Optimal Two-Stage Sampling for Mean Estimation in Multilevel Populations when Cluster Size is Informative.
Supplementary Material 1

Francesco Innocenti\textsuperscript{1*}, Math J.J.M. Candel\textsuperscript{1}, Frans E.S. Tan\textsuperscript{1}, Gerard J.P. van Breukelen\textsuperscript{1,2}

\textsuperscript{1} Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, the Netherlands.
\textsuperscript{2} Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands.
\*Correspondence to: Francesco Innocenti, Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, P. Debyeplein 1, 6229 HA Maastricht, the Netherlands. E-mail: francesco.innocenti@maastrichtuniversity.nl

Contents

1. A Simulation Study for the Evaluation of the Accuracy of the Approximations for $E(\hat{\mu})$ and $V(\hat{\mu})$ under TSS2 and TSS3 ................................................................. 2
1.1. Cluster size distributions. ................................................................. 3
1.2. Simulation procedure. ................................................................. 5
1.3. Simulation results. ................................................................. 7
2. Optimal Design and Relative Efficiencies for a given budget ................................................................. 8
2.1. Sampling Variances. ................................................................. 8
2.2. Derivation of the Optimal Design. ................................................................. 11
2.3. Relative Efficiencies. ................................................................. 13
3. Local Optimality Problem and Maximin Design ................................................................. 22
3.1. Derivation of the Maximin Design. ................................................................. 22
3.2. A general program in \texttt{R} to find the maximin parameter values. ................................................................. 28
3.3. Accounting for the approximations in the maximin design for TSS2 and TSS3. ................................................................. 28
4. Sample Size Calculation for Cross-Population Comparisons ................................................................. 30
References ................................................................. 32

List of Tables

Table S.1. Components of $E(\hat{\mu})$ and $V(\hat{\mu})$ under TSS2 and TSS3. ................................................................. 4
Table S.2. Summary of the Cluster Size Distributions. ................................................................. 5
**Table S.3.** Relative Bias for $E(V(\mu|N))$ under TSS3 and TSS2, for $n=20$. ........................................... 9

**Table S.4.** Relative Bias for $V(E(\mu|N))$ and $V(\hat{\mu})$ under TSS3, for $n=20$. ........................................... 9

**Table S.5.** Relative Bias for $V(E(\mu|N))$ and $V(\hat{\mu})$ under TSS2, for $n=20$. ........................................... 10

**Table S.6.** Relative Bias for $V(\hat{\mu})$ under TSS2 and TSS3, for different values of $n$. ........................................... 10

**Table S.7.** Real Cluster Size Distributions. ........................................... 14

**Table S.8.** Robustness of the optimal design against misspecification of $\psi$. ........................................... 16

**Table S.9.** Maximin budget split. ........................................... 32

**List of Figures**

**Figure S.1.** Summary of the Cluster Size Distributions. ........................................... 6

**Figure S.2.** Optimal number of individuals per cluster for TSS1. ........................................... 14

**Figure S.3.** Optimal number of individuals per cluster for TSS1. ........................................... 14

**Figure S.4.** Optimal number of individuals per cluster for TSS2. ........................................... 15

**Figure S.5.** Budget $C$ for the Optimal TSS1. ........................................... 17

**Figure S.6.** Budget $C$ for the Optimal TSS2. ........................................... 18

**Figure S.7.** Budget $C$ for the Optimal TSS3. ........................................... 19

**Figure S.8.** Relative Efficiency of TSS1 versus SRS. ........................................... 23

**Figure S.9.** Relative Efficiency of TSS2 versus SRS. ........................................... 24

**Figure S.10.** Relative Efficiency of TSS3 versus SRS. ........................................... 25

**Figure S.11.** Relative Efficiency of TSS2 versus TSS1. ........................................... 26

**Figure S.12.** Relative Efficiency of TSS3 versus TSS1. ........................................... 27

**Figure S.13.** Ratio of $V(\hat{\mu})_A$ to $V(\hat{\mu})_L$ ........................................... 29

**1 A Simulation Study for the Evaluation of the Accuracy of the Approximations for $E(\hat{\mu})$ and $V(\hat{\mu})$ under TSS2 and TSS3**

A simulation study was performed to find how many clusters $k$ must be sampled to obtain a fairly unbiased estimator of $\mu$ under TSS2 and TSS3 (Table 1, main text), and to evaluate the accuracy of the approximations in the expressions for the marginal variances $V(\hat{\mu})$ under TSS2 and TSS3 (Table 1, main text). Specifically, these expressions have been derived in Innocenti et al. [1] by applying the Delta method [2], which linearizes a nonlinear function (e.g. $E(\hat{\mu})$ and $V(\hat{\mu})$) through a Taylor series approximation (for details, see Appendix B in [1]). Thus, a simulation study was needed to assess the accuracy of these approximations, which depends on sample size $k$, but also on the distribution of cluster size in the population, the intraclass correlation coefficient $\rho$, and the correlation between cluster effect and cluster size $\rho_{\mu N}$. Note that $E(\hat{\mu}) = E(E(\hat{\mu}|N))$ and $V(\hat{\mu}) = E(V(\hat{\mu}|N)) + V(E(\hat{\mu}|N))$, and each component of $E(\hat{\mu})$ and $V(\hat{\mu})$ is shown in Table S.1 From the expression for $V(\hat{\mu})$ in Table S.1 (i.e. as sum of the last two rows) one can obtained the equation for $V(\hat{\mu})$ in Table 1 (main text) by replacing $\sigma_N^2 = \rho \sigma_Y^2$, $\sigma^2 = (1-\rho) \sigma_Y^2$, and $\alpha_Y^2 = \left(\frac{\rho_{\mu N}}{1-\rho_{\mu N}}\right) \sigma_N^2 = \psi \rho \sigma_Y^2$, and with some rearrangements (see section 2.1).
Six cluster size distributions were considered: a negative binomial, a discrete uniform, a four-parameter beta, a bimodal distribution, the distribution of general practice patient list size in England, and the distribution of public high school size in Italy. These distributions are described in the next paragraph. From these distributions, ten scenarios were chosen in order to cover a broad range of values for the coefficient of variation $\tau_N \in [0.026, 0.996]$, skewness $\zeta_N \in [-0.496, 2.12]$ and kurtosis $\eta_N \in [1.677, 14.549]$ of cluster size. These scenarios are summarized in Table S.2 and plotted in Figure S.1. Note that, in accordance with assumption 2 (i.e. $\frac{\bar{n}}{N} \rightarrow 0$), the average cluster size in the population $\bar{N}$ was taken at least equal to 400, so that for $\bar{n} = 20$ a negligible sampling fraction is obtained at the second stage. The value of $\bar{n} = 20$ was chosen because it was the median number of individuals sampled per cluster in 21 school-based surveys reviewed by Shackleton et al. [3]. The true values for those parameters in $E(\bar{N})$ and $V(\bar{N})$ (Table S.1, last three rows) that are not related to cluster size distribution were $\bar{N}_0 = 10$, $\sigma_0 = 0.5$, $\rho = \{0.05, 0.30\}$ and $\sigma^2 = \sigma^2 + \sigma^2 = \{5, 0.833\}$ (which leads to the first and second value of $\rho$, respectively), and $\rho_{2N} = \{0.25, 0.5, 0.75\}$. Finally, the values for the sample sizes were $k = \{20, 30, 40\}$ and $n = 20$. Thus, the simulation study involved $10 \times 2 \times 3 \times 3 = 180$ scenarios in total. However, some additional scenarios with $n = \{10, 50, 100\}$ (values that are in line with the sample sizes reported by Shackleton et al. [3]) were explored for some cluster size distributions in order to study the effect of $n$ on the accuracy of the approximations.

1.1 Cluster size distributions

The cluster size distributions used in the simulation study are presented in detail below.

**Negative Binomial.** It was chosen because it is a positively skewed distribution that allows to consider a broad range of values for the coefficient of variation of cluster size. The probability mass function is

$$Pr(X = x; r, \pi) = \frac{\Gamma(\frac{r+x}{\pi})}{\Gamma(r)} \frac{\pi^r}{(1-\pi)^x},$$

with mean and variance $\theta_X = \frac{r(\pi-1)}{\pi}$ and $\sigma^2_X = \frac{r(\pi-1)}{\pi^2}$, respectively. However, the following re-parametrization ([2], p. 96) makes easier to choose and interpret the moments of the Negative Binomial distribution. Substituting $\theta_X$ and $\pi = \frac{\theta_X}{\theta_X + 1}$ in $\sigma^2_X$, we obtain $\sigma^2_X = \theta_X \left(1 + \frac{\theta_X}{\theta_X + 1}\right)$. Hence, the coefficient of variation is $\tau_X = \sqrt{\frac{1}{\theta_X} + \frac{1}{\pi}}$, the skewness is $\zeta_X = \frac{2 - \pi}{\sqrt{\pi(1-\pi)}} = \frac{1}{\sqrt{\theta_X}} \left(\sqrt{\theta_X} + 2 \sqrt{\frac{\theta_X}{\pi}}\right)$, and the kurtosis is $\eta_X = 3 + \frac{\theta^2}{\pi(1-\pi)} = 3 + \frac{\theta}{\pi} + \frac{1}{\tau_X^2}$ ([4], p.140). To prevent sampling clusters with size equal to 0, the cluster size distribution is then given by the following transformation: $N_j = X_j + 2$, where $X_j$ is the Negative Binomial described previously. Thus, $\theta_N = \theta_X + 2$, $\sigma_N^2 = \sigma_X^2$, $\tau_N = \frac{\sigma}{\theta_X + 2}$, $\zeta_N := \frac{\sqrt{\theta_X + 2}}{\sqrt{\theta_X}} \zeta_X$, $\eta_N := \frac{\sqrt{\theta_X + 2}}{\sqrt{\theta_X}} \eta_X$. For this distribution, two scenarios were considered: (i) $\tau_N = 0.5$, $\theta_N = 402$, $\zeta_N = 1$, and $\eta_N = 4.5$, obtained by setting $r = 4$, (ii) $\tau_N = 0.996$, $\theta_N = 402$, $\zeta_N = 2$, and $\eta_N = 9$, obtained with $r = 1$. These two scenarios were chosen to assess the effect of an increase of the coefficient of variation, skewness, and kurtosis, on the bias of the first-order Taylor series approximations.

**Discrete Uniform.** A Discrete Uniform distribution on the interval $[2, 8]$ was chosen as an example of symmetric distribution. The parameters of this distribution are $\theta_N = \frac{a+2}{2}$, $\sigma_N^2 = \frac{\left(\frac{q-1}{2}\right)^2-1}{12}$, $\tau_N = \frac{1}{\sqrt{12}} \sqrt{\left(\frac{\left(\frac{q-1}{2}\right)^2-1}{(q+2)^2}\right)} \in [0, 0.575]$, $\zeta_N = 0$, and $\eta_N = 3 - \frac{6(q-1)^2+1}{5(q-1)^2-1}$. The considered scenario was for the interval $[2, 800]$, so that $\theta_N = 401$, $\tau_N = 0.575$, $\zeta_N = 0$, and $\eta_N = 1.8$. The interval $[2, 800]$ was chosen to have $\theta_N = 401$, so that for $\bar{n} = 20$ a negligible sampling fraction is obtained at the second stage (i.e. $\bar{n}/\theta_N \rightarrow 0$, see assumption 2 in the main text).

**Four-parameter Beta.** The Four-parameter Beta distribution is a generalization of the Beta distribution which extends its range to the interval $[\min, \max]$, where $\min$ and $\max$ can be different from 0 and 1, respectively. The probability density function is

$$f(N_j) = \frac{(N_j - \min)^{a-1} (\max - N_j)^{b-1}}{B(a, b) (\max - \min)^{a+b-1}},$$

where $a > 0$, $b > 0$, $\min < \max$, and $B(a, b)$ is the Beta function. The moments of this distribution are $\theta_N = \min + \frac{a}{a+b} (\max - \min)$, $\sigma_N^2 = \frac{ab}{(a+b)^2(a+b+1)} (\max - \min)^2$, $\zeta_N = \frac{2(b-a)}{\sqrt{a+b+1}} \sqrt{\frac{\max - \min}{a+b+1}}$, and $\eta_N = \frac{a+b+1}{ab(a+b+2)} (2(a+b)^2 + ab(a+b-6)).$
Thus, the coefficient of variation is \( \tau_N = \sqrt{\frac{\hat{\mu} \cdot \phi}{(\hat{\mu}^2 + \phi^2) \cdot (\frac{1}{2})}} \). This distribution was chosen because it can have negative, positive, and zero skewness. It is negatively skewed for \( a > b > 1 \). Unfortunately, one cannot have \( \zeta_N < 0 \) (i.e. \( b < a \)) and \( \tau_N > 0.577 \) at the same time. Hence, the following scenarios were considered: (i) \( \tau_N = 0.54, \theta_N = 420, \zeta_N = -0.082, \eta_N = 1.832, \) obtained with \( a = 1.1, b = 1, \min = 2, \max = 800; \) (ii) \( \tau_N = 0.266, \theta_N = 534, \zeta_N = -0.468, \eta_N = 2.625, \) obtained with \( a = 4, b = 2, \min = 2, \max = 800; \) (iii) \( \tau_N = 0.445, \theta_N = 401, \zeta_N = 0, \) \( \eta_N = 2.143, \) obtained with \( a = b = 2, \min = 2, \max = 800. \) The first and second scenarios were chosen to assess the effect of a small and large negative skewness, respectively, on the bias of the first-order Taylor series approximations. The third scenario was chosen to have a symmetric unimodal distribution (see Figure S.1, second row, left panel). Note that a scenario with a positive skewness is not considered for this distribution because covered by the Negative Binomial. Furthermore, note that the Discrete Uniform is a special case of the Four-parameter Beta, for \( a = b = 1 \) and the rounding off due to the discrete nature of cluster size.

**Bimodal.** The bimodal distribution was obtained as a mixture of two distributions: a Four-parameter Beta (i.e. \( X_1 \sim \text{Beta}(\min = 2, \max = 1000, a = 2, b = 4) \)) with mean \( \theta_1 \) and variance \( \sigma_1^2 \) and a Normal density (i.e. \( X_2 \sim N(\theta_2 = 1000, \sigma_2 = 50) \)). This mixture is motivated by the need to prevent negative cluster sizes, which are allowed by the Normal distribution but not by the Four-parameter Beta distribution. If we denote by \( w_1 \) and \( w_2 = 1 - w_1 \) the weights of the mixture, then the moments of the mixture \( N_j = w_1 X_{1j} + w_2 X_{2j} \) are (derived from Table 2.2. p14, p35 in [4]) \( \theta_N = \theta_1 \times w_1 + \theta_2 \times w_2, \sigma_N^2 = (w_1 \sigma_1^2 + w_2 \sigma_2^2) + (w_1 \theta_1^2 + w_2 \theta_2^2) - (\theta_1 w_1 + \theta_2 w_2)^2, \zeta_N = \frac{E(N_j - \theta_N)^4}{\sigma_N^4} = w_1[(\theta_1 - \theta_N)^4 + (\theta_2 - \theta_N)\frac{\theta_1^2 - \theta_2^2}{\sigma_1^2} + \frac{\theta_1^2 + \theta_2^2}{\sigma_2^2}] - (\theta_1 w_1 + \theta_2 w_2)^2 \cdot \frac{1}{\sigma_1^2 + \sigma_2^2}. \)

\( \eta = \frac{E(N_j - \theta_N)^4}{\sigma_N^4} = \frac{1}{\sigma_1^2 + \sigma_2^2}[(w_1 \theta_1^4 + w_2 \theta_2^4) - (\theta_1 w_1 + \theta_2 w_2)^2 \cdot \frac{1}{\sigma_1^2 + \sigma_2^2}] \). Two scenarios were considered: (i) \( \theta_N = 401, \tau_N = 0.654, \zeta_N = 0.889, \eta_N = 3.576, \) obtained by setting \( w_1 = 0.9, \) and \( w_2 = 0.1; \) (ii) \( \theta_N = 701, \tau_N = 0.504, \zeta_N = -0.496, \eta_N = 1.677, \) obtained with \( w_1 = 0.45, \) and \( w_2 = 0.55. \) The first scenario can be compared with the first Negative Binomial in terms of bimodality vs unimodality, given that both distributions are positively skewed, have positive excess kurtosis and similar coefficients of variation (see Figure S.1). The second scenario was chosen to have a negatively skewed bimodal distribution.

**The General Practice Patient List Size distribution in England.** In October 2017, 58,719,921 \( = N_{\text{pop}} \) patients were registered at 7,353 = \( K \) general practices in England [5]. The parameters of the distribution of patient list size for general practices are \( \theta_N = 7986, \tau_N = 0.633, \zeta_N = 2.12, \) and \( \eta_N = 14.549. \)

**The Public High School Size distribution in Italy.** 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 [6]. The parameters of the distribution of high school size are \( \theta_N = 403, \tau_N = 0.912, \zeta_N = 1.256, \) and \( \eta_N = 4.315. \)

| \( \mu \) | TSS2 | TSS3 |
|---|---|---|
| \( \sum_{j=1}^{p} \rho_{N_j} \beta_j \) | \( \sum_{j=1}^{p} \rho_{N_j} \beta_j \) | \( \sum_{j=1}^{p} \rho_{N_j} \beta_j \) |
| \( E(\mu|\mathbf{N}) \) | \( \beta_0 + \alpha_1 \left( N (\frac{C V^2}{N} + 1) - \theta_N \right) \) | \( \beta_0 + \alpha_1 \left( N (\frac{C V^2}{N} + 1) - \theta_N \right) \) |
| \( V(\mu|\mathbf{N}) \) | \( \frac{\beta_0 + \alpha_1 (C V^2 + 1) \sigma_N^2 + \sigma_1^2}{n k} \) | \( \frac{\beta_0 + \alpha_1 (C V^2 + 1) \sigma_N^2 + \sigma_1^2}{n k} \times (\frac{C V^2}{N} + 1) \) |
| \( E(V(\mu|\mathbf{N})) \) | \( \frac{n k}{n k} \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) | \( \frac{n k}{n k} \times \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) |
| \( V(E(V(\mu|\mathbf{N})) \) | \( \frac{\alpha_1^2}{4} \sigma_1^2 \left( \frac{C V^2}{N} + 1 \right) \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) | \( \frac{\alpha_1^2}{4} \sigma_1^2 \left( \frac{C V^2}{N} + 1 \right) \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) |
| \( \frac{\alpha_1^2}{4} \sigma_1^2 \left( \frac{C V^2}{N} + 1 \right) \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) | \( \frac{\alpha_1^2}{4} \sigma_1^2 \left( \frac{C V^2}{N} + 1 \right) \left( \frac{\beta_0 + \alpha_1 (\frac{C V^2}{N} + 1)}{\sigma_N^2 + \sigma_1^2} \right) \) |
|
|---|
| **Table S.2.** Summary of the Cluster Size Distributions. |

| Positive Skewness | Negative Binomial |
|-------------------|-------------------|
| $\zeta_N > 0$     | $\theta_N = 402, \tau_N = 0.5, \zeta_N = 1, and \eta_N = 4.5$ |
| **Negative Binomial** | $\theta_N = 402, \tau_N = 0.996, \zeta_N = 2, and \eta_N = 9$ |
| **Bimodal**       | $\theta_N = 401, \tau_N = 0.654, \zeta_N = 0.889, and \eta_N = 3.576$ |

| Symmetric | $\zeta_N = 0$ |
|-----------|----------------|
| Discrete Uniform on [2,800], | $\theta_N = 401, \tau_N = 0.575, \zeta_N = 0, and \eta_N = 1.8$ |
| **Four-parameter Beta on [2,800],** | $\theta_N = 401, \tau_N = 0.445, \zeta_N = 0, \eta_N = 2.143$ |

| Negative Skewness | $\zeta_N < 0$ |
|-------------------|----------------|
| **Four-parameter Beta on [2,800],** | $\theta_N = 420, \tau_N = 0.54, \zeta_N = -0.082, \eta_N = 1.832$ |
| **Bimodal**       | $\theta_N = 534, \tau_N = 0.266, \zeta_N = -0.468, \eta_N = 2.625$ |

1.2 Simulation procedure

To generate the true values of $E(\bar{\mu})$ and $V(\bar{\mu})$ under TSS2 and TSS3, and compare with the approximations in Table S.1 (computed by plugging the parameters and sample sizes into $E(\bar{\mu})$ and $V(\bar{\mu})$ as given in Table S.1, respectively, fourth row and sum of last two rows), the following procedure was performed.

**Step 1.** Sample $k$ clusters from the population, that is, generate $k$ cluster sizes $(N = (N_1, \ldots, N_k)^T)$ from the cluster size distribution (see Table S.2 and Figure S.1) (outer loop). Note that non-integer cluster sizes were rounded off.

**Step 2.** (a) Generate $k$ cluster effects $v_j$ from $N(0, \sigma^2)$. (b) For each sample cluster, sample from $N(0, \sigma^2)$ $n$ individuals (with their $e_i, s$) for TSS2, and $n_j = pN_j$ individuals for TSS3. (c) Generate the $Y$’s for the $nk$ individuals in the sample with the equation $y_{ij} = \beta_0 + \alpha_i (N_j - \theta_N) + v_j + e_{ij}$. (d) Compute $\hat{\mu}$ (recall that $\hat{\mu}_{TSS3} = \mu_{TSS2} = \frac{\sum_{j=1}^{k} N_j \bar{Y}_j}{\sum_{j=1}^{k} N_j}$). Repeat Step 2 500 times ($s = 1, \ldots, 500$) (inner loop).

**Step 3.** Denote by $\hat{\mu}$, the estimate of $\mu$ obtained at the $s$-th iteration of Step 2. Compute $E(\hat{\mu} | N) = \frac{\sum_{j=1}^{k} \mu_j}{\sum_{j=1}^{k} n_j}$ and $V(\hat{\mu} | N) = \frac{\sum_{j=1}^{k} (\mu_j - E(\mu_j))^2}{\sum_{j=1}^{k} n_j}$. Repeat Steps 1-3 5000 times ($l = 1, \ldots, 5000$).

**Step 4.** Denote by $E(\bar{\mu} | N)_l$ and $V(\bar{\mu} | N)_l$ the values of the conditional mean and variance obtained at the $l$-th iteration of Step 3. Compute $E(\mu | N)_G.V. = \frac{\sum_{l=1}^{500} E(\bar{\mu} | N)_l}{5000}$, $E(\mu | N)_G.V. = \frac{\sum_{l=1}^{500} V(\mu | N)_l}{5000}$, $V(\mu | N)_G.V. = \frac{\sum_{l=1}^{500} (E(\mu | N)_l - E(\mu | N))_G.V.}{4999}$, and $V(\mu | N)_G.V. = E(\mu | N)_G.V. + V(\mu | N)_G.V.$, where the subscript $G.V.$ is short for “generated value”, as opposed to the values to which these $G.V.$s will be compared to evaluate the accuracy of the approximations that, instead, are obtained by plugging the parameters and sample sizes into the equations given in Table S.1.

Note that the $k$ cluster effects ($v_j$) are generated in the inner loop (i.e. step 2.a) instead of the outer loop (i.e. step 1), because $E(\mu | N)$ and $V(\mu | N)$ (obtained in step 3) are conditional on cluster size only (see Table S.1, second and third row). Instead, generating cluster effects already in step 1 would entail a conditioning of $E(\mu | N)$ and $V(\mu | N)$ also on

\[ \tilde{N} = \frac{\sum_{j=1}^{k} N_j}{k}, \text{ and } S_N = \sqrt{\frac{\sum_{j=1}^{k}(N_j - \tilde{N})^2}{k}}, \]  

$\tau_N = \frac{\alpha_N}{\bar{\alpha}_N}$ is the population coefficient of variation of cluster size, $\zeta_N = E\left[\left(N - \tilde{N}\right)^3\right]$ is the skewness and $\eta_N = E\left[\left(N - \tilde{N}\right)^4\right]$ is the kurtosis of cluster size distribution. Furthermore, for TSS2 $n = E(\tilde{n}) = pE(\tilde{N}) = p\bar{\eta}_N$. 

\[ \begin{align*} 
\text{Derivations are given in [1]. Note } N &= (N_1, \ldots, N_k)^T, CV_N = \frac{S_N}{\tilde{N}} \text{ is the sample coefficient of variation of cluster size, where } \\
&\bar{N} = \frac{\sum_{j=1}^{k} N_j}{k} \text{ and } S_N = \sqrt{\frac{\sum_{j=1}^{k}(N_j - \bar{N})^2}{k}}, \text{ and } \tau_N = \frac{\alpha_N}{\bar{\alpha}_N} \text{ is the population coefficient of variation of cluster size, } \\
&\zeta_N = E\left[\left(N - \bar{N}\right)^3\right] \text{ is the skewness and } \eta_N = E\left[\left(N - \bar{N}\right)^4\right] \text{ is the kurtosis of cluster size distribution. Furthermore, for TSS2 } n = E(\tilde{n}) = pE(\tilde{N}) = p\bar{\eta}_N. 
\end{align*} \]

Optimal TSS for Mean Estimation in Multilevel Populations when Cluster Size is Informative 5
Summary of the ten cluster size distributions used in the simulation study. Note that the Discrete Uniform distribution is shown in the panel of the three Four-parameter Beta distributions, since it is a special case of the Four-parameter Beta distribution for \( a = b = 1 \) and the rounding off due to the discrete nature of cluster size.

**Fig. S.1.** Summary of the ten cluster size distributions used in the simulation study. Note that the Discrete Uniform distribution is shown in the panel of the three Four-parameter Beta distributions, since it is a special case of the Four-parameter Beta distribution for \( a = b = 1 \) and the rounding off due to the discrete nature of cluster size.
cluster effect $v_j$. This simulation study was performed using version 3.4.3 of R [7], and the R code is given in section 1 of the Supplementary Material 2.

The accuracy of $E(\hat{\mu}), E(V(\hat{\mu}|N)), V(\hat{\mu}|N)$, and $V(\hat{\mu}) = E(V(\hat{\mu}|N)) + V(E(\hat{\mu}|N))$ as given in Table S.1, was evaluated by computing the following Relative Biases (R.B.):

- Marginal Expectation (i.e. unbiasedness of $\hat{\mu}_{TSS3}$ and $\hat{\mu}_{TSS2}$): $R.B. (E(\hat{\mu}))_1 = \frac{E(\hat{\mu})_{D.M} - \mu}{\mu}$, where $E(\hat{\mu})_{D.M}$ is the equation based on the Delta method (Table S.1, fourth row) and $\mu = \beta_0 + \alpha_1 \theta_N$ is computed from the input values, and $R.B. (E(\hat{\mu}))_2 = \frac{E(\hat{\mu})_{G.V} - E(\hat{\mu})_{G.V}}{E(\hat{\mu})_{G.V}}$, where $E(\hat{\mu})_{G.V.} = \frac{\sum_{i=1}^{2000} E(\hat{\mu})_i}{2000}$ is the value of $E(\hat{\mu})$ generated from the algorithm described previously.

- Expectation of the conditional variance: $R.B. (E(V(\hat{\mu}|N))) = \frac{E(V(\hat{\mu}|N))_{D.M} - E(V(\hat{\mu}|N))_{G.V.}}{E(V(\hat{\mu}|N))_{G.V.}}$, where $E(V(\hat{\mu}|N))_{D.M.}$ is given in the fifth row of Table S.1 and $E(V(\hat{\mu}|N))_{G.V.} = \frac{\sum_{i=1}^{2000} V(\hat{\mu}|N)_i}{2000}$, where $V(\hat{\mu}|N)_{G.V.}$ is obtained from the algorithm described previously.

- Variance of the conditional expectation: $R.B. (V(E(\hat{\mu}|N))) = \frac{V(E(\hat{\mu}|N))_{D.M} - V(E(\hat{\mu}|N))_{G.V.}}{V(E(\hat{\mu}|N))_{G.V.}}$, where $V(E(\hat{\mu}|N))_{D.M.}$ is given in the last row of Table S.1 and $V(E(\hat{\mu}|N))_{G.V.} = \frac{\sum_{i=1}^{2000} [V(\hat{\mu}|N)_i - (E(\hat{\mu}|N))_{G.V.}]^2}{4999}$.

- Marginal Variance: $R.B. (V(\hat{\mu})) = \frac{V(\hat{\mu})_{D.M} - V(\hat{\mu})_{G.V.}}{V(\hat{\mu})_{G.V.}}$, where $V(\hat{\mu})_{D.M.} = E(V(\hat{\mu}|N))_{D.M.} + V(E(\hat{\mu}|N))_{D.M.}$ (i.e. the sum of the last two rows in Table S.1), and $V(\hat{\mu})_{G.V.} = E(V(\hat{\mu}|N))_{G.V.} + V(E(\hat{\mu}|N))_{G.V.}$ (i.e. the sum of $E(V(\hat{\mu}|N))_{G.V.}$ with $V(E(\hat{\mu}|N))_{G.V.}$ as obtained from the algorithm described previously).

### 1.3 Simulation Results

The results of the simulation study are summarised in Tables S.3-S.6. Specifically, Table S.3 shows the relative bias of $E(V(\hat{\mu}|N))$ both for TSS3 and TSS2, while the relative biases of $V(E(\hat{\mu}|N))$ and $V(\hat{\mu})$ are given in Table S.4 for TSS3, and in Table S.5 for TSS2. For each cluster size distribution, the relative biases are shown in Tables S.3-S.5 for each combination (columns) of $\rho = \{0.05, 0.30\}$ and $\rho_{uN} = \{0.25, 0.50\}$ ($\rho_{uN} = 0.75$ is not tabulated here because, probably, less realistic, but it will be commented upon anyway), and for the smallest and largest considered $k$ (rows). With the same structure as Tables S.3-S.5, Table S.6 shows, for some cluster size distributions, the relative bias for $V(\hat{\mu})$ both for TSS2 and TSS3 for a given $k$ and different values of $n$. The complete results of the simulation study can be provided by the authors upon request.

The relative biases for $E(\hat{\mu})$ (i.e. $R.B. (E(\hat{\mu}))_1$ and $R.B. (E(\hat{\mu}))_2$ are not tabulated, as they were smaller than 0.5% in all the considered scenarios, with a maximum (in absolute value) of 0.3%, which occurred under the Negative Binomial distribution with $\tau_N = 0.996$, when $\rho_{uN} = 0.75$ and $k = 20$ (for both TSS2 and TSS3, and any value of $\rho$ and $n$). This means that the mean estimators for TSS2 and TSS3 (see Table S.1, first row) are already quite unbiased for $k = 20$ and need no further consideration.

The relative bias for $E(V(\hat{\mu}|N))$ is shown in Table S.3 both for TSS2 and TSS3 (for two distributions the results for $k = 100$ are shown, because for $k = 40$ $R.B. (V(\hat{\mu}))$ was too large). The relative bias for $E(V(\hat{\mu}|N))$ is always smaller than 3% in absolute value, with a maximum of 2.6% that occurred under the distribution of Italian high school size for $k = 20$, $\rho_{uN} = 0.75$, and $\rho = 0.05$. Furthermore, this relative bias tends to decrease as $k$ increases and to have a negative sign. In Table S.3, there seem to be no major differences across different values of $\rho$ for a given $\rho_{uN}$, or across different values of $\rho_{uN}$ for a given $\rho$, or across sampling schemes (TSS2, TSS3). Finally, there are no clear patterns related to the cluster size distribution.

The relative bias for $V(E(\hat{\mu}|N))$ is shown in Table S.4 for TSS3, and in Table S.5 for TSS2. The bias for $V(E(\hat{\mu}|N))$ is larger than that for $E(V(\hat{\mu}|N))$, and tends to be negative in sign, except for the two Negative Binomial distributions and the distribution of general practice list size in England. The relative bias decreases as $k$ increases for most distributions, with the exception of the Negative Binomial with $\tau_N = 0.5$, the Four-parameter Beta with $\zeta_N = 0$, and the Bimodal distribution with $\zeta_N = 0.889$. For $\rho_{uN} = 0.25$, the relative bias decreases as $\rho$ goes from 0.05 to 0.3, with the exception of the two Negative Binomial distributions, and the distribution of general practice list size in England. In contrast, for $\rho_{uN} = 0.5$, the pattern relative to $\rho$ is not clear. For a given $\rho$, the relative bias decreases as $\rho_{uN}$ goes from 0.25 to 0.5, except for the Two Negative Binomial distributions, and the distribution of general practice list size in England. From $\rho_{uN} = 0.5$ to $\rho_{uN} = 0.75$, the relative bias decreases or remains the same for most distributions, except for the Four-parameter Beta with $\tau_N = 0.266$ and the Bimodal distribution with $\zeta_N = -0.496$. Finally, the patterns of this relative bias under TSS3 are the same as those under TSS2.

The relative bias for $V(\hat{\mu})$ is shown in Table S.4 for TSS3, and in Table S.5 for TSS2. This relative bias decreases as $k$ increases for most distributions, with the exception of the Negative Binomial distribution with $\tau_N = 0.5$ and
the Bimodal distribution with \( \tau_N \approx 0.654 \). The sign of the bias is negative for all distributions but the two Negative Binomial distributions and the distribution of general practice size in England. For \( k = 20 \) and \( \rho \leq 0.3 \), under most distributions the relative bias is smaller (in absolute value) than 2.5% for \( \rho_{uN} \leq 0.5 \), and smaller than 5% for \( \rho_{uN} = 0.75 \). The two exceptions are the Negative Binomial distribution with \( \tau_N = 0.996 \) and the English general practice size distribution, for which \( k = 100 \) to have a relative bias smaller than or equal to 6% for \( \rho_{uN} \leq 0.5 \), and 8.5% for \( \rho_{uN} = 0.75 \). Note that all distributions with a small relative bias (i.e. 5%) for \( V(\hat{\mu}) \) are characterised by values of skewness and kurtosis that are fairly close (say, \( \pm 1.5 \)) to those of the Normal distribution (i.e. \( \xi_N = 0 \) and \( \eta_N = 3 \)), while the Negative Binomial with \( \tau_N = 0.996 \), and the distribution of general practice size in England have both large skewness (\( \xi_N \approx 2 \) and especially large kurtosis (excess kurtosis \( \eta_N - 3 \geq 6 \)). Now, the Delta method (2, Theorem 5.5.28 p. 245), used to derive \( V(\hat{\mu}) \) under TSS2 and TSS3, implies that \( \hat{\mu} = \sum_{j=1}^k w_j \hat{y}_j \) is asymptotically Normally distributed with mean \( \mu \) and variance \( V(\hat{\mu}) = \sum_{j=1}^k w_j^2 V(\hat{y}_j) \). Thus, cluster size distributions with large skewness and/or excess kurtosis (\( \eta_N - 3 \)) can make the convergence to the Normal distribution more problematic, and then the approximation less accurate, unless a large \( k \) (here, 100) is sampled.

The results in Tables S.3-S.5 discussed above were obtained with \( n = 20 \). To explore the effect of \( n \) on the accuracy of the approximation for \( V(\hat{\mu}) \), additional values of \( n \) were considered: \( n = \{10, 50, 100\} \). Note from Table S.1 that \( E(V(\hat{\mu} | N)) \) is a (decreasing) function of \( n \) unlike \( V(E(\hat{\mu} | N)) \). Thus, as \( n \to 0 \), \( V(E(\hat{\mu} | N)) \to \infty \) and then the relative bias of \( V(\hat{\mu}) \) moves towards that of \( V(E(\hat{\mu} | N)) \). In contrast, as \( n \to \infty \), \( V(E(\hat{\mu} | N)) \) decreases (thus approaching \( \frac{\sigma_0^2}{\tau} \times (E(CV_0^2) + 1) \), see Table S.1, third row), so that the relative bias of \( V(\hat{\mu}) \) moves towards the relative bias of \( V(E(\hat{\mu} | N)) \). Tables S.6 (first three rows) gives the results of the simulations for \( k = 20 \) and \( n = \{10,50\} \), under the two Negative Binomial distributions and the Bimodal distribution with \( \tau_N = 0.504 \), for both TSS2 and TSS3, \( \rho = \{0.05, 0.30\} \), and \( \rho_{uN} = \{0.25, 0.50\} \). In all scenarios, the relative bias of \( V(\hat{\mu}) \) remains below 2.5% across the different values of \( n \) \( = \{10,50\} \) for the Negative Binomial with \( \tau_N = 0.5 \) and the Bimodal distribution with \( \tau_N = 0.504 \), while it increases as \( n \) increases for the Negative Binomial with \( \tau_N = 0.996 \). To see whether \( k = 100 \) remains a safe sample size for the Negative Binomial with \( \tau_N = 0.996 \) and the English general practice size distribution, additional simulations for these two distributions were run for \( k = 100 \) and \( n = 100 \). The results are given in the last two rows of Table S.6, where it can be seen that the relative bias of \( V(\hat{\mu}) \) increases as \( n \) increases (mostly when \( \rho = 0.05 \)) and in any case it remains \( \leq 6\% \) for \( \rho_{uN} \leq 0.5 \). When \( \rho = 0.30 \) the relative bias is stable across different values of \( n \) and \( \leq 6\% \) (for \( \rho_{uN} \leq 0.5 \)), which can be explained by noting that if \( \rho = 0.30 \) then \( E(V(\hat{\mu} | N)) \) is already close to \( \frac{\sigma_0^2}{\tau} \times (E(CV_0^2) + 1) \) at \( n = 20 \).

To summarize, sampling \( k = 20 \) clusters guarantees fairly unbiased estimates of \( \mu \) under TSS2 and TSS3 independently of the cluster size distribution, and fair accuracy (i.e. relative bias \( \leq 5\% \)) of the variances under TSS2 and TSS3 when \( \rho_{uN} \leq 0.75 \), \( \rho \leq 0.3 \), and the skewness and kurtosis of cluster size distribution are relatively close (say, \( \pm 1.5 \)) to those of the Normal distribution. For cluster size distributions with more extreme skewness and kurtosis at least \( k = 100 \) clusters must be sampled in order to achieve a reasonable accuracy (i.e. relative bias \( \leq 6\% \)) of the sampling variances in Table S.1 (sum of the last two rows), for \( \rho_{uN} \leq 0.5 \) and \( \rho \leq 0.3 \). Finally, these two lower-bounds for \( k \) remain safe across different values of \( n \).

## 2 Optimal Design and Relative Efficiencies for a given budget

### 2.1 Sampling Variances

**Rewritings of the sampling variances from Table 1 in [1] to Table 1 in the main text.**

Recall that \( \rho = \frac{\sigma_0^2}{\sigma_1^2} \) and that \( \sigma_0^2 \sigma_2^2 = \psi \sigma_1^2 \). From Table 1 (Innocenti et al. [1]), we have that

\[
V_{TSS1}(\hat{\mu}) = \frac{n \sigma_0^2 + \sigma_1^2}{nk} + \sigma_0^2 \frac{\psi}{\sigma_2^2} \frac{\xi_N}{k} (\xi_N + \tau_N) + 1 \approx \frac{n \sigma_0^2 + \sigma_1^2}{nk} + \psi \sigma_1^2 \frac{\xi_N}{k} (\xi_N + \tau_N) + 1 = \frac{\sigma_1^2}{nk} \left[ 1 + \rho (n-1) + n \psi \sigma_1^2 \frac{\xi_N}{k} (\xi_N + \tau_N) + 1 \right] = \frac{\sigma_1^2}{nk} \left[ 1 + \rho \left[ (n-1) + n \psi (\xi_N + \tau_N) + 1 \right] \right] \quad 0 \leq (\xi_N + \tau_N) \leq \frac{1}{\rho} \left[ 1 + \rho \left( n-1 + n \psi (\xi_N + \tau_N) + 1 \right) \right].
\]
Table S.3. Relative Bias for R. B. (E(V (µ | N))) under TSS3 and TSS2, for n = 20.

| Cluster Size Distribution | τN | k | ρoN = 0.25 | ρoN = 0.5 | ρoN = 0.25 | ρoN = 0.5 | ρoN = 0.25 | ρoN = 0.5 |
|----------------------------|----|----|-------------|-------------|-------------|-------------|-------------|-------------|
| Negative Binomial          | 0.50 | 20 | 0.1%        | -0.1%       | 0.1%        | -0.1%       | 0.0%        | -0.1%       |
|                            | 0.996 | 20 | 0.2%        | -0.2%       | 0.2%        | -0.2%       | 0.0%        | -0.2%       |
| GP Patient list size in England | 0.633 | 20 | 1.6%        | 1.4%        | 1.5%        | 1.6%        | 0.0%        | 0.1%        |
| High School size in Italy  | 0.912 | 20 | -1.6%       | -1.7%       | -1.7%       | -1.6%       | -2.5%       | -1.8%       |
| Discrete Uniform on [2, 800] | 0.575 | 20 | -1.6%       | -1.6%       | -1.6%       | -1.6%       | -1.6%       | -1.6%       |
| Four-parameter Beta        | 0.54 | 20 | -1.4%       | -1.3%       | -1.2%       | -1.3%       | -1.3%       | -1.3%       |
|                            | 0.266 | 20 | -0.2%       | -0.2%       | -0.2%       | -0.2%       | -0.2%       | -0.2%       |
|                            | 0.445 | 20 | -0.7%       | -0.8%       | -0.8%       | -0.5%       | -0.9%       | -0.9%       |
| Bimodal                    | 0.654 | 20 | -0.1%       | -0.2%       | -0.3%       | -0.3%       | -0.8%       | -0.9%       |
|                            | 0.504 | 20 | -1.5%       | -1.5%       | -1.7%       | -1.5%       | -1.5%       | -1.4%       |

Table S.4. Relative Bias for V (E (µ | N)) and V (µ) under TSS3, for n = 20.

| Cluster Size Distribution | Parameters | k | ρoN = 0.25 | ρoN = 0.5 | ρoN = 0.25 | ρoN = 0.5 | ρoN = 0.25 | ρoN = 0.5 | ρoN = 0.5 |
|----------------------------|------------|----|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Negative Binomial          | θN = 402, τN = 0.5, ζN = 1, ηN = 4.5 | 20 | -1.2%       | 0.5%        | 1.2%        | 1.6%        | 0.1%        | 0%          | 0.3%        |
|                            | θN = 402, τN = 0.996, ζN = 2, ηN = 9 | 20 | 25.1%      | 26%         | 27.3%       | 27.4%       | 2.7%        | 4.3%        | 9.6%        |
| GP Patient list size in England | θN = 7986, τN = 0.633, ζN = 2.12, ηN = 14.549 | 20 | 43.9%      | 43.8%       | 44.8%       | 44.8%       | 6.2%        | 8.8%        | 17.8%       |
| High School size in Italy  | θN = 403, τN = 0.912, ζN = 1.256, ηN = 4.315 | 20 | -3%         | -2.1%       | -0.2%       | -0.4%       | -1.7%       | -1.8%       | -1.3%       |
| Discrete Uniform on [2, 800] | θN = 401, τN = 0.575, ζN = 0, ηN = 1.8 | 20 | -13.9%      | -9.7%       | -6.7%       | -5.4%       | -1.9%       | -1.9%       | -2.1%       |
| Four-parameter Beta        | θN = 420, τN = 0.54, ζN = 0.082, ηN = 1.832 | 20 | -10.6%      | -8.5%       | -5.4%       | -4.5%       | -1.6%       | -1.6%       | -1.6%       |
|                            | θN = 534, τN = 0.266, ζN = -0.468, ηN = 2.625 | 20 | -6.1%       | -4.6%       | -0.5%       | -0.2%       | -0.3%       | -0.4%       | -0.3%       |
| Bimodal                    | θN = 401, τN = 0.445, ζN = 0, ηN = 2.143 | 20 | -5.8%       | -3.8%       | -1.2%       | 0.1%        | -0.9%       | -1%         | -0.9%       |
|                            | θN = 701, τN = 0.504, ζN = -0.496, ηN = 1.677 | 20 | -14.8%      | -7.4%       | -1.1%       | 0.5%        | -1.7%       | -1.7%       | -1.6%       |
Table S.5. Relative Bias for $V(\hat{\mu} | N)$ and $V(\hat{\mu})$ under TSS2, for $n = 20$.

| Cluster Size Distribution | Parameters | $R.B. (V(\hat{\mu} | N))$ | $R.B. (V(\hat{\mu}))$ |
|---------------------------|------------|--------------------------|--------------------------|
|                           | $k$        | $p_{\hat{\mu}N} = 0.25$ | $p_{\hat{\mu}N} = 0.5$ |
|                           | $p = 0.05$ | $p = 0.3$                | $p = 0.05$                |
| Negative Binomial         | $\theta_N = 402$, $\tau_N = 0.5$, $\zeta_N = 1$, $\eta_N = 4.5$ | 20 | -1.4% | 0.5%  | 1.3%  | 1.6%  | -0.5%  | 0.2%  | 0.5%  |
|                           |            | 40 | -2.9% | 4.2%  | 4.2%  | 0.1%  | 0.2%  | 0.9%  | 1.4%  |
|                           | $\theta_N = 402$, $\tau_N = 0.996$, $\zeta_N = 2$, $\eta_N = 9$ | 100 | 25.5% | 25.9% | 27.2% | 27.3% | 2%    | 4.3%  | 10.3% | 13.5% |
| GP Patient list size in England | $\theta_N = 7986$, $\tau_N = 0.633$, $\zeta_N = 2.12$, $\eta_N = 14.549$ | 20 | 44%  | 43.6% | 44.9% | 44.8% | 5.3%  | 8.7%  | 18.2% | 23.9% |
| High School size in Italy | $\theta_N = 403$, $\tau_N = 0.912$, $\zeta_N = 1.256$, $\eta_N = 4.315$ | 20 | -2.3% | -1.8% | -0.3% | -0.5% | -2.5% | -1.8% | -1.9% | -1.3% |
| Discrete Uniform on $[2,800]$ | $\theta_N = 401$, $\tau_N = 0.575$, $\zeta_N = 0$, $\eta_N = 1.8$ | 20 | -11.9% | -9.7% | -5.9% | -5.2% | -1.8% | -1.9% | -2%   | -2.2% |
| Four-parameter Beta       | $\theta_N = -0.982$, $\tau_N = 1.832$ | 20 | -10.1% | -8.5% | -5.4% | -4.4% | -1.7% | -1.6% | -1.7% | -1.8% |
|                           | $\theta_N = 534$, $\tau_N = 0.266$, $\zeta_N = -0.468$, $\eta_N = 2.625$ | 20 | -6.4% | -4.3% | -0.6% | -0.4% | -0.4% | -0.4% | -0.4% | -0.2% |
|                           | $\theta_N = 401$, $\tau_N = 0.445$, $\zeta_N = 0$, $\eta_N = 2.143$ | 20 | -5.8% | -4%   | -0.9% | 0.2%  | -1%   | -1%   | -0.9% | -0.5% |
| Bimodal                   | $\theta_N = 401$, $\tau_N = 0.654$, $\zeta_N = 0.889$, $\eta_N = 3.578$ | 20 | -5%   | -2.2% | -2%   | -1.2% | -1%   | -1%   | -0.5% | -1.2% |
|                           | $\theta_N = 701$, $\tau_N = 0.504$, $\zeta_N = -0.496$, $\eta_N = 1.677$ | 20 | -12.3% | -7.7% | -0.8% | 0.9%  | -1.7% | -1.7% | -1.5% | -1.2% |

Table S.6. Relative Bias for $V(\hat{\mu})$ under TSS2 and TSS3, for different values of $n$.

| Cluster Size Distribution | Parameters | $R.B. (V(\hat{\mu}))$ under TSS3 | $R.B. (V(\hat{\mu}))$ under TSS2 |
|---------------------------|------------|------------------------|------------------------|
|                           | $k$        | $n$                           | $p_{\hat{\mu}N} = 0.25$ | $p_{\hat{\mu}N} = 0.5$ |
|                           | $p = 0.05$ | $p = 0.3$                | $p = 0.05$                |
| Negative Binomial         | $\theta_N = 402$, $\tau_N = 0.5$, $\zeta_N = 1$, $\eta_N = 4.5$ | 20 | 10 | -0.1% | 0.1% | 0.1% | 0.4% | -0.8% | -0.1% | -0.2% | 0.3% |
|                           |            | 50 | 0.1% | 0.0% | 0.4% | 0.6% | -0.1% | 0.0% | 0.3% | 0.6% |
|                           | $\theta_N = 402$, $\tau_N = 0.996$, $\zeta_N = 2$, $\eta_N = 9$ | 20 | 10 | 1.9% | 3.8% | 7.3% | 12.7% | 0.6% | 3.6% | 7.8% | 13% |
| Bimodal                   | $\theta_N = 701$, $\tau_N = 0.504$, $\zeta_N = -0.496$, $\eta_N = 1.677$ | 20 | 10 | -1.7% | -1.9% | -1.5% | -1.4% | -1.6% | -1.8% | -1.5% | -1.4% |
|                           |            | 50 | -1.6% | -1.6% | -1.4% | -1.3% | -1.7% | -1.6% | -1.4% | -1.3% |
| Negative Binomial         | $\theta_N = 402$, $\tau_N = 0.996$, $\zeta_N = 2$, $\eta_N = 9$ | 100 | 20 | 13% | 1.5% | 3.5% | 4.9% | 1.2% | 1.6% | 4% | 5% |
|                           |            | 100 | 20 | 1.3% | 1% | 4.8% | 5% | 1.4% | 1.7% | 4.9% | 5.1% |
| GP Patient list size in England | $\theta_N = 7986$, $\tau_N = 0.633$, $\zeta_N = 2.12$, $\eta_N = 14.549$ | 20 | 20 | 1.6% | 2.4% | 4.5% | 5.9% | 1.2% | 2.4% | 4.5% | 6% |
|                           |            | 100 | 22.2% | 2.4% | 5.8% | 6% | 2.3% | 2.4% | 5.8% | 6% |

\[ V_{TSS2}(\hat{\mu}) = \frac{p_{\theta_N} \sigma_\theta^2 \left( \frac{t_N + 1 \tau_N}{t_N + 1} \right) + \sigma^2}{p_{\theta_N} k} + \frac{\alpha_\theta^2 \sigma_\theta^2 \left( \frac{t_N + 1 \tau_N}{t_N + 1} \right)^2 + \sigma^2}{k} \]

\[ \cdot \frac{1 + \rho \left[ n \left( \frac{t_N + 1 \tau_N}{t_N + 1} \right) - 1 \right] + n \psi \rho \left( \frac{k + 1}{k} \right)^2 \eta_N - \frac{k + 1}{k} + \tau_N (\tau_N - 2 \xi_N) + 2 \left( \frac{k + 1}{k} \right) \tau_N (\tau_N - 1) + 1}{p_{\theta_N} k} \]
\[
\begin{align*}
\text{Optimal TSS for Mean Estimation in Multilevel Populations when Cluster Size is Informative 11}
\end{align*}
\]
The optimal variance is obtained by plugging into $\text{Var}(\hat{\mu})$ the optimal sample sizes, as follows

$$Var(\hat{\mu})^* = \frac{n^*V(u) + V(\varepsilon)}{n^*k^*} = \frac{\left(\sqrt{c_2V(\varepsilon)} c_1 + V(\varepsilon)\right)
\left(c_2 + \sqrt{c_2V(\varepsilon)}\right)}{c_2 + \sqrt{c_2V(\varepsilon)}} = \frac{\left(\sqrt{c_2V(u) + c_1V(\varepsilon)}\right)^2}{C}.
$$

Following these steps, the sampling variance, the optimal design, and the optimal variance are shown for each sampling scheme.

**TSS1:** The sampling variance of the unbiased estimator under TSS1 is (Innocenti et al. [1])

$$V_{TSS1}(\hat{\mu}) = n\left(\sigma^2 + \frac{\alpha_1^2\sigma_0^2}{\tau_N(\zeta_N - \tau_N)}\right) + \sigma^2\frac{n}{nk},$$

where $V_{TSS1}(u) = \sigma^2 + \frac{\alpha_1^2\sigma_0^2}{\tau_N(\zeta_N - \tau_N)}$. Thus, the optimal design is given by

$$n^* = \left(\frac{c_1}{c_2}\right) \left(\frac{\sigma^2 + \frac{\alpha_1^2\sigma_0^2}{\tau_N(\zeta_N - \tau_N)} + \sigma^2\frac{n}{nk}}{\left(c_2 + \sqrt{c_2V(u)}\right)^2}\right)\text{,}$$

and $k^* = \frac{C}{c_1(c_2 + n^*)}$. Hence, $\alpha_1^2\sigma_0^2 = \left(\frac{c_1}{c_2}\right) \left(\frac{\rho^2}{1 - \rho^2}\right)$ and $\alpha_2^2 = \psi\alpha_1^2$ (see [1]). The optimal variance follows

$$V_{TSS1}(\hat{\mu})^* = \frac{\left(\sqrt{c_1\left(\frac{1}{\tau_N}\right)} + \sqrt{c_1\sigma^2\left(\frac{1}{\tau_N}\right)}\right)^2}{C} = \frac{c_1\sigma^2}{C} \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right)^2.$$

To examine the relation between $\rho$ and the budget $C$, in the main text we defined $g(\rho, \psi) = c_1 \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right)^2$ (see section 3.2, main text). We have that

$$\frac{\partial g(\rho, \psi)}{\partial \rho} = 2c_1 \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right) \frac{\left(\frac{c_1}{c_2}\right) \left(\frac{\rho^{2}}{1 - \rho^{2}}\right) - 1}{\sqrt{2(1 - \rho)}}.$$

Hence, $g(\rho, \psi)$, and thus $C$ (see section 3.2, main text), is an increasing function of $\rho$ if $\frac{\partial g(\rho, \psi)}{\partial \rho} > 0$. Since $\eta_2$ is sufficiently large such that $\frac{c_1}{c_2 + n^*} > 0.5$, and $r_\rho > 1$. The latter is the case if $\zeta_N \geq \tau_N - \frac{1}{\sqrt{2}}$, and $\rho < \frac{c_1(1 + \psi(\zeta_N - \tau_N))}{c_1(1 + \psi(\zeta_N - \tau_N)) + 1}$. Note that

$$V_{TSS2}(\hat{\mu}) = \left(\frac{c_1}{c_2 + n^*}\right)^2 \left(\frac{\rho^2}{1 - \rho^2}\right) + \alpha_1^2\sigma_0^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right) + \alpha_2^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right) + \frac{1}{\tau_N^2}\left(\frac{\rho^2}{1 - \rho^2}\right).$$

Assume that $k$ is sufficiently large such that $\frac{c_1}{c_2 + n^*} \approx 0$, $\frac{k}{k - 1} \approx 1$, and $\frac{k - 1}{k - 1} \approx 1$, then

$$V_{TSS2}(\hat{\mu}) = n\left(\frac{c_1}{c_2 + n^*}\right)^2 \left(\frac{\rho^2}{1 - \rho^2}\right) + \alpha_1^2\sigma_0^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right) + \alpha_2^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right) + \frac{1}{\tau_N^2}\left(\frac{\rho^2}{1 - \rho^2}\right).$$

The optimal design is given by

$$n^* = \frac{\left(\frac{c_1}{c_2}\right) \left(\frac{\rho^2}{1 - \rho^2}\right) + \alpha_1^2\sigma_0^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right) + \alpha_2^2\left(\frac{1}{\tau_N^2(\zeta_N - \tau_N)} + \frac{1}{\tau_N^2}\right)}{\left(c_2 + \sqrt{c_2V(u)}\right)^2}.$$

The optimal variance follows

$$V_{TSS2}(\hat{\mu})^* = \frac{\left(\sqrt{c_1\left(\frac{1}{\tau_N}\right)} + \sqrt{c_1\sigma^2\left(\frac{1}{\tau_N}\right)}\right)^2}{C} = \frac{c_1\sigma^2}{C} \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right)^2.$$

To examine the relation between $\rho$ and the budget $C$, in the main text we defined $g(\rho, \psi) = c_1 \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right)^2$ (see section 3.2, main text). We have that

$$\frac{\partial g(\rho, \psi)}{\partial \rho} = 2c_1 \left(\sqrt{c_1\rho(1 + \psi(\zeta_N - \tau_N))} + \sqrt{1 - \rho}\right) \frac{\left(\frac{c_1}{c_2}\right) \left(\frac{\rho^{2}}{1 - \rho^{2}}\right) - 1}{\sqrt{2(1 - \rho)}}.$$

Hence, $g(\rho, \psi)$, and thus $C$ (see section 3.2, main text), is an increasing function of $\rho$ if $\frac{\partial g(\rho, \psi)}{\partial \rho} > 0$. Since
Relative Efficiencies

The relative efficiency of the optimal designs of two sampling schemes for a given budget is defined as the ratio of their optimal variances \( V(\hat{\mu})^* \) given in Table 2 (main text). Figures S.8, S.9, and S.10 show, for a given research budget, the relative efficiency of the optimal TSS1, the optimal TSS2, and the optimal TSS3, respectively, versus SRS, as a function of \( \rho \), for different values of \( c_r \) and \( \psi \), and different cluster size distributions. Figures S.11 and S.12 show the relative efficiency, for a given research budget, of the optimal TSS2 and the optimal TSS3, respectively, versus the optimal TSS1, as a function of \( \rho \), for different values of \( c_r \) and \( \psi \), and different cluster size distributions. For each relative efficiency in Table 3 (main text), the ratio depending on the features of cluster size distribution is studied.

2.3 Relative Efficiencies
Fig. S.2. Optimal number of individuals per cluster \( n^* \) under TSS1, as a function of \( \rho \), for different values of \( c_r \) and \( \psi \) (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that \( \psi = 0.35 \) corresponds to \( \rho uN = \pm 0.51 \).

### Table S.7. Real Cluster Size Distributions.

| Cluster Size distribution | \( \theta_N \) | \( \tau_N \) | \( \zeta_N \) | \( \eta_N \) |
|---------------------------|----------------|-------------|-------------|-------------|
| General Practice List size distribution in England, 2017. [5] | 7.986 | 0.633 | 2.12 | 14.549 |
| Public High School size distribution in Italy, 2016/2017. [6] | 403 | 0.912 | 1.256 | 4.315 |
| Public High School size distribution in France, 2014/2015. [9] | 764 | 0.621 | 0.886 | 3.582 |
| Public Lower Secondary School size distribution in Italy, 2016/2017. [6] | 225 | 0.789 | 1.351 | 5.303 |
| Public Lower Secondary School size distribution in France, 2014/2015. [9] | 493 | 0.387 | 0.63 | 5.47 |
| Public Primary School size distribution in Italy, 2016/2017. [6] | 171 | 0.761 | 1.451 | 5.740 |
| Public Primary School size distribution in France, 2014/2015. [9] | 135 | 0.71 | 1.045 | 4.084 |
Fig. S.3. Optimal number of individuals per cluster $n^* = \rho^* \theta_N$ under TSS2, as a function of $\rho$, for different values of $c_r$ and $\psi$ (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that $\psi = 0.35$ corresponds to $\rho_{inN} = \pm 0.51$. 
**Optimal Number of Individuals per Cluster for TSS3**

**Fig. S.4.** Optimal number of individuals per cluster $n^*$ under TSS3, as a function of $\rho$, for different values of $c_r$ and $\psi$ (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that $\psi = 0.35$ corresponds to $\rho_uN = \pm 0.51$.

**Table S.8.** Robustness of the optimal design against misspecification of $\psi$, assuming the general practice list size distribution in England, $\rho = 0.05$, $c_r = 10$, and $C/c_1 = 1000$.

|       | TSS1 | TSS2 | TSS3 |
|-------|------|------|------|
|       | $\psi = 0$ | $\psi = 1/3$ | $\psi = 0$ | $\psi = 1/3$ | $\psi = 0$ | $\psi = 1/3$ |
| $n^*$ | 13.78 | 10.74 | 11.65 | 7.01 | 13.78 | 8.30 |
| $k^*$ | 42.04 | 48.22 | 46.2 | 58.79 | 42.04 | 54.65 |
| $Var(\hat{\mu})/\sigma^2$ if $\psi = 1/3$ | 0.00360 | 0.00354 | 0.00595 | 0.00559 | 0.00690 | 0.00647 |
| $Var(\hat{\mu}|\psi = 1/3)$ | 0.983 | | | | | |
| $Var(\hat{\mu}|\psi = 0)$ | | | | | | |
| $\psi = 0$ | $\psi = 1$ | $\psi = 0$ | $\psi = 1$ | $\psi = 0$ | $\psi = 1$ |
| $n^*$ | 13.78 | 8.04 | 11.65 | 4.65 | 13.78 | 5.50 |
| $k^*$ | 42.04 | 55.44 | 46.2 | 68.27 | 42.04 | 64.52 |
| $Var(\hat{\mu})/\sigma^2$ if $\psi = 1$ | 0.00514 | 0.00478 | 0.01129 | 0.00944 | 0.01276 | 0.01057 |
| $Var(\hat{\mu}|\psi = 1)$ | 0.930 | | | | | |
| $Var(\hat{\mu}|\psi = 0)$ | | | | | | |

Note: $\beta = 0.445$, $c_r = 0$, $\eta_N = 2.143$ for TSS1; $\beta = 0.54$, $c_r = -0.082$, $\eta_N = 1.832$ for TSS2; $\beta = 0.504$, $c_r = 0.496$, $\eta_N = 1.677$ for bimodal.
Fig. S.5. Budget $C$ needed for the optimal TSS1 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 $\psi$, for different values of $\rho$ and $c_r$ (curves) with $c_1 = 10$, and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1.

Note that $\psi \in [0, 1.3]$ corresponds to $\rho_{\eta N} \in [-0.75, +0.75]$.

To see whether
\[
\frac{1 + \rho \psi (\zeta_N - \tau_N) + 1}{\sqrt{c_r \rho (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] + \sqrt{1 - \rho})}} \leq [1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1]]
\]
in $RE$ (TSS1 vs SRS) is smaller than one, we need to show that the difference between denominator and numerator is bigger than zero:
\[
\left( \sqrt{c_r \rho (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] + \sqrt{1 - \rho})} \right)^2 - [1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1]] =
\]
\[
c_r \rho (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] + (1 - \rho) + 2 \sqrt{c_r \rho (1 - \rho) (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1]) - 1 - \rho [\tau_N (\zeta_N - \tau_N) + 1] =
\]
\[
= \rho (c_r - 1) \{ 1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] \} + 2 \sqrt{c_r \rho (1 - \rho) (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1])\}
\]
Note that $c_r \rho (1 - \rho) > 0$, since that $\rho \in (0, 1)$ and $c_r = \frac{c_r}{c_1} > 1$, and $\sqrt{c_r \rho (1 - \rho) (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1])} \geq 0$ holds when $(1 + \psi [\tau_N (\zeta_N - \tau_N) + 1]) \geq 0$ that, in turn, holds for $\zeta_N \geq \frac{-1 - (1 - \frac{1}{\tau_N})}{\frac{\psi}{\tau_N}} = \tau_N - \frac{1}{\psi} - \frac{1}{\tau_N \psi}$. Therefore, for $\zeta_N \geq \tau_N - \frac{1}{\psi} - \frac{1}{\tau_N \psi}$ we have that $(1 + \psi [\tau_N (\zeta_N - \tau_N) + 1]) \geq 0$, and then
\[
\left( \sqrt{c_r \rho (1 + \psi [\tau_N (\zeta_N - \tau_N) + 1] + \sqrt{1 - \rho})} \right)^2 -
Budget C for the Optimal TSS2

Fig. S.6. Budget C needed for the optimal TSS2 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 $\psi$, for different values of $\rho$ and $c_r$ (curves) with $c_1 = 10$, and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that $\psi \in [0, 1.3]$ corresponds to $\rho_N \in [-0.75, +0.75]$. 
Fig. S.7. Budget $C$ needed for the optimal TSS3 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 $\psi$, for different values of $\rho$ and $c_r$ (curves) with $c_1 = 10$, and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that $\psi \in [0, 1]$ corresponds to $\rho_{\text{N}} \in [-0.75, +0.75]$.

$[1 + \rho \psi [\zeta_N - \tau_N]] \geq 0$, which implies that $\frac{1 + \rho \psi [\zeta_N - \tau_N + 1]}{\sqrt{c_r \rho [1 + \psi (\zeta_N - \tau_N) + 1]} + \sqrt{1 - \rho}} \leq 1$, and then SRS is more efficient than TSS1 provided that $\left(\frac{c_{\text{SRS}}}{c_1}\right) = \left(\frac{C}{C - c_0}\right) = 1$. Note that $\zeta_N \geq \tau_N - \frac{1}{\tau_N} - \frac{1}{\psi \tau_N}$ and $\left(\frac{c_{\text{SRS}}}{c_1}\right) = \left(\frac{C}{C - c_0}\right) = 1$ is a sufficient condition for $RE$ (TSS1 vs SRS) $\leq 1$, but it is not a necessary condition. Indeed, $RE < 1$ for the Italian high school sizes distribution (which satisfies $\zeta_N \geq \tau_N - \frac{1}{\tau_N} - \frac{1}{\psi \tau_N}$) when $\left(\frac{c_{\text{SRS}}}{c_1}\right) = 4$ and $\left(\frac{C}{C - c_0}\right) = 1.25$

Assuming $\rho = 0.2$ and $c_r = 50$ (for $\psi = 0$ then $RE \approx 0.3$, for $\psi = 0.35$ then $RE \approx 0.25$).

To see whether $\left(\frac{c_{\text{SRS}}}{c_1} + 1 + \psi (\tau_N + \tau_N (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N)) + \sqrt{1 - \rho}\right)^2$ in $RE$ (TSS2 vs SRS) is smaller than one, we need to show that the difference between numerator and denominator is smaller than zero:

$$1 + \rho \psi [\zeta_N - \tau_N + 1] - \left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^3 + \tau_N \eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N) + 1]} + \sqrt{1 - \rho}\right)^2 =$$
\[
= 1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1] - \{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] + (1 - \rho) + \\
+ 2 \sqrt{c_r \rho (1 - \rho)} [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] \} = \\
= \rho \{1 - c_r (\tau_N^2 + 1) + \psi [\tau_N (\zeta_N - \tau_N) + 1 - c_r (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] \} + \\
- 2 \sqrt{c_r \rho (1 - \rho)} [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)].
\]

Note that \(1 - c_r (\tau_N^2 + 1) \leq 0 \) since \(c_r = \frac{c}{c_1} \geq 1 \) and \(\tau_N^2 \geq 0\), and \(\sqrt{c_r \rho (1 - \rho)} [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] \geq 0\) that holds \(\forall \zeta_N \in \mathbb{R}\) as shown as follows.

Note that \(c_r \rho (1 - \rho) \geq 0\). Since \(\eta \geq \zeta_N + 1 [8]\), we have that \(\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) \geq \tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\zeta_N - 2) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) = \psi \tau_N^2 \zeta_N^2 + 2 \zeta_N \tau_N (1 - \tau_N^2) \zeta_N + \{\tau_N^2 + 1 + (1 - \tau_N^2)^2 \psi \} \geq 0 \) \(\forall \zeta_N \in \mathbb{R}\), since solving that inequality with respect to \(\zeta_N\) we have that \(\Delta = b^2 - 4ac = (2 \psi \tau_N (1 - \tau_N^2))^2 - 4 \psi \tau_N^2 \{\tau_N^2 + 1 + (1 - \tau_N^2)^2 \psi \} \leq 0\). Therefore, in order to see whether

\[
1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1] - \left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] + \sqrt{1 - \rho}}\right)^2 \leq 0,
\]

we just need to find when \(\tau_N (\zeta_N - \tau_N) + 1 - c_r (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) \leq 0\). Recall that \(\eta \geq \zeta_N + 1 [8]\), so

\[
\tau_N (\zeta_N - \tau_N) + 1 - c_r (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) \leq 0,
\]

\[
\tau_N (\zeta_N - \tau_N) + 1 - c_r (\tau_N^2 + \tau_N^2 (\zeta_N - 2) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) = \\
- c_r \zeta_N \tau_N^2 + (\tau_N - 2 \zeta_N (1 - \tau_N^2) c_r) \zeta_N + (1 - \tau_N)^2 (1 - c_r (1 - \tau_N^2)) \leq 0,
\]

and solving with respect to \(\zeta_N\) we have that \(\Delta = b^2 - 4ac = (\tau_N - 2 \zeta_N (1 - \tau_N^2) c_r)^2 + 4 \tau_N (1 - \tau_N^2) (1 - c_r (1 - \tau_N^2)) = \tau_N^2 + 4 \tau_N^2 (1 - \tau_N^2)^2 c_r^2 - 4 \tau_N (1 - \tau_N^2) + 4 \tau_N (1 - \tau_N^2)^2 c_r^2 = \tau_N^2 + 4 \tau_N^2 (1 - \tau_N^2)^2 c_r^2 - 4 \tau_N (1 - \tau_N^2) + 4 \tau_N (1 - \tau_N^2)^2 c_r^2 = \tau_N^2 + 4 \tau_N^2 (1 - \tau_N^2)^2 c_r^2 - 4 \tau_N (1 - \tau_N^2) + 4 \tau_N (1 - \tau_N^2)^2 c_r^2 = \tau_N^2 + 4 \tau_N^2 (1 - \tau_N^2)^2 c_r^2 - 4 \tau_N (1 - \tau_N^2) + 4 \tau_N (1 - \tau_N^2)^2 c_r^2 = \tau_N^2 \)

so that for \(\zeta_N \leq \tau_N - \frac{1}{\sqrt{\tau_N}} \sqrt{\frac{1 - c_r (1 - \tau_N^2)}{c_r}} \zeta_N^2 \leq \tau_N + \frac{1}{\sqrt{\tau_N}} - \frac{1}{\sqrt{\tau_N}}\) we have that \(0 \geq - c_r \zeta_N \tau_N^2 + (\tau_N - 2 \zeta_N (1 - \tau_N^2) c_r) \zeta_N + (1 - c_r (1 - \tau_N^2)) \geq \tau_N (\zeta_N - \tau_N) + 1 - c_r (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1) \geq 0\). Hence, for a cluster size distribution such that \(\zeta_N \leq \tau_N - \frac{1}{\sqrt{\tau_N}} \sqrt{\frac{1 - c_r (1 - \tau_N^2)}{c_r}} \leq \tau_N + \frac{1}{\sqrt{\tau_N}} - \frac{1}{\sqrt{\tau_N}}\), we have that

\[
\frac{1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1]}{\left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] + \sqrt{1 - \rho}}\right)^2} \leq 1,
\]

that is, SRS is more efficient than TSS2 if \(\left(\frac{c_{\text{ave}}}{c_1}\right) = \left(\frac{c}{c_{\text{ave}} - c_0}\right) = 1\). Note that this result also holds for a Normal distribution (i.e. \(\zeta_N = 0\) and \(\eta = 3\)), as shown below:

\[
\frac{1 + \rho \psi [\tau_N (\zeta_N - \tau_N) + 1]}{\left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^2 + \tau_N^2 (\eta - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1)] + \sqrt{1 - \rho}}\right)^2} = \\
\frac{1 + \rho \psi [1 - \tau_N^2]}{\left(\sqrt{c_r \rho [\tau_N^2 + 1 + \psi (\tau_N^2 + 1)] + \sqrt{1 - \rho}}\right)^2} \leq 1.
\]

Note that \(\zeta_N \leq \tau_N - \frac{1}{\sqrt{\tau_N}} \) or \(\zeta_N \geq \tau_N + \frac{1}{\sqrt{\tau_N}} - \frac{1}{\sqrt{\tau_N}}\) or \(N_j \sim N (\theta_N, \sigma^2)\), and \(\left(\frac{c_{\text{ave}}}{c_1}\right) = \left(\frac{c}{c_{\text{ave}} - c_0}\right) = 1\) is a sufficient condition for \(RE\) (TSS2 vs SRS) \(\leq 1\), but it is not a necessary condition. Indeed, \(RE < 1\) for the Italian high school sizes distribution (which satisfies \(\zeta_N \geq \tau_N + \frac{1}{\sqrt{\tau_N}} - \frac{1}{\sqrt{\tau_N}}\)) when \(\left(\frac{c_{\text{ave}}}{c_1}\right) = 4\) and \(\left(\frac{c}{c_{\text{ave}} - c_0}\right) = 1.25\) assuming \(\rho = 0.2\) and \(c_r = 50\) (for \(\psi = 0\) then \(RE \approx 0.19\), for \(\psi = 0.55\) then \(RE \approx 0.14\)).
c For $RE$ (TSS3 vs SRS) the same result obtained for $RE$ (TSS2 vs SRS) applies since the two relative efficiencies only differ in the denominator, where the component $\sqrt{(1 - \rho)}$ is inflated by $(\tau_0^2 + 1) \geq 1$. Likewise, the condition given for TSS2 is sufficient for $RE$ (TSS3 vs SRS) $\leq 1$, but not a necessary condition since, in the same example given for TSS2, $RE \approx 0.17$ ($\psi = 0$) and $RE \approx 0.12$ ($\psi = 0.35$).

d To see whether or when TSS1 is more efficient than TSS2, that is, $RE$ (TSS2 vs TSS1) $= \frac{\tau_{TSS1}(\hat{\alpha})}{\tau_{TSS2}(\hat{\alpha})}$ $\leq 1$, we need to show that the difference between the numerator and denominator is smaller than zero:

$$\left( \sqrt{c_r \rho (1 + \psi(\tau_N (\zeta_N - \tau_N) + 1))} + \sqrt{1 - \rho} \right)^2$$

$$= c_r \rho (1 + \psi(\tau_N (\zeta_N - \tau_N) + 1)) (1 - \rho) + 2c_{r_r} \rho (1 - \rho) (1 + \psi(\tau_N (\zeta_N - \tau_N) + 1)) +$$

$$- c_r \rho \left( \tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \right) - (1 - \rho) +$$

$$- 2 \sqrt{c_r \rho (1 - \rho)} \left( \tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \right) =$$

$$= -c_r \rho \tau_0^2 + c_r \rho \psi \left( \tau_N (\zeta_N - \tau_N) + 1 - (\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \right) +$$

$$+ 2 \sqrt{c_r \rho (1 - \rho)} \left( \sqrt{1 + \psi(\tau_N (\zeta_N - \tau_N) + 1)} - \sqrt{\tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1)} \right).$$

Since $c_r > 1$, $\rho \in (0, 1)$, and $\tau_N \geq 0$, we need to find when $\left\{ \tau_N (\zeta_N - \tau_N) + 1 - (\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \right\} \leq 0$ and $\left\{ \sqrt{1 + \psi(\tau_N (\zeta_N - \tau_N) + 1)} - \sqrt{\tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1)} \right\} \leq 0$. Let us first focus on the latter inequality. Recall from (a) and (b) that $1 + \psi(\tau_N (\zeta_N - \tau_N) + 1) \geq 0$ for $\zeta_N \geq \tau_N - \frac{1}{\tau_N}$ $- \frac{1}{\tau_N}$, while $\tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \geq 0 \forall \zeta_N \in \mathbb{R}$. Now, let us find when the following inequality holds $1 + \psi(\tau_N (\zeta_N - \tau_N) + 1) \leq \tau_0^2 + 1 + \psi(\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1)$, which with some rearrangements gives $\psi \left\{ \tau_N (\zeta_N - \tau_N) + 1 - (\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \right\} - \tau_0^2 \leq 0$. Thus, to see whether or when TSS1 is more efficient than TSS2, we need to find when

$$\tau_N (\zeta_N - \tau_N) + 1 - (\tau_0^2 + \tau_0^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_0^2) + 1) \leq 0.$$
for a cluster size distribution such that \( \tau_N - \frac{1}{\tau_N} - \frac{1}{\psi N} \leq \zeta_N \leq \tau_N - \frac{1}{\psi N} \sqrt{\psi N} \geq \tau_N \), which means that \( \text{RE} (\text{TSS2 vs TSS1}) \leq 1 \). Note that \( \text{RE} (\text{TSS2 vs TSS1}) \leq 1 \) for a Normal distribution (i.e. \( \zeta_N = 0 \) and \( \eta_N = 3 \)), since we have that

\[
\text{RE} (\text{TSS2 vs TSS1}) = \left( \frac{\sqrt{c_\rho[1 + \psi(1 - \tau_N^2)]} + \sqrt{1 - \rho}}{\sqrt{c_\rho[1 + \psi(1 - \tau_N^2)]} + \sqrt{1 - \rho}} \right)^2 \leq 1.
\]

Note that \( \tau_N - \frac{1}{\psi N} - \frac{1}{\psi N} \leq \zeta_N \leq \tau_N - \frac{1}{\psi N} \sqrt{\psi N} \geq \tau_N \) or \( \zeta_N \geq \tau_N \) or \( N_j \sim N (\theta_N, \sigma_N^2) \) is a sufficient condition for \( \text{RE} (\text{TSS2 vs TSS1}) \leq 1 \), but not a necessary condition since \( \text{RE} < 1 \) for the four-parameter Beta distribution with \( \zeta_N = 0 \) (which does not satisfy the condition) in Figure S.11.

e For \( \text{RE} (\text{TSS3 vs TSS1}) \) the same result obtained for \( \text{RE} (\text{TSS2 vs TSS1}) \) applies since the two relative efficiencies only differ in the denominator, where the component \( \sqrt{(1 - \rho)} \) is inflated by \( (\tau_N^2 + 1) \geq 1 \). Like for TSS2, \( \tau_N - \frac{1}{\psi N} - \frac{1}{\psi N} \leq \zeta_N \leq \tau_N - \frac{1}{\psi N} \sqrt{\psi N} \geq \tau_N \) or \( \zeta_N \geq \tau_N \) or \( N_j \sim N (\theta_N, \sigma_N^2) \) is a sufficient but not necessary condition for \( \text{RE} (\text{TSS3 vs TSS1}) \leq 1 \), since \( \text{RE} < 1 \) for the four-parameter Beta distribution with \( \zeta_N = 0 \) (which does not satisfy the condition) in Figure S.12.

3 Local Optimality Problem and Maximin Design

3.1 Derivation of the Maximin Design

To find the maximin design, or more precisely, to find the values of the nuisance parameters that maximize the sampling variance of the mean estimator (see step 3 of the maximin procedure in section 4, main text), we need to take the partial derivative of the sampling variance with respect to each nuisance parameter in turn.

**TSS1:** Taking the partial derivatives of \( V_{\text{TSS1}} (\hat{\mu}) \) (Table 1, main text) with respect to \( \rho \) and \( \psi \), we have that

\[
\frac{\partial V_{\text{TSS1}} (\hat{\mu})}{\partial \rho} = \frac{\sigma_N^2}{n k} \left\{ n \left[ 1 + \psi \left( \tau_N (\zeta_N - \tau_N) + 1 \right) \right] - 1 \right\} \geq 0,
\]

for cluster size distributions such that \( \zeta_N \geq 1 - n \{ 1 + \psi (1 - \tau_N^2) \} \psi N \tau_N \), and

\[
\frac{\partial V_{\text{TSS1}} (\hat{\mu})}{\partial \psi} = \frac{\sigma_N^2}{k} \left[ \tau_N (\zeta_N - \tau_N) + 1 \right] \geq 0,
\]

for cluster size distributions such that \( \zeta_N \geq \tau_N - \frac{1}{\psi N} \psi N \tau_N \). Since \( \tau_N - \frac{1}{\psi N} \psi N \tau_N \leq \tau_N - \frac{1}{\psi N} \), we can conclude that for cluster size distributions such that \( \zeta_N \geq \tau_N - \frac{1}{\psi N} \), \( V_{\text{TSS1}} (\hat{\mu}) \) is an increasing function of \( \rho \) and \( \psi \). Note that, in turn, \( \psi \) is an increasing function of \( \rho_{\text{un}}^2 \) with a minimum in \( \rho_{\text{un}} = 0 \). Although \( \tau_N \) and \( \zeta_N \) are known before drawing a TSS1 sample (Table 1, main text), for completeness we also give the partial derivatives of \( V_{\text{TSS1}} (\hat{\mu}) \) with respect to these two parameters:

\[
\frac{\partial V_{\text{TSS1}} (\hat{\mu})}{\partial \zeta_N} = \frac{\sigma_N^2}{k} \psi N \tau_N \geq 0, \quad \text{and} \quad \frac{\partial V_{\text{TSS1}} (\hat{\mu})}{\partial \tau_N} = \frac{\sigma_N^2}{k} \psi (\zeta_N - 2 \tau_N),
\]

so \( V_{\text{TSS1}} (\hat{\mu}) \) is an increasing function of \( \zeta_N \), and also of \( \tau_N \) but only for positively skewed distributions such that \( \zeta_N > 2 \tau_N \).

**TSS3:** Taking the partial derivatives of \( V_{\text{TSS3}} (\hat{\mu}) \) (equation (4)) with respect to \( \rho \), \( \psi \), \( \zeta_N \), \( \eta_N \), and \( \tau_N \) we have that

\[
- \frac{\partial V_{\text{TSS3}} (\hat{\mu})}{\partial \rho} = \frac{\sigma_N^2}{n k} \left\{ (\tau_N^2 + 1) (n - 1) + n \psi \left( \tau_N^2 + \tau_N (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1 \right) \right\} \geq 0 \forall \zeta_N \in \mathbb{R}, \text{as shown below.}
\]

Note that

\[
\frac{\sigma_N^2}{n k} \left\{ (\tau_N^2 + 1) (n - 1) + n \psi \left( \tau_N^2 + \tau_N (\eta_N - 3) + 2 \zeta_N \tau_N (1 - \tau_N^2) + 1 \right) \right\} \geq 0 \Leftrightarrow
\]

\[
\zeta_N \geq 1 - \frac{1}{\psi N} \psi N \tau_N \geq \tau_N - \frac{1}{\psi N} \sqrt{\psi N} \geq \tau_N \text{ for a cluster size distribution such that } \zeta_N \geq \tau_N - \frac{1}{\psi N} \sqrt{\psi N} \geq \tau_N.]
Fig. S.8. Relative Efficiency of the optimal TSS1 versus SRS, for a given research budget and assuming \( \left( \frac{c_{\text{srs}}}{c_{1}} \right) = \left( \frac{C}{C_{0}} \right) = 1 \) (values greater than 1 give a higher RE of TSS1 versus SRS), as a function of \( \rho \), for different values of \( c_r \) and \( \psi \) (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that \( \psi = 0 \) corresponds to \( \rho_{\mu N} = \pm 0.51 \).

\[
\left[ (\tau_{N}^2 + 1)(n - 1) + n\psi \left( \tau_{N}^4 + \tau_{N}^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_{N}^2) + 1 \right) \right] \geq 0.
\]

Recall that \( \eta_N \geq \zeta_N^2 + 1 \), then

\[
\left[ (\tau_{N}^2 + 1)(n - 1) + n\psi \left( \tau_{N}^4 + \tau_{N}^2 (\eta_N - 3) + 2\zeta_N \tau_N (1 - \tau_{N}^2) + 1 \right) \right] \geq

\left[ (\tau_{N}^2 + 1)(n - 1) + n\psi \left( \tau_{N}^4 + \tau_{N}^2 (\zeta_N^2 - 2) + 2\zeta_N \tau_N (1 - \tau_{N}^2) + 1 \right) \right] =

= n\psi \tau_{N}^4 \zeta_N + 2n\psi \tau_N (1 - \tau_{N}^2) \zeta_N + \left( (\tau_{N}^2 + 1)(n - 1) + n\psi (1 - \tau_{N}^2) \right) \geq 0
\]

(where \( n\psi \tau_{N}^2 > 0 \)) and solving the latter inequality with respect to \( \zeta_N \) we have that \( \Delta = \left( 2n\psi \tau_N (1 - \tau_{N}^2) \right)^2 - 4n\psi \tau_{N}^2 \left( (\tau_{N}^2 + 1)(n - 1) + n\psi (1 - \tau_{N}^2) \right)^2 < 0 \), thus

\[
0 \leq n\psi \tau_{N}^2 \zeta_N + 2n\psi \tau_N (1 - \tau_{N}^2) \zeta_N + \left( (\tau_{N}^2 + 1)(n - 1) + n\psi (1 - \tau_{N}^2) \right) =
\]
Fig. S.9. Relative Efficiency of the optimal TSS2 versus SRS, for a given research budget and assuming \( \left( \frac{c_{rs}}{c_s} \right) = 1 \) (values greater than 1 give a higher RE of TSS2 versus SRS), as a function of \( \rho \), for different values of \( c_r \) and \( \psi \) (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that \( \psi = 0.35 \) corresponds to \( \rho_{IN} = \pm 0.51 \).
Fig. S.10. Relative Efficiency of the optimal TSS3 versus SRS, for a given research budget and assuming 
\( \left( \frac{c_{SRS}}{c} \right) = 1 \) (values greater than 1 give a higher RE of TSS3 versus SRS), as a function of \( \rho \), for different values of \( c_r \) and \( \psi \) (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that \( \psi = 0 \) corresponds to \( \rho_{uN} = \pm 0.51 \).

\[
\begin{align*}
\text{RE(TSS3 vs SRS)} & = \left( \frac{\sigma^2}{\sigma^2_Y} \right) \\
& = \left\{ \left( \frac{\sigma^2}{\sigma^2_Y} \right) \right\} = 1 \\
& \leq \left( \frac{\sigma^2}{\sigma^2_Y} \right) \\
\text{Bimodal: } & = \left( \frac{\sigma^2}{\sigma^2_Y} \right) \\
\text{Beta: } & = \left( \frac{\sigma^2}{\sigma^2_Y} \right)
\end{align*}
\]

\[
\begin{align*}
\partial V_{TSS3}(\hat{\mu}) & = \frac{\partial V_{TSS3}(\hat{\mu})}{\partial \psi} = \sigma^2 \rho k \left( \psi \tau_N + \rho \left( (n-1) + n \psi (\tau_N + \tau_N^2 (\xi_N - 3) + 2 \xi_N (1 - \xi_N^2) + 1) \right) \right) \\
& \geq 0 \forall \xi_N \in \mathbb{R}, \text{ as shown in section 2.1.}
\end{align*}
\]

\[
\begin{align*}
\partial V_{TSS3}(\hat{\mu}) & = \frac{\partial V_{TSS3}(\hat{\mu})}{\partial \tau_N} = 2 \sigma^2 \rho k \psi \tau_N (1 - \tau_N^2) \geq 0 \text{ for } \tau_N \leq 1.
\end{align*}
\]

\[
\begin{align*}
\partial V_{TSS3}(\hat{\mu}) & = \frac{\partial V_{TSS3}(\hat{\mu})}{\partial \eta_N} = \sigma^2 \rho k \psi \tau_N \eta_N \geq 0.
\end{align*}
\]

\[
\begin{align*}
\partial V_{TSS3}(\hat{\mu}) & = \frac{\partial V_{TSS3}(\hat{\mu})}{\partial \zeta_N} = \frac{\sigma^2 \rho}{4k} \left[ 2 \tau_N + \rho \left\{ 2 (n-1) + n \psi (4 \tau_N + 2 \tau_N (\eta_N - 3) + 2 \xi_N (1 - \xi_N^2) + 1) \right\} \right] \\
& = \frac{\sigma^2 \rho}{4k} \left\{ 4n \psi \rho \tau_N^2 - 6n \psi \xi_N \rho \tau_N^2 + 2 \tau_N \left\{ 1 + \rho (n-1) + n \psi (\eta_N - 3) + 2n \psi \rho \xi_N \right\} \right\}, \text{ which is a cubic equation in } \\
\text{tau_N and function of } n, \psi, \rho, \xi_N, \text{ and } \eta_N, \text{ which can then have several patterns depending on these parameters.}
\end{align*}
\]
**TSS2:** Taking the partial derivatives of $V_{TSS2} (\hat{\mu})$ (equation (3)) with respect to $\rho$, $\psi$, $\zeta_N$, $\eta_N$, and $\tau_N$ we have that

$$- \frac{\partial V_{TSS2} (\hat{\mu})}{\partial \rho} = \frac{\sigma^2}{\mu} \left\{ n \left( \left( \tau_N^2 + 1 \right) + \psi \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \right) - 1 \right\} \geq 0 \forall \zeta_N \in \mathbb{R},$$

as shown below. Recall that $\eta_N \geq \zeta_N + 1$, then

$$\left\{ n \left( \left( \tau_N^2 + 1 \right) + \psi \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \right) - 1 \right\} \geq
$$

$$\left\{ \left( \tau_N^2 + 1 \right) + \psi \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \right\} - 1 =
$$

$$n \psi \tau_N^2 \zeta_N^2 + 2 n \psi \tau_N \left( 1 - \tau_N^2 \right) \zeta_N + n \left( \tau_N^2 + 1 + \psi \left( 1 - \tau_N^2 \right)^2 \right) - 1 \geq 0$$

and solving the inequality with respect to $\zeta_N$ follows that $\Delta = \left[ 2 n \psi \tau_N (1 - \tau_N^2)^2 \right]^2 - 4 n \psi \tau_N^2 \left( \left( \tau_N^2 + 1 + \psi \left( 1 - \tau_N^2 \right)^2 \right) - 1 \right] = -4 n \psi \tau_N^2 \left( \left( \tau_N^2 + 1 + \psi \left( 1 - \tau_N^2 \right)^2 \right) - 1 \right)$

since $n \geq 1$ and $\tau_N \geq 0$. Hence, $0 \leq \left\{ n \left( \left( \tau_N^2 + 1 \right) + \psi \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \right) - 1 \right\} \leq
$$

$$\left\{ \left( \tau_N^2 + 1 \right) + \psi \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \right\} - 1.$$

$$- \frac{\partial V_{TSS2} (\hat{\mu})}{\partial \psi} = \frac{\sigma^2}{\mu} \left( \tau_N^2 + \tau_N^2 \left( \eta_N - 3 \right) + 2 \zeta_N \tau_N \left( 1 - \tau_N^2 \right) + 1 \right) \geq 0 \forall \zeta_N \in \mathbb{R},$$

as shown in section 2.1.

$$- \frac{\partial V_{TSS2} (\hat{\mu})}{\partial \tau_N} = 2 \frac{\sigma^2 \rho}{\mu} \psi \tau_N \left( 1 - \tau_N^2 \right) \geq 0 \text{ for } \tau_N \leq 1.$$
Fig. S.12. Relative Efficiency of the optimal TSS3 versus the optimal TSS1, for a given research budget, as a function of $\rho$, for different values of $c_r$ and $\psi$ (curves), and different cluster size distributions (panels). The cluster size distributions are shown in Figure S.1. Note that $0.4 < RE < 0.9$, and that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

$$-rac{\partial V_{TSS3}(\hat{\mu})}{\partial \eta} = \frac{\sigma^2}{k} \psi \tau_N^2 \geq 0.$$  

$$-rac{\partial V_{TSS2}(\hat{\mu})}{\partial \zeta_N} = \frac{\sigma^2}{k} \{2 \tau_N + \psi (4 \tau_N^3 + 2 \tau_N (\eta_N - 3) + 2 \zeta_N - 6 \zeta_N \tau_N^2) \},$$  

which is a cubic equation in $\tau_N$ and function of $\psi$, $\zeta_N$, and $\eta_N$, which can then have several patterns depending on these parameters.

Note that $V_{TSS3}(\hat{\mu})$ and $V_{TSS2}(\hat{\mu})$ are decreasing functions of $\zeta_N$ for $\tau_N > 1$. Thus, if the range of plausible values for $\tau_N$ is $(1, L]$ (with $L > 1$), then in order to derive the maximin design, take the lower-bound of the range for $\zeta_N$ instead of the upper-bound, as suggested in section 4 of the main text. If the range of plausible values for $\tau_N$ is $[L_1, L_2]$ (with $L_1 < 1 < L_2$), to find the worst-case values of $\zeta_N$ and $\tau_N$, search numerically for that combination of values of $\zeta_N$ and $\tau_N$ that maximizes $V(\hat{\mu})$ within the ranges of plausible values for $\zeta_N$ and $\tau_N$, given the chosen values (i.e. the upper-bounds) for $\rho$, $\psi$, and $\eta_N$. This numerical evaluation is implemented in an R function given in section 2 in the Supplementary Material 2.
3.2 A general program in R to find the maximin parameter values

The sampling variances on which the optimal TSS2 and TSS3 are based are equations (3) and (4), respectively. As shown in the previous paragraph, $V_{TSS3}(\hat{\mu})$ and $V_{TSS2}(\hat{\mu})$ are increasing functions of $\rho$, $\psi$, $\eta$, and (for $\tau_N \leq 1$) also $\xi_N$. Thus, to derive the maximin design, one only needs to plug into the optimal design equations (Table 2, main text) the largest plausible value for each of these parameters. In contrast, to find the worst-case value for $\tau_N$, a numerical evaluation is needed. In the situation for which the plausible range for $\tau_N$ is $[L_1, L_2]$ (with $L_1 < 1 < L_2$), $V_{TSS3}(\hat{\mu})$ and $V_{TSS2}(\hat{\mu})$ are no longer monotone functions of $\xi_N$ then, for this case, a numerical evaluation is needed to find the worst-case values for $\tau_N$ and $\xi_N$.

Such numerical evaluation can be performed with two R functions, called worst_case_tau_TSS3 and worst_case_tau_TSS2, given in section 2 of the Supplementary Material 2. These functions return the worst-case value for $\tau_N$ and $\xi_N$, and the corresponding maximum variance value for TSS3 and TSS2, respectively. Both functions need as input: the largest plausible values for the intraclass correlation $\rho$ (rho), the parameter $\psi$ (psi), and the kurtosis of cluster size distribution $\eta_N$ (eta), the lower-bound and the upper-bound of the range of plausible values for the skewness $\xi_N$ (zeta_LB and zeta_UB, respectively) and the coefficient of variation $\tau_N$ (tau_LB and tau_UB, respectively) of cluster size distribution, and finally, the (expected value of the average) number of individuals per cluster $n$. Note that $\sigma^2_{\hat{\mu}}$ is just a multiplicative factor in the expressions for $V_{TSS3}(\hat{\mu})$ and $V_{TSS2}(\hat{\mu})$, thus assuming other values than 1 for $\sigma^2_{\hat{\mu}}$ does not change the result for $\tau_N$. Unfortunately, $n$ is also in the numerator of equations (3) and (4), so a value for $n$ is still required as input for the R functions.

Using these two R functions (Supplementary Material 2, section 2), a numerical evaluation was performed to search for the worst-case $\tau_N$ within the range $\tau_N \in [0, 1]$, given several upper-bounds for $\rho$, $\xi_N$, $\eta_N$, and $\psi$, and several values for $n$. Specifically, $\rho = \{0.1, 0.3, 0.5\}$, $\psi = \{0.067, 0.35, 1.29\}$, $\xi_N = \{-3, -2, -1, 0, 1, 2, 3\}$, and $\eta_N = \{1.5, 3, 4.5, 6, 9, 12, 15\}$, which involves a total of $3 \times 3 \times 31 = 279$ scenarios since $\eta_N \geq \xi_N^2 + 1$ [8] and $n = \{5, 50\}$. In all scenarios, the worst-case value of $\tau_N$ within the range $[0, 1]$ was the upper-bound $\tau_N = 1$. Using the two R functions in section 2 in the Supplementary Material 2, one can explore broader ranges for $\tau_N$.

3.3 Accounting for the approximations in the maximin design for TSS2 and TSS3

As mentioned in section 3.1 of the main text, the maximin design for TSS2 and TSS3 depends on two approximations: the first-order Taylor series approximation evaluated in section 1 and the large $k$ approximation in equations (3) and (4). Denote by $V(\hat{\mu})_A$ the expression for the variance based on the first-order Taylor series approximation (i.e. Table S.1), by $k_A$ the $k$ such that $V(\hat{\mu})_A$ is accurate (i.e. bias $\leq 5\%$), by $V(\hat{\mu})_L$ the variance based on the two approximations (i.e. equations (3) and (4)), and by $k_L$ the $k$ such that $0.95 \leq \frac{V(\hat{\mu})_A}{V(\hat{\mu})_L} \leq 1.05$, that is, the second approximation is fairly accurate. Recall, from section 1, that $k_A = 20$ under cluster size distributions with skewness and kurtosis close to (say, $\pm 1.5$) those of the Normal distribution, and $k_A = 100$ under distributions with extreme moments.

A potential limitation of the maximin design for TSS2 and TSS3 is that the maximin $k^{MD}$ can be in the interval $[k_A, k_L]$ and then the large $k$ approximation is biased. To address this issue, $\frac{V(\hat{\mu})_A}{V(\hat{\mu})_L}$ was numerically evaluated (both for TSS2 and TSS3) to search for $k_L$ under several combinations of worst-case values for $\rho$, $\psi$, $\tau_N$, $\xi_N$, and $\eta_N$. Specifically, $\rho = \{0.05, 0.1, 0.25\}$, $\psi = \{0.067, 0.35\}$, $\tau_N = \{0.5, 0.75, 1\}$, while for $\xi_N$ and $\eta_N$ two sets of values were considered, respectively, $\xi_N = \{0.5, 1, 1.5\}$ and $\eta_N = \{1.5, 3, 4.5\}$ giving $3 \times 2 \times 3 \times 6 = 108$ scenarios, and $\xi_N = \{2, 2.5, 3\}$ and $\eta_N = \{6, 9, 12, 15\}$ giving $3 \times 2 \times 3 \times 9 = 162$ scenarios (recall that $\eta_N \geq \xi_N^2 + 1$). The R code to perform this numerical evaluation is given in section 3 of the Supplementary Material 2. In all scenarios $k_A$ was equal to $k_L$, alleviating the concerns about this issue. The results of this numerical evaluation are plotted in Figure S.13 for three worst-case scenarios (rows). The first two rows show the results for a scenario of small–moderate skewness and kurtosis (i.e. $\xi_N = \{0.5, 1, 1.5\}$ and $\eta_N = \{1.5, 3, 4.5\}$), so that $k_A = 20$ according to the results of the simulation study. The bottom row, instead, covers the case of large skewness and kurtosis (i.e. $\xi_N = \{2, 2.5, 3\}$ and $\eta_N = \{6, 9, 12, 15\}$), and then $k_A = 100$, based on the simulation study. The second and third row show the results for the scenario with the largest values for each parameter, while the first row gives the results for the scenario with the smallest values for each parameter.
Fig. S.13. Ratio of the expression for the variance of the TSS3 (left column) and TSS2 (right column) based on the first-order Taylor series approximation, $V(\hat{\mu})$, to the expression based on the large $k$ approximation and the first-order Taylor series approximation, $V(\hat{\mu})_1$, as a function of the number of clusters $k$, for the worst-case scenario given by $\rho$, $\psi$, $\tau_N$, $\zeta_N$, and $\eta_N$ (rows), and for different values of $n$ (curves). Note that $k_A$, the required number of clusters for $V(\hat{\mu})_1$ to be accurate, is 20 in the first two rows, and 100 in the bottom row.
4 Sample Size Calculation for Cross-Population Comparisons

Van Breukelen and Candel [10] derive optimal and maximin sample sizes for cluster randomized trials allowing costs and variances to vary across treated and control arms. They also give a procedure to derive maximin sample sizes and maximin budget split between arms for the desired power level and the chosen type I error rate. The sampling variance of $\hat{\mu}$ under TSS1 for non-informative cluster size $\psi = 0$ (Table 1, main text) is proportional to the sampling variance of the treatment effect estimator in a cluster randomized trial (eq. (3) in [10]). The results of Van Breukelen and Candel [10] are here extended to TSS1 under informative cluster size ($\psi \neq 0$) and non-comparable cluster size distributions. Since the French and the Italian samples are independent, we have that $V (\hat{\mu}_F - \hat{\mu}_I) = V (\hat{\mu}_F) + V (\hat{\mu}_I)$, where $V (\hat{\mu})$ is given in Table 1 (main text). Hence, equation (7a) of Van Breukelen and Candel [10] is extended to the following equation

$$V (\hat{\mu}_F - \hat{\mu}_I) = V (\hat{\mu}_F) + V (\hat{\mu}_I) = \frac{\sigma_{\hat{\mu},F}^2}{n_F k_F} \left[ 1 + \rho_F \left( (n_F - 1) + n_F \psi_F (\tau_{\psi,F} (\zeta_{\psi,F} - \tau_{\psi,F}) + 1) \right) \right]$$

$$+ \frac{\sigma_{\hat{\mu},I}^2}{n_I k_I} \left[ 1 + \rho_I \left( (n_I - 1) + n_I \psi_I (\tau_{\psi,I} (\zeta_{\psi,I} - \tau_{\psi,I}) + 1) \right) \right].$$

(S.1)

Note that for $\psi_F = \psi_I = 0$ equation (S.1) reduces to equation (7a) in [10]. Let $C_F = k_F (c_{2,F} + c_{1,F} n_F)$ be the cost constraint for the French sample, and $C_I = k_I (c_{2,I} + c_{1,I} n_I)$ be the Italian cost constraint, where $C_F$ and $C_I$ are the budget for France and Italy, respectively. Minimizing $V (\hat{\mu}_F - \hat{\mu}_I)$ subject to the two cost constraints gives, per country, the equations for the optimal $n^*$ and $k^*$ given in Table 2 in the main text (TSS1, row), where each quantity has the subscript of the corresponding nation (i.e. “F” for France, and “I” for Italy). For $\psi_F = \psi_I = 0$ (or when the cluster size distribution and $\alpha$ are the same in both populations, so that $\mu_F - \mu_I = \beta_{0,F} - \beta_{0,I}$) these equations reduce to equations (7b) and (7c) in [10]. Plugging into equation (S.1) the optimal design per country gives

$$V (\hat{\mu}_F - \hat{\mu}_I) = \frac{g_F (\rho_F, \psi_F) \sigma_{\hat{\mu},F}^2}{C_F} + \frac{g_I (\rho_I, \psi_I) \sigma_{\hat{\mu},I}^2}{C_I},$$

(S.2)

which generalizes equation (7d) in [10] to $\psi \neq 0$, and where

$$g_F (\rho_F, \psi_F) = c_{1,F} \left( \sqrt{c_{2,F} \rho_F (1 + \psi_F (\tau_{\psi,F} (\zeta_{\psi,F} - \tau_{\psi,F}) + 1)) + \sqrt{1 - \rho_F}} \right)^2,$$

and

$$g_I (\rho_I, \psi_I) = c_{1,I} \left( \sqrt{c_{2,I} \rho_I (1 + \psi_I (\tau_{\psi,I} (\zeta_{\psi,I} - \tau_{\psi,I}) + 1)) + \sqrt{1 - \rho_I}} \right)^2.$$

The design can be further optimized by constraining the total budget, instead of each separate budget, and deriving the optimal budget split between France and Italy. Denote by $C$ the total budget of the study and by $f$ the fraction of that budget allocated to the French sample, so that $C_F = f C$ and $C_I = (1 - f) C$. Plugging $C_F = f C$ and $C_I = (1 - f) C$ into equation (S.2) and minimizing (S.2) as a function of $f$ gives the optimal budget split

$$\frac{f^*}{1 - f^*} = \left( \frac{\sigma_{\hat{\mu},F}}{\sigma_{\hat{\mu},I}} \right) \left( \frac{\sqrt{g_F (\rho_F, \psi_F)}}{\sqrt{g_I (\rho_I, \psi_I)}} \right),$$

(S.3)

which reduces to equation (8) in [10] when $\psi_F = \psi_I = 0$. Note that $f^* = \left( 1 + \left( \frac{f^*}{1 - f^*} \right)^{-1} \right)^{-1}$ and $1 - f^* = \left( 1 + \frac{f^*}{1 - f^*} \right)^{-1}$, and then plugging these equations into (S.2) with $C_F = f^* C$ and $C_I = (1 - f^*) C$ gives the optimal variance

$$V (\hat{\mu}_F - \hat{\mu}_I)^* = \left( \frac{\sigma_{\hat{\mu},F} \sqrt{g_F (\rho_F, \psi_F)} + \sigma_{\hat{\mu},I} \sqrt{g_I (\rho_I, \psi_I)}}{C} \right)^2,$$

(S.4)

which extends equation (9) in [10] to the case of $\psi \neq 0$. Thus, the optimal sample sizes $(n^*,k^*)$ per country are given in the TSS1 row of Table 2 (main text), provided that $C$ in the $k^*$ equation is replaced with $C_F = f^* C$ for France and $C_I = (1 - f^*) C$ for Italy.
The optimal design and the optimal budget split depend on several unknown parameters (i.e., $\rho_F$, $\psi_F$, $\rho_I$, $\psi_I$, and $\sigma_{r,F}^2$). A maximin approach is adopted to overcome such local optimality. The four steps to obtain the maximin design are given in section 4 of [10], or similarly, in section 4 of the main text. However, we still need to extend equation (10) in [10] to derive the maximin budget split $f_{MD}^{\rho}$. Denote by $V_{\text{max}} \geq \sigma_{y,F}^2 + \sigma_{y,I}^2$ the largest plausible upper-bound for $\sigma_{y,F}^2 + \sigma_{y,I}^2$, and $\frac{\sigma_{r,F}}{\sigma_{r,I}} \in \left[\frac{1}{q}, q\right]$ the range of values for the ratio $\frac{\sigma_{r,F}}{\sigma_{r,I}}$ based on $q \geq 1$. Recall from section 3 that $V(\hat{\mu})$ for TSS1 is an increasing function of $\rho$ (at least up to $\rho \leq 0.5$) and $\psi$. Hence, plugging into (S.2) the largest plausible value for these two parameters, $\rho$ (max) and $\psi$ (max), respectively, we have that

$$
\max V(\hat{\mu}_F - \hat{\mu}_I) = \frac{g_F(\rho(\text{max}), \psi(\text{max}))}{\sqrt{\psi_C} \frac{\sigma_{y,F}^2}{\sigma_{y,I}^2}} + \frac{g_I(\rho(\text{max}), \psi(\text{max}))}{(1 - f) C} \left(V_{\text{max}} - \sigma_{y,F}^2 \right) = \frac{g_F(\rho(\text{max}), \psi(\text{max}))}{\sqrt{\psi_C} \frac{\sigma_{y,F}^2}{\sigma_{y,I}^2}} + \frac{g_I(\rho(\text{max}), \psi(\text{max}))}{(1 - f) C} \left(V_{\text{max}} - \sigma_{y,F}^2 \right) \tag{S.5}
$$

where $h = \sqrt{\frac{g_F(\rho(\text{max}), \psi(\text{max}))}{g_I(\rho(\text{max}), \psi(\text{max}))} \frac{\sigma_{y,F}^2}{\sigma_{y,I}^2}}$ Equation (S.5) is a function of $f$ and $\sigma_{y,F}^2$, so that to find the maximin budget split $f_{MD}^{\rho}$ and the minimum maximum variance (i.e. the variance of the maximin design), one needs to (i) maximize equation (S.5) as a function of $\sigma_{y,F}^2$ within its range $\sigma_{y,F}^2 \in \left[\frac{V_{\text{max}}}{1 + q^2}, \frac{q^2 V_{\text{max}}}{1 + q^2}\right]$, which is obtained by combining $\frac{1}{q} \leq \frac{\sigma_{r,F}}{\sigma_{r,I}} \leq q$ with $V_{\text{max}} \geq \sigma_{y,F}^2 + \sigma_{y,I}^2$; (ii) minimize that maximum of equation (S.5) as a function of $f = \frac{f}{(1 - f)}$. Note that equation (S.5) has the same form of equation (10) in [10] and is not a function of $\psi$ (fixed at $\psi(\text{max})$). Hence, following the steps given in [10]'s Appendix B one obtains the results in Table S.9, provided that their $p$ is replaced with $h = \sqrt{\frac{g_F(\rho(\text{max}), \psi(\text{max}))}{g_I(\rho(\text{max}), \psi(\text{max}))} \frac{\sigma_{y,F}^2}{\sigma_{y,I}^2}}$ and their $g_c(\rho(\text{max}), \psi(\text{max}))$ is replaced with $g_I(\rho(\text{max}), \psi(\text{max})) = c_{1,1} \left(\sqrt{c_{2,1} \rho(\text{max})} (1 + \psi(\text{max}) [\xi_{N,F}(\xi_{N,F} - \xi_{N,I}) + 1]) + \sqrt{1 - \rho(\text{max})} \right)^2$. Table S.9 extends Van Breukelen and Candel [10]'s Table 1. Thus, the maximin sample sizes ($n_{MD}^p$, $n_{MD}^p$) per country are obtained by plugging into the TSS1 sample size equations in Table 2 (main text): (i) the sampling costs ($c_1$, $c_2$) per country, (ii) the values of $\tau_N$ and $\zeta_N$ per cluster size distribution, (iii) the values $\rho(\text{max})$ and $\psi(\text{max})$, and (iv) $C_F = f_{MD}^{\rho} C_F$ for France and $C_I = (1 - f_{MD}^{\rho}) C_I$ for Italy, where $f_{MD}^{\rho}$ is determined according to Table S.9. Section 4 of the Supplementary Material 2 provides an R code to compute the maximin sample sizes per country and the maximin budget split for the desired power level. The steps of the R code are as follows:

1. Specify $c_{1,F}, c_{1,I}, c_{2,F}, c_{2,I}, \rho(\text{max}), \psi(\text{max}), \mu_F - \mu_I, V_{\text{max}}$ (i.e. the maximum sum of outcome variances in both populations), and $\left[\frac{1}{q}, q\right]$.
2. Compute $V(\hat{\mu}_F - \hat{\mu}_I)$ for the desired power level $1 - \gamma$, the chosen type I error rate $\alpha$, and the smallest relevant difference $\mu_F - \mu_I$, according to $V(\hat{\mu}_F - \hat{\mu}_I) = \left(\frac{\mu_F - \mu_I}{z_{q - th}}\right)^2$, where $z_q$ is the $q - th$ percentile of the standard normal distribution.
3. Compute the maximin $n_{MD}^p$ and $n_{MD}^p$ by plugging into the $n^p$ equation of Table 2 (TSS1 row, main text): $c_{r,F} = \frac{c_2,F}{c_1,F}$ and $c_{r,I} = \frac{c_2,I}{c_1,I}$, respectively, $\rho = \rho(\text{max}), \psi = \psi(\text{max})$, and the values of $\tau_N$ and $\zeta_N$ of the corresponding cluster size distributions.
4. Compute $h$.
5. Let $f_{MD}^{\rho}$ be the fraction of the budget allocated to France under the maximin budget split, then $1 - f_{MD}^{\rho}$ is the fraction allocated to Italy. Compute the maximin budget split $C_F = \frac{f_{MD}^{\rho}}{(1 - f_{MD}^{\rho})}$ as given in Table S.9 using $h$ and $\left[\frac{1}{q}, q\right]$.
6. Compute the total budget $C$ by equating the maximum variance for the maximin design as given in Table S.9 with $V(\hat{\mu}_F - \hat{\mu}_I)$ as computed in step 2.
7. Using $C$ from step 6 and $f_{MD}^{\rho} = \frac{C_F}{C_I}$ from step 5, compute $C_F = C f_{MD}^{\rho}$ and $C_I = (1 - f_{MD}^{\rho}) C$. 

Optimal TSS for Mean Estimation in Multilevel Populations when Cluster Size is Informative 31
8. Finally, compute the maximin $k^*_{MD}$ and $k^*_{MD}$ by plugging into the $k^*$ equation in Table 2 (TSS1 row, main text), $c_{r,F}, c_{1,F}, n^*_{MD}, C_F$ for $k^*_{MD}$, and $c_{r,I}, c_{1,I}, n^*_{MD}, C_I$ for $k^*_{MD}$.

| Relation of $h$ to $q$ | Maximin budget split $\frac{f_{MD}}{1-q/h}$ | Maximum $V(\hat{\mu}_F - \hat{\mu}_I)$ for the Maximin Design |
|------------------------|------------------------------------------|---------------------------------------------------------------|
| $\frac{1}{q} \leq h \leq q$ | $h^2$ | $g(|\rho(\text{max}), \psi(\text{max})|)V_{\text{max}} \times (1 + h^2)$ |
| $h > q$ | $hq$ | $g(|\rho(\text{max}), \psi(\text{max})|)V_{\text{max}} \times (hq + 1)^2$ |
| $h < \frac{1}{q}$ | $\frac{h}{q}$ | $g(|\rho(\text{max}), \psi(\text{max})|)V_{\text{max}} \times \left(\frac{h+q}{q^2+1}\right)$ |

Table S.9. Maximin budget split and maximum $V(\hat{\mu}_F - \hat{\mu}_I)$ for heterogeneous costs and variances given a fixed maximum total variance $V_{\text{max}}$ and a fixed total budget $C$, as a function of the relation between $q$ and $h$.

References

1. 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. *Stat Med* 2019; 38: 1817-1834.
2. Casella G, Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove: Duxbury; 2002.
3. 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.
4. Forbes C, Evans M, Hastings N, Peacock B. *Statistical Distributions*. 4th ed. Hoboken: John Wiley & Sons; 2011.
5. Salt K. Patients registered at a GP practice. October 2017; Special Topic-Practice list size comparison 2017. Health and Social Care Information Centre. 2017. https://files.digital.nhs.uk/publication/b/k/gp-reg-prac-topic-int-ict-17.pdf. Accessed November 8, 2017.
6. Direzione generale per i contratti, gli acquisti e per i sistemi informativi e la statistica. Studenti per anno di corso e fascia di età: scuola statale. DGCASIS, 2018. http://dati.istruzione.it/opendata/opendata/catalogo/elements1/?area=Studenti. Accessed May 1, 2018.
7. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
8. Pearson K. Editorial note to 'Inequalities for moments of frequency functions and for various statistical constants'. *Biometrika* 1929; 21: 361–375.
9. Ministère de l’Éducation nationale, de la Jeunesse et de la Vie associative. L’Éducation nationale publie les effectifs des établissements scolaires. MENJVA; 2015. https://www.data.gouv.fr/fr/posts/education-nationale-publie-les-effectifs-des-établissements-scolaires/. Accessed October 11, 2018.
10. Van Breukelen GJP, Candel MJJM. Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Stat Med* 2018; 37: 3027-3046.