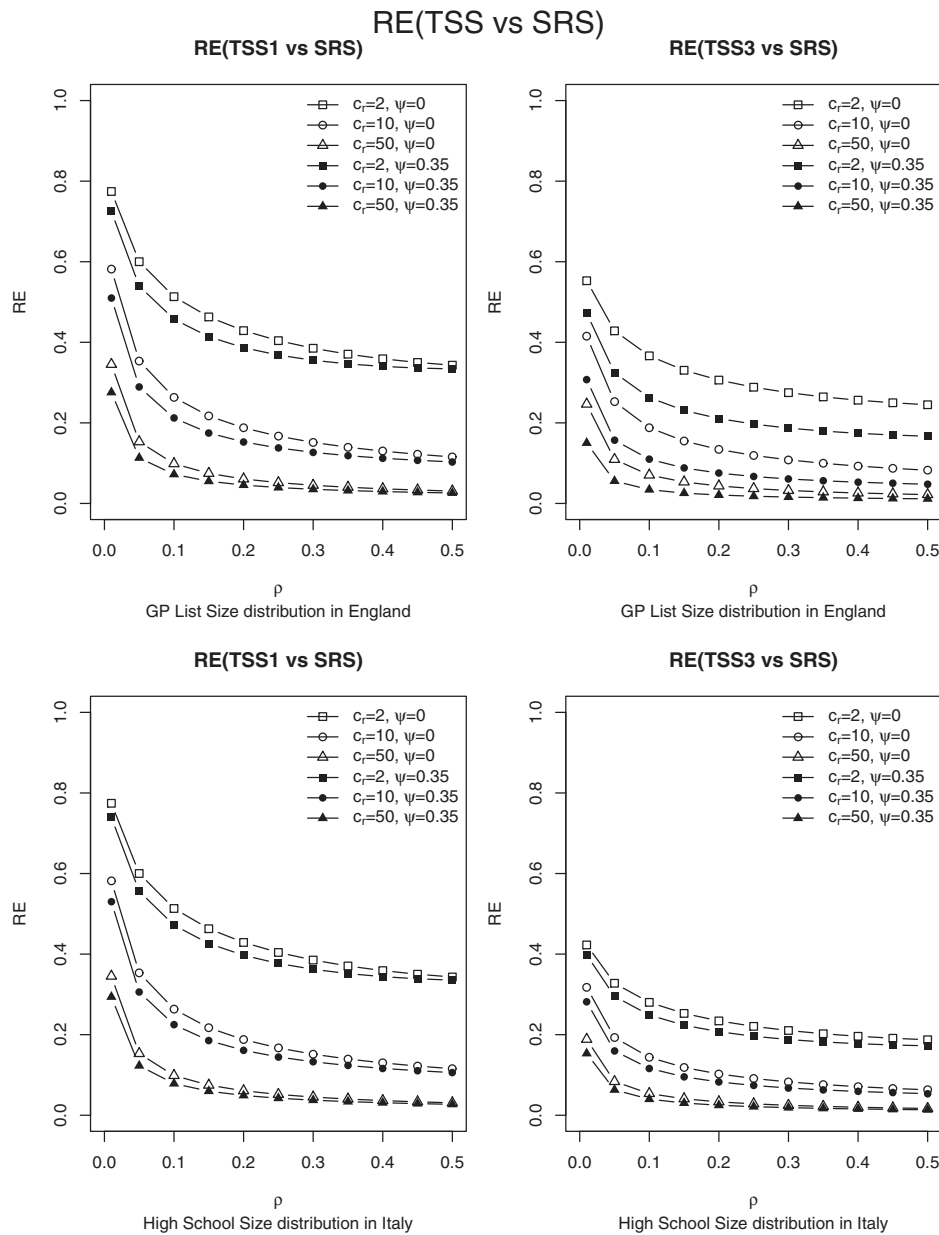


Figure 3. Relative efficiency of the optimal TSS1 versus SRS (left column), and of the optimal TSS3 versus SRS (right column), for a given research budget C and assuming $(c_{srs}/c_1) = (C/(C-c_0)) = 1$ (values greater than 1 give a higher RE of TSS versus SRS), as a function of ρ , for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

(S.M.1) for other four cluster size distributions. For $\psi = 0$, these reduce to $RE(TSS2 \text{ vs } TSS1) = \frac{(\sqrt{c_r \rho} + \sqrt{1-\rho})^2}{(\sqrt{c_r \rho (\tau_N^2 + 1)} + \sqrt{1-\rho})^2}$ and $RE(TSS3 \text{ vs } TSS1) = \frac{1}{\tau_N^2 + 1}$, which are both decreasing functions of τ_N , but $RE(TSS2 \text{ vs } TSS1)$ also decreases as ρ and/or c_r increases. For $\psi \neq 0$, the patterns are the same as before with two major differences. First, both REs decrease as η_N increases (Table 3). Second, for $\psi = 0.35$, both REs differ at most 6% from their values at $\psi = 0$ (Figure 4, and Figures S.11–S.12 in S.M.1), except for the English general practice (GP) list size distribution that, having an extreme kurtosis (i.e. $\eta_N = 14.55$), shows a drop in RE (compared with the case $\psi = 0$) larger than 20%. Note that TSS1 is the most efficient design in Figure 4 and Figures S.11–S.12 (S.M.1).

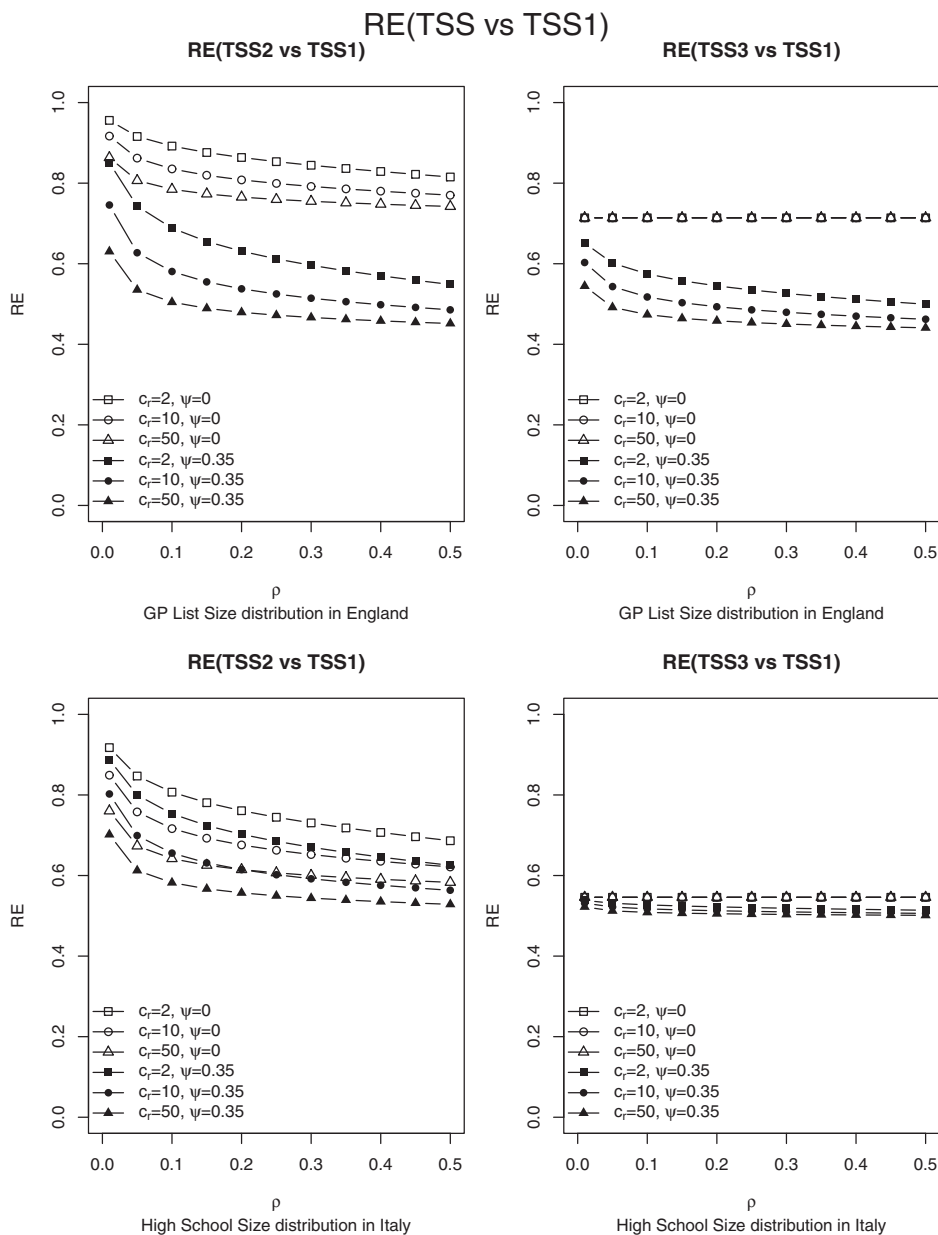


Figure 4. Relative efficiency of the optimal TSS2 versus the optimal TSS1 (left column), and of the optimal TSS3 versus the optimal TSS1 (right column), for a given research budget, as a function of ρ , for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.I (S.M.I). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

4 Maximin design

In section 3.2 it has been noticed that the optimal designs in Table 2 require a priori knowledge of some nuisance parameters (i.e. ρ , τ_N , ζ_N , η_N , and ψ). This is known as the *local optimality problem* in optimal design literature.^{20,21} Basically, this means that the optimal design is optimal only for certain values of these nuisance parameters. In this paper, the local optimality problem is solved taking a maximin approach.^{20–22} This approach has been applied in several contexts, such as longitudinal studies,^{23–25} fMRI experiments,²⁶ cluster randomized and multicentre trials,^{27–29} cost-effectiveness studies,^{30,31} life-event studies,³² test construction,³³ and biological and pharmacological studies.^{34–37} The maximin approach is composed of the following steps:

1. Define the parameter space, that is, for each unknown parameter (i.e. ρ , τ_N , ζ_N , η_N , and ψ) determine the range of plausible values (e.g. $\rho \in [0, 0.30]$).

2. Define the design space, that is, the set of all candidate designs (n, k) . In this step, one can rule out those designs that are unfeasible in practice (e.g. too many clusters to cover relative to the time available for data collection), thus preventing sample size adjustments afterwards.
3. For each design (n, k) in the design space, find those values of the nuisance parameters which minimize the efficiency $V(\hat{\mu})^{-1}$ (and thus maximize $V(\hat{\mu})$) within the range of their plausible values, as defined in step 1.
4. Choose the design that maximizes the minimum efficiency obtained in step 3. In other words, choose those values of k and n that minimize $V(\hat{\mu})$ given the worst-case values of the nuisance parameters chosen in step 3.

The resulting design is called the *maximin design*, which is the optimal design for the worst-case scenario, as defined by that set of parameter values chosen in step 3. The advantage of the maximin design is that it not only maximizes the efficiency and the power in the worst-case scenario, but it also guarantees at least that same efficiency and power level for all the other parameter values within the parameter space. Indeed, $V(\hat{\mu})$ is smaller and the power for hypothesis testing on μ is larger, for all other parameter values than for the worst-case values chosen in step 3, given any fixed sample size (i.e. k and n).

Following the four steps above, we now explain how to find the maximin design for each sampling scheme. The optimal design for TSS1 depends on ρ , τ_N , ζ_N , and ψ . However, to draw a TSS1 sample we need to know the cluster size distribution in the population anyway, which means that τ_N and ζ_N are also known before sampling. Thus, for TSS1, only ρ and ψ are unknown. The maximin design for TSS1 is obtained by plugging into the optimal sample sizes equations (Table 2, TSS1 row) the largest realistic values of ρ and ψ (for proofs, see section 3.1 in S.M.1). Unlike for TSS1, when sampling with TSS2 or TSS3 the researcher needs no prior knowledge of the whole cluster size distribution. Indeed, if such information is available, sampling with TSS1 is a better choice (Table 3). The maximin design for TSS2 and TSS3 is obtained by plugging into the optimal design equations (Table 2) the upper-bounds of the ranges for ρ , ζ_N , η_N , and ψ , and the worst-case value of τ_N (for proofs, see section 3.1 in S.M.1). The latter value can be obtained with an R function given in S.M.2 (section 2), which searches numerically for the value of τ_N that maximizes $V(\hat{\mu})$ (i.e. equations (3) and (4)) within its range of plausible values, given the worst-case values for ρ , ζ_N , η_N , and ψ . For several upper-bounds for ρ , ζ_N , η_N , and ψ , a numerical evaluation was performed and this always gave $\tau_N = 1$ as worst-case value of τ_N within the range $[0,1]$ (for details, see section 3.2 in S.M.1).

To be on the safe side in sample size planning, one can assume for ρ the parameter range $[0, 0.10]$ in health and medical research,^{38,39} and $[0, 0.25]$ in educational research.^{18,40} Lacking empirical evidence for ψ or ρ_{uN} , we propose $\psi \in [0, 0.35]$, which corresponds to $\rho_{uN} \in [-0.51, 0.51]$. The range $\tau_N \in [0,1]$ can be justified by considering Table S.7 (S.M.1), and the extreme cases of an exponential cluster size distribution, for which $\tau_N = 1$, and of a binary distribution with half of all clusters having size 2 and the other half having size $2\theta_N - 2$, for which $\tau_N \approx 1$. Finally, for ζ_N and η_N , the ranges $\zeta_N \in [0.5, 2]$ and $\eta_N \in [3, 15]$ can be chosen based on Table S.7 (S.M.1). Since $V(\hat{\mu})$ under TSS2 and TSS3 is an increasing function of ζ_N and η_N (at least if $\tau_N \leq 1$, which will usually hold), assuming positive skewness and positive excess kurtosis (i.e. $\eta_N - 3 > 0$) is a safe choice.

As mentioned in section 3.1, the optimal design for TSS2 and TSS3 depends on two approximations: the first-order Taylor series approximation used to derive $V(\hat{\mu})$ for TSS2 and TSS3 in Table 1, and the large k approximation to simplify the equations in Table 1 into equations (3) and (4). Since the maximin design is the optimal design for the worst-case scenario, the same approximations also underlie the maximin design. Based on the simulation study and the numerical evaluation discussed in S.M.1 (sections 1 and 3.3), it turned out that each approximation induces a bias of at most 5% in the $V(\hat{\mu})$ used to derive the optimal/maximin design if the optimal/maximin $k^{MD} \geq 20$, or, for $\zeta_N \geq 2$ and $\eta_N \geq 9$, $k^{MD} \geq 100$. Since $V(\hat{\mu}) \propto \frac{1}{k}$, a simple solution is to increase the maximin k^{MD} with 10% to ensure sufficient power at the expense of a 10% higher budget C . However, if the maximin $k^{MD} < 20$ or (for $\zeta_N \geq 2$ and $\eta_N \geq 9$) $k^{MD} < 100$ both approximations are biased by more than 5%. A solution is to first increase C such that maximin $k^{MD} \geq 20$ or (for $\zeta_N \geq 2$ and $\eta_N \geq 9$) $k^{MD} \geq 100$, and then further increase C by 10%.

5 Sample size calculation for cross-population comparisons

The results of the previous sections allow to efficiently plan a survey not only for estimating a mean, but also for comparing different populations, if the samples are independent. An example of such a study is the ESPAD study,⁶ which compares substance use among 15–16-year-old students across 35 European countries. For a fixed separate budget per population, the optimal design per population is given in Table 2 and the maximin design in section 4. However, the design can be further optimized by constraining the total budget (i.e. the sum of the

separate budgets) instead of each separate budget and finding the optimal (or maximin) budget split between populations (for details, see section 4 of S.M.1). For the case of comparing two populations, this optimization was formalized into a procedure to compute maximin sample sizes per population and the maximin budget split between populations, obtained by extending Van Breukelen and Candel²⁸ to TSS1 with informative cluster sizes and different cluster size distributions per population. This procedure for comparing two populations is implemented in an R code given in section 4 in S.M.2. To use this program, the researcher needs to specify c_1 and c_2 per population, τ_N and ζ_N of the cluster size distribution of each population, the largest plausible values for ρ and ψ , a range for the ratio of the outcome standard deviations (σ_y) between the two populations, the smallest difference $\mu_F - \mu_I$ that is worthwhile being detected, the maximum sum of outcome variances in both populations V_{\max} , the power level $1 - \gamma$, and the type I error rate α . The R code (S.M.2, section 4) returns the maximin sample sizes per population and the maximin budget split. The steps of this procedure are given in S.M.1 (section 4). This procedure is presented only for TSS1, because it is the most efficient sampling scheme for many cluster size distributions.

Let us demonstrate the procedure with the following example. Suppose that we want to plan a survey to estimate and compare the average alcohol consumption among high school students between France and Italy. Similar to the ESPAD study,⁶ alcohol consumption Y_{ij} is measured as the average volume of ethanol (in centilitres) consumed on the last drinking day. Based on adolescent health literature, at the design stage, school size (i.e. total number of students) can be assumed to be informative, that is, related to alcohol consumption. Indeed, it has been found that school size and school connectedness, broadly defined as the degree of belonging at school, are inversely related,^{12,13} as well as school connectedness and alcohol use.¹¹ TSS1 is the most efficient two-stage sampling scheme for both high school size distributions (this can be verified by checking the conditions in the rightmost column of Table 3, with the numbers given in the second and third row of Table S.7 of S.M.1), and so it is chosen for both populations. Suppose that we want to test the null hypothesis H_0 that $\mu_F = \mu_I$ against the alternative hypothesis H_1 that $\mu_F \neq \mu_I$, where μ_F and μ_I are the population means of alcohol consumption in France and Italy, respectively. Since the French and the Italian samples are independent, we can apply the procedure above to determine how many schools and how many students per school one has to sample per country, and how to split the total budget between countries.

The results are shown in Table 4 for four different cost scenarios. Two largest plausible values are assumed for ρ and ψ , respectively, $\rho(\max) = \{0.1, 0.2\}$ and $\psi(\max) = \{0, 0.35\}$. This combination of costs and model parameters (Table 4, first six columns) gives a total of $4 \times 2 \times 2 = 16$ scenarios, each corresponding to a row in Table 4. The seventh column in Table 4 gives the maximin budget split $\frac{C_F}{C_I}$ (i.e. the ratio of the budget for France, C_F , to that for Italy, C_I), and from the eighth to the eleventh column the maximin sample sizes per country are shown. Finally, the rightmost column of Table 4 shows the total budget required to detect a standardized difference of medium size ($d = \frac{\mu_F - \mu_I}{\sqrt{V_{\max}}} = 0.5$), with 90% power using a two-tailed test with $\alpha = 0.05$. From Table 4, it can be seen that the maximin n^{MD} per country is an increasing function of c_r , a decreasing function of ρ and ψ , and is inversely related to the maximin k^{MD} . Furthermore, the maximin budget split $\frac{C_F}{C_I} = 1$ only for $\psi = 0$ and homogeneous costs ($c_{1,F} = c_{1,I}$ and $c_{2,F} = c_{2,I}$). In all other scenarios $\frac{C_F}{C_I} < 1$, meaning that more budget is allocated to the Italian sample than to the French sample. Given that $\rho(\max)$ and $\psi(\max)$ are the same for both countries, $\frac{C_F}{C_I} < 1$ because (i) sampling a student is more expensive in Italy than in France ($c_{1,F} < c_{1,I}$), or (ii) sampling a school is more expensive in Italy than in France ($c_{2,F} < c_{2,I}$), or (iii) only for $\psi = 0.35$, the school size distribution in Italy is such that $\tau_N(\zeta_N - \tau_N)$ is larger than in France (see Tables S.7 and S.9 of S.M.1). Finally, the total budget C required for the desired power is larger for $\psi = 0.35$ than for $\psi = 0$ (Table 4, rightmost column), suggesting that ignoring informative cluster size at the design stage has the consequence of determining a research budget which is too low for the desired power level. Specifically, informative cluster size requires C to increase with 23–32% depending on the scenario (the larger ρ and/or $c_{2,I}$, the larger this relative increase, see Table 4, rightmost column).

6 Discussion

To estimate an overall mean, two-stage sampling is a logistically convenient way to collect data from a multilevel population. In practice, resources (time and money) for sampling are limited. Thus, this paper presents optimal sample sizes per design stage that either maximize the precision of the population mean estimate for the available research budget, or minimize the research budget for the required precision for estimation. Such optimal designs were derived for three TSS schemes: sampling clusters with probability proportional to cluster size, and then the

Table 4. Maximin design (n_F^{MD} , k_F^{MD} , n_I^{MD} , k_I^{MD}) and budget C needed to detect a standardized difference of medium size ($d = 0.5$) with a power of 90% using a two-tailed test with $\alpha = 0.05$ and assuming $\frac{\sigma_{y,F}}{\sigma_{y,I}} \in [\frac{1}{3}, 3]$, as a function of the maximum ψ , the maximum ρ , the cost per individual in France $c_{1,F}$ and in Italy $c_{1,I}$, and the cost for sampling a cluster in France $c_{2,F}$ and in Italy $c_{2,I}$.

$\psi(max)$	$\rho(max)$	$c_{1,F}$	$c_{2,F}$	$c_{1,I}$	$c_{2,I}$	Maximin budget		n_F^{MD}	n_I^{MD}	k_F^{MD}	k_I^{MD}	C	
						split	$\frac{c_F}{c_I}$						
0	0.1	10	200	10	200	1		13.42	13.42	14.04	14.04	9386.54	
		10	200	20	200	0.74		13.42	9.49	14.04	16.38	11077.36	
		10	200	10	400	0.64		13.42	18.97	14.04	12.39	12002.01	
		10	200	20	400	0.50		13.42	13.42	14.04	14.04	14079.82	
	0.2	10	200	10	200	1		8.94	8.94	24.33	24.33	14084.50	
		10	200	20	200	0.79		8.94	6.32	24.33	27.44	16002.68	
		10	200	10	400	0.60		8.94	12.65	24.33	22.13	18692.58	
		10	200	20	400	0.50		8.94	8.94	24.33	24.33	21126.75	
	0.35	0.1	10	200	10	200	0.98		11.31	11.10	18.52	19.08	11734.95
			10	200	20	200	0.74		11.31	7.85	18.52	21.91	13620.31
			10	200	10	400	0.61		11.31	15.70	18.52	17.09	15317.98
			10	200	20	400	0.49		11.31	11.10	18.52	19.08	17670.90
0.2		10	200	10	200	0.97		7.54	7.40	32.58	33.63	18187.89	
		10	200	20	200	0.79		7.54	5.23	32.58	37.39	20365.46	
		10	200	10	400	0.57		7.54	10.47	32.58	30.97	24601.40	
		10	200	20	400	0.49		7.54	7.40	32.58	33.63	27402.39	

same number of individuals per cluster (TSS1); sampling clusters with equal probability, and then the same percentage of individuals per cluster (TSS2); and sampling clusters with equal probability, and then the same number of individuals per cluster (TSS3).

The optimal sample size equations were derived allowing cluster size to be informative, that is, to be related to the outcome variable of interest. It turned out that the optimal designs given in Table 2 are quite robust against misspecification of the degree of informativeness of cluster size ψ . As shown in section 3.2 and in Table S.8 (S. M.1), the relative efficiency of the optimal TSS1 assuming $\psi = 0$ (i.e. non-informative cluster size) versus the optimal TSS1 assuming $\psi > 0$ (i.e. informative cluster size), when the true $\psi > 0$ was close to one. Nevertheless, ignoring informative cluster size is risky for two reasons. First, assuming $\psi = 0$ one would be tempted to combine the unweighted average of cluster means with TSS3, because this strategy (i.e. combination of sampling scheme and estimator) is unbiased and efficient for $\psi = 0$. However, this strategy is biased and inefficient if the true $\psi > 0$. Thus, assuming $\psi > 0$ is always prudent because it leads to combining the unweighted average of cluster means with TSS1, that is, choosing a strategy which is unbiased and highly efficient both for informative and non-informative cluster size. Second, assuming $\psi = 0$ can lead to underestimating the research budget for the desired power level, because the research budget is an increasing function of ψ (see Figure 2, and Table 4, rightmost column). This applies not only to TSS1, but also if, because of practical constraints, one has to choose TSS2 or TSS3 as a sampling scheme. For these two reasons, we recommend assuming $\psi > 0$ at the design stage of the survey.

The optimal designs of the three TSS schemes were compared with each other and with SRS under the constraint of a fixed budget. In contrast to what was the case under the constraint of a fixed total sample size,¹⁵ SRS can be less efficient than TSS, because it is more expensive to construct a sampling frame of all individuals in the population than of those from the selected clusters only ($c_0 > 0$), and because it is more costly to sample and measure geographically dispersed individuals than those that are grouped in a natural cluster (e.g. school, general practice) ($c_{SRS} > c_1$). Under informative cluster size, the optimal TSS1 was shown to be the most efficient sampling scheme for many cluster size distributions, followed by TSS2, and then TSS3. We thus recommend TSS1, provided all cluster sizes are known before sampling.

The optimal design depends on several unknown parameters (i.e. the intraclass correlation ρ , the informativeness parameter ψ , and the cluster size distribution's coefficient of variation τ_N , skewness ζ_N , and kurtosis η_N). To address this issue the maximin approach was proposed. For the considered TSS schemes, this strategy consists of plugging the worst-case value for each unknown parameter into the optimal design equations in Table 2. For ρ , ψ , and η_N , the largest plausible value is the worst-case value. If all plausible values for $\tau_N \leq 1$, then the largest plausible value for ζ_N is also the worst-case value. The worst-case value for τ_N can be obtained with an R code,

given in S.M.2 (section 2). However, a numerical evaluation showed that if the largest plausible value for τ_N is 1, this is the worst-case value for τ_N . The R code also returns the worst-case value for ζ_N in the rather unrealistic case that some plausible values for $\tau_N > 1$. The maximin approach has the advantages of being relatively simple to implement, and being robust against misspecification of the unknown parameters by maximizing the minimum efficiency over the ranges of their plausible values. An alternative approach is to obtain estimates of the nuisance parameters from a pilot study and use these in the sample size calculation. However, ρ risks to be underestimated (and thus the main survey to be under-powered), unless the pilot study samples a large number of clusters and of individuals per cluster, which means a sizeable portion of the limited resources for the main survey has to be devoted to the pilot study.⁴¹ The underestimation is likely to be even more severe for skewness and kurtosis, given that their traditional estimators are biased downwards unless the sample size is large or (only for the skewness) cluster size is normally distributed.⁴² For all these reasons, we recommend the maximin approach. Relatedly, to improve the planning of future surveys, empirical studies should report values of these nuisance parameters like in Table S.7 (S.M.1).

The results of this paper also allow to efficiently plan surveys for comparing different populations, provided the samples are independent. For TSS1, a procedure to derive maximin sample sizes and maximin budget split between populations was obtained by extending Van Breukelen and Candel's²⁸ findings to informative cluster size. Analogous extensions for TSS2 and TSS3 could be explored. However, when either cluster size is non-informative ($\psi = 0$), or the cluster size distribution as well as the informativeness parameter α_1 is the same in both populations (e.g. treated and control groups in a cluster randomized trial), we have that $\mu_F - \mu_I = \beta_{0,F} - \beta_{0,I}$ (see equation (2)) and then the equations given in this paper reduce to simpler expressions as also derived by Van Breukelen and Candel²⁸ (i.e. those for TSS1 with $\psi = 0$).

Finally, in this paper the model-based approach to survey sampling was adopted. However, the results of this paper are valid also under the design-based approach, provided model (1) and assumption 4 hold and inference is then based on the sampling scheme.¹⁵ Future research could extend the results of this paper by considering dichotomous outcomes, three-level populations, and by deriving the optimal design for longitudinal studies to monitor trends.


Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Francesco Innocenti  <https://orcid.org/0000-0001-6113-8992>

Math JJM Candel  <https://orcid.org/0000-0002-2229-1131>

Gerard JP van Breukelen  <https://orcid.org/0000-0003-0949-0272>

Supplemental material

Supplemental material for this article is available online.

References

1. Cochran WG. *Sampling techniques*. 3rd ed. New York: John Wiley & Sons, 1977.
2. Lohr SL. *Sampling: design and analysis*. 2nd ed. Boston: Brooks/Cole, 2010.
3. Chambers RL, Clark RG. *An introduction to model-based survey sampling with applications*. Oxford: Oxford University Press, 2012.
4. Patton GC, Hibbert M, Rosier MJ, et al. Patterns of common drug use in teenagers. *Aust N Z J Public Health* 1995; **19**: 393–399.
5. Warren CW, Riley L, Asma S, et al. Tobacco use by youth: a surveillance report from the Global Youth Tobacco Survey project. *Bull World Health Organ* 2000; **78**: 868–876.
6. ESPAD Group. *Results from the European school survey project on alcohol and other drugs*. ESPAD Report 2015. Luxembourg, 2016.

7. DeFrances CJ, Lucas CA, Buie VC, et al. 2006 National Hospital discharge survey. National Health Statistics Report. Report no. 5, 30 July 2008, Hyattsville.
8. Jones A. The National Nursing Home survey: 1999 summary. *Vital Health Stat 13* 2002; **152**: 1–116.
9. Seaman S, Pavlou M and Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Stat Med* 2014; **33**: 5371–5387.
10. Nevalainen J, Datta S and Oja H. Inference on the marginal distribution of clustered data with informative cluster size. *Stat Pap* 2014; **55**: 71–92.
11. Resnick MD, Bearman PS, Blum RW, et al. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *JAMA* 1997; **278**: 823–832.
12. McNeely CA, Nonnemaker JM and Blum RW. Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *J Sch Health* 2002; **72**: 138–146.
13. Thompson DR, Iachan R, Overpeck M, et al. 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**: 379–386.
14. Kelly E and Stoye G. *Does GP practice size matter? GP practice size and the quality of primary care*. Institute for Fiscal Studies. Report no. R101, November 2014, London.
15. Innocenti F, Candel MJJM, Tan FES, et al. Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Stat Med* 2019; **38**: 1817–1834.
16. R Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2017.
17. Goldstein H. *Multilevel Statistical Models*. 4th ed. Chichester: John Wiley & Sons, 2011.
18. Shackleton N, Hale D, Bonell C, et al. Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM Popul Health* 2016; **2**: 217–225.
19. Moerbeek M, Van Breukelen GJP and Berger MPF. Design issues for experiments in multilevel populations. *J Educ Behav Stat* 2000; **25**: 271–284.
20. Atkinson AC, Donev AN and Tobias RD. *Optimum experimental designs, with SAS*. Oxford: Oxford University Press, 2007.
21. Berger MPF and Wong WK. *An introduction to optimal designs for social and biomedical research*. Chichester: John Wiley & Sons, 2009.
22. Wong WK. A unified approach to the construction of minimax designs. *Biometrika* 1992; **79**: 611–619.
23. Ouwens JNM, Tan FES and Berger MPF. Maximin D-optimal design for longitudinal mixed effects models. *Biometrics* 2002; **58**: 735–741.
24. Winkens B, Schouten HJA, Van Breukelen GJP, et al. Optimal designs for clinical trials with second-order polynomial treatment effects. *Stat Meth Med Res* 2007; **16**: 523–537.
25. Tekle FB, Tan FES and Berger MPF. Maximin D-Optimal designs for binary longitudinal responses. *Comput Stat Data Anal* 2008; **52**: 5253–5262.
26. Maus B, Van Breukelen GJP, Goebel R, et al. Robustness of optimal design of fMRI experiments with application of a genetic algorithm. *NeuroImage* 2010; **49**: 2433–2443.
27. Candel MJJM and Van Breukelen GJP. Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Stat Meth Med Res* 2015; **24**: 557–573.
28. Van Breukelen GJP and Candel MJJM. Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Stat Med* 2018; **37**: 3027–3046.
29. Wu S, Wong WK and Crespi CM. Maximin optimal designs for cluster randomized trials. *Biometrics* 2017; **73**: 916–926.
30. Manju MA, Candel MJJM and Berger MPF. Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Stat Med* 2014; **33**: 2538–2553.
31. Manju MA, Candel MJJM and Berger MPF. Optimal and maximin sample sizes for multicentre cost-effectiveness trials. *Stat Meth Med Res* 2015; **24**: 513–539.
32. Tan FES. Conditions for D_A -maximin marginal designs for generalized linear mixed models to be uniform. *Commun Stat Theory Meth* 2010; **40**: 255–266.
33. Berger MPF, King CYJ and Wong WK. Minimax D-Optimal designs for item response theory models. *Psychometrika* 2000; **65**: 377–390.
34. King CYJ and Wong WK. Minimax D-optimal designs for the logistic model. *Biometrics* 2000; **56**: 1263–1267.
35. Dette H and Biedermann S. Robust and efficient designs for the Michaelis-Menten Model. *J Am Stat Assoc* 2003; **98**: 679–686.
36. Dette H, Lopez IM, Ortiz Rodriguez IM, et al. Maximin efficient design of experiment for exponential regression models. *J Stat Plan Inference* 2006; **136**: 4397–4418.
37. Pronzato L and Walter E. Robust experiment design via maximin optimization. *Math Biosci* 1988; **89**: 161–176.
38. Adams G, Gulliford MC, Ukoumunne OC, et al. Patterns of intracluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004; **57**: 785–794.

39. Eldridge SM, Ashby D, Feder GS, et al. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials* 2004; **1**: 80–90.
40. Hedges LV and Hedberg EC. Intraclass correlation values for planning group-randomized trials in education. *Educ Eval Policy Anal* 2007; **29**: 60–87.
41. Eldridge SM, Costelloe CE, Kahan BC, et al. How big should the pilot study for my cluster randomised trial be? *Stat Meth Med Res* 2016; **25**: 1039–1056.
42. Joanes DN and Gill CA. Comparing measures of sample skewness and Kurtosis. *J R Stat Soc D* 1998; **47**: 183–189.

Appendix

Notation

Section	Symbol	Definition
2	K	number of clusters in the population
	j	index for clusters
	N_j	size of cluster j in the population
	$N_{pop} = \sum_{j=1}^K N_j$	population size
	k	number of clusters in the sample
	$\theta_N = \frac{N_{pop}}{K}$	population mean of cluster size
	σ_N^2	population variance of cluster size
	$\tau_N = \frac{\sigma_N}{\theta_N}$	population coefficient of variation of cluster size
	$\zeta_N = \frac{E[(N_j - \theta_N)^3]}{\sigma_N^3}$	population skewness of cluster size
	$\eta_N = \frac{E[(N_j - \theta_N)^4]}{\sigma_N^4}$	population kurtosis of cluster size
	$\bar{N} = \frac{\sum_{j=1}^k N_j}{k}$	average population size of the sampled clusters
	m	number of individuals sampled with SRS
	i	index for individuals
	Y_{ij}	outcome variable of interest
	ε_{ij}	effect of individual i in cluster j
	σ_ε^2	population variance of ε_{ij}
	ν_j	component of cluster effect that does not depend on cluster size
	σ_ν^2	population variance of ν_j
	β_0	average of all cluster-specific means in the population
	α_0	intercept of the relation between cluster effect and cluster size
	α_1	slope of the relation between cluster effect and cluster size
	$u_j = \alpha_0 + \alpha_1 N_j + \nu_j$	effect of cluster j
	$E(u_j) = 0$	and $V(u_j) = \sigma_u^2 = \sigma_\nu^2 + \alpha_1^2 \sigma_N^2$ = population mean and variance of u_j
	μ	average of all individual outcomes in the population
	$\hat{\mu}$	population mean estimator
	$V(\hat{\mu})$	sampling variance of $\hat{\mu}$
	$\rho_{uN} = \frac{E[u_j(N_j - \theta_N)]}{\sqrt{\sigma_\nu^2 + \alpha_1^2 \sigma_N^2} \sigma_N}$	correlation between u_j and N_j
	$\psi = \left(\frac{\rho_{uN}^2}{1 - \rho_{uN}^2} \right)$	degree of informativeness of cluster size
	$\sigma_y^2 = \sigma_\nu^2 + \sigma_\varepsilon^2$	total unexplained outcome variance
	$\rho = \frac{\sigma_\nu^2}{\sigma_y^2}$	intraclass correlation coefficient
π_j	inclusion probability of cluster j	
π_{ij}	conditional inclusion probability of individual i	
n_j	number of individuals sampled per cluster for TSS2	
$p = \frac{n_j}{N_j}$	proportion of individuals sampled per cluster for TSS2	

	$\bar{n} = \frac{\sum_{j=1}^k n_j}{k}$	average sample size of the sampled clusters
	n	number of individuals sampled per cluster for TSS1 and TSS3.
		Expected value of \bar{n} for TSS2
3.1	π_i	inclusion probability of individual i under SRS
	C	budget for sampling and measuring
	c_2	(average) cost for sampling a cluster
	c_1	(average) cost for sampling an individual from a sampled cluster
	c_0	extra-cost due to constructing the sampling frame for SRS compared with the sampling frame for TSS
	c_{srs}	(average) cost for sampling an individual directly from the population
	$V(\hat{\mu})^*$	sampling variance of $\hat{\mu}$ under the optimal design
	n^*	optimal number of individuals per cluster
	k^*	Optimal number of clusters
	$p^* = \frac{n^*}{\theta_N}$	optimal proportion of individuals to sample per cluster for TSS2
3.2	$c_r = \frac{c_2}{c_1}$	cluster-to-individual cost ratio
	α	type I error rate
	γ	type II error rate
	z_q	qth percentile of the standard normal distribution
	$g(\rho, \psi)$	numerator of $V(\hat{\mu})^*$ in Table 2 excluding σ_y^2
	μ_0	value of μ under H_0
	$d_0 = \frac{\mu - \mu_0}{\sigma_y}$	standardized difference for one-sample t-test
3.3	$RE(D1 \text{ vs } D2) = \frac{V_{D2}(\hat{\mu})^*}{V_{D1}(\hat{\mu})^*}$	relative efficiency of optimal design $D1$ versus optimal design $D2$
4	k^{MD}	maximin number of clusters
5	$\mu_F, \mu_I, \sigma_{y,F}^2, \sigma_{y,I}^2$	population mean and total unexplained outcome variance in France (F), and Italy (I)
	$V_{\max} \geq \sigma_{y,F}^2 + \sigma_{y,I}^2$	maximum plausible upper-bound for $\sigma_{y,F}^2 + \sigma_{y,I}^2$
	$\rho(\max)$ and $\psi(\max)$	largest plausible values assumed for ρ and ψ
	$d = \frac{\mu_F - \mu_I}{\sqrt{\frac{V_{\max}}{2}}}$	standardized difference for unpaired two-sample t-test
	n^{MD}	maximin number of individuals per cluster