Efficient designs for mean estimation in multilevel populations and test norming

Efficient designs for mean estimation in multilevel populations and test norming

Francesco Innocenti

Printing: ProefschriftMaken || www.proefschriftmaken.nl

ISBN 978-94-6423-180-9

Efficient designs for mean estimation in multilevel populations and test norming

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday 20 May 2021, at 12.00 hours

by

Francesco Innocenti

Supervisors:

Prof. dr. G.J.P. van Breukelen

Dr. M.J.J.M. Candel

Co-supervisor:

Dr. F.E.S. Tan

Assessment Committee:

Prof. dr. R.M.M. Crutzen (chair)

Prof. dr. P. Goos (KU Leuven and Antwerp University)

Dr. S. Jolani

Dr. ir. M. Moerbeek (Utrecht University)

Prof. dr. H. Verbeek

The research presented in this thesis was conducted at CAPHRI Care and Public Health Research Institute, Department of Methodology and Statistics, of Maastricht University. CAPHRI participates in the Netherlands School of Public Health and Care Research CaRe.

Contents

Contents i

1 Introduction 1

1.1 Survey sampling for mean estimation in multilevel populations 2
1.2 Normative studies for deriving reference values 5
1.3 Model-based approach 9
1.3.1 Models for two-stage sampling 9
1.3.2 Models for norming 10
1.4 Optimal design 11
1.5 Maximin design 13
1.5.1 Local optimality problem 13
1.5.2 Model-dependence 14
1.6 Outline of the thesis 15

2 Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative 17

2.1 Introduction 19
2.2 Assumptions and sampling schemes 21
2.3 Definition and estimation of the population mean μ 24
2.4 Relative efficiencies of TSS schemes versus SRS and each other 27
2.5 Design-based inference for two-stage sampling when cluster size is informative 33
2.6 Application to two real cluster size distributions 38
2.7 Discussion 41
Appendix A: Derivation of the population mean μ 43
Appendix B: Results of Table 2.1 46
Appendix C: Notation, and table of contents of the online supplementary material 49

3 Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative 51

3.1 Introduction 53
3.2 Assumptions, sampling schemes and mean estimators 55
3.3 Optimal design and relative efficiencies for a given budget 61
3.3.1 Optimal design 61
3.3.2 Effect of cluster size informativeness on the optimal design and study budget needed 64
3.3.3 Relative efficiencies for a given budget 67
3.4 Maximin design 73
3.5 Sample size calculation for cross-population comparisons 75

3.6 Discussion 79
Appendix: Notation, and table of contents of the online supplementary materials 82

4 Sample size calculation and optimal design for regression-based norming of tests and questionnaires 87

4.1 Introduction 89
4.2 Models for norming and variances of the norm statistics 91
4.2.1 Models for norming and norm statistics 91
4.2.2 Variances of the norm statistics 94
4.3 Optimal and robust design 97
4.3.1 Designs for optimizing precision of Z-score and PR-score estimation 97
4.3.2 Efficiency of the optimal design when there is uncertainty about the "true" model 103
4.4 Sample size calculation 108
4.4.1 Sample size calculation for hypothesis testing 108
4.4.2 Sample size calculation for precision of norms estimation 113
4.5 Application 114
4.6 Discussion 116
Appendix A: Derivations of the sampling variances of the Z-score and PR-score estimators 119
Appendix B: Summary of the results of the simulation studies 121
Appendix C: Derivations of the sample size calculation formulas (i.e. equations (4.14) and (4.15)) 123
Appendix D: Table of contents of the online supplementary materials 126

5 Sample size calculation and optimal design for multivariate regression-based norming 127

5.1 Introduction 129
5.2 Multivariate regression-based norming 131
5.2.1 Multivariate regression models for norming 131
5.2.2 Two alternative approaches to multivariate regression-based norming 133
5.2.3 Sampling variances of the norm statistics 138
5.3 Optimal and robust design 141
5.3.1 Designs for optimizing precision of norms estimation 141
5.3.2 Robustness of the optimal design against misspecification of the regression model 145
5.4 Sample size calculation for the Mahalanobis distance-based approach 147
5.5 Application 151
5.6 Discussion 154
Appendix A: Derivations of the results in Table 5.2 158
Appendix B: Derivations of the sample size calculation formula for the Mahalanobis distance-based approach 163
Appendix C: Notation, and table of contents of the online supplementary materials 165

6 Discussion 167

6.1 Guidelines 168

6.2 Ideas for future research 172

7 Summary 175

8 Scientific and social impact of the thesis 179

Samenvatting 183

References 187

Acknowledgments 200

About the author 202

Online supplements (not part of this thesis, but available upon request):

Supplementary material for chapter 2

Supplementary materials 1 and 2 for chapter 3

Supplementary materials A and B for chapter 4

Supplementary materials A and B for chapter 5

Chapter 1

Introduction

This thesis deals with efficient designs for two types of observational studies, that is, surveys for mean estimation in multilevel populations, and normative studies for estimating reference values (or norms) to compare individuals with the reference population. Examples of the first type of studies are school-based surveys for monitoring substance use among adolescents, such as the European School Survey Project on Alcohol and Other Drugs (ESPAD Group, 2016), and national surveys for estimating the average length of stay for discharges from hospitals, such as the National Hospital Discharge Survey (DeFrances et al., 2008). Examples of normative studies are the development of norms for patients' orientation toward chronic pain as measured by the Pain Catastrophizing and the Internal Control subscales of the Pain Cognition List (Van Breukelen & Vlaeyen, 2005), and the development of norms for information processing speed and associated brain dysfunction as measured by the oral and written version of the Letter Digit Substitution test (Van der Elst et al., 2006a). Surveys for mean estimation are important because they allow researchers to compare different populations with respect to their means (e.g. comparing European countries in terms of average alcohol consumption among adolescents, like in the ESPAD study). Normative studies are important because they allow clinicians to interpret individuals' performance on a test by comparing their scores with those of their peers (e.g. individuals with the same age, sex, and education) in the reference population, and then to make decisions about, for instance, clinical treatments. Given the practical importance of these two types of studies, estimates of population means or of reference values should be precise. This goal can be attained by a careful design of the study. Specifically, maximum precision in mean or norms estimation is achieved by drawing a sample

as prescribed by the optimal design, where the definition of optimal design differs in the two types of studies. Furthermore, to derive an optimal design a statistical model for describing the population under study should be assumed, and this entails a dependency of the optimal design on the assumed model. Before introducing the different definitions of optimal design, the models used in this thesis, and the approach adopted to overcome the model dependency of the optimal design, the position of this thesis with respect to the literature for each type of study is explained first.

1.1 Survey sampling for mean estimation in multilevel populations

Multilevel populations arise when individuals (e.g. students, patients) are nested within clusters (e.g. schools, hospitals). For these populations, there are two population means that can be the target of inference: the average of all individual outcomes ignoring cluster membership (i.e. first, pooling all students from all schools in the population, and then computing the mean alcohol consumption), and the average of all cluster means (i.e. first, computing the mean alcohol consumption within each school, and then taking the average of all school means). The first population mean can be interpreted as the expected outcome for an individual randomly sampled from the population ignoring cluster membership, so it should be the target of inference when the main interest is on individuals and not on the multilevel structure of the population (e.g. the researcher wants to estimate the average alcohol consumption among Dutch adolescents and the fact that these are nested within schools is not of scientific interest, but it can facilitate the collection of the sample as will be explained later). The second population mean can be interpreted as the expected outcome for the average individual from the average cluster, thus it should be the target of inference when the multilevel structure of the population is of scientific interest (e.g. the researcher wants to estimate the average alcohol consumption for the average Dutch adolescent from the average school, so that school averages can be compared with this population mean). This thesis focuses on the first population mean, as typically done in survey sampling literature (see, for instance, Cochran (1977), Lohr (2010), and Valliant et al. (2000)).

In drawing the sample to estimate either of these two population means cluster membership can be either ignored or taken into account, as shown in Figure 1.1. In the first case, a simple random sample (SRS) of individuals is directly drawn from the population. In the second case, a sample of clusters is drawn first, and then individuals within each selected

cluster are sampled. This latter sampling scheme is called two-stage sampling (TSS). In practice, clusters vary in size (e.g. number of students enrolled in a school, number of patients admitted to a hospital), and this leads to (at least) three alternative TSS schemes: (i) sampling clusters with probability proportional to cluster size and then sampling the same number of individuals from each selected cluster (TSS1); (ii) sampling clusters with equal probability and then sampling per selected cluster a number of individuals proportional to cluster size (TSS2); (iii) sampling clusters with equal probability and then sampling the same number of individuals per selected cluster (TSS3). These three TSS schemes are illustrated in Figure 1.2.

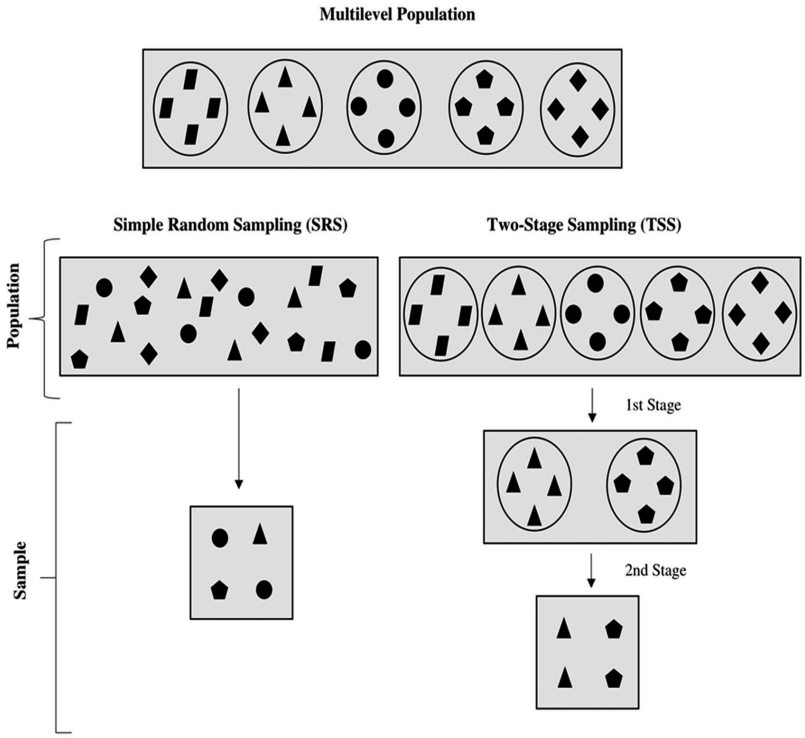


Figure 1.1. Simple random sampling of individuals (left side), and two-stage sampling (right side).

In practice cluster size can not only vary, but also can be related to the outcome variable of interest, and in this case is said to be informative. For instance, cluster size is informative when the amount of alcohol consumed by an adolescent is related to the number of students enrolled in the school, as small schools might provide a more supportive environment. Another example can be the length of stay in a hospital for a patient that might be shorter for patients admitted to hospitals with fewer patients, because healthcare workers are not overwhelmed and

can dedicate enough time to each patient. In the informative cluster size literature (Nevalainen et al., 2014; Panageas et al., 2007; Seaman et al., 2014), the main focus has been on how to handle informative cluster size when the target of inference is the association between the outcome variable and some covariates. For instance, Seaman et al. (2014) have reviewed several methods available in the literature to make cluster-specific inferences with Generalized Linear Mixed Models and population-average inferences with Generalized Estimating Equations when cluster size is informative. This thesis focuses on unbiased and efficient estimation of the average of all individual outcomes (as opposed to the average of all cluster means) with the three aforementioned TSS schemes, and on sample size planning for these sampling schemes, when cluster size is informative.

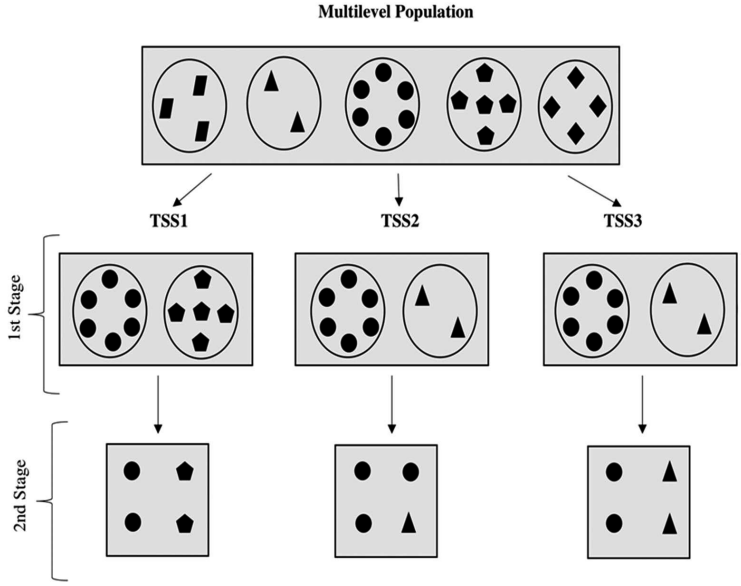


Figure 1.2. The three two-stage sampling schemes considered in this thesis. TSS1 samples clusters with probability proportional to cluster size, and the same number of individuals per selected cluster. TSS2 samples clusters with equal probability, and the same percentage of individuals per selected cluster. TSS3 samples clusters with equal probability, and the same number of individuals per selected cluster.

Survey sampling literature (Chambers & Clark, 2012; Cochran, 1977; Lohr, 2010; Särndal et al, 1992; Sukhatme, 1954; Valliant et al, 2000) has dealt with the issue of estimating the average of all individual outcomes with SRS, and with TSS when there is either no cluster size variation or cluster size varies but is non-informative. In this setting, it has also been established that, under the constraint of a fixed total sample size (i.e. fixed total number of

individuals in the sample), SRS is more efficient than TSS, and sampling clusters with probability proportional to size tends to be more efficient than equal probability sampling. Furthermore, optimal sample size equations for TSS (i.e. the combination of number of clusters and number of individuals per selected cluster that minimizes the sampling variance of the population mean estimator under the constraint of a fixed budget for sampling and measuring) are available in the literature when there is no cluster size variation (i.e. when TSS1, TSS2, and TSS3 coincide), which are similar to those for cluster randomized trials with homogeneous costs and variances (see, for instance, Moerbeek et al., 2000).

This thesis extends the results available in survey sampling literature to informative cluster size, as follows: (i) conditions under which the two types of population means in a multilevel population coincide are derived, (ii) the unbiased (or approximately unbiased) estimators of the average of all individual outcomes for the three aforementioned TSS schemes are given, (iii) conditions under which TSS1 is more efficient than TSS2 and TSS3, under the constraint of a fixed total sample size, are given, (iv) optimal sample sizes equations for TSS1, TSS2, and TSS3 are derived, and consequences of ignoring cluster size informativeness at the design phase of the study are investigated, (v) conditions under which TSS is more efficient than SRS, and TSS1 is more efficient than TSS2 and TSS3, under the constraint of a fixed budget for sampling and measuring (as opposed to a fixed total sample size), are derived, (vi) an approach is proposed to deal with uncertainty about model parameters needed for sample size planning, and (vii) a sample size calculation procedure is proposed for comparing two populations in terms of their population means.

1.2 Normative studies for deriving reference values

In normative studies, a representative sample of individuals is drawn from the reference population (e.g. adults aged 18 to 80 years), a test or questionnaire is administered to the participants and, based on their scores, reference values (or norms) are estimated in order to be able to compare future subjects with the reference population. Reference values can be provided in terms of several types of norm statistic (Oosterhuis et al., 2017): mean test score and standard deviation, percentiles (i.e. the test score value below which a given percentage of individuals falls), percentile rank scores (i.e. the percentage of individuals with a test score equal to or lower than a certain value), and Z-scores (i.e. how many standard deviations the individual's test score is below or above the average). There are two approaches to norming: the traditional

approach and the regression-based approach. The traditional approach consists of first splitting the sample drawn for norming into subgroups based on some relevant demographic factors (e.g. age and sex), and then computing the norm statistics of interest within each subgroup. This is illustrated in the top figure of Figure 1.3, with, for instance, an equal number of persons per age group, and the same sex distribution per age group. The regression-based approach consists of two steps (Van Breukelen & Vlaeyen, 2005): first, a regression of the test score on some relevant predictors is performed, and then norm statistics are estimated from the cumulative distribution of the standardized residuals obtained from the model. The regression-based approach has three main advantages over the traditional approach. First, it uses the whole sample to establish norms instead of norming per subgroup, thereby increasing the precision of the norms. Second, it allows researchers to identify which independent variables (e.g. demographic factors) affect the test score, thereby increasing the validity of the norms. Third, under the assumption of a specific regression model for relating the test score to relevant predictors, it is possible to express the sampling variance of the norm statistic of interest as a function of the joint distribution of the predictors (e.g. the age distribution per sex, and the sex distribution) in the normative sample. This allows to find the joint distribution of the predictors that minimizes the sampling variance of the norm statistic and thus maximizes the precision of the norms under the assumed regression model. This is illustrated in Figure 1.3, where the middle figure shows the optimal distribution of age and sex under the assumption of a linear age effect on the test score, and the bottom figure for a quadratic age effect on the test score. This joint distribution will be called the optimal design for the normative study. A limitation of the regression-based approach is that the validity of the norms depends on whether the model assumptions are met, for instance about the linearity of an age effect, or the homoscedasticity of residual variances.

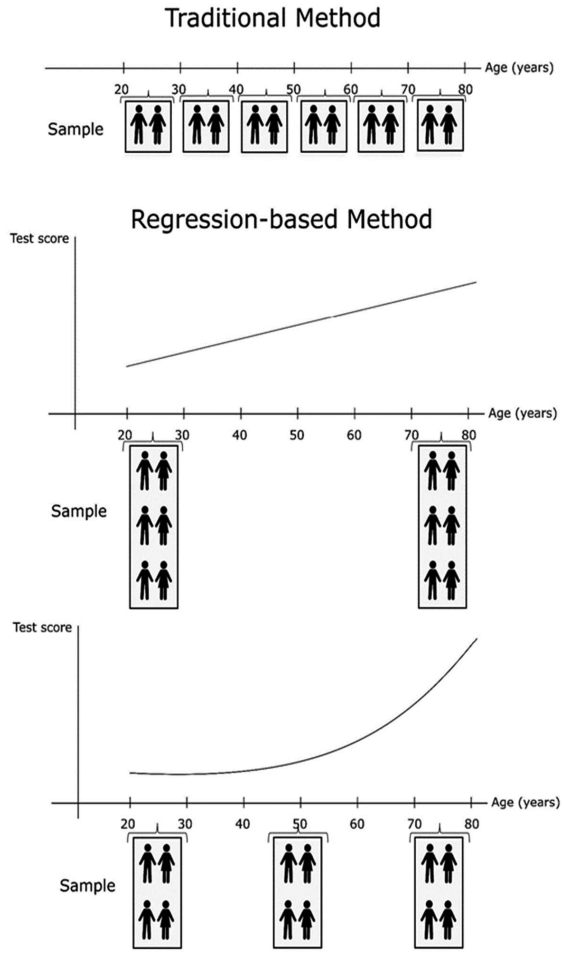


Figure 1.3. Illustration of the two approaches to norming: the traditional approach (top figure), which splits the sample in subgroups based on demographic factors (i.e. age and sex), and the regression-based approach (two bottom figures), which under the assumption of a model relating the test score to relevant predictors (e.g. age and sex) allows to derive the joint distribution of the predictors in the sample that maximizes precision of norms estimation (i.e. the optimal design). The figure in the middle shows the optimal distribution of age and sex under the assumption of a linear age effect, while the bottom figure shows it for a quadratic age effect.

To maximize precision of norms estimation, the size and the design of the sample on which the norms are based should be carefully planned. Oosterhuis et al. (2016) have provided sample size requirements for percentile estimation under both traditional and regression-based norming. However, these sample size requirements were based on a simulation study and thus limited to the considered scenarios. Furthermore, no equations to derive the optimal design (i.e.

variance formula for the percentile estimator) were given in Oosterhuis et al. (2016). Oosterhuis et al. (2017) have derived variance formulas for several norm statistics under the traditional norming approach, but this approach requires larger sample sizes than the regression-based approach (Oosterhuis et al, 2016), and has the limitation that, without the assumption of a model, the optimal design cannot be derived. Another practical problem in norming, besides sample size calculation, is the fact that normative studies often derive norms for several outcome variables with the same sample. Van der Elst et al. (2017) have extended the regression-based approach to norming of several tests by proposing to use multivariate regression instead of fitting a univariate regression model for each test separately. However, in Van der Elst et al. (2017)'s multivariate regression-based approach, once the relevant predictors have been identified, and the parameters of the multivariate regression model have been estimated, each test is normed separately like in univariate regression-based norming. In other words, this approach targets the performance of an individual on each single test, and gives no guidelines on how to combine all norm statistic values obtained for an individual, one per test, in order to evaluate his/her overall performance across all administered tests.

In this thesis, the results available in the literature for univariate and multivariate regression-based norming are extended as follows: (i) a new multivariate regression-based approach is proposed that combines the several test scores obtained for an individual in the Mahalanobis distance, which is used as a measure of the overall performance across all administered tests, thus taking into account the correlation between these test scores, (ii) sampling variance formulas are derived for three types of norm statistics, namely Z-score, percentile rank score, and Mahalanobis distance, (iii) based on these variance formulas, optimal designs are derived for five regression models representing the reference population, (iv) efficient designs that are robust against misspecification of the regression model are provided, and (v) procedures are proposed to determine the required sample size for the optimal design of the normative sample such that individuals' positions relative to the derived norms can be assessed with pre-specified power and precision.

In the next section, the model-based approach adopted in this thesis is motivated, and the models considered in this work are briefly introduced.

1.3 Model-based approach

In this thesis a model-based approach is adopted, that is, the population under study is described by a statistical model. This is not the only possible approach to inference. As seen in the previous section, the traditional approach to norming did not involve models. Likewise, in the survey sampling literature the dominant paradigm is the design-based approach, in which inference is not based on a model for the outcome variable of interest (which is assumed to be a fixed unknown quantity in the design-based approach), but on the distribution of the inclusion indicator over repeated samples with a probability sampling design, where the inclusion indicator is a binary variable indicating whether the person is included into the sample or not. A model-free approach, such as the design-based approach to survey sampling and the traditional approach to norming, has the advantage of being robust in the sense of not being dependent on model assumptions. However, robustness comes at the price of efficiency (Little, 2004). Specifically, assuming a model allows to find the design that maximizes precision of estimation, that is, the optimal design. Furthermore, the assumption of a model facilitates the comparison between sampling schemes in survey sampling, and the identification of relevant predictors of the outcome variable of interest in normative studies. Since the main objective of this thesis is to improve the design of surveys and normative studies in order to obtain precise estimates, the model-based approach is adopted. Strategies to find robust designs (i.e. designs that are a trade-off between robustness and efficiency) will be discussed in section 1.5. In the next sections, the models considered in this thesis are presented.

1.3.1 Models for two-stage sampling

Suppose that a researcher is interested in estimating the mean alcohol consumption among students nested within schools. The outcome variable is quantitative (i.e. alcohol consumption) and measured at the individual (e.g. student) level. Students are sampled with two-stage sampling (i.e. first, schools are sampled, and then students within the selected schools are drawn), thus sampling error occurs at each design level. This is taken into account by assuming the following two-level random intercept model for the alcohol consumption y_{ij} of the i-th student from the j-th school

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}$$

where β_0 is the average of all school-specific means, u_j is the random effect of school j, and ε_{ij} is the residual (i.e. individual deviation from the school average) of student i nested within school j. The school and the student effects are assumed to be unrelated. The student residual ε_{ij} is normally distributed with zero mean and variance σ_ε^2. When cluster size is non-informative, the school effect u_j is also normally distributed with zero mean and variance σ_u^2. However, in this thesis school size is allowed to be informative. Specifically, a linear relation between the school effect u_j and the school size N_j is assumed, that is, $u_j = \gamma\left(N_j - \theta_N\right) + v_j$, where γ is the slope of this relation, θ_N is the mean school size in the population, and v_j is the random component of the school effect that does not depend on school size. This latter component v_j is assumed to be normally distributed with zero mean, and variance σ_v^2. Hence, the conditional distribution of the school effect u_j given the school size N_j is a normal distribution with mean $\gamma\left(N_j - \theta_N\right)$ and variance σ_v^2.

1.3.2 Models for norming

Univariate regression-based norming. Suppose that a researcher wants to derive norms for the oral version of the Letter Digit Substitution Test (LDST) for assessing information processing speed in adults aged 20 to 80 years. A representative sample is drawn from the reference population. The reference population can be modeled with a multiple linear regression model, as follows

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i = \boldsymbol{x}_i{}'\boldsymbol{\beta} + \varepsilon_i$$

where y_i is the test score of the i-th participant, $\boldsymbol{x}_i = [x_{1i}, \ldots, x_{ki}]'$ is a vector of the participant's scores on the k predictors (e.g. the participant's age and sex), $\boldsymbol{\beta} = [\beta_0, \beta_1 \ldots, \beta_k]'$ is the vector of regression coefficients of the predictors, and ε_i is the residual of the i-th participant that is assumed to follow a normal distribution with zero mean and variance σ_ε^2.

Multivariate regression-based norming. Suppose now that the researcher wants to compare subjects' oral and written performance on the LDST with the reference population. The oral and written versions of the LDST are normed using the same sample of participants. For participant *i*, the model for the test score on the oral version of the LDST is

$$y_{1i} = \beta_{01} + \beta_{11} x_{1i} + \ldots + \beta_{k1} x_{ki} + \varepsilon_{1i},$$

and that for the written version of the LDST is

$$y_{2i} = \beta_{02} + \beta_{12}x_{1i} + \ldots + \beta_{k2}x_{ki} + \varepsilon_{2i}.$$

To take into account the correlation between the test scores of the same individual (i.e. between y_{1i} and y_{2i}), these two models are combined in the following multivariate linear regression model

$$\boldsymbol{y}_i = \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{y}_i = [y_{1i}, y_{2i}]'$ is the vector with the two test scores of the i-th participant, $\boldsymbol{B} = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \vdots & \vdots \\ \beta_{k1} & \beta_{k2} \end{bmatrix}$ is the matrix of the regression coefficients for both test scores (note that each column contains the regression coefficients for one test score), $\boldsymbol{x}_i = [x_{1i}, \ldots, x_{ki}]'$ is the vector with predictor scores, and $\boldsymbol{\varepsilon}_i = [\varepsilon_{1i}, \varepsilon_{2i}]'$ is the vector of residuals of the i-th participant, which is assumed to follow a multivariate normal distribution with zero means vector and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$, where σ_1^2 is the variance of ε_{1i}, σ_2^2 is the variance of ε_{2i}, and $\sigma_{12} = \sigma_{21}$ is the covariance between ε_{1i} and ε_{2i}. Note that scores of the same participant can be correlated, while scores of different participants are unrelated

In the next section, the optimality criteria used to find the optimal sampling design for mean estimation respectively test norming in this thesis are presented.

1.4 Optimal design

As explained in the previous section, the first step to derive an optimal design is to specify a model for the population under study. The next step is to choose an optimality criterion that defines the optimal design. Since the aim is to maximize precision of estimation, the optimality criterion used in this thesis is the sampling variance (i.e. squared standard error) of the estimator of interest (e.g. mean estimator, Z-score estimator). The optimal design is thus defined as the design that minimizes this sampling variance. However, how this minimization is done depends on the type of study.

Optimal two-stage sampling for mean estimation in multilevel populations. The sampling variance $V(\hat{\mu})$ of the population mean estimator $\hat{\mu}$ is a decreasing function of the sample size at each design stage (i.e. number of clusters and number of individuals per cluster). Clearly, taking each of these sample sizes as large as possible minimizes the sampling variance $V(\hat{\mu})$ and thereby maximizes the precision of $\hat{\mu}$. However, resources (i.e. time and money) are limited in practice, and this can be taken into account by minimizing $V(\hat{\mu})$ subject to a cost constraint. Thus, the optimal design for mean estimation with a two-stage sampling scheme is defined as the combination of number of clusters and number of individuals per selected cluster that minimizes $V(\hat{\mu})$ subject to the cost constraint.

Optimal design for maximizing precision of norms estimation. A design ξ is defined as a joint distribution of the predictors in the normative sample, given the sample size. Under the assumption of a regression model like those given in section 1.3.2, the sampling variance of the norm statistic of interest (i.e. Z-score, percentile rank score, Mahalanobis distance) is a function of the distribution of the scores on the predictors (e.g. age and sex) in the normative sample. The optimal design is then defined as the joint distribution of the predictors' scores that minimizes the sampling variance of the norm statistic over the set of all possible joint distributions of the predictors' scores, that is, over the design region. The sampling variance of the norm statistic turns out to depend on the design ξ of the normative sample only through the so-called standardized prediction variance $d(\boldsymbol{X}, \xi)$. This is the variance of a predicted value for a subject, multiplied by the sample size and divided by the error variance. The optimal design can then be obtained simply by minimizing the standardized prediction variance.

The G-optimality criterion is an optimality criterion that targets the standardized prediction variance $d(\boldsymbol{X}, \xi)$. Specifically, the G-optimality criterion defines the optimal design as the design of the normative sample that minimizes the maximum of $d(\boldsymbol{X}, \xi)$, where the maximum is taken over all possible combinations of the predictor values (e.g. all possible combinations of age and sex) for which predictions can be done. This corresponds to minimizing the maximum of the sampling variance of the norm statistic over the design region. However, when the residuals $\boldsymbol{\varepsilon}$ have equal variance-covariance matrix $\boldsymbol{\Sigma}$, G-optimality is equivalent to D-optimality (Wong, 1995), which defines the optimal design as the design that minimizes the determinant of the variance-covariance matrix of the regression coefficients estimators (i.e. $\widehat{\boldsymbol{\beta}}$ for univariate norming, $\widehat{\boldsymbol{B}}$ for multivariate norming). For the multivariate regression-based approach, another optimality criterion is furthermore introduced. Under the

multivariate regression-based approach proposed by Van der Elst et al. (2017), multiple norm statistic values are obtained for each individual. Since this approach does not take into account the correlation between them, a suitable optimality criterion is A-optimality, that is, to minimize the sum of the sampling variances of the norm statistics for the same individual. Since these sampling variances are functions of the standardized prediction variance, also in this case the optimal design can be obtained with the G-optimality criterion.

In the next section, in order to cope with model uncertainty in the design stage, strategies to find robust designs are presented.

1.5 Maximin design

Two issues can arise with the optimal designs presented in the previous section. First, optimal designs for survey sampling can require prior knowledge of some model parameter values (e.g. variances or correlations), which is known as the local optimality problem in optimal design literature (Atkinson et al., 2007; Berger & Wong, 2009; Goos & Jones, 2011). Second, the optimal design of a normative study depends on the assumed model, but at the design phase of a study there is often uncertainty about the "true" model (e.g. whether the age effect on a test score is linear or not). Strategies to find designs that are robust against misspecification of the unknown parameter values or of the best model are presented in the next sections. The use of these robust designs (instead of the optimal design) in sample size calculation procedures allows to find a compromise between efficiency and robustness.

1.5.1 Local optimality problem

The local optimality problem occurs when the optimal design is optimal only for certain values of the model parameters. This is the case, for instance, for the optimal designs derived for the three aforementioned TSS schemes, because these optimal designs depend on the prior knowledge of some features of the cluster size distribution in the population, the degree of informativeness of cluster size, and the correlation between the outcomes for two individuals belonging to the same cluster. The approach adopted in this thesis to overcome this issue is the maximin approach (Atkinson et al., 2007; Berger & Wong, 2009; Wong, 1992). This approach has been applied in several contexts, such as longitudinal studies (Ouwens et al,., 2002; Tekle et al., 2008; Winkens et al., 2007), fMRI experiments (Maus et al., 2010), cluster randomized and multicentre trials (Candel & Van Breukelen, 2015; Van Breukelen & Candel, 2018; Wu et

al., 2017), cost-effectiveness studies (Manju et al., 2014; Manju et al., 2015), life-event studies (Safarkhani et al., 2014; Tan, 2010), test construction (Berger et al., 2000), and biological and pharmacological studies (Dette & Biedermann, 2003; Dette et al., 2006; King & Wong, 2000; Pronzato & Walter, 1988), because it has the advantage of being relatively simple to implement. An alternative approach to deal with the local optimality problem is the Bayesian approach, which assumes a prior distribution for the model parameters and derives the optimal design by averaging over the prior information (Abebe et al., 2015; Chaloner, 1984; Chaloner & Verdinelli, 1995; Dette, 1996; Goos et al., 2010; Han & Chaloner, 2004; Yu et al., 2008).

The local optimality problem is solved by the maximin approach as follows: First, a range of plausible values is defined for each unknown model parameter. Second, for each feasible design (i.e. combination of number of clusters and number of individuals per selected cluster), the values of the model parameters that minimize the efficiency $V(\hat{\mu})^{-1}$ are searched within the ranges of plausible values defined in the first step. Third and last, the design that maximizes the minimum efficiency obtained in the second step is chosen. That design is called the maximin design, and is the optimal design for the worst-case scenario. By maximizing the minimum efficiency over the ranges of plausible parameter values, the maximin design is robust against misspecification of the unknown parameters.

1.5.2 Model-dependence

A limitation of the optimal design for a normative study is that it depends on the assumed model, for instance on whether we assume the age effect on a test score to be linear or not, or to differ between males and females or not. This is important because, in practice, there is uncertainty about the "true" model. The maximin approach can also be applied to overcome this issue. In this setting, the first step is to choose a criterion for defining a design as the most robust design against misspecification of the model. Two possible choices are the efficiency criterion (i.e. the reciprocal of the sampling variance), and the relative efficiency (i.e. the ratio of the sampling variances of two competing designs). The second step is to define a set of plausible models, and to find the optimal design for each of these models. The third step is to find, for each design, the minimum efficiency value or the minimum relative efficiency value (i.e. relative to the optimal design for the model under consideration) across all plausible models. The most robust design is the design which maximizes the minimum efficiency or the minimum relative efficiency across all feasible designs. The resulting design is called the absolute maximin design

under the efficiency criterion, and the RE maximin design under the relative efficiency criterion.

1.6 Outline of the thesis

The chapters in this thesis can be read as self-contained articles. The notation can vary between chapters, but each chapter has a table summarizing the notation to increase readability. The topics of each chapter are as follows.

Chapter 2 deals with unbiased and efficient estimation of the average of all individual outcomes in a multilevel population for the sampling schemes SRS of individuals, TSS1, TSS2, and TSS3 when cluster size is informative. The unbiased (or approximately unbiased) estimator of the population mean is given for each sampling scheme. Furthermore, to establish which is the most efficient sampling scheme, the three TSS schemes are compared with each other and with SRS under the constraint of a fixed total sample size (i.e. number of individuals). These results are obtained under the assumption of a model for the multilevel population, that is, under the model-based approach. Since in survey sampling literature the dominant paradigm is the design-based approach, the two approaches are compared in the considered setting of estimating the average of all individual outcomes for each of the four considered sampling schemes when cluster size is informative.

Chapter 3 extends the results obtained in chapter 2 by deriving the optimal design for TSS1, TSS2 and TSS3, that is, the number of clusters and number of individuals per selected cluster that minimizes the sampling variance of the population mean estimator subject to a budget constraint instead of a constraint on the total sample size. Relatedly, the effects of ignoring informative cluster size on the optimal design are investigated. Furthermore, under the constraint of a fixed budget for sampling and measuring, the optimal designs for the three TSS schemes are compared in terms of efficiency with each other and with SRS. Since the optimal design depends on unknown model parameters, maximin designs are derived to overcome this dependency. Finally, a procedure is proposed for computing maximin sample sizes to compare the means of two populations.

Chapter 4 deals with sample size calculation and optimal design for univariate regression-based norming. Sampling variance formulas are derived for two norm statistics often used in practice: the Z-score and percentile rank score. Based on these variance formulas,

optimal designs are obtained for five regression models with a quantitative and a qualitative predictor, differing in whether they allow for interaction and nonlinearity. Since at the design phase of a study there is uncertainty about the "true" norming model, designs robust against misspecification of the model are derived based on two criteria: efficiency, and relative efficiency. Also a procedure is proposed to determine the required sample size for the optimal design of the normative sample.

In chapter 5, a new approach to multivariate regression-based norming is proposed. This approach differs from that available in the literature in that it combines all the scores obtained for an individual through the Mahalanobis distance, instead of providing separate norms for each outcome variable. Sampling variance formulas are derived for the two norm statistics used in the two multivariate regression-based approaches, that is, the Z-score and Mahalanobis distance. Furthermore, for both multivariate regression-based approaches optimal designs are derived for the multivariate version of the five regression models considered in chapter 4. To deal with the uncertainty about the "true" model, also robust designs are obtained based on the efficiency and the relative efficiency criteria. Finally, a sample size calculation procedure is proposed only for the Mahalanobis distance-based approach, because it is not hampered by multiple testing issues.

In chapter 6, some practical guidelines for planning surveys and normative studies are given, and ideas for future research are outlined. Each chapter of the thesis is summarized in chapter 7, and the scientific and social impact of the results of this thesis are discussed in chapter 8. For the sake of brevity, online supplementary materials for chapters 2, 3, 4, and 5 are not included into this thesis but are available upon request. A table of contents of the online supplementary materials can be found at the end of each chapter.

Chapter 2

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

This chapter has been published in *Statistics in Medicine* (2019), 38: 1817-1834, with co-authors Math J.J.M. Candel, Frans E.S. Tan, and Gerard J.P. van Breukelen

Abstract

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

Keywords: design-based inference; hierarchical population; informative cluster size; model-based inference; two-stage sampling

2.1 Introduction

Hierarchical or multilevel populations arise when individuals or micro-units are nested within clusters or macro-units (Goldstein, 2011; Snijders & Bosker, 2012). Considering, for the sake of simplicity, only populations with two levels of nesting, examples include patients clustered in general practices, elderly people nested in nursing homes, and students grouped in schools. In these populations the overall mean of an outcome variable (e.g. cholesterol level, blood pressure, body mass index) can be defined in two ways: as the mean of all individuals in the population ignoring cluster membership (i.e. first, pooling all patients from all clusters in the population, and then computing the average cholesterol level), or as the mean of all cluster-specific means (i.e. first, computing the mean cholesterol level within each cluster, and then averaging all the cluster-specific means). These two definitions coincide only under special conditions, as will be seen later, but this paper focuses on the first definition only. Related to these two definitions, is the concept of informative cluster size.

When clusters vary in size in the population (e.g. small versus large general practices), cluster sizes can be seen as realizations of a random variable (Van Breukelen et al., 2007), and the outcome variable of interest may be related to cluster size (e.g. surgeons operating on many patients might have better performances than those operating on fewer patients, see Panageas et al., 2007). If this is the case, then cluster size is said to be informative (Seaman et al., 2014). Nevalainen et al. (2014) describe and give practical examples of three data-generating mechanisms that can lead to informative cluster size. Briefly, a latent variable (e.g. the competence of the surgeon) influences cluster size (e.g. the number of patients) and the outcome variable (e.g. success of the operation) at the same time; or cluster size affects the outcome variable (e.g. surgeons become better by practice); or vice versa, the outcome variable affects cluster size (e.g. better surgeons get more referrals). Relatedly, Seaman et al. (2014) point out that the standard methods to analyse clustered data, namely Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equations (GEE), implicitly assume that cluster size is unrelated to the outcome variable, and discuss different methods to handle informative cluster size for cluster-specific inference with GLMM and population-average inference with GEE.

The topic of this paper is the unbiased and efficient estimation of the population mean in the presence of informative cluster size. To estimate the population mean, individuals can be sampled either with Simple Random Sampling (SRS), that is, directly from the population, or with Two-Stage Sampling (TSS), that is, sampling first clusters and then individuals within the

sampled clusters (Cochran, 1977; Lohr, 2010; Särndal et al., 1992). Given cluster size variation in the population, at least three alternative TSS schemes can be considered:

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

In order to evaluate each sampling scheme in terms of unbiasedness and efficiency of mean estimation, it is useful to distinguish two approaches to inference in survey sampling literature (Skinner & Wakefield, 2017): the design-based paradigm (Cochran, 1977; Lohr, 2010; Särndal et al., 1992), and the model-based approach (Chambers & Clark, 2012; Little, 2004; Valliant et al., 2000). In the design-based approach, the outcome value for each unit (e.g. patient) in the population is assumed to be a fixed unknown quantity. The random variable is then the *inclusion indicator*, that is, the variable that states whether or not a unit is included into the sample. Thus, inference is based on the distribution of the inclusion indicator over repeated samples with a probability sampling design. In contrast, the model-based approach assumes that the outcome value in the real finite population is a realization of a stochastic model, representing a hypothetical infinite population. Inference is then based on the probabilistic model. As long as the assumptions of the model are met, model-based inference can then ignore the sampling scheme and condition on the observed sample (Chambers & Clark, 2012; Little, 2004; Lohr, 2010; Skinner & Wakefield, 2017). However, if the model residuals (i.e. the stochastic part) are correlated with the variables which determine the sampling probabilities (and then with the sampling probabilities themselves), the sampling design is said to be informative (Little, 2004; Pfeffermann, 1993; Pfeffermann et al., 1998; Skinner & Wakefield, 2017; Snijders & Bosker, 2012, p. 222; Sudgen & Smith, 1984). When this is the case, model-based analysis is biased, unless the sampling design is taken into account (Snijders & Bosker, 2012, p. 237). In the multilevel modelling literature, many authors have investigated unbiased estimation when two-stage sampling with unequal sampling probabilities is informative, but they assumed non-informative cluster size (Asparouhov, 2006; Grilli & Pratesi, 2004; Koziol et al., 2017; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006). In this paper this sampling scheme is informative due to the cluster size being informative.

In this paper, cluster size is treated as a random variable and assumed to be informative, but the special case of non-informative cluster size will also be covered briefly. Furthermore, a simple hierarchical linear model (Goldstein, 2011; Snijders & Bosker, 2012), for the outcome variable in the population, is assumed and used to define the parameter of interest (i.e. the population mean). We thus adopt a model-based approach but will also make a comparison with design-based inference. It will be shown that the type of analysis (i.e. unweighted versus weighted analysis) needed for unbiased estimation of the population mean depends on the chosen sampling scheme. Furthermore, the three aforementioned TSS schemes will be compared with each other and with SRS in terms of their efficiency under the constraint of a fixed total sample size. It will also be shown that their relative efficiencies depend on the cluster size distribution.

The rest of the paper is organized as follows. In section 2.2, the assumptions on which our findings are based and the considered sampling schemes are presented in more detail. In section 2.3, the population mean is derived under a linear mixed model for a two-level hierarchical population with varying and informative cluster size. Furthermore, section 2.3 deals with the estimation of the population mean under different sampling schemes, presenting both the expectation and sampling variance of the estimator under each scheme. In section 2.4, the three TSS schemes are compared with each other and with SRS in terms of efficiency for a given total sample size (number of individuals). In section 2.5, the relative efficiencies of the three TSS schemes are derived under the design-based approach, and compared with those obtained under the model-based framework. The results of this paper are applied in section 2.6 to two real populations: high schools in Italy, and general practices in England. Some final remarks are offered in section 2.7. The online Supplementary Material contains part of the derivations of the equations given in this paper, as well as additional tables and figures.

2.2 Assumptions and sampling schemes

The structure of the data is hierarchical with two levels of nesting (e.g. pupils are nested within schools, patients within general practitioners (GPs)).The results of this paper are based on the following assumptions (the notation is summarized in Table 2.A in appendix C):

Assumption 1: The population is composed of K clusters (e.g. schools, GPs) and each cluster j contains N_j individuals (e.g. students, patients), that is, clusters are allowed to have different

sizes. The total number of individuals in the population (i.e. the population size) is $N_{pop} = \sum_{j=1}^{K} N_j$.

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

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

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij} \tag{2.1}$$

where $u_j | N_j \sim N\left(\gamma\left(N_j - \theta_N\right), \sigma_v^2\right)$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $u_j \perp \varepsilon_{ij}$, and γ and σ_v^2 will be defined in the next assumption. Note that multilevel models, such as equation (2.1), are not only a standard procedure for modelling hierarchical populations (Goldstein, 2011; Snijders & Bosker, 2012), but also a natural way for taking into account the clustering induced by Two-Stage Sampling in a model-based approach (Chambers & Clark, 2012, p. 65; Goldstein, 2011, pp. 212-213; Little, 2004; Lohr, 2010, pp. 200, 262-264; Makela et al., 2018; Skinner & Wakefield, 2017; Snijders & Bosker, 2012, pp. 218, 223; Valliant et al., 2000, p. 256; Zheng & Little, 2004).

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

In order to deal with cluster size variation and informative cluster size in estimating the population mean (i.e. the average of all individual outcomes), three competing TSS schemes

are considered, which will be compared with SRS of individuals and with each other, under the constraint that all sampling schemes have the same total sample size.

Two-Stage Sampling 1 (TSS1):

Stage 1: Sample k clusters with probability proportional to cluster size N_j, that is, $\frac{N_j}{\sum_{j=1}^{K} N_j}$ is the probability of cluster *j* being sampled if one cluster is randomly sampled, and so the inclusion probability for the *j*-th cluster, that is, the probability that cluster *j* is sampled given a total of k sampled clusters, is $\pi_j = 1 - \left(1 - \frac{N_j}{\sum_{j=1}^{K} N_j}\right)^k$ (Särndal et al., 1992, p. 51). If $\frac{N_j}{\sum_{j=1}^{K} N_j} \to 0, \ \forall\, j = 1, \dots, K$, then $\pi_j \approx \frac{kN_j}{\sum_{j=1}^{K} N_j}$; this approximation will be used.

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

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

Two-Stage Sampling 2 (TSS2):

Stage 1: Sample k clusters with Simple Random Sampling (SRS), that is, $\pi_j = \frac{k}{K}, \ \forall j = 1, \dots, K$.

Stage 2: Sample the same percentage of individuals per cluster p, that is, the number of individuals sampled per cluster (i.e. n_j) is proportional to the cluster size in the population (i.e. N_j), and so $\pi_{i|j} = \frac{n_j}{N_j} = p \ \ \forall i = 1, \dots, N_j$ and $\forall j = 1, \dots, K$.

Under this sampling scheme the unconditional probability of being included into the sample is the same for all individuals, that is, $\pi_{ij} = \pi_j \pi_{i|j} = \frac{k}{K} \frac{n_j}{N_j} = \frac{k}{K} p$. In contrast to what was the case for TSS1, we now need to know only the cluster sizes for the sampled clusters before sampling individuals from those sampled clusters.

Two-Stage Sampling 3 (TSS3):

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

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

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

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

2.3 Definition and estimation of the population mean μ

To find the population mean $E(Y_{ij})$ and variance $V(Y_{ij})$, defined from model (2.1) as the marginal expectation and variance of Y_{ij} over cluster effect u_j and individual effect ε_{ij}, the marginal expectation and variance of cluster effect u_j (i.e. $E(u_j)$ and $V(u_j)$, respectively) are needed. If cluster size is non-informative (i.e. $\gamma = 0$ in assumption 4), then $E(u_j) = 0$ and $V(u_j) = \sigma_v^2$ leading to $E(Y_{ij}) = \beta_0$ and $V(Y_{ij}) = \sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2$. In contrast, if cluster size is informative (i.e. $\gamma \neq 0$ in assumption 4), $E(u_j) = 0$ or $E(u_j) \neq 0$ depending on the sampling scheme. To prevent misunderstanding, note that the cluster effect u_j in the population does not depend on the sampling design, and its marginal distribution in the population is $f(u_j) = \int f(u_j|N_j) f(N_j)\, dN_j$ (where $f(.)$ indicates a probability density function). Nevertheless, the sampling design determines the cluster effect sampling distribution, which is, for a sample of

size one, equal to $\int f(u_j|N_j)f(N_j)\,dN_j$ if clusters are sampled with equal probabilities, and equal to $\int \left(\frac{N_j}{\theta_N}\right) f(u_j|N_j)f(N_j)\,dN_j$, if clusters are sampled with probabilities proportional to their size.

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

$$\text{(a)}\ \ E_{\text{TSS2/TSS3}}(u_j) = 0, \quad \text{and} \quad \text{(b)}\ V_{\text{TSS2/TSS3}}(u_j) = \sigma_v^2 + \gamma^2\sigma_N^2 = \sigma_u^2. \qquad (2.2)$$

Note that $\gamma^2\sigma_N^2$ is the component of $V_{\text{TSS2/TSS3}}(u_j)$ explained by N_j, and σ_v^2 is the unexplained variance of u_j. Hence, the following expression for $E(Y_{ij})$ comes from model (2.1) and equation (2.2.a)

$$E_{\text{TSS2/TSS3}}(Y_{ij}) = \beta_0, \qquad (2.3)$$

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

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

$$\text{(a)}\ E_{\text{SRS/TSS1}}(u_j) = \gamma\theta_N\tau_N^2, \text{ and (b)}\ V_{\text{SRS/TSS1}}(u_j) = \sigma_v^2 + \gamma^2\sigma_N^2[\tau_N(\zeta_N - \tau_N) + 1] \qquad (2.4)$$

where $\tau_N = \frac{\sigma_N}{\theta_N}$ and $\zeta_N = \frac{E\left[\left(N_j-\theta_N\right)^3\right]}{\sigma_N^3}$ are the coefficient of variation and the skewness of cluster size distribution in the population, respectively. Note that $V_{\mathrm{SRS/TSS1}}(u_j) = V_{\mathrm{TSS2/TSS3}}(u_j)$ if one of the following conditions holds: (i) $\tau_N = 0$ (i.e. no cluster size variation), (ii) $\gamma = 0$ (i.e. cluster size is non-informative), (iii) $\zeta_N = \tau_N$ (e.g. N_j is Poisson distributed, see Table S.M.1 in the Supplementary Material). Likewise, $E_{\mathrm{SRS/TSS1}}(u_j) = E_{\mathrm{TSS2/TSS3}}(u_j)$ if either condition (i) or (ii) holds. Thus, from model (2.1) and equation (2.4.a) the population mean that we here want to estimate follows

$$E_{\mathrm{SRS/TSS1}}(Y_{ij}) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu. \tag{2.5}$$

This mean can be interpreted as the expected outcome for an individual randomly sampled from the population ignoring cluster membership by SRS. Note that the two definitions of $E(Y_{ij})$ in equations (2.3) and (2.5) coincide if either clusters have the same size in the population (i.e. $\tau_N = 0$) or cluster size is not related to the outcome (i.e. $\gamma = 0$). Given the focus of this paper on μ, model (2.1) can be rewritten from equation (2.5) as follows

$$y_{ij} = \mu + b_j + \varepsilon_{ij}, \tag{2.6}$$

where $b_j = u_j - \gamma\theta_N\tau_N^2 = u_j - E_{\mathrm{SRS/TSS1}}(u_j)$ (see equation (2.4.a)) with $E_{\mathrm{SRS/TSS1}}(b_j) = 0$ and $V_{\mathrm{SRS/TSS1}}(b_j) = V_{\mathrm{SRS/TSS1}}(u_j)$ (see equation (2.4.b)).

To estimate μ unbiasedly, the weight of a cluster should be proportional to its size, either in the sampling scheme or in the estimator (for details, see appendices A and B). For each sampling scheme, the first row of Table 2.1 presents the unbiased or approximately unbiased (i.e. for k sufficiently large) estimator of μ under model (2.6), the second and third row present the conditional expectation and variance of $\hat{\mu}$, the fourth row gives the marginal expectation of $\hat{\mu}$, and the last two rows show the two components of the marginal variance of $\hat{\mu}$ (i.e. $Var(\hat{\mu}) = E\left(V(\hat{\mu}|\boldsymbol{N}_*)\right) + V\left(E(\hat{\mu}|\boldsymbol{N}_*)\right)$, where $\boldsymbol{N}_* = \boldsymbol{N} = (N_1, \dots, N_k)^T$ under TSS and $\boldsymbol{N}_* = \boldsymbol{N}_{SRS} = \left(N_1, \dots, N_{k_{SRS}}\right)^T$ under SRS) (for proofs, see appendix B). As the first row of Table 2.1 shows, the estimator of μ is a weighted sum of cluster means in each sampling scheme, but the weights differ between schemes. Under SRS k_{SRS} clusters are indirectly sampled from the population and large clusters have higher chance of being sampled, thus the unweighted estimator is unbiased for μ (recall that from assumption 2, $\frac{m}{N_{pop}} \to 0$ which implies that $k_{SRS} \to m$). Under

TSS1 clusters are sampled with probabilities proportional to their size, and so μ is estimated unbiasedly by the unweighted average of cluster means. Under TSS3 and TSS2 cluster means must be weighted by cluster size (i.e. N_j in TSS3, and also in TSS2 since $n_j = pN_j$) in the analysis, because clusters are weighted equally by these sampling designs, that is, all clusters have equal sampling probability (for details, see appendix B). An exception to this is the special case of non-informative cluster size (i.e. $\gamma = 0$), in which the two definitions of population means coincide (i.e. $\mu = \beta_0$). It then follows that $E(u_j) = 0$ for any sampling scheme (see appendix A), and from model (2.1), then results that $E(\bar{y}_j) = \beta_0$. Thus, any estimator of $\mu = \beta_0$ of the form $\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{y}_j}{\sum_{j=1}^{k} w_j}$ is unbiased then, although some weights w_j are more efficient than others (Searle & Pukelsheim, 1986; Van Breukelen & Candel, 2012).

2.4 Relative efficiencies of TSS schemes versus SRS and each other

Under the constraint of a fixed total sample size (i.e. $m = \bar{n}k$), the efficiency of the three TSS schemes can be investigated by computing their *relative efficiencies*, defined as the ratio of the sampling variances of $\hat{\mu}$ under two competing sampling schemes (i.e. the variances obtained as the sum of the last two rows of Table 2.1). For instance, the relative efficiency of TSS1 versus SRS is defined as the ratio of $V(\hat{\mu})$ for SRS to $V(\hat{\mu})$ for TSS1 (i.e. $RE(TSS1\ vs\ SRS) = V(\hat{\mu}_{SRS})/V(\hat{\mu}_{TSS1})$). The relative efficiencies are given in Table 2.2 (for proof, see section 2 of the Supplementary Material), while the relative efficiency of TSS2 versus TSS1 is plotted in Figure 2.1. As shown by Table 2.2, the numerator and denominator of the relative efficiency are both a weighted sum of two components, respectively $E(V(\hat{\mu}|\boldsymbol{N}))$ and $V(E(\hat{\mu}|\boldsymbol{N}))$ from last two rows of Table 2.1, with weights determined by the correlation between cluster effect and cluster size $corr(u_j, N_j)$. The component $E(V(\hat{\mu}|\boldsymbol{N}))$ with weight $\left(1 - corr(u_j, N_j)^2\right)$ depends on the intraclass correlation $\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2}$, the coefficient of variation of cluster size τ_N, and the average number of individuals sampled per cluster $\bar{n}$. The other component, $V(E(\hat{\mu}|\boldsymbol{N}))$, weighted by $corr(u_j, N_j)^2$, is a function of the coefficient of variation τ_N, the skewness ζ_N, and (for TSS2 and TSS3 only) the kurtosis η_N of cluster size distribution. Denote by ω the relative efficiency under non-informative cluster size (i.e. $RE = \omega$ if $corr(u_j, N_j) = 0$), and by λ the relative efficiency under a perfect linear relation between u_j and N_j (i.e. $RE = \lambda$ if $corr(u_j, N_j)^2 = 1$). These two extremes can be derived directly from Table 2.2 and Figure

2.1, which plots the RE against $corr(u_j, N_j)$. Therefore, the RE moves from ω to λ as $corr(u_j, N_j)$ moves from zero to one. For small to moderate correlations (say, $|corr(u_j, N_j)| < 0.7$), ω receives more weight in the relative efficiency. If ω and λ are both smaller than or equal to one, the relative efficiency is also smaller than or equal to one. Now, the ω's shown in Table 2.2 are all smaller than one, which entails the following ordering of the sampling schemes in terms of efficiency based on ω (from most to least efficient): SRS, TSS1, TSS2, and TSS3. Under a perfect linear relation between cluster effect and cluster size (i.e. $corr(u_j, N_j)^2 = 1$), $RE = \lambda$ and SRS is more efficient than TSS1, while TSS2 and TSS3 are equally efficient. Furthermore, TSS1 is more efficient than TSS2 and TSS3 (i.e. $\lambda \leq 1$) if one of the following conditions is met (for proofs, see section 2 of the Supplementary Material): the cluster size distribution is positively skewed (i.e. $\zeta_N > 0$) with $\tau_N \in [0, \zeta_N]$, or is symmetric (i.e. $\zeta_N = 0$) with $\tau_N \in [0,1]$ and $k \in \left[1, \frac{(2-\tau_N^2)+\sqrt{2-\tau_N^2}}{(1-\tau_N^2)}\right]$, or is Normal. Thus, for any value of $corr(u_j, N_j)$ the ordering of the sampling schemes in terms of efficiency based on $V(\hat{\mu})$ is (from most to least efficient): SRS, TSS1, TSS2, and TSS3. However, if none of the aforementioned conditions is met, λ might be bigger than one and then, to see whether TSS1 is more efficient than TSS2 and TSS3, the relative efficiency must be computed for the specific cluster size distribution.

Given that $RE = \omega$ if $corr(u_j, N_j) = 0$ and ω has more weight than λ in the RE for $|corr(u_j, N_j)| < 0.7$, it is useful to have a closer look at the patterns of the ω's shown in Table 2.2. First, the ω of any TSS scheme versus SRS is a decreasing function of the intraclass correlation ρ, the average number of individuals sampled per cluster $\bar{n}$, and (only for TSS2 and TSS3) of the coefficient of variation of cluster size τ_N. Second, $\omega(TSS2vsTSS1)$, $\omega(TSS3vsTSS1)$, and $\omega(TSS3vsTSS2)$ are decreasing functions of the coefficient of variation of cluster size τ_N. Third, as the intraclass correlation ρ and/or the average number of individuals sampled per cluster $\bar{n}$ increase, TSS2 moves away from TSS1 and towards TSS3 in terms of efficiency as expressed by ω (see Figure 2.2).

When the outcome variable is unrelated to the cluster size (i.e. $\gamma = 0$ and so also $corr(u_j, N_j) = 0$), the population mean μ is equal to β_0 as shown in section 2.3. In this special case, any estimator of μ of the form $\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{y}_j}{\sum_{j=1}^{k} w_j}$ is unbiased. However, some weights are more

efficient than others. For TSS2, weighting cluster means by their inverse variance (i.e. $w_j = Var(\bar{y}_j)^{-1} = \left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n_j}\right)^{-1}$, where $\sigma_u^2 = \sigma_v^2$ since $\gamma = 0$) is optimal, and unweighted analysis (i.e. $w_j = 1$) is more or less efficient than cluster size weighting (i.e. $w_j = pN_j$), depending on the intraclass correlation ρ and the average cluster size in the sample (Searle & Pukelsheim, 1986; Van Breukelen et al., 2007). The conditional variance of the optimal estimator is $Var\left(\frac{\sum_{j=1}^{k} \bar{y}_j Var(\bar{y}_j)^{-1}}{\sum_{j=1}^{k} Var(\bar{y}_j)^{-1}} \middle| \boldsymbol{N}\right) = \left(\sum_{j=1}^{k} \frac{pN_j}{pN_j\sigma_u^2 + \sigma_\varepsilon^2}\right)^{-1}$ (Van Breukelen et al., 2007, eq. (6)). Under TSS1 and TSS3, the same number of individuals is sampled per cluster (i.e. $n_j = n, \forall j = 1, \ldots, k$), so the estimator with $w_j = Var(\bar{y}_j)^{-1}$ reduces to $\frac{\sum_{j=1}^{k} \bar{y}_j}{k}$. Thus, for TSS1 and TSS3, $w_j = 1$ is optimal and its sampling variance is given in the fifth row of the TSS1 column in Table 2.1 (for proof, see appendix B or section 2.3 of the Supplementary Material), so TSS1 and TSS3 are equally efficient then, given equal weighting of cluster means, but TSS3 is more practical because, unlike TSS1, it does not require the knowledge of all cluster sizes in the population. The optimal estimator of TSS2 is less efficient than that of TSS3 and TSS1 (i.e. $RE\left(\frac{\sum_{j=1}^{k} \bar{y}_j Var(\bar{y}_j)^{-1}}{\sum_{j=1}^{k} Var(\bar{y}_j)^{-1}} vs \frac{\sum_{j=1}^{k} \bar{y}_j}{k}\right) \leq 1$, for proof see section 2.3 of the Supplementary Material). Therefore, TSS3 combined with $\frac{\sum_{j=1}^{k} \bar{y}_j}{k}$ is the best strategy to estimate μ if cluster size is not informative. To prevent misunderstanding, note that the ordering of sampling schemes in this last paragraph only holds if non-informative cluster size is combined with optimal weighting of cluster means. Those weights differ from the ones in Table 2.1 first row, on which Table 2.2 and Figures 2.1 and 2.2 are based, and which are needed for unbiased estimation of the population mean if cluster size is informative.

Table 2.1. Estimators of the population mean $\mu = \beta_0 + \gamma\theta_N\tau_N^2$: conditional and marginal expectations and variances.

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

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

Table 2.2. Relative efficiencies of TSS schemes versus SRS and each other.

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

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

Relative efficiency of TSS2 versus TSS1 under the model-based approach

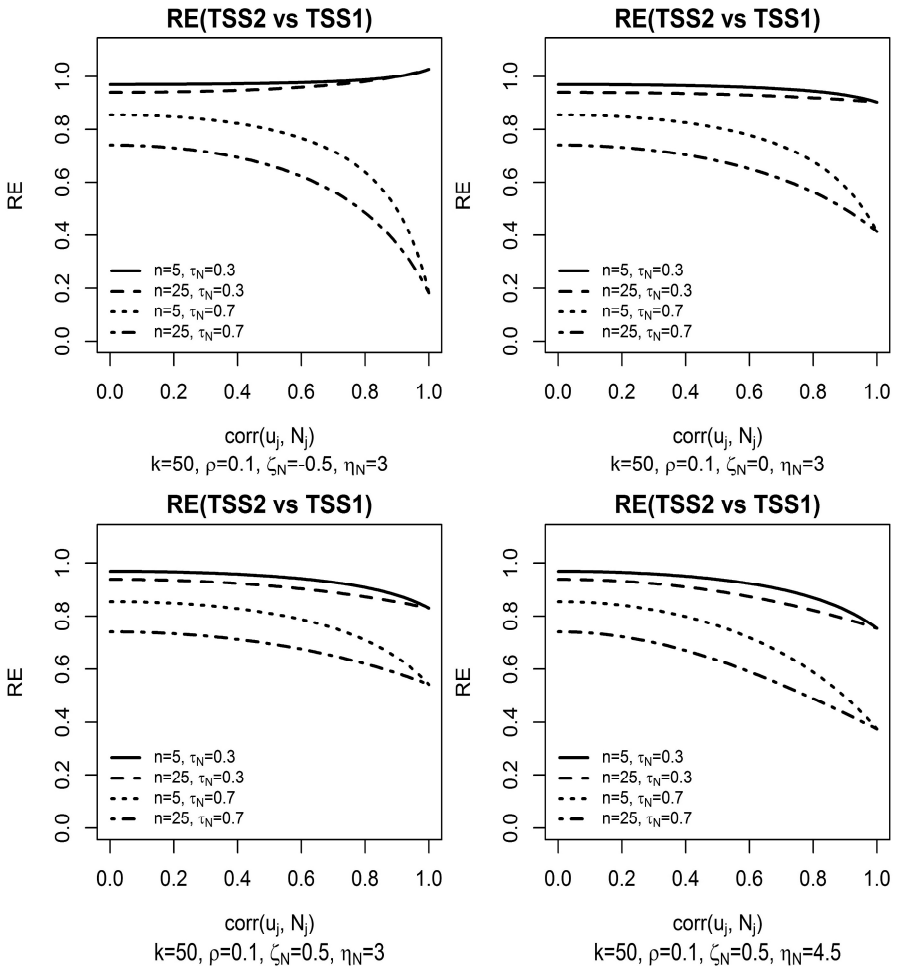


Figure 2.1. Model-based relative efficiency of TSS2 versus TSS1, for a given total sample size $\bar{n}k$, as a function of the (absolute value of the) correlation between cluster effect and cluster size (i.e. $corr(u_j, N_j)$), for different values of the average number of individuals sampled per cluster (i.e. $\bar{n}$) and of the coefficient of variation of cluster size (i.e. τ_N) (curves), and different cluster size distributions (panels). The values of the relative efficiency at $corr(u_j, N_j) = 0$ and $corr(u_j, N_j) = 1$ refer to ω and λ, respectively.

Relative efficiency of TSS3 versus TSS2 under the model-based approach and non-informative cluster size

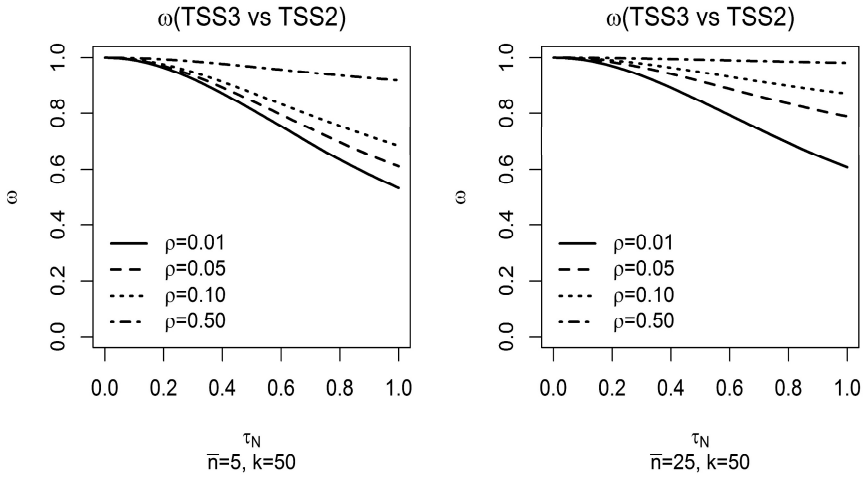


Figure 2.2. Model-based relative efficiencies of TSS3 versus TSS2, for a given total sample size $\bar{n}k$ and non-informative cluster size (i.e. $\gamma = 0$), as a function of the coefficient of variation of cluster size (i.e. τ_N), for different values of the intraclass correlation (i.e. ρ) (curves) and for different average numbers of individuals sampled per cluster (i.e. $\bar{n}$) (panels).

2.5 Design-based inference for two-stage sampling when cluster size is informative

The aim of this section is to study the relative efficiencies of the three TSS schemes compared with SRS and with each other under the design-based approach. It is important to emphasize that the inferential framework of this section is different from the model-based approach adopted in the rest of the paper. So far, the outcome variable Y_{ij} and cluster size N_j were both seen as random variables, and inference was based on the probability distribution of Y_{ij} given in model (2.1). In contrast, in the design-based approach (i.e. this section), the outcome variable Y_{ij} and cluster size N_j are fixed quantities, the inclusion indicator is the only random variable (e.g. for cluster j, it is defined as $I_j = 1$ if cluster j is included into the sample, which occurs with probability π_j, and $I_j = 0$ otherwise), and inference is based on the probability distribution induced by the sampling scheme.

The notation of this section remains the same as before with the important distinction that all population quantities here must be interpreted as relating to the finite population. Thus,

the two types of population means can be expressed as $\mu = \frac{\sum_{j=1}^{K} N_j \bar{Y}_j}{\sum_{j=1}^{K} N_j}$ and $\beta_0 = \frac{\sum_{j=1}^{K} \bar{Y}_j}{K}$, respectively, where $\bar{Y}_j$ is the mean of all N_j individuals within cluster j. Furthermore, in the population the outcome variable for the i-th individual within the j-th cluster can be decomposed (combining model (2.1) with assumption 4) as follows

$$Y_{ij} = \beta_0 + \gamma\left(N_j - \theta_N\right) + \nu_j + \varepsilon_{ij}, \tag{2.7}$$

where ν_j is the cluster effect with $E\left(\nu_j\right) = \frac{\sum_{j=1}^{K} \nu_j}{K} = 0$ and $V\left(\nu_j\right) = \frac{\sum_{j=1}^{K} \nu_j^2}{K} = \sigma_\nu^2$, and $\nu_j \perp N_j$, while ε_{ij} is the individual effect with $E\left(\varepsilon_{ij}\right) = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}}{N_j} = 0, V\left(\varepsilon_{ij}\right) = \frac{\sum_{i=1}^{N_j} \varepsilon_{ij}^2}{N_j} = \sigma_\varepsilon^2$, and $\nu_j \perp \varepsilon_{ij}$, which entails that $\bar{Y}_j$ here represents $\beta_0 + u_j$ in model (2.1). Note that, in this section, no distributional assumptions are made for equation (2.7), all quantities (i.e. Y_{ij}, N_j, ν_j, and ε_{ij}) are just fixed constants, the only random variable is the inclusion indicator and its probability distribution is the foundation of inference. From equation (2.7), it follows that $\mu = \beta_0 + \gamma\theta_N\tau_N^2$, an expression that is similar to equation (2.5) but refers to the finite population (for proof, see section 3 of the Supplementary Material). Hence, under both inferential paradigms, the two population means coincide (i.e. $\mu = \beta_0$) only if either there is no cluster size variation in the population (i.e. $\tau_N = 0$), or cluster size is non-informative (i.e. $\gamma = 0$).

For each sampling scheme, Table 2.3 shows in the first row the estimator of the population mean μ, in the second row the sampling variance of $\hat{\mu}$ as available in the design-based literature (Cochran, 1977; Lohr, 2010; Särndal et al., 1992; Sukhatme, 1954), and in the third row again the sampling variance of $\hat{\mu}$ but under the assumption that equation (2.7) describes the outcome variable Y_{ij} in the population (for proofs, see section 3 of the Supplementary Material). For large enough k (say, $k \geq 30$), the model-based variances given in Table 2.1 are equal to the design-based variances given in the third row of Table 2.3. Furthermore, the estimators of Table 2.3 are the same as those of the model-based approach (i.e. Table 2.1, first row). The estimators under SRS and TSS1 are unbiased (Cochran, 1977, p. 308; Lohr, 2010, p. 236), while the estimator under TSS2 and TSS3, the so-called ratio estimator, is only approximately unbiased (Lohr, 2010, p. 186; Sukhatme, 1954, pp. 323-324) and then the number of sampled clusters k is assumed to be large enough to neglect this bias. It is important to emphasize that, under the design-based paradigm, the properties of an estimator (i.e. approximate unbiasedness, variance as given in the second row of Table 2.3) are

based only on the sampling scheme (Lohr, 2010, p. 147; Särndal et al., 1992, p. 239). The assumption that the outcome variable is described by equation (2.7) (i.e. Table 2.3, third row) is needed to allow a fair comparison with the results obtained under the model-based approach. However, the assumption of a model, like equation (2.7), to evaluate competing sampling schemes is appropriate under the design-based framework, provided that inference is then based on the sampling scheme only (Cochran, 1977, p. 256; Hansen et al., 1983; Lohr, 2010, p. 205; Smith, 1994).

Similarly to section 2.4, the relative efficiency of two competing sampling schemes is defined as the ratio of their variances (as given in the third row of Table 2.3). For large enough k (say, $k \geq 30$), it turns out that these relative efficiencies (given in Table S.M.2 and shown in Figures S.M.1-2 of the Supplementary Material) are approximately equal to those shown in Table 2.2 because the variances in Table 2.1 and those in the third row of Table 2.3 are approximately equal. The only distinction to be made is that $corr(u_j, N_j)$ is replaced with the correlation between cluster mean and cluster size $corr(\bar{Y}_j, N_j)$. Like in section 2.4, numerator and denominator of the relative efficiency are both made up of two components, weighted by $corr(\bar{Y}_j, N_j)^2$ and $\left(1 - corr(\bar{Y}_j, N_j)^2\right)$, respectively, and only the component weighted by $corr(\bar{Y}_j, N_j)^2$ depends on the skewness and kurtosis of the cluster size distribution. The extreme cases of the relative efficiency, namely under non-informative cluster size and a perfect relation between cluster mean and cluster size, are denoted by ω and λ, respectively. The patterns and the ordering of the relative efficiencies are then those of section 2.4. Specifically, for any value of $corr(\bar{Y}_j, N_j)$, SRS is the most efficient sampling scheme, followed by TSS1 (under the conditions given in section 2.4), TSS2, and finally TSS3.

To conclude, even though the mathematical foundations of the two inferential approaches are different, in the considered setting, they yield almost the same results: the population mean estimators are the same, as well as the relative efficiencies, provided that k is large enough and equation (2.7) holds in the population. An advantage of the design-based approach is robustness because the unbiasedness and the variance of a design-based estimator do not depend on the assumptions of a model. Nevertheless, the model-based approach has a practical advantage when designing a survey, more specifically for choosing a sampling scheme and computing the sample size. The sampling variances in Table 2.1 (last two rows) and Table 2.3 (last row), and the relative efficiencies in Table 2.2, all based on equation (2.7), can be

obtained by specifying the intraclass correlation ρ, the correlation $corr(u_j, N_j)$, and four parameters of cluster size distribution (i.e. θ_N, τ_N, ζ_N, and η_N). In contrast, the sampling variances in Table 2.3 (second row) from the design-based approach require the knowledge of cluster size N_j and cluster mean $\bar{Y}_j$ for all the K clusters in the population. If that information were available, then the population mean μ would also be known, making the survey superfluous.

Table 2.3. Population mean μ estimator and sampling variance per sampling scheme under the design-based approach.

	SRS	TSS1	TSS2	TSS3
$\hat{\mu}$	$\frac{\sum_{i=1}^{m} y_i}{m}$ (Cochran, 1977, p. 22; Lohr, 2010, eq. (2.8), p. 35; Sukhatme, 1954, eq. (5), p. 21)	$\frac{\sum_{j=1}^{k} \bar{y}_j}{k}$ (Cochran, 1977, eq. (11.39), p. 308; Lohr, 2010, p. 236; Sukhatme, 1954, eq. (2), p. 359)	$\frac{\sum_{j=1}^{k} N_j \bar{y}_j}{\sum_{j=1}^{k} N_j}$ (Cochran, 1977, eq. (11.25), p. 303; Lohr, 2010, eq. (5.26), p. 186; Sukhatme, 1954, eq. (76), p. 317)	$\frac{\sum_{j=1}^{k} N_j \bar{y}_j}{\sum_{j=1}^{k} N_j}$ (Cochran, 1977, eq. (11.25), p. 303; Lohr, 2010, eq. (5.26), p. 186; Sukhatme, 1954, eq. (76), p. 317)
$\mathrm{Var}(\hat{\mu})$	$\frac{\sum_{i=1}^{N_{pop}} (Y_i - \mu)^2}{m(N_{pop} - 1)}$ (Cochran, 1977, eq. (2.8), p. 23; Lohr, 2010, eq. (2.9), p. 36; Sukhatme, 1954, eq. (39), p. 29)	$\frac{1}{k}\sum_{j=1}^{K} \frac{N_j}{N_{pop}} (\bar{Y}_j - \mu)^2 + \frac{1}{k}\sum_{j=1}^{K} \frac{N_j}{N_{pop}} \frac{\sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_j)^2}{n(N_j - 1)}$ (Cochran, 1977, eq. (11.33), p. 307; Sukhatme, 1954, eq. (14), p. 362)	$\frac{\sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 (\bar{Y}_j - \mu)^2}{k(K-1)} + \frac{1}{kK}\sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \frac{\sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_j)^2}{n_j(N_j - 1)}$ (Cochran, 1977, eq. (11. 27), p. 304; Sukhatme, 1954, eq. (96), p. 325)	$\frac{\sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 (\bar{Y}_j - \mu)^2}{k(K-1)} + \frac{1}{kK}\sum_{j=1}^{K} \left(\frac{N_j}{\theta_N}\right)^2 \frac{\sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_j)^2}{n(N_j - 1)}$ (Cochran, 1977, eq. (11. 27), p. 304; Sukhatme, 1954, eq. (96), p. 325)
$\mathrm{Var}(\hat{\mu})$ under eq. (2.7)	$\frac{\sigma_v^2 + \sigma_\varepsilon^2 + \gamma^2 \sigma_N^2 [\tau_N(\zeta_N - \tau_N) + 1]}{m}$	$\frac{n\sigma_v^2 + \sigma_\varepsilon^2}{nk} + \gamma^2 \frac{\sigma_N^2 [\tau_N(\zeta_N - \tau_N) + 1]}{k}$	$\frac{p\theta_N(\tau_N^2 + 1)\sigma_v^2 + \sigma_\varepsilon^2}{p\theta_N k} + \gamma^2 \frac{\sigma_N^2(\tau_N^4 + \tau_N^2(\eta_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1)}{k}$	$(\tau_N^2 + 1)\frac{(n\sigma_v^2 + \sigma_\varepsilon^2)}{nk} + \gamma^2 \frac{\sigma_N^2(\tau_N^4 + \tau_N^2(\eta_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1)}{k}$

Note: $m = \bar{n}k$ is the number of individuals sampled with SRS, k is the number of clusters sampled with a TSS scheme, and $\bar{n} = \frac{\sum_{j=1}^{k} n_j}{k}$. For any TSS scheme, we assume $\frac{k}{K} \to 0$ and $\frac{\bar{n}}{\theta_N} \to 0$ or sampling with replacement at each stage, and for SRS $\frac{m}{N_{pop}} \to 0$ or sampling with replacement. In the third row, the outcome variable is assumed to be described by equation (2.7). For large enough k, the variances in the third row are equal to those in the last two rows of Table 2.1. Note that $\zeta_N = \left(\frac{1}{\sigma_N^3}\right)\left(\frac{\sum_{j=1}^{K}(N_j - \theta_N)^3}{K}\right)$ is the skewness, and $\eta_N = \left(\frac{1}{\sigma_N^4}\right)\left(\frac{\sum_{j=1}^{K}(N_j - \theta_N)^4}{K}\right)$ is the kurtosis of cluster size distribution in the population. Derivations are given in section 3 of the Supplementary Material.

2.6 Application to two real cluster size distributions

With the aim of planning a survey to estimate the population mean μ of a quantitative outcome variable Y_{ij} in a two-level population, we want to establish whether TSS1 is more efficient than TSS2 for the population under study and assess its efficiency gain relative to TSS2. The outcome variable Y_{ij} is assumed to be decomposed as shown in equation (2.7), but the analysis is carried out for both the model-based and the design-based approach. Two real cluster size distributions are considered: the distribution of public high school size in Italy, and the distribution of patient list size for general practices in England.

School size and alcohol consumption. In adolescent health literature it has been shown that greater connection between students and school (e.g. positive relations with teachers and peers, participation in school activities) is associated with less emotional distress, substance consumption (e.g. alcohol, cigarettes, marijuana), violence, and suicidal intentions (Resnick et al., 1997). Furthermore, it has been found that *school connectedness* and school size are inversely related (McNeely et al., 2002; Thompson et al., 2006), which suggests that school size can be informative for health risk behaviours in adolescents. Suppose that we want to estimate the average weekly alcohol consumption (in litres) among high school students in Italy. According to the Italian Ministry of Education (DGCASIS, 2018), in the school year 2016/2017 in Italy there were $6{,}235 = K$ public high schools with a total of $2{,}515{,}060 = N_{pop}$ students enrolled. The distribution of public high school size in Italy (with parameters $\theta_N = 403$, $\tau_N = 0.912$, $\zeta_N = 1.256$, and $\eta_N = 4.315$) is plotted in Figure 2.3 (first column, first row). The first row of Figure 2.3 also shows the relative efficiency of TSS2 versus TSS1, for a sample of $50 = k$ schools and $20 = \bar{n}$ students per school, as a function of the (absolute value of the) correlation between school size and school specific-mean, for different values of the intraclass correlation, under the model-based (second column) and the design-based approach (third column). As can be seen from Figure 2.3, under both inferential approaches TSS1 is more efficient than TSS2 and allows a sizeable efficiency gain (about 15%) even for non-informative school size and a small intraclass correlation ($\rho = 0.01$).

Patient list size for general practices and government expenditure on health. According to Eurostat (Eurostat, 2018), in 2016 health was the second largest area of government expenditure in the United Kingdom with a share of 7.6% of the Gross Domestic Product (GDP). Spending for hospital services represented the largest component of the

government expenditure on health, with a share of 5.7% of the GDP (Eurostat, 2018). In reducing such costs, general practices can play a role by effectively treating those conditions, which can lead to avoidable hospitalisations (e.g. influenza, diabetic complications). Kelly and Stoye (2014) have found that small practices (defined as those with three or fewer full-time equivalent (FTE) practitioners) had higher rates of hospitalisations for such preventable conditions in 2010/2011 in England. This suggests that patient list size can be informative for government expenditure on health, given that patients per general practice were proportional to the number of FTE practitioners (see figure 2.6 and table 2.3 in Kelly & Stoye, 2014). Suppose we want to estimate the average per capita government expenditure on health in England. According to the Health and Social Care Information Centre (Salt, 2017), in October 2017, 58,719,921$= N_{pop}$ patients were registered at 7,353 $= K$ general practices in England. The distribution of patient list size for general practices in England (with parameters $\theta_N = 7{,}986$, $\tau_N = 0.633$, $\zeta_N = 2.12$, and $\eta_N = 14.549$) is plotted in Figure 2.3 (first column, second row). The second row of Figure 2.3 shows the relative efficiency of TSS2 versus TSS1, for a sample of $50 = k$ practices and $20 = \bar{n}$ patients per practice, as a function of the (absolute value of the) correlation between patient list size and general practice specific-mean, for different values of the intraclass correlation, under the model-based (second column) and the design-based approach (third column). As shown in the second row of Figure 2.3, TSS1 is more efficient than TSS2 under both inferential paradigms and its efficiency gain increases as the intraclass correlation and/or the correlation between patient list size and the general practice specific-mean increase.

To conclude, the two examples show that TSS2 leads to important efficiency losses relative to TSS1, and that in planning a survey it is more practical to use variances based on a model, like those given in Table 2.1 or third row of Table 2.3, than the design-based variances in the second row of Table 2.3, which require the prior knowledge of all cluster sizes N_j as well as all cluster means $\bar{Y}_j$ in the population.

Relative efficiency of TSS2 versus TSS1 for two real cluster size distributions

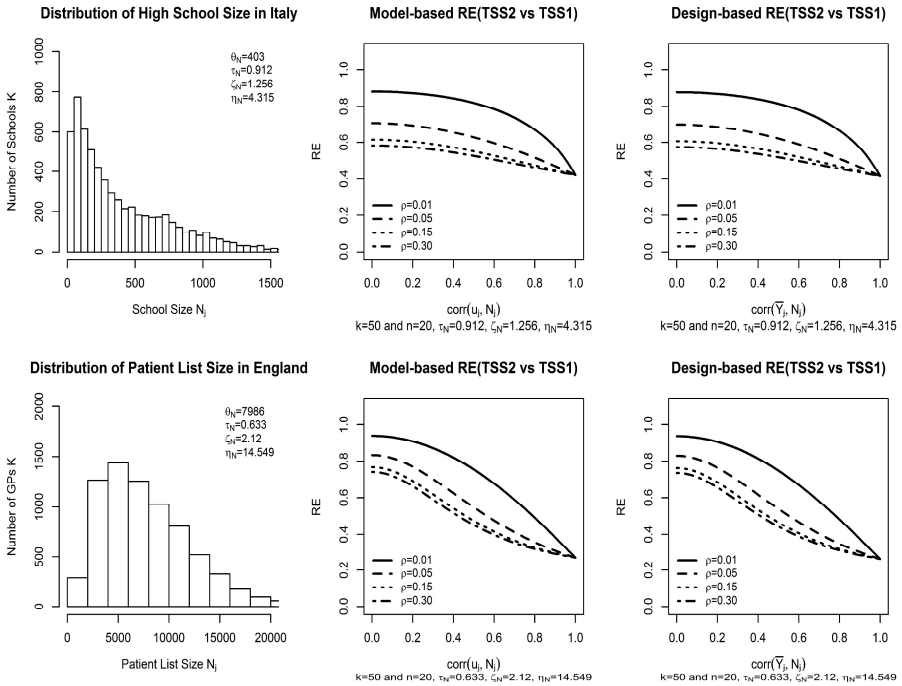


Figure 2.3. First column: Distribution of public high school size in Italy (first row), distribution of patient list size for general practices in England (second row). Second column: Model-based relative efficiency of TSS2 versus TSS1, as a function of the (absolute value of the) correlation between cluster effect and cluster size (i.e. $corr(u_j, N_j)$), for different values of the intraclass correlation coefficient ρ (curves). Third column: Design-based relative efficiency of TSS2 versus TSS1, as a function of the (absolute value of the) correlation between cluster mean and cluster size (i.e. $corr(\bar{Y}_j, N_j)$), for different values of the intraclass correlation coefficient ρ (curves).

2.7 Discussion

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

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

Although design-based inference has the advantage of being robust against violations of the model assumptions, comparing the four sampling schemes in terms of their relative efficiencies, as well as sample size planning, can only be done taking a model-based approach. Sample size planning within the design-based approach would require knowledge of the size and outcome mean of all clusters in the population (see Table 2.3, second row), which in turn would imply that the population mean is already known. Furthermore, models are also needed to deal with missing data and measurement error (Särndal et al., 1992).

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

Appendix A: Derivation of the population mean μ

Assuming that $u_j = \gamma\left(N_j - \theta_N\right) + \nu_j$ (i.e. assumption 4) and then plugging this equation into model (2.1) give

$$y_{ij} = \beta_0 + \gamma\left(N_j - \theta_N\right) + \nu_j + \varepsilon_{ij}. \qquad (2.A.1)$$

Now, before deriving the population mean μ under model (2.1) with informative cluster size as in (2.A.1), we will first show how the sampling scheme affects the sampling distribution of cluster size, which, in turn, influences the cluster effect sampling distribution if cluster size is informative.

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

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

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

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

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

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

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

$$E_{\mathrm{SRS/TSS1}}(N_j) = E_{\mathrm{TSS2/TSS3}}(N_j)(\tau_N^2 + 1) \tag{2.A.3a}$$

$$V_{\mathrm{SRS/TSS1}}(N_j) = V_{\mathrm{TSS2/TSS3}}(N_j)[\tau_N(\zeta_N - \tau_N) + 1]. \tag{2.A.3b}$$

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

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

$$E(u_j|N_j) = E(\gamma(N_j - \theta_N) + v_j|N_j) = \gamma(N_j - \theta_N) \tag{2.A.4a}$$

$$V(u_j|N_j) = V(\gamma(N_j - \theta_N) + v_j|N_j) = \sigma_v^2. \tag{2.A.4b}$$

In contrast, the marginal expectation (i.e. $E(u_j) = E\left(E(u_j|N_j)\right)$) and the marginal variance (i.e. $V(u_j) = E\left(V(u_j|N_j)\right) + V\left(E(u_j|N_j)\right)$) of u_j are affected by the sampling scheme because they are obtained by integrating $E(u_j|N_j)$ and $V(u_j|N_j)$ over the cluster size sampling distribution. Thus, if clusters are weighted equally in the cluster size sampling distribution (i.e. under TSS2 or TSS3), it follows from (2.A.2) that

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

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

that is, equations (2.2.a) and (2.2.b), respectively. In contrast, if each cluster of size N_j is weighted by the factor $\frac{N_j}{\theta_N}$ in the cluster size sampling distribution (i.e. under SRS or TSS1), it follows from (2.A.3) that

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

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

that is, equations (2.4.a) and (2.4.b), respectively. The two definitions of population means (i.e. equations (2.3) and (2.5)) now follow from $E_{\mathrm{TSS2/TSS3}}(u_j)$ and $E_{\mathrm{SRS/TSS1}}(u_j)$, respectively, given model (2.1).

Appendix B: Results of Table 2.1

The following facts will be used in this appendix. First, $\boldsymbol{N} = (N_1, \dots, N_k)^T$ denotes the vector of the cluster sizes of the k clusters drawn with TSS, $\boldsymbol{N}_{SRS} = \left(N_1, \dots, N_{k_{SRS}}\right)^T$ is the vector of the cluster sizes of the k_{SRS} clusters indirectly sampled with SRS, while $\boldsymbol{N}_*$ is used when the sampling scheme is not specified (i.e. $\boldsymbol{N}_* = \boldsymbol{N}$ for TSS and $\boldsymbol{N}_* = \boldsymbol{N}_{SRS}$ for SRS). Second, note that the four estimators in the first row of Table 2.1 are all of the form $\hat{\mu} = \frac{\sum_{j=1}^{k_*} w_j \bar{y}_j}{\sum_{j=1}^{k_*} w_j}$, where $k_* = k$ for the three TSS schemes and $k_* = k_{SRS}$ for SRS (recall that $\frac{m}{N_{pop}} \to 0$ (i.e. assumption 2), which entails that $k_{SRS} \to m$). Third, from equation (2.A.1) we have that $\bar{Y}_j = \beta_0 + \gamma\left(N_j - \theta_N\right) + \nu_j + \bar{\varepsilon}_j$.

Conditional expectations and unbiasedness

The conditional expectation of any estimator in Table 2.1 has the form $E(\hat{\mu}|\boldsymbol{N}_*) = \frac{\sum_{j=1}^{k_*} w_j E(\bar{y}_j|\boldsymbol{N}_*)}{\sum_{j=1}^{k_*} w_j}$, where $E\left(\bar{y}_j|\boldsymbol{N}_*\right) = \beta_0 + \gamma\left(N_j - \theta_N\right)$. Thus, the second row of Table 2.1 follows: $E(\hat{\mu}_{TSS1}|\boldsymbol{N}) = \beta_0 + \gamma(\bar{N} - \theta_N)$ since $w_j = 1$, where $\bar{N} = \frac{\sum_{j=1}^{k} N_j}{k}$; $E(\hat{\mu}_{TSS3}|\boldsymbol{N}) = \beta_0 + \gamma(\bar{N}(CV_N^2 + 1) - \theta_N)$ since $w_j = N_j$ and $\frac{\sum_{j=1}^{k} N_j^2}{\sum_{j=1}^{k} N_j} = \frac{S_N^2 + \bar{N}^2}{\bar{N}} = \bar{N}(CV_N^2 + 1)$, where $S_N^2 = \frac{\sum_{j=1}^{k}\left(N_j - \bar{N}\right)^2}{k}$; $E(\hat{\mu}_{TSS2}|\boldsymbol{N}) = E(\hat{\mu}_{TSS3}|\boldsymbol{N})$ because $w_j = n_j = pN_j$; $E(\hat{\mu}_{SRS}|\boldsymbol{N}_{SRS}) = \beta_0 + \gamma(\bar{N}_{SRS} - \theta_N)$ because $w_j = 1$, where $\bar{N}_{SRS} = \frac{\sum_{j=1}^{k_{SRS}} N_j}{k_{SRS}}$.

To prove the unbiasedness of the four estimators (fourth row of Table 2.1), we need to derive their marginal expectation, that is, integrating the conditional expectation over the cluster size sampling distribution. Thus, equation (2.A.3a) implies that $\hat{\mu}_{TSS1}$ and $\hat{\mu}_{SRS}$ are unbiased because $E(\hat{\mu}_{TSS1}) = E_{\text{TSS1}}\left(E(\hat{\mu}_{TSS1}|\boldsymbol{N})\right) = \beta_0 + \gamma(E_{\text{TSS1}}(\bar{N}) - \theta_N) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu$ and $E(\hat{\mu}_{SRS}) = E_{\text{SRS}}\left(E(\hat{\mu}_{SRS}|\boldsymbol{N}_{SRS})\right) = \beta_0 + \gamma(E_{\text{SRS}}(\bar{N}_{SRS}) - \theta_N) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu$. In contrast, $\hat{\mu}_{TSS3}$ and $\hat{\mu}_{TSS2}$ are asymptotically unbiased, because $E(\hat{\mu}_{TSS3}) = E(\hat{\mu}_{TSS2}) = E_{TSS3}\left(E(\hat{\mu}_{TSS3}|\boldsymbol{N})\right) = E_{TSS2}\left(E(\hat{\mu}_{TSS2}|\boldsymbol{N})\right) = \beta_0 + \gamma\left(E_{TSS2/TSS3}\left(\frac{S_N^2 + \bar{N}^2}{\bar{N}}\right) - \theta_N\right) \approx \beta_0 + \gamma\theta_N\left(\frac{k-1}{k}\right)\tau_N^2 \approx \mu$,

where $E_{TSS2/TSS3}\left(\frac{S_N^2+\bar{N}^2}{\bar{N}}\right) = E\left(\frac{S_N^2+\bar{N}^2}{\bar{N}}\right) = E\left(\bar{N}(CV_N^2+1)\right) \approx \theta_N\left(\left(\frac{k-1}{k}\right)\tau_N^2+1\right)$ comes from (i) $E(S_N^2) = \left(\frac{k-1}{k}\right)\sigma_N^2$, and (ii) the multivariate version of the delta method (Casella & Berger, 2002, pp. 241-242). To better understand why the unweighted average of cluster means is an unbiased estimator of μ under SRS and TSS1 but biased under TSS2 and TSS3, note that (i) $E\left(\sum_{j=1}^{k_*}\frac{\bar{y}_j}{k_*}\middle|\boldsymbol{N}_*\right) = \beta_0 + \gamma(\bar{N}_* - \theta_N)$ for any sampling scheme, and (ii) $E(N_j)$ depends on the sampling scheme (see equations (2.A.2a) and (2.A.3a)), and so

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

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

Conditional variances

The conditional variance of any estimator in Table 2.1 has the form $V(\hat{\mu}|\boldsymbol{N}_*) = \frac{\sum_{j=1}^{k_*} w_j^2 V(\bar{y}_j|N_*)}{\left(\sum_{j=1}^{k_*} w_j\right)^2}$. Furthermore, note that under TSS1 and TSS3 n individuals are sampled per cluster and so $V\left(\bar{y}_j\middle|\boldsymbol{N}\right) = \sigma_v^2 + \frac{\sigma_\varepsilon^2}{n}$, whereas under TSS2 n_j individuals are sampled per cluster, then $V\left(\bar{y}_j\middle|\boldsymbol{N}\right) = \sigma_v^2 + \frac{\sigma_\varepsilon^2}{n_j}$. Under SRS k_{SRS} clusters are sampled indirectly from the population, but given that $k_{SRS} \to m$ (which follows from $\frac{m}{N_{pop}} \to 0$ in assumption 2) we have that $V\left(\bar{y}_j\middle|\boldsymbol{N}_{SRS}\right) = \sigma_v^2 + \sigma_\varepsilon^2$. Thus, the third row of Table 2.1 follows: $V(\hat{\mu}_{TSS1}|\boldsymbol{N}) = \frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk}$ since $w_j = 1$; $V(\hat{\mu}_{TSS3}|\boldsymbol{N}) = \frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk} \times (CV_N^2+1)$ since $w_j = N_j$ and $\frac{\sum_{j=1}^{k} N_j^2}{\left(\sum_{j=1}^{k} N_j\right)^2} = \frac{S_N^2+\bar{N}^2}{k\bar{N}^2} = \frac{(CV_N^2+1)}{k}$; $V(\hat{\mu}_{TSS2}|\boldsymbol{N}) = \frac{p\bar{N}(CV_N^2+1)\sigma_v^2+\sigma_\varepsilon^2}{p\bar{N}k}$ since $w_j = n_j = pN_j$ and $\frac{\sum_{j=1}^{k} n_j^2}{\left(\sum_{j=1}^{k} n_j\right)^2} = \frac{\sum_{j=1}^{k} N_j^2}{\left(\sum_{j=1}^{k} N_j\right)^2}$; and $V(\hat{\mu}_{SRS}|\boldsymbol{N}_{SRS}) = \frac{\sigma_v^2+\sigma_\varepsilon^2}{m}$ since $w_j = 1$.

Marginal variances

Recall that the marginal variance is defined as $V(\hat{\mu}) = E\big(V(\hat{\mu}|\boldsymbol{N}_*)\big) + V\big(E(\hat{\mu}|\boldsymbol{N}_*)\big)$. From equation (2.A.3b) follows that $V(\hat{\mu}_{TSS1}) = E_{TSS1}\left(\frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk}\right) + V_{TSS1}\big(\beta_0 + \gamma(\overline{N} - \theta_N)\big) = \frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk} + \gamma^2\frac{V_{\mathrm{TSS1}}(N)}{k} = \frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk} + \gamma^2\frac{\sigma_N^2[\tau_N(\zeta_N-\tau_N)+1]}{k}$, and that $V(\hat{\mu}_{SRS}) = E_{SRS}\left(\frac{\sigma_v^2+\sigma_\varepsilon^2}{m}\right) + V_{SRS}\big(\beta_0 + \gamma(\overline{N}_{SRS} - \theta_N)\big) = \frac{\sigma_v^2+\sigma_\varepsilon^2}{m} + \gamma^2\frac{V_{\mathrm{SRS}}(N)}{m} = \frac{\sigma_\varepsilon^2+\sigma_v^2+\gamma^2\sigma_N^2[\tau_N(\zeta_N-\tau_N)+1]}{m}$. The derivation of the marginal variances of TSS3 and TSS2 requires more steps. The first component of $V(\hat{\mu})$ (fifth row of Table 2.1) for TSS3 and TSS2 are, respectively, $E\big(V(\hat{\mu}_{TSS3}|\boldsymbol{N})\big) = \frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk} \times (E(CV_N^2) + 1) \approx \left(\frac{n\sigma_v^2+\sigma_\varepsilon^2}{nk}\right) \times \left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)$, and $\mathrm{E}\big(V(\hat{\mu}_{TSS2}|\boldsymbol{N})\big) = \frac{(E(CV_N^2)+1)\sigma_v^2}{k} + \frac{\sigma_\varepsilon^2}{pk}E\left(\frac{1}{\overline{N}}\right) \approx \frac{p\theta_N\left(\frac{k(\tau_N^2+1)}{\tau_N^2+k}\right)\sigma_v^2+\sigma_\varepsilon^2}{p\theta_N k}$, where both $E(CV_N^2) + 1 = E\left(\frac{S_N^2}{\overline{N}^2}\right) + 1 \approx \frac{E(S_N^2)}{E(\overline{N}^2)} + 1 = \frac{\left(\frac{k-1}{k}\right)\sigma_N^2}{\frac{\sigma_N^2}{k}+\theta_N^2} + 1 = \frac{k(\tau_N^2+1)}{\tau_N^2+k}$ and $E\left(\frac{1}{\overline{N}}\right) \approx \frac{1}{\theta_N}$ follow from the delta method (Casella & Berger, 2002, pp. 241-242). The second component of $V(\hat{\mu})$ (sixth row of Table 2.1) is the same under TSS2 and TSS3 since $E(\hat{\mu}_{TSS2}|\boldsymbol{N}) = E(\hat{\mu}_{TSS3}|\boldsymbol{N})$ (see Table 2.1, second row), then $V\big(E(\hat{\mu}_{TSS3}|\boldsymbol{N})\big) = V\big(E(\hat{\mu}_{TSS2}|\boldsymbol{N})\big) = \gamma^2 V_{TSS2/TSS3}\left(\frac{S_N^2+\overline{N}^2}{\overline{N}}\right) \approx \gamma^2\frac{\sigma_N^2}{k}\left[\left(\frac{k-1}{k}\right)^2\tau_N^2\left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N)\right) + 2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N - \tau_N) + 1\right]$, which is derived as follows. To apply the delta method, compute the first derivatives of $g(S_N^2, \overline{N}) = \frac{S_N^2+\overline{N}^2}{\overline{N}}$ at $\big(E(S_N^2), E(\overline{N})\big)^T$:

$$\left.\frac{\partial g(S_N^2,\overline{N})}{\partial S_N^2}\right|_{S_N^2=\left(\frac{k-1}{k}\right)\sigma_N^2,\overline{N}=\theta_N} = \frac{1}{\theta_N}, \quad \left.\frac{\partial g(S_N^2,\overline{N})}{\partial \overline{N}}\right|_{S_N^2=\left(\frac{k-1}{k}\right)\sigma_N^2,\overline{N}=\theta_N} = 1 - \left(\frac{k-1}{k}\right)\tau_N^2.$$

Then, plug these derivatives into equation (5.5.9) in Casella and Berger (2002, p. 242):

$$\mathrm{Var}\big(g(S_N^2,\overline{N})\big) \approx \frac{1}{\theta_N^2}Var(S_N^2) + \left(1 - \left(\frac{k-1}{k}\right)\tau_N^2\right)^2 Var(\overline{N}) + 2\frac{1}{\theta_N}\left(1 - \left(\frac{k-1}{k}\right)\tau_N^2\right)Cov(S_N^2,\overline{N}).$$

Finally, in the previous expression replace $Var(S_N^2)$, $Var(\overline{N})$, and $Cov(S_N^2,\overline{N})$ with $Var(S_N^2) = \left(\frac{k-1}{k}\right)^2 Var\left(\frac{\sum_{j=1}^k (N_j-\overline{N})^2}{k-1}\right) = \left(\frac{k-1}{k}\right)^2\frac{\sigma_N^4}{k}\left(\eta_N - \frac{k-3}{k-1}\right)$ (Theorem 2, p. 229, in Mood et al., 1974), where $\eta_N = E\left[\left(\frac{N_j-\theta_N}{\sigma_N}\right)^4\right]$ is the kurtosis of cluster size distribution, $Var(\overline{N}) = \frac{\sigma_N^2}{k}$, and $Cov(S_N^2,\overline{N}) = \left(\frac{k-1}{k}\right)\frac{\sigma_N^3\zeta_N}{k}$ (Zhang, 2007).

Appendix C: Notation, and table of contents of the online supplementary material

Table 2.A summarizes the notation used in this chapter, while Table 2.B gives the table of contents of the online supplementary material, and the link to download it.

Table 2.A. Notation.

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

Table 2.B. Table of contents of the supplementary material, and link to download it.

Contents

1. Expectation and variance of cluster size sampling distribution under SRS or TSS1
2. Relative efficiencies of TSS schemes versus SRS and each other (i.e. section 2.4)
   2.1. Relative efficiencies
   2.2. Extremes of the relative efficiencies
   2.3. Relative efficiencies of the optimal estimators for non-informative cluster size
3. Design-based inference: results of section 2.5

The online supplementary material can be downloaded at the following link:

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Fsim.8070&file=SIM_8070-Supp-0001-SIM8070_online_Supplementary_Material.pdf

Chapter 3

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

This chapter has been published online in *Statistical Methods in Medical Research* (17 Sep. 2020), doi:10.1177/0962280220952833, with co-authors Math J.J.M. Candel, Frans E.S. Tan, and Gerard J.P. van Breukelen

Abstract

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

Keywords: cross-national comparisons; informative cluster size; maximin design; optimal design; sample size calculation; two-stage sampling

3.1 Introduction

For the purpose of estimating the mean or prevalence of an outcome variable (e.g. alcohol consumption or smoking) in a hierarchical population (e.g. students within schools, patients within general practices), or of comparing subpopulations with respect to such a mean or prevalence, it is often convenient, for economic or logistic reasons, to sample in two stages: first, clusters (e.g. schools, general practices) are sampled and then individuals (e.g. students, patients) are drawn from the sampled clusters (Chambers & Clark, 2012; Cochran, 1977; Lohr, 2010). Examples of these multi-stage sampling designs include school-based surveys for monitoring substance use among adolescents (ESPAD Group, 2016; Patton et al., 1995; Warren et al., 2000), and national surveys for estimating the average length of stay for discharges from hospitals (DeFrances et al., 2008), or nursing homes (Jones, 2002). The topic of this paper is the efficient design of Two-Stage Sampling (TSS) schemes for estimating the mean of a quantitative outcome variable in a two-level population.

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

Additionally to cluster size variation, further complications arise with informative cluster sizes, that is, when cluster size is related to the outcome of interest (Nevalainen et al., 2014; Seaman et al., 2014). For instance, cluster size is informative when the amount of alcohol consumed by an adolescent is related to the number of students enrolled in the school, as small schools might provide a more supportive environment (McNeely et al., 2002; Resnick et al., 1997; Thompson et al., 2006), or when the number of patients registered to a general practice affects its efficacy in preventing expensive hospitalisations (Kelly & Stoye, 2014), thus impacting on public expenditure on health per patient. Informative cluster sizes not only can have direct policy implications, such as introducing a limit to school or general practice size, they also have consequences for statistical data analysis and sample size planning. In informative cluster size literature (see the review by Seaman et al. (2014), and references

therein), the main focus has been on how to handle informative cluster size when the target of inference is the association between the outcome variable and some covariates (e.g. a risk factor). For instance, Seaman et al. (2014) have discussed several methods to make cluster-specific inferences with Generalized Linear Mixed Models and population-average inferences with Generalized Estimating Equations when cluster size is informative. Innocenti et al. (2019), instead, have investigated a different topic: the implications of informative cluster size for unbiased and efficient estimation of a population mean in surveys conducted with the three aforementioned TSS schemes. The present paper is also about mean estimation for these three TSS schemes when cluster size is informative, but focuses instead on sample size planning, and the consequences of informative cluster size for the required sample sizes and budget.

Innocenti et al. (2019)'s results are the starting point of this paper and therefore summarized here. First, there are two definitions of overall mean in a two-level population, namely the average of all individual outcomes and the average of all cluster-specific means. These two definitions coincide only if cluster sizes are either equal or non-informative. Second, when cluster size is informative, estimation of the mean of all individual outcomes (i.e. the definition used in this paper) is unbiased under TSS1 with the unweighted average of cluster means, and asymptotically unbiased under TSS2 and TSS3 with the average of cluster means weighted by cluster size. In contrast, when cluster size is non-informative, the unweighted average of cluster means is unbiased for all sampling schemes, but optimally efficient for TSS1 and TSS3 only. Third, under the constraint of a fixed total sample size, SRS is more efficient than any TSS scheme, TSS3 is the least efficient TSS scheme, and TSS1 is the most efficient for many cluster size distributions. Indeed, when cluster size is informative, the relative efficiency of these sampling schemes depends on some features of the cluster size distribution in the population, such as the coefficient of variation, the skewness, and the kurtosis. However, when cluster size is non-informative, TSS1 and TSS3 are equally efficient and outperform TSS2. Fourth, the two inferential paradigms in survey sampling, namely the model-based (Chambers & Clark, 2012) and the design-based approach (Cochran, 1977; Lohr, 2010), give similar results in terms of unbiased and efficient estimation of the average of all individual outcomes with the three aforementioned TSS schemes, at least if the model assumptions are met. Furthermore, sample size planning and sampling schemes comparisons, which are the topics of this paper, are much more feasible with the assumption of a model for the outcome variable of interest (Innocenti et al., 2019). For these two reasons, the model-based approach is adopted here.

This work extends the results of Innocenti et al. (2019) in the following ways. First, for each of the three aforementioned TSS schemes, the optimal design is derived. Here, the optimal design is defined as that design (i.e. number of clusters and number of individuals per cluster) that minimizes the sampling variance of the population mean estimator subject to a cost constraint. Second, the three optimal TSS schemes are compared with SRS and with each other under the constraint of a fixed budget. Third, to take care of uncertainty with respect to model parameters and distributional features of cluster size, as a practical alternative, maximin designs are derived. Fourth, sample size calculations for making comparisons between populations are derived and illustrated.

This paper is structured as follows. In section 3.2, the assumptions of this paper are presented, as well as the sampling schemes and the corresponding mean estimators. Furthermore, the findings of a simulation study to assess the accuracy of some results in Innocenti et al. (2019) that are relevant to the present paper are summarized. In section 3.3, the optimal design for each TSS scheme is derived, and these optimal TSS designs are compared with each other and with SRS for a fixed budget. Furthermore, the consequences of ignoring informative cluster size at the design phase of a study are investigated. Section 3.4 deals with the maximin approach, that is, a strategy to solve the dependency of the optimal design on unknown nuisance parameters. Section 3.5 provides a procedure for computing sample sizes for surveys aimed to make cross-population comparisons, and the procedure is illustrated in planning a survey for comparing the average alcohol consumption among adolescents in France and Italy. Section 3.6 offers some final remarks. The mathematical derivations of the results, the description of the simulation study discussed in section 3.2, and additional figures and tables can be found in the Supplementary Material 1 (S.M.1). The Supplementary Material 2 (S.M.2) provides the R (R Core Team, 2020) code of the simulation study and other R codes to apply some of the mathematical results of this paper.

3.2 Assumptions, sampling schemes and mean estimators

The results of Innocenti et al. (2019) and this paper are based on the following assumptions (the notation used in the main text is summarized in Table 3.A in the appendix).

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

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

Assumption 3: The outcome variable Y_{ij} is quantitative (e.g. alcohol consumption) and measured at the individual level. Further, Y_{ij} shows variation at the cluster level as well as at the individual level. Therefore, sampling error occurs at each design level. This is taken into account by assuming the following two-level random intercept model for the outcome of the i-th individual from the j-th cluster (Chambers & Clark, 2012; Goldstein, 2011)

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij} \tag{3.1}$$

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

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

Innocenti et al. (2019) show that β_0 in model (3.1) is the average of all cluster-specific means in the population, and differs from the average of all individual outcomes in the population μ, unless cluster size is non-informative ($\alpha_1 = 0$) or constant across clusters, as can be seen from the following expression

$$\mu = \beta_0 + \alpha_1\theta_N\tau_N^2, \tag{3.2}$$

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

population of individuals (which yields μ, see Innocenti et al., 2019). This paper focuses on μ.

With the aim of estimating μ, the three aforementioned TSS schemes are studied in this paper. For each of these TSS schemes and SRS, Table 3.1 summarizes the sampling procedure (i.e. sample size and inclusion probability per design stage) and the required knowledge before sampling. Furthermore, Table 3.1 shows the population mean estimator $\hat{\mu}$ and the sampling variance $V(\hat{\mu})$ for each sampling scheme. Denote by $\rho_{uN} = \frac{E\left[\left(u_j - E(u_j)\right)\left(N_j - E(N_j)\right)\right]}{\sigma_u \sigma_N} = \frac{E\left[u_j\left(N_j - \theta_N\right)\right]}{\sigma_u \sigma_N}$ the correlation between u_j and N_j, where $E(u_j) = 0$ and $V(u_j) = \sigma_u^2 = \sigma_v^2 + \alpha_1^2 \sigma_N^2$, and by $\psi = \left(\frac{\rho_{uN}^2}{1-\rho_{uN}^2}\right)$ the degree of informativeness of cluster size. From Table 3.1, note that $\bar{n} = \frac{\sum_{j=1}^{k} n_j}{k} = n$ for TSS1 and TSS3, while $\bar{n} = \frac{\sum_{j=1}^{k} n_j}{k} = p\frac{\sum_{j=1}^{k} N_j}{k} = p\bar{N}$ for TSS2, where $\bar{N}$ is the average population size of the k sampled clusters (not to be confused with $\bar{n}$, that is, the average sample size of the sampled clusters) . Furthermore, for TSS2 $n = E(\bar{n}) = pE(\bar{N}) = p\theta_N$. The sampling variances in Table 3.1 are functions of the total unexplained outcome variance $\sigma_y^2 = \sigma_v^2 + \sigma_\varepsilon^2$, the intraclass correlation coefficient $\rho = \frac{\sigma_v^2}{\sigma_y^2} \in [0,1]$, the sample sizes (k, n), the parameter ψ, and some features of the cluster size distribution in the population: the coefficient of variation τ_N, the skewness ζ_N, and (for TSS2 and TSS3 only) the kurtosis η_N. When cluster size is non-informative ($\psi = 0$), $V(\hat{\mu})$ depends only on σ_y^2, ρ, k, n, and (for TSS2 and TSS3 only) τ_N. The estimators $\hat{\mu}$ associated with SRS and TSS1 are unbiased, and their sampling variances $V(\hat{\mu})$ are exact expressions (Innocenti et al., 2019).

The estimators associated with TSS2 and TSS3 are only asymptotically unbiased, and the corresponding sampling variances are based on first-order Taylor series approximations (Innocenti et al., 2019). The accuracy of these approximations was evaluated through a simulation study discussed in supplementary material S.M.1 (section 1), but the main findings are summarized here. Sampling $k = 20$ clusters guarantees nearly unbiased estimates of μ under TSS2 and TSS3 independently of the cluster size distribution, and fair accuracy (i.e. bias $\leq 5\%$) of the variances in Table 3.1 (TSS2 and TSS3 row) when $|\rho_{uN}| \leq 0.75$, $\rho \leq 0.3$, and ζ_N and η_N are relatively close (say, ± 1.5) to those of the Normal distribution (i.e. $\zeta_N = 0$ and $\eta_N = 3$). However, for cluster size distributions with extreme skewness and kurtosis (e.g. $\zeta_N \geq 2$ and $\eta_N \geq 9$) at least $k = 100$ clusters must be sampled to achieve a reasonable accuracy (i.e. bias $\leq$6%) of the sampling variances in Table 3.1, for $\rho_{uN} \leq 0.5$ and $\rho \leq 0.3$. Furthermore, the

simulations showed that the two lower-bounds for k (i.e. 20, and 100) guarantee the corresponding accuracy level across different values for n (at least for $2 \leq n \leq 100$). To contextualize these two lower-bounds for k, in a school-based survey for studying substance use among adolescents in 21 European countries, Shackleton et al. (2016) have reported that, across countries, $k \in [36,531]$ (Median$= 123$) and $\bar{n} \in [5.92,119.62]$ (Median$= 20.74$).

Table 3.1. Sampling schemes, required prior knowledge, population mean estimators, and sampling variances.

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

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

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

3.3 Optimal design and relative efficiencies for a given budget

3.3.1 Optimal design

For any sampling scheme, the precision of the estimator $\hat{\mu}$, and thus also the width of a confidence interval for μ and the statistical power for testing a hypothesis on μ, depends on the number of clusters and on the sample size per cluster (Table 3.1). This raises the question of the best combination of sample sizes at each design stage (i.e. sampling many clusters versus sampling many individuals per cluster). Define the optimal design as that design (i.e. number of clusters and number of individuals per cluster), which minimizes $V(\hat{\mu})$ subject to a cost constraint, given that time and budget are limited in practice. For TSS, the cost constraint is assumed to be $C = k(c_2 + c_1 n)$, where C is the budget for sampling and measuring (excluding costs for constructing the sampling frame and other costs not related to sample size). From now on C is called the *research budget*. Furthermore, c_2 is the average cost for sampling a cluster, c_1 is the average cost for sampling an individual from a sampled cluster, and $(c_2 + c_1 n)$ is the cost per cluster including the costs for sampling n individuals from that cluster (recall that for TSS2 $n = p\theta_N$). For SRS, the cost constraint is $C = c_0 + c_{srs} m$, where m is the number of individuals to sample, c_{srs} is the average cost for sampling an individual directly from the population, and c_0 represents the extra-cost due to constructing the sampling frame for a SRS compared with the sampling frame for a TSS.

For each TSS scheme the optimal design (i.e. the optimal sample sizes k^* and n^*) for estimating μ and the optimal variance $V(\hat{\mu})^*$ (i.e. $V(\hat{\mu})$ under the optimal design) are given in Table 3.2 (for proofs, see section 2.2 of S.M.1). For TSS2, one can obtain the optimal proportion of individuals to sample per cluster p^* from the optimal n^*, by dividing n^* as given in Table 3.2 (TSS2 row) by θ_N. The optimal TSS2 and TSS3 designs depend on two approximations of $V(\hat{\mu})$: the first-order Taylor approximation mentioned in section 3.2 and evaluated in S.M.1 (section 1), which underlies the equations in Table 3.1, and an approximation based on large k (i.e. k such that $\frac{\tau_N^2}{k} \approx 0$, $\frac{k-1}{k} \approx 1$, and $\frac{k-3}{k-1} \approx 1$) to simplify the expressions in Table 3.1. These two approximations give (for details, see section 2.1 of S.M.1)

$$V_{TSS2}(\hat{\mu}) \approx \frac{\sigma_y^2}{nk}\left\{1 + \rho\left[n\left((\tau_N^2 + 1) + \psi(\tau_N^4 + \tau_N^2(\eta_N - 3) + 2\zeta_N\tau_N(1 - \tau_N^2) + 1)\right) - 1\right]\right\} \quad (3.3)$$

and

$$V_{TSS3}(\hat{\mu}) \approx \frac{\sigma_y^2}{nk}\{\tau_N^2 + 1 + \rho[(\tau_N^2 + 1)(n-1) + n\psi(\tau_N^4 + \tau_N^2(\eta_N - 3) + 2\zeta_N\tau_N(1-\tau_N^2) + 1)]\}, \quad (3.4)$$

where for TSS2 $n = p\theta_N$. Recall from section 3.2 that, for TSS2 and TSS3, k must be large anyway, because the estimators $\hat{\mu}_{TSS2}$ and $\hat{\mu}_{TSS3}$ given in Table 3.1 are only asymptotically unbiased. As a special case, $\psi = 0$ gives the optimal design and optimal variance for non-informative cluster size (for which case $\beta_0 = \mu$), which under TSS1 coincide with the equations available for cluster randomized trials (for instance, see Moerbeek et al., 2000). There is no such equivalence under TSS2 due to sample size variation between clusters, and under TSS3 due to weighting cluster means by cluster size if informative cluster size is assumed in the design phase. Indeed, under non-informative cluster size, no weighting is needed under TSS3 (Innocenti et al., 2019), and then the optimal design equations for TSS1 apply to TSS3 as well.

Note from Table 3.2 that the optimal number of clusters k^* and the optimal number of individuals per cluster n^* are inversely related, and that n^* is an increasing function of the cluster-to-individual cost ratio $c_r = \frac{c_2}{c_1} > 1$ and a decreasing function of ρ and ψ. These relations between the optimal design and c_r, ρ, and ψ hold, under TSS1, for $\zeta_N > \tau_N - \frac{1}{\tau_N}$, and always under TSS2 and TSS3 (for proof, see section 2.1 of S.M.1). The condition $\zeta_N > \tau_N - \frac{1}{\tau_N}$ is met by all the distributions in Tables S.2 and S.7 (S.M.1). Hence, this condition is assumed to be satisfied when considering results for TSS1 in the sequel.

Table 3.2. Optimal design and optimal variance $V(\hat{\mu})^*$ for each sampling scheme.

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

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

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

The optimal number of individuals per cluster n^* for TSS1 and TSS3 is plotted in Figure 3.1, for two real-life cluster size distributions: the general practice list size distribution in England, and the public high school size distribution in Italy (both distributions are shown in Figure S.1, S.M.1). The behaviour of n^* for other cluster size distributions is shown in Figures S.2 and S.4 (S.M.1) for TSS1 and TSS3, respectively, and in Figure S.3 (S.M.1) for TSS2. In most scenarios in Figure 3.1 and Figures S.2-S.4 (S.M.1), the difference between n^* for $\psi = 0.35$ (i.e. $\rho_{uN} = \pm 0.51$) and n^* for $\psi = 0$ (i.e. $\rho_{uN} = 0$) is small, which means that the ratio of $V(\hat{\mu})$ under the design assuming $\psi = 0.35$ to $V(\hat{\mu})$ under the design assuming $\psi = 0$, when the true $\psi = 0.35$, is close to 1. So, the optimal designs in Table 3.2 are quite robust against misspecification of ψ, in the sense of being efficient relative to the optimal design for the true ψ and given a fixed research budget C. However, ignoring informativeness can lead to serious underestimation of the sampling variance of the mean estimator, and thereby also of the budget needed, as will be seen below. Further, the optimal design depends not only on ψ, but also on ρ and the cluster size distribution $(\tau_N, \zeta_N, \eta_N)$. That dependence will be addressed in section 3.4.

An example will now show that (a) given a study budget, the optimal design is robust against misspecification of cluster size informativeness, but (b) the budget needed is very sensitive to misspecification. Suppose we plan a survey to estimate μ in the population of all patients of all general practices in England. The parameters of the general practice patient list size distribution are $\tau_N = 0.633$, $\zeta_N = 2.12$, and $\eta_N = 14.549$ (Table S.2, S.M.1). Furthermore, suppose that $\rho = 0.05$, $c_r = 10$, and $C/c_1 = 1000$. The optimal TSS1 samples $n^* = 10.74$ individuals and $k^* = 48.22$ clusters assuming $\psi = 1/3$, and $n^* = 13.78$ and $k^* = 42.04$ assuming $\psi = 0$ (see Table 3.2, TSS1 row). If the true $\psi = 1/3$, $V(\hat{\mu}) = \sigma_y^2 \times 0.00354$ for the design correctly assuming $\psi = 1/3$, and $V(\hat{\mu}) = \sigma_y^2 \times 0.00360$ for the design incorrectly assuming $\psi = 0$ (see variance equation in Table 3.1, TSS1 row), giving a variance ratio $0.00354/0.00360 = 0.983$. Additional results for TSS1, TSS2, and TSS3 are given in Table S.8 (S.M.1), which shows that even in some more extreme cases (e.g. $\psi = 1$, i.e. $\rho_{uN} = \pm 0.707$) the variance ratio still exceeds 0.8. The example given here and those in Table S.8 (S.M.1) show that the optimal designs in Table 3.2 are quite robust against misspecification of ψ, in the sense of being efficient relative to the optimal design for the true ψ and given a fixed research budget C.

However, ignoring informativeness can lead to serious underestimation of the budget needed. Suppose one wants to test the null hypothesis H_0 that $\mu = \mu_0$ against the alternative hypothesis H_1 that $\mu \neq \mu_0$. The budget that guarantees the desired power level $1-\gamma$ for the chosen type I error rate α, is then obtained by equating $V(\hat{\mu})^*$ in Table 3.2 with $\left(\frac{\mu-\mu_0}{z_{1-\gamma}+z_{1-\frac{\alpha}{2}}}\right)^2$, where z_q is the qth percentile of the standard normal distribution. This gives $C = \frac{g(\rho,\psi)\left(z_{1-\gamma}+z_{1-\frac{\alpha}{2}}\right)^2}{d_0^2}$, where $g(\rho,\psi)$ is the numerator of $V(\hat{\mu})^*$ in Table 3.2 excluding σ_y^2, and $d_0 = \frac{\mu-\mu_0}{\sigma_y}$ is the standardized difference between true mean and mean according to H_0. Since $g(\rho,\psi)$ is an increasing function of c_1, c_2, and ψ, the required budget C for the desired power level also increases with c_1, c_2, and ψ. Likewise, C increases with ρ, at least up to $\rho = 0.5$ (for proofs, see section 2.2 in S.M.1). The required budget C to detect a standardized difference of medium size ($d_0 = 0.5$), with 90% power and two-tailed $\alpha = 0.05$, is plotted in Figure 3.2 for TSS1 and TSS3, as function of ψ, for the general practice list size distribution in England and the public high school size distribution in Italy, and assuming $c_1 = 10$. As can be seen in Figure 3.2, the research budget C is not robust against misspecification of ψ. For example, the required budget C for the optimal TSS1, assuming the English general practice list size distribution, $c_r = 30$, $c_1 = 10$, and $\rho = 0.10$ (Figure 3.2, left column, first row), is underestimated by 29% if one incorrectly assumes $\psi = 0$ when the true $\psi = 0.35$. The required budget C is also shown, for other cluster size distributions, in Figures S.5 and S.7 (S.M.1) for TSS1 and TSS3, respectively, and in Figure S.6 (S.M.1) for TSS2. These figures show that C increases with ρ, c_2, and ψ, and that the impact of the cluster size distribution on C becomes more relevant as ψ increases. Hence, ignoring informative cluster size at the design phase of the survey can lead to underestimating the required budget for the chosen effect size and desired power level. Finally, for the desired power level, the required budget is smallest with the optimal TSS1, and largest with the optimal TSS3.

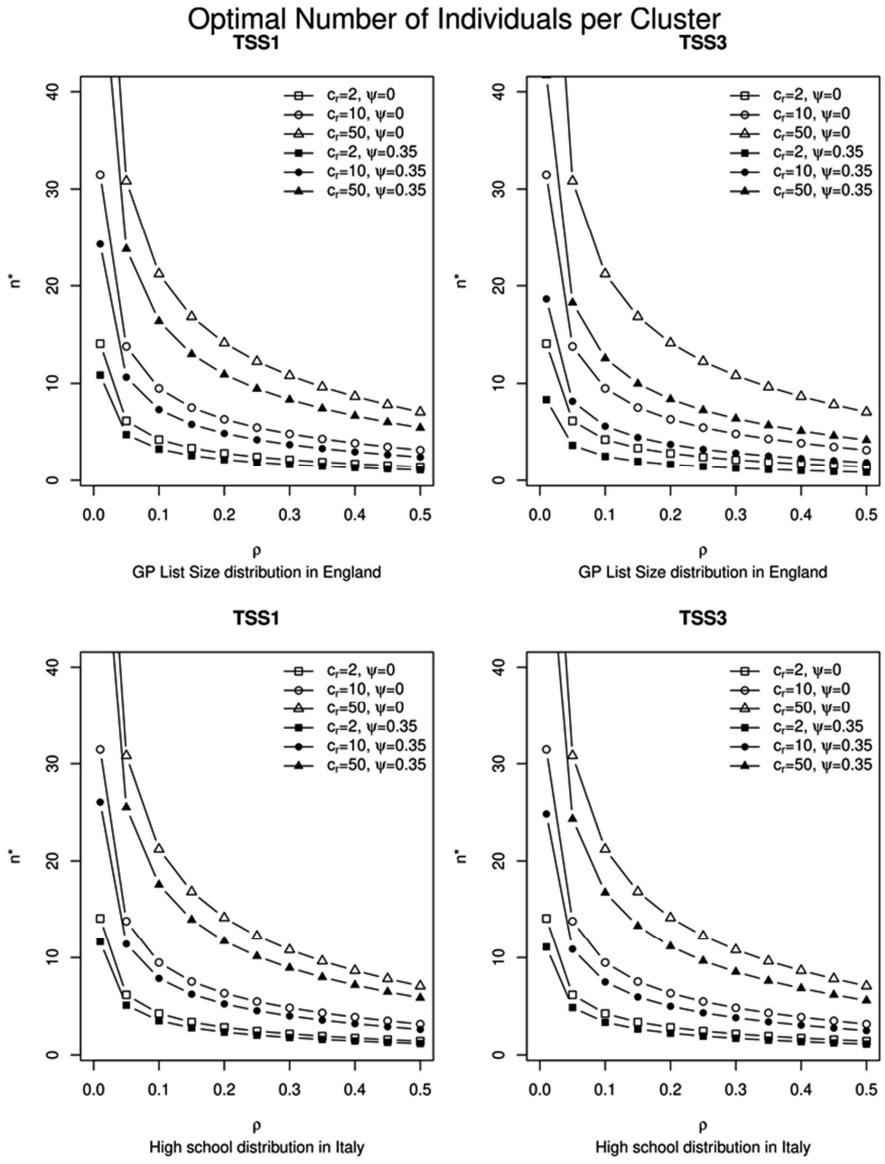


Figure 3.1. Optimal number of individuals per cluster n^* under TSS1 (left column), and TSS3 (right column), as a function of ρ, for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

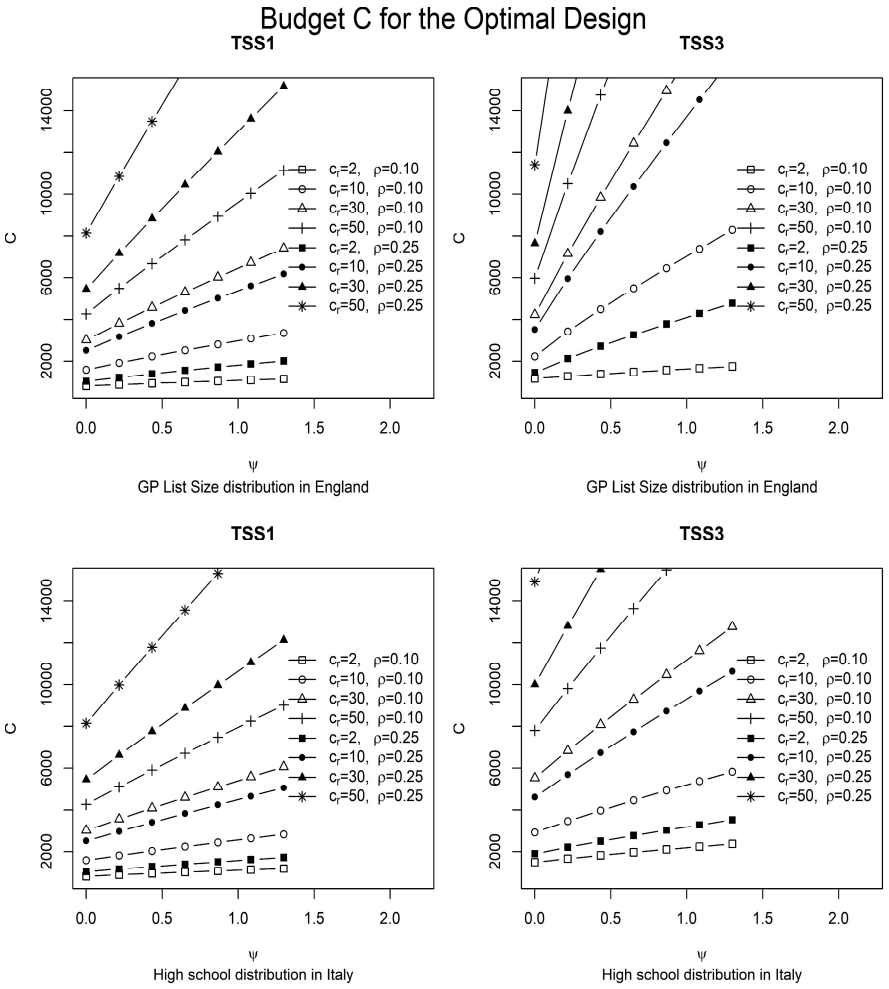


Figure 3.2. Budget C needed for the optimal design to detect a standardized difference between hypothesized and true population mean of medium size ($d_0 = 0.5$), with 90% power using a two-tailed test with $\alpha = 0.05$, as a function of ψ, for different values of ρ and c_r (curves) with $c_1 = 10$, different sampling schemes (columns), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi \in [0,1.3]$ corresponds to $\rho_{uN} \in [-0.75, +0.75]$.

3.3.3 Relative efficiencies for a given budget

We now compare the efficiency of the optimal designs in Table 3.2 with each other and with SRS, under the constraint of a fixed research budget. The relative efficiency (RE) of the optimal designs for two sampling schemes is defined as the ratio of their optimal variances $V(\hat{\mu})^*$ in Table 3.2, more specifically, $RE(D1\ vs\ D2) = \frac{V_{D2}(\hat{\mu})^*}{V_{D1}(\hat{\mu})^*}$. These REs are shown in Table 3.3 (for

proofs, see section 2.3, S.M.1), which also gives the sufficient (but not necessary) conditions under which each RE is smaller than one.

The RE of a TSS scheme compared with SRS (Table 3.3, first three rows) is composed of three ratios. The first ratio is a function of ρ, ψ, τ_N, ζ_N, η_N, and c_r, and is always smaller than one for $\psi = 0$, and also for $\psi \neq 0$ at least under the conditions for ζ_N given in the rightmost column of Table 3.3 (for proofs, see section 2.3, S.M.1). The other two components of $RE(TSS\ vs\ SRS)$ are the ratio $\frac{c_{srs}}{c_1}$, for the costs per individual in SRS relative to TSS, and the budget ratio $\frac{C}{C-c_0}$. Since sampling an individual directly from the population will be more expensive than sampling an individual after having sampled the cluster to which he/she belongs (i.e. $c_{srs} > c_1$), and constructing the sampling frame for a SRS has extra-costs compared with constructing the sampling frame for a TSS (i.e. $c_0 > 0$), the ratios $\frac{c_{srs}}{c_1}$ and $\frac{C}{C-c_0}$ will always be at least one and often larger than one. As a result, the RE can become larger than one, implying that SRS can be less efficient than TSS under the constraint of a fixed budget.

The REs of the optimal TSS1 and TSS3 versus SRS are shown in Figure 3.3, for the general practice list size distribution in England and the public high school size distribution in Italy, and assuming $\left(\frac{c_{srs}}{c_1}\right) = \left(\frac{C}{C-c_0}\right) = 1$ (note that values greater than 1 give a higher RE of TSS versus SRS). Further, Figures S.8-S.10 (S.M.1) show the REs of the optimal TSS1, TSS2, and TSS3 versus SRS for other cluster size distributions. For $\psi = 0$, the RE of any optimal TSS versus SRS is a decreasing function of (i) c_r (Table 3.3), (ii) ρ (at least for $\rho \leq 0.5$, see Figure 3.3, and Figures S.8-S.10 in S.M.1), and (iii), only for TSS2 and TSS3, τ_N (Table 3.3). For $\psi \neq 0$, the patterns remain almost the same as before and the REs also do not seem to vary much across cluster size distributions (Figure 3.3, and Figures S.8-S.10 in S.M.1).

The REs of the three TSS schemes compared with each other (Table 3.3, last three rows) are functions of ρ, c_r, ψ, τ_N, ζ_N, and η_N. The optimal TSS2 is more efficient than the optimal TSS3 since $RE(TSS3\ vs\ TSS2) < 1$ (unless $\tau_N = 0$, Table 3.3, or $\rho_{uN} \approx \pm 1$, Innocenti et al. (2019), since in both cases $RE(TSS3\ vs\ TSS2) = 1$). The REs of TSS2 and TSS3 versus TSS1 are smaller than one, and so the optimal TSS1 is the most efficient TSS scheme, at least for cluster size distributions satisfying the conditions in Table 3.3 (rightmost column), such as all distributions in Table S.7 (S.M.1). For other cluster size distributions, one must compute the RE for that particular distribution to see whether $RE < 1$. However, for $\psi = 0$, the REs in the

last three rows of Table 3.3 are all smaller than one for any cluster size distribution, making TSS1 the most efficient TSS scheme, followed by TSS2. Note that this only holds if informative cluster size ($\psi \neq 0$) is assumed at the design stage, such that in TSS3 cluster means are weighted by cluster size to estimate μ (Table 3.1). If non-informative cluster size ($\psi = 0$) is assumed already in the design stage, then no weighting is needed for TSS3 (Innocenti et al., 2019), and TSS3 then is as efficient as TSS1.

The RE of the optimal TSS2 and TSS3 versus the optimal TSS1 are shown in Figure 3.4, for the general practice list size distribution in England and the public high school size distribution in Italy, and in Figures S.11-S.12 (S.M.1) for other four cluster size distributions. For $\psi = 0$, these reduce to $RE(TSS2\ vs\ TSS1) = \frac{\left(\sqrt{c_r\rho}+\sqrt{1-\rho}\right)^2}{\left(\sqrt{c_r\rho(\tau_N^2+1)}+\sqrt{1-\rho}\right)^2}$ and $RE(TSS3\ vs\ TSS1) = \frac{1}{\tau_N^2+1}$, which are both decreasing functions of τ_N, but $RE(TSS2\ vs\ TSS1)$ also decreases as ρ and/or c_r increases. For $\psi \neq 0$, the patterns are the same as before with two major differences. First, both REs decrease as η_N increases (Table 3.3). Second, for $\psi = 0.35$, both REs differ at most 6% from their values at $\psi = 0$ (Figure 3.4, and Figures S.11-S.12 in S.M.1), except for the English general practice (GP) list size distribution that, having an extreme kurtosis (i.e. $\eta_N = 14.55$), shows a drop in RE (compared with the case $\psi = 0$) larger than 20%. Note that TSS1 is the most efficient design in Figure 3.4 and Figures S.11-S.12 (S.M.1).

Table 3.3. Relative efficiencies of TSS schemes versus SRS and each other for a given budget.

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

Note: Derivations are given in section 2.3 in supplementary material S.M.1. Recall that $V(\hat{\mu})^*$ is the optimal variance in Table 3.2, $c_r = \frac{c_2}{c_1} > 1$, $\psi = \left(\frac{\rho_{uN}^2}{1-\rho_{uN}^2}\right)$, $\left(\frac{c_{srs}}{c_1}\right) \geq 1$ and $\left(\frac{C}{C-c_0}\right) \geq 1$. The conditions for ζ_N in the rightmost column are valid for $\psi \neq 0$ and are satisfied by all distributions in Table S.7 (S.M.1).

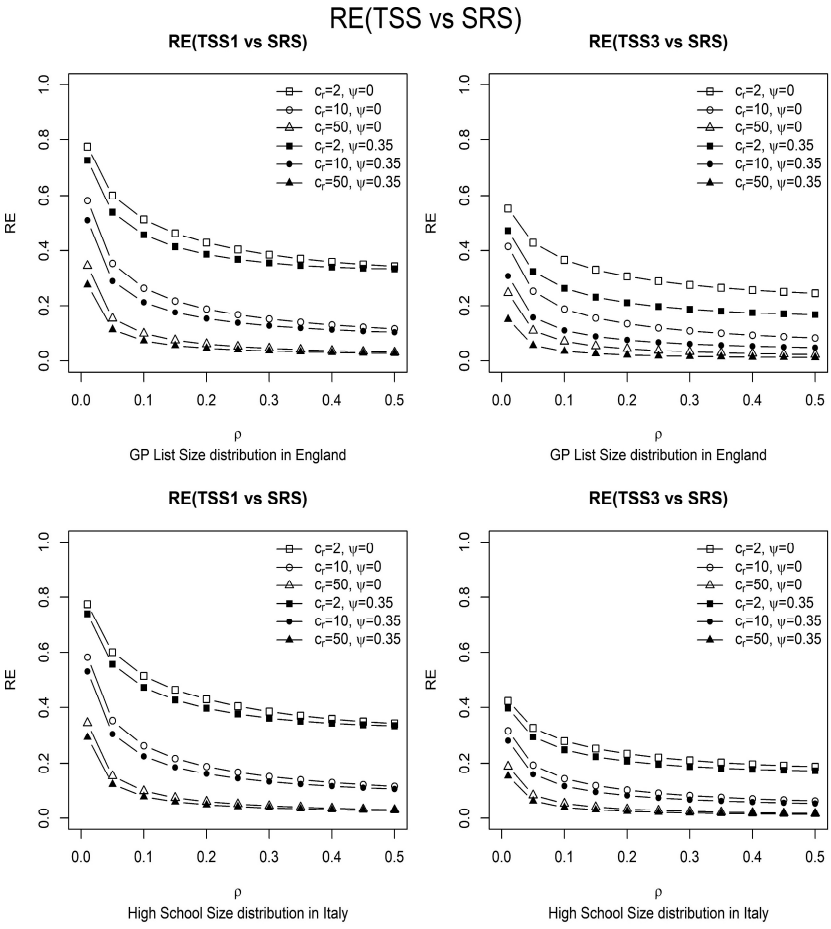


Figure 3.3. Relative efficiency of the optimal TSS1 versus SRS (left column), and of the optimal TSS3 versus SRS (right column), for a given research budget C and assuming $(c_{srs}/c_1) = (C/(C - c_0)) = 1$ (values greater than 1 give a higher RE of TSS versus SRS), as a function of ρ, for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

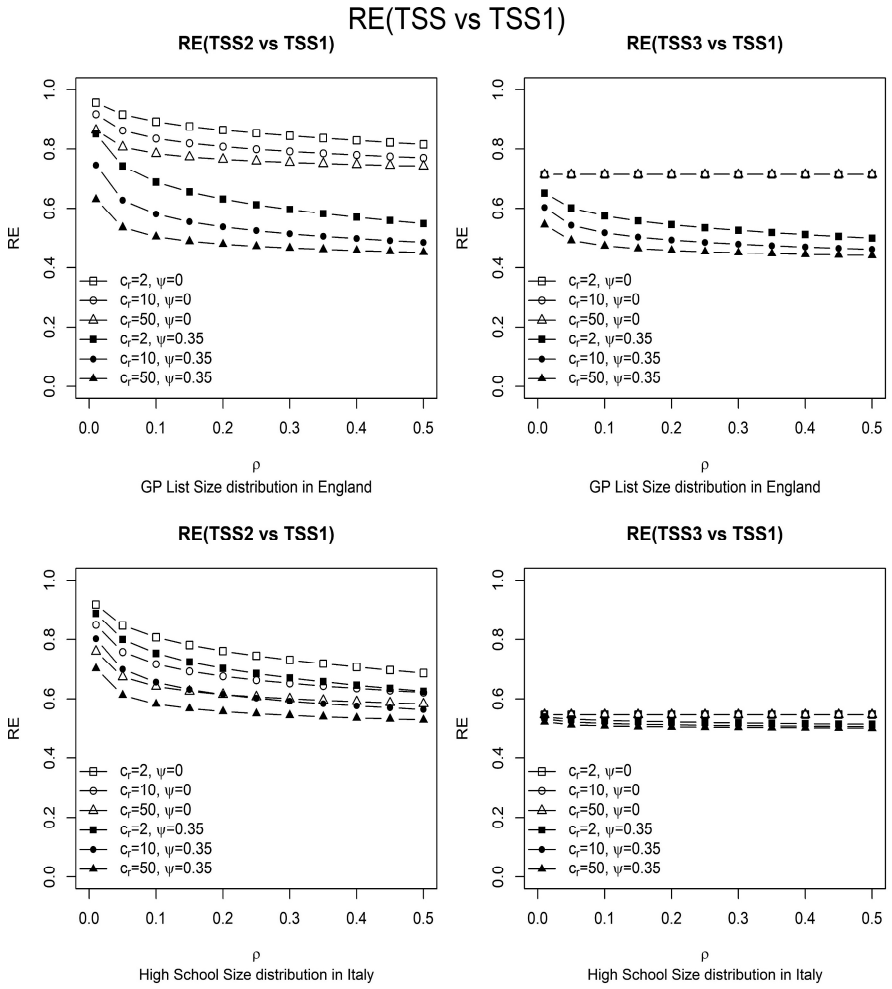


Figure 3.4. Relative efficiency of the optimal TSS2 versus the optimal TSS1 (left column), and of the optimal TSS3 versus the optimal TSS1 (right column), for a given research budget, as a function of ρ, for different values of c_r and ψ (curves), and different cluster size distributions (rows). The cluster size distributions are shown in Figure S.1 (S.M.1). Note that $\psi = 0.35$ corresponds to $\rho_{uN} = \pm 0.51$.

3.4 Maximin design

In section 3.3.2 it has been noticed that the optimal designs in Table 3.2 require a priori knowledge of some nuisance parameters (i.e. ρ, τ_N, ζ_N, η_N, and ψ). This is known as the *local optimality problem* in optimal design literature (Atkinson et al., 2007; Berger & Wong, 2009). Basically, this means that the optimal design is optimal only for certain values of these nuisance parameters. In this paper, the local optimality problem is solved taking a maximin approach (Atkinson et al., 2007; Berger & Wong, 2009; Wong, 1992). This approach has been applied in several contexts, such as longitudinal studies (Ouwens et al,., 2002; Tekle et al., 2008; Winkens et al., 2007), fMRI experiments (Maus et al., 2010), cluster randomized and multicentre trials (Candel & Van Breukelen, 2015; Van Breukelen & Candel, 2018; Wu et al., 2017), cost-effectiveness studies (Manju et al., 2014; Manju et al., 2015), life-event studies (Tan, 2010), test construction (Berger et al., 2000), and biological and pharmacological studies (Dette & Biedermann, 2003; Dette et al., 2006; King & Wong, 2000; Pronzato & Walter, 1988). The maximin approach is composed of the following steps:

1. Define the parameter space, that is, for each unknown parameter (i.e. ρ, τ_N, ζ_N, η_N, and ψ) determine the range of plausible values (e.g. $\rho \in [0,0.30]$).
2. Define the design space, that is, the set of all candidate designs (n, k). In this step, one can rule out those designs that are unfeasible in practice (e.g. too many clusters to cover relative to the time available for data collection), thus preventing sample size adjustments afterwards.
3. For each design (n, k) in the design space, find those values of the nuisance parameters which minimize the efficiency $V(\hat{\mu})^{-1}$ (and thus maximize $V(\hat{\mu})$) within the range of their plausible values, as defined in step 1.
4. Choose that design that maximizes the minimum efficiency obtained in step 3. In other words, choose those values of k and n that minimize $V(\hat{\mu})$ given the worst-case values of the nuisance parameters chosen in step 3.

The resulting design is called the *maximin design*, which is the optimal design for the worst-case scenario, as defined by that set of parameter values chosen in step 3. The advantage of the maximin design is that it not only maximizes the efficiency and the power in the worst-case scenario, but it also guarantees at least that same efficiency and power level for all the other parameter values within the parameter space. Indeed, $V(\hat{\mu})$ is smaller and the power for

hypothesis testing on μ is larger, for all other parameter values than for the worst-case values chosen in step 3, given any fixed sample size (i.e. k and n).

Following the four steps above, we now explain how to find the maximin design for each sampling scheme. The optimal design for TSS1 depends on ρ, τ_N, ζ_N, and ψ. However, to draw a TSS1 sample we need to know the cluster size distribution in the population anyway, which means that τ_N and ζ_N are also known before sampling. Thus, for TSS1, only ρ and ψ are unknown. The maximin design for TSS1 is obtained by plugging into the optimal sample sizes equations (Table 3.2, TSS1 row) the largest realistic values of ρ and ψ (for proofs, see section 3.1 in S.M.1). Unlike for TSS1, when sampling with TSS2 or TSS3 the researcher needs no prior knowledge of the whole cluster size distribution. Indeed, if such information is available, sampling with TSS1 is a better choice (Table 3.3). The maximin design for TSS2 and TSS3 is obtained by plugging into the optimal design equations (Table 3.2) the upper-bounds of the ranges for ρ, ζ_N, η_N, and ψ, and the worst-case value of τ_N (for proofs, see section 3.1 in S.M.1). The latter value can be obtained with an R function given in S.M.2 (section 2), which searches numerically for the value of τ_N that maximizes $V(\hat{\mu})$ (i.e. equations (3.3) and (3.4)) within its range of plausible values, given the worst-case values for ρ, ζ_N, η_N, and ψ. For several upper-bounds for ρ, ζ_N, η_N, and ψ, a numerical evaluation was performed and this always gave $\tau_N = 1$ as worst-case value of τ_N within the range $[0,1]$ (for details, see section 3.2 in S.M.1).

To be on the safe side in sample size planning, one can assume for ρ the parameter range $[0, 0.10]$ in health and medical research (Adams et al., 2004; Eldridge et al., 2004), and $[0, 0.25]$ in educational research (Hedges & Hedberg, 2007; Shackleton et al., 2016). Lacking empirical evidence for ψ or ρ_{uN}, we propose $\psi \in [0, 0.35]$, which corresponds to $\rho_{uN} \in [-0.51, 0.51]$. The range $\tau_N \in [0,1]$ can be justified by considering Table S.7 (S.M.1), and the extreme cases of an exponential cluster size distribution, for which $\tau_N = 1$, and of a binary distribution with half of all clusters having size 2 and the other half having size $2\theta_N - 2$, for which $\tau_N \approx 1$. Finally, for ζ_N and η_N the ranges $\zeta_N \in [0.5, 2]$ and $\eta_N \in [3, 15]$ can be chosen based on Table S.7 (S.M.1). Since $V(\hat{\mu})$ under TSS2 and TSS3 is an increasing function of ζ_N and η_N (at least if $\tau_N \leq 1$, which will usually hold), assuming positive skewness and positive excess kurtosis (i.e. $\eta_N - 3 > 0$) is a safe choice.

As mentioned in section 3.3.1, the optimal design for TSS2 and TSS3 depends on two approximations: the first-order Taylor series approximation used to derive $V(\hat{\mu})$ for TSS2 and TSS3 in Table 3.1, and the large k approximation to simplify the equations in Table 3.1 into

equations (3.3) and (3.4). Since the maximin design is the optimal design for the worst-case scenario, the same approximations also underlie the maximin design. Based on the simulation study and the numerical evaluation discussed in S.M.1 (sections 1 and 3.3), it turned out that each approximation induces a bias of at most 5% in the $V(\hat{\mu})$ used to derive the optimal/maximin design if the optimal/maximin $k^{MD} \geq 20$, or, for $\zeta_N \geq 2$ and $\eta_N \geq 9$, $k^{MD} \geq 100$. Since $V(\hat{\mu}) \propto \frac{1}{k}$, a simple solution is to increase the maximin k^{MD} with 10% to ensure sufficient power at the expense of a 10% higher budget C. However, if the maximin $k^{MD} < 20$ or (for $\zeta_N \geq 2$ and $\eta_N \geq 9$) $k^{MD} < 100$ both approximations are biased by more than 5%. A solution is to first increase C such that maximin $k^{MD} \geq 20$ or (for $\zeta_N \geq 2$ and $\eta_N \geq 9$) $k^{MD} \geq 100$, and then further increase C by 10%.

3.5 Sample size calculation for cross-population comparisons

The results of the previous sections allow to efficiently plan a survey not only for estimating a mean, but also for comparing different populations, if the samples are independent. An example of such a study is the ESPAD study (ESPAD Group, 2016), which compares substance use among 15/16 year old students across 35 European countries. For a fixed separate budget per population, the optimal design per population is given in Table 3.2 and the maximin design in section 3.4. However, the design can be further optimized by constraining the total budget (i.e. the sum of the separate budgets) instead of each separate budget and finding the optimal (or maximin) budget split between populations (for details, see section 4 of S.M.1). For the case of comparing two populations, this optimization was formalized into a procedure to compute maximin sample sizes per population and the maximin budget split between populations, obtained by extending Van Breukelen and Candel (2018) to TSS1 with informative cluster sizes and different cluster size distributions per population. This procedure for comparing two populations is implemented in an R code given in section 4 in S.M.2. To use this program, the researcher needs to specify c_1 and c_2 per population, τ_N and ζ_N of the cluster size distribution of each population, the largest plausible values for ρ and ψ, a range for the ratio of the outcome standard deviations (σ_y) between the two populations, the smallest difference $\mu_F - \mu_I$ that is worthwhile being detected, the maximum sum of outcome variances in both populations V_{max}, the power level $1 - \gamma$, and the type I error rate α. The R code (S.M.2, section 4) returns the maximin sample sizes per population and the maximin budget split. The steps of this procedure

are given in S.M.1 (section 4). This procedure is presented only for TSS1, because it is the most efficient sampling scheme for many cluster size distributions.

Let us demonstrate the procedure with the following example. Suppose that we want to plan a survey to estimate and compare the average alcohol consumption among high school students between France and Italy. Similar to the ESPAD study (ESPAD Group, 2016), alcohol consumption Y_{ij} is measured as the average volume of ethanol (in centilitres) consumed on the last drinking day. Based on adolescent health literature, at the design stage, school size (i.e. total number of students) can be assumed to be informative, that is, related to alcohol consumption. Indeed, it has been found that school size and school connectedness, broadly defined as the degree of belonging at school, are inversely related (McNeely et al., 2002; Thompson et al., 2006), as well as school connectedness and alcohol use (Resnick et al., 1997). TSS1 is the most efficient two-stage sampling scheme for both high school size distributions (this can be verified by checking the conditions in the rightmost column of Table 3.3, with the numbers given in the second and third row of Table S.7 of S.M.1), and so it is chosen for both populations. Suppose that we want to test the null hypothesis H_0 that $\mu_F = \mu_I$ against the alternative hypothesis H_1 that $\mu_F \neq \mu_I$, where μ_F and μ_I are the population means of alcohol consumption in France and Italy, respectively. Since the French and the Italian samples are independent, we can apply the procedure above to determine how many schools and how many students per school one has to sample per country, and how to split the total budget between countries.

The results are shown in Table 3.4 for four different cost scenarios. Two largest plausible values are assumed for ρ and ψ, respectively, $\rho(max) = \{0.1,0.2\}$ and $\psi(max) = \{0,0.35\}$. This combination of costs and model parameters (Table 3.4, first six columns) gives a total of $4 \times 2 \times 2 = 16$ scenarios, each corresponding to a row in Table 3.4. The seventh column in Table 3.4 gives the maximin budget split $\frac{C_F}{C_I}$ (i.e. the ratio of the budget for France, C_F, to that for Italy, C_I), and from the eighth to the eleventh column the maximin sample sizes per country are shown. Finally, the rightmost column of Table 3.4 shows the total budget required to detect a standardized difference of medium size ($d = \frac{\mu_F - \mu_I}{\sqrt{\frac{V_{max}}{2}}} = 0.5$), with 90% power using a two-tailed test with $\alpha = 0.05$. From Table 3.4, it can be seen that the maximin n^{MD} per country is an increasing function of c_r, a decreasing function of ρ and ψ, and is inversely related to the maximin k^{MD}. Furthermore, the maximin budget split $\frac{C_F}{C_I} = 1$ only for $\psi = 0$ and

homogeneous costs ($c_{1,F} = c_{1,I}$ and $c_{2,F} = c_{2,I}$). In all other scenarios $\frac{C_F}{C_I} < 1$, meaning that more budget is allocated to the Italian sample than to the French sample. Given that $\rho(max)$ and $\psi(max)$ are the same for both countries, $\frac{C_F}{C_I} < 1$ because (i) sampling a student is more expensive in Italy than in France ($c_{1,F} < c_{1,I}$), or (ii) sampling a school is more expensive in Italy than in France ($c_{2,F} < c_{2,I}$), or (iii), only for $\psi = 0.35$, the school size distribution in Italy is such that $\tau_N(\zeta_N - \tau_N)$ is larger than in France (see Tables S.7 and S.9 of S.M.1). Finally, the total budget C required for the desired power is larger for $\psi = 0.35$ than for $\psi = 0$ (Table 3.4, rightmost column), suggesting that ignoring informative cluster size at the design stage has the consequence of determining a research budget which is too low for the desired power level. Specifically, informative cluster size requires C to increase with 23-32% depending on the scenario (the larger ρ and/or $c_{2,I}$, the larger this relative increase, see Table 3.4, rightmost column).

Table 3.4. Maximin design (n_F^{MD}, k_F^{MD}, n_I^{MD}, k_I^{MD}) and budget C needed to detect a standardized difference of medium size ($d = 0.5$) with a power of 90% using a two-tailed test with $\alpha = 0.05$ and assuming $\frac{\sigma_{y,F}}{\sigma_{y,I}} \in \left[\frac{1}{3}, 3\right]$, as a function of the maximum ψ, the maximum ρ, the cost per individual in France $c_{1,F}$ and in Italy $c_{1,I}$, and the cost for sampling a cluster in France $c_{2,F}$ and in Italy $c_{2,I}$.

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

3.6 Discussion

To estimate an overall mean, Two-Stage Sampling is a logistically convenient way to collect data from a multilevel population. In practice, resources (time and money) for sampling are limited. Thus, this paper presents optimal sample sizes per design stage that either maximize the precision of the population mean estimate for the available research budget, or minimize the research budget for the required precision for estimation. Such optimal designs were derived for three TSS schemes: sampling clusters with probability proportional to cluster size, and then the same number of individuals per cluster (TSS1); sampling clusters with equal probability, and then the same percentage of individuals per cluster (TSS2); sampling clusters with equal probability, and then the same number of individuals per cluster (TSS3).

The optimal sample size equations were derived allowing cluster size to be informative, that is, to be related to the outcome variable of interest. It turned out that the optimal designs given in Table 3.2 are quite robust against misspecification of the degree of informativeness of cluster size ψ. As shown in section 3.3.2 and in Table S.8 (S.M.1), the relative efficiency of the optimal TSS1 assuming $\psi = 0$ (i.e. non-informative cluster size) versus the optimal TSS1 assuming $\psi > 0$ (i.e. informative cluster size), when the true $\psi > 0$, was close to one. Nevertheless, ignoring informative cluster size is risky for two reasons. First, assuming $\psi = 0$ one would be tempted to combine the unweighted average of cluster means with TSS3, because this strategy (i.e. combination of sampling scheme and estimator) is unbiased and efficient for $\psi = 0$. However, this strategy is biased and inefficient if the true $\psi > 0$. Thus, assuming $\psi > 0$ is always prudent because it leads to combining the unweighted average of cluster means with TSS1, that is, choosing a strategy which is unbiased and highly efficient both for informative and non-informative cluster size. Second, assuming $\psi = 0$ can lead to underestimating the research budget for the desired power level, because the research budget is an increasing function of ψ (see Figure 3.2, and Table 3.4, rightmost column). This applies not only to TSS1, but also if, because of practical constraints, one has to choose TSS2 or TSS3 as a sampling scheme. For these two reasons, we recommend assuming $\psi > 0$ at the design stage of the survey.

The optimal designs of the three TSS schemes were compared with each other and with SRS under the constraint of a fixed budget. In contrast to what was the case under the constraint of a fixed total sample size (Innocenti et al., 2019), SRS can be less efficient than TSS, because

it is more expensive to construct a sampling frame of all individuals in the population than of those from the selected clusters only ($c_0 > 0$), and because it is more costly to sample and measure geographically dispersed individuals than those that are grouped in a natural cluster (e.g. school, general practice) ($c_{srs} > c_1$). Under informative cluster size, the optimal TSS1 was shown to be the most efficient sampling scheme for many cluster size distributions, followed by TSS2, and then TSS3. We thus recommend TSS1, provided all cluster sizes are known before sampling.

The optimal design depends on several unknown parameters (i.e. the intraclass correlation ρ, the informativeness parameter ψ, and the cluster size distribution's coefficient of variation τ_N, skewness ζ_N, and kurtosis η_N). To address this issue the maximin approach was proposed. For the considered TSS schemes, this strategy consists of plugging the worst-case value for each unknown parameter into the optimal design equations in Table 3.2. For ρ, ψ, and η_N, the largest plausible value is the worst-case value. If all plausible values for $\tau_N \leq 1$, then the largest plausible value for ζ_N is also the worst-case value. The worst-case value for τ_N can be obtained with an R code, given in S.M.2 (section 2). However, a numerical evaluation showed that if the largest plausible value for τ_N is 1, this is the worst-case value for τ_N. The R code also returns the worst-case value for ζ_N in the rather unrealistic case that some plausible values for $\tau_N > 1$. The maximin approach has the advantages of being relatively simple to implement, and being robust against misspecification of the unknown parameters by maximizing the minimum efficiency over the ranges of their plausible values. An alternative approach is to obtain estimates of the nuisance parameters from a pilot study and use these in the sample size calculation. However, ρ risks to be underestimated (and thus the main survey to be under-powered), unless the pilot study samples a large number of clusters and of individuals per cluster, which means a sizeable portion of the limited resources for the main survey has to be devoted to the pilot study (Eldridge et al., 2016). The underestimation is likely to be even more severe for skewness and kurtosis, given that their traditional estimators are biased downwards unless the sample size is large or (only for the skewness) cluster size is normally distributed (Joanes & Gill, 1998). For all these reasons, we recommend the maximin approach. Relatedly, to improve the planning of future surveys, empirical studies should report values of these nuisance parameters like in Table S.7 (S.M.1).

The results of this paper also allow to efficiently plan surveys for comparing different populations, provided the samples are independent. For TSS1, a procedure to derive maximin

sample sizes and maximin budget split between populations was obtained by extending Van Breukelen and Candel (2018)'s findings to informative cluster size. Analogous extensions for TSS2 and TSS3 could be explored. However, when either cluster size is non-informative ($\psi = 0$), or the cluster size distribution as well as the informativeness parameter α_1 are the same in both populations (e.g. treated and control groups in a cluster randomized trial), we have that $\mu_F - \mu_I = \beta_{0,F} - \beta_{0,I}$ (see equation (3.2)) and then the equations given in this paper reduce to simpler expressions as also derived by Van Breukelen and Candel (2018) (i.e. those for TSS1 with $\psi = 0$).

Finally, in this paper the model-based approach to survey sampling was adopted. However, the results of this paper are valid also under the design-based approach, provided model (3.1) and assumption 4 hold and inference is then based on the sampling scheme (Innocenti et al., 2019). Future research could extend the results of this paper by considering dichotomous outcomes, three-level populations, and by deriving the optimal design for longitudinal studies to monitor trends.

Appendix: Notation, and table of contents of the online supplementary materials

The notation used in the main text is listed in Table 3.A, which provides for each symbol the section where it was introduced for the first time, and its definition in words. Table 3.B shows the table of contents of the online supplementary materials, and the link to download them.

Table 3.A. Notation used in the main text.

Section	Symbol	Definition
3.2	K	Number of clusters in the population
	j	Index for clusters
	N_j	Size of cluster j in the population
	$N_{pop} = \sum_{j=1}^{K} N_j$	Population size
	k	Number of clusters in the sample
	$\theta_N = \frac{N_{pop}}{K}$	Population mean of cluster size
	σ_N^2	Population variance of cluster size
	$\tau_N = \frac{\sigma_N}{\theta_N}$	Population coefficient of variation of cluster size
	$\zeta_N = \frac{E\left[\left(N_j - \theta_N\right)^3\right]}{\sigma_N^3}$	Population skewness of cluster size
	$\eta_N = \frac{E\left[\left(N_j - \theta_N\right)^4\right]}{\sigma_N^4}$	Population kurtosis of cluster size
	$\bar{N} = \frac{\sum_{j=1}^{k} N_j}{k}$	Average population size of the sampled clusters
	m	Number of individuals sampled with SRS
	i	Index for individuals
	Y_{ij}	Outcome variable of interest
	ε_{ij}	Effect of individual i in cluster j
	σ_ε^2	Population variance of ε_{ij}
	ν_j	Component of cluster effect that does not depend on cluster size
	σ_ν^2	Population variance of ν_j

Section	Symbol	Definition
	β_0	Average of all cluster-specific means in the population
	α_0	Intercept of the relation between cluster effect and cluster size
	α_1	Slope of the relation between cluster effect and cluster size
	$u_j = \alpha_0 + \alpha_1 N_j + \nu_j$	Effect of cluster j
	$E(u_j) = 0$ and $V(u_j) = \sigma_u^2 = \sigma_\nu^2 + \alpha_1^2\sigma_N^2$	Population mean and variance of u_j
	μ	Average of all individual outcomes in the population
	$\hat{\mu}$	Population mean estimator
	$V(\hat{\mu})$	Sampling variance of $\hat{\mu}$
	$\rho_{uN} = \frac{E[u_j(N_j - \theta_N)]}{\sqrt{\sigma_\nu^2 + \alpha_1^2\sigma_N^2}\sigma_N}$	Correlation between u_j and N_j
	$\psi = \left(\frac{\rho_{uN}^2}{1-\rho_{uN}^2}\right)$	Degree of informativeness of cluster size
	$\sigma_y^2 = \sigma_\nu^2 + \sigma_\varepsilon^2$	Total unexplained outcome variance
	$\rho = \frac{\sigma_\nu^2}{\sigma_y^2}$	Intraclass correlation coefficient
	π_j	Inclusion probability of cluster j
	$\pi_{i\|j}$	Conditional inclusion probability of individual i
	n_j	Number of individuals sampled per cluster for TSS2
	$p = \frac{n_j}{N_j}$	Proportion of individuals sampled per cluster for TSS2
	$\bar{n} = \frac{\sum_{j=1}^{k} n_j}{k}$	Average sample size of the sampled clusters
	n	Number of individuals sampled per cluster for TSS1 and TSS3. Expected value of $\bar{n}$ for TSS2
	π_i	Inclusion probability of individual i under SRS
3.3.1	C	Budget for sampling and measuring
	c_2	(Average) cost for sampling a cluster
	c_1	(Average) cost for sampling an individual from a sampled cluster

Section	Symbol	Definition
	c_0	Extra-cost due to constructing the sampling frame for SRS compared with the sampling frame for TSS
	c_{srs}	(Average) cost for sampling an individual directly from the population
	$V(\hat{\mu})^*$	Sampling variance of $\hat{\mu}$ under the optimal design
	n^*	Optimal number of individuals per cluster
	k^*	Optimal number of clusters
	$p^* = \frac{n^*}{\theta_N}$	Optimal proportion of individuals to sample per cluster for TSS2
	$c_r = \frac{c_2}{c_1}$	Cluster-to-individual cost ratio
3.3.2	α	Type I error rate
	γ	Type II error rate
	z_q	qth percentile of the standard normal distribution
	$g(\rho, \psi)$	Numerator of $V(\hat{\mu})^*$ in Table 3.2 excluding σ_y^2
	μ_0	Value of μ under H_0
	$d_0 = \frac{\mu - \mu_0}{\sigma_y}$	Standardized difference for one-sample t-test
3.3.3	$RE(D1\ vs\ D2) = \frac{V_{D2}(\hat{\mu})^*}{V_{D1}(\hat{\mu})^*}$	Relative efficiency of optimal design $D1$ versus optimal design $D2$
3.4	k^{MD}	Maximin number of clusters
3.5	$\mu_F, \mu_I, \sigma_{y,F}^2, \sigma_{y,I}^2$	Population mean and total unexplained outcome variance in France (F), and Italy (I)
	$V_{max} \geq \sigma_{y,F}^2 + \sigma_{y,I}^2$	Maximum plausible upper-bound for $\sigma_{y,F}^2 + \sigma_{y,I}^2$
	$\rho(max)\ and\ \psi(max)$	Largest plausible values assumed for ρ and ψ
	$d = \frac{\mu_F - \mu_I}{\sqrt{\frac{V_{max}}{2}}}$	Standardized difference for unpaired two-sample t-test
	n^{MD}	Maximin number of individuals per cluster

Table 3.B. Table of contents of the online supplementary materials, and link to download them.

	Contents
Supplementary Material 1 (S.M.1)	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 1.1. Cluster size distributions 1.2. Simulation procedure 1.3. Simulation results 2. Optimal design and relative efficiencies for a given budget 2.1. Sampling variances 2.2. Derivation of the optimal design 2.3. Relative efficiencies 3. Local optimality problem and maximin design 3.1. Derivation of the maximin design 3.2. A general program in R to find the maximin parameter values 3.3. Accounting for the approximations in the maximin design for TSS2 and TSS3 4. Sample size calculation for cross-population comparisons
Supplementary Material 2 (S.M.2)	1. R code of the simulation study 2. R code to find the maximin parameter values when estimating one population mean 3. R code to numerically evaluate the ratio between $V(\hat{\mu})_{\Delta}$ and $V(\hat{\mu})_{L}$ (i.e. section 3.3, Supplementary Material 1) 4. R code to compute the maximin TSS1 design for comparing two population means

Both supplementary materials can be downloaded at the following link:

https://journals.sagepub.com/doi/suppl/10.1177/0962280220952833

Chapter 4

Sample size calculation and optimal design for regression-based norming of tests and questionnaires

This chapter has been accepted for publication in *Psychological Methods*, with co-authors Frans E.S. Tan, Math J.J.M. Candel, and Gerard J.P. van Breukelen.

Chapter 5

Sample size calculation and optimal design for multivariate regression-based norming

This chapter has been submitted for publication with co-authors Math J.J.M. Candel, Frans E.S. Tan, and Gerard J.P. van Breukelen

Chapter 6

Discussion

The focus of this thesis has been on the efficient design of two types of observational studies, namely surveys for estimating means in multilevel populations and normative studies to derive norms for tests and questionnaires. As seen in the previous chapters, these two types of studies are quite different, and their design poses different challenges. However, two common challenges can be pointed out.

The first common challenge comes from the use of models. On the one hand, models are extremely useful in planning these studies because they allow to derive designs that yield maximum efficiency (i.e. the optimal design). On the other hand, the choice of the model is crucial, because a design that is optimal under a certain model can be sub-optimal under other models. For instance, two-stage sampling scheme TSS3 is optimal under non-informative cluster size but is the least efficient TSS scheme under a linear relation between cluster size and cluster effect (see chapter 3). As another example, the balanced three equidistant age levels design is optimal for test norming in case of a quadratic age effect but sub-optimal for a linear effect (see chapters 4 and 5). Hence, one has to find a trade-off between efficiency and robustness in planning these studies. In this thesis, designs that are a trade-off between efficiency and robustness have been derived with the maximin approach. This strategy is relatively simple to implement, and maximizes efficiency (or relative efficiency) under the worst-case scenario.

The second common challenge is the derivation of the sampling variance of the statistic of interest (i.e. mean estimator in TSS, Z-score, percentile rank score, and Mahalanobis distance) needed to derive the optimal design. Most of the sampling variance formulas in this

thesis were derived with the delta method, that is, were based on approximations, and simulation studies were needed to assess the bias induced by these approximations (see chapters 3, 4, and 5). From these simulation studies, lower-bounds for the sample size were obtained, that is, values of the sample size below which the sampling variance formula is not accurate (i.e. the bias is too large). These lower-bounds for the sample size should be taken into account in sample size calculation.

In the next section, some practical guidelines for designing surveys and normative studies are presented.

6.1 Guidelines

The main results of this thesis can be summarized in terms of a few practical guidelines on how to design each type of study.

Guidelines for designing a survey for mean estimation in multilevel populations

1. To decide on the best two-stage sampling scheme: Is the cluster size distribution in the population known before sampling?
    a. Yes, then sample with TSS1, and estimate the population mean with the unweighted average of cluster means.
    b. No. Based on prior knowledge, can informative cluster size be ruled out?
        i. Yes, then sample with TSS3, and estimate the population mean with the unweighted average of cluster means.
        ii. No, then sample with TSS2, and estimate the population mean with the average of cluster means weighted by cluster size.
2. To compute the required sample size (i.e. number of clusters, number of individuals per cluster): Use the maximin design instead of the optimal design. For each TSS scheme, detailed guidelines on the range of plausible values for each unknown model parameter can be found in section 3.4.

These guidelines are shown in the form of a decision tree in Figure 6.1

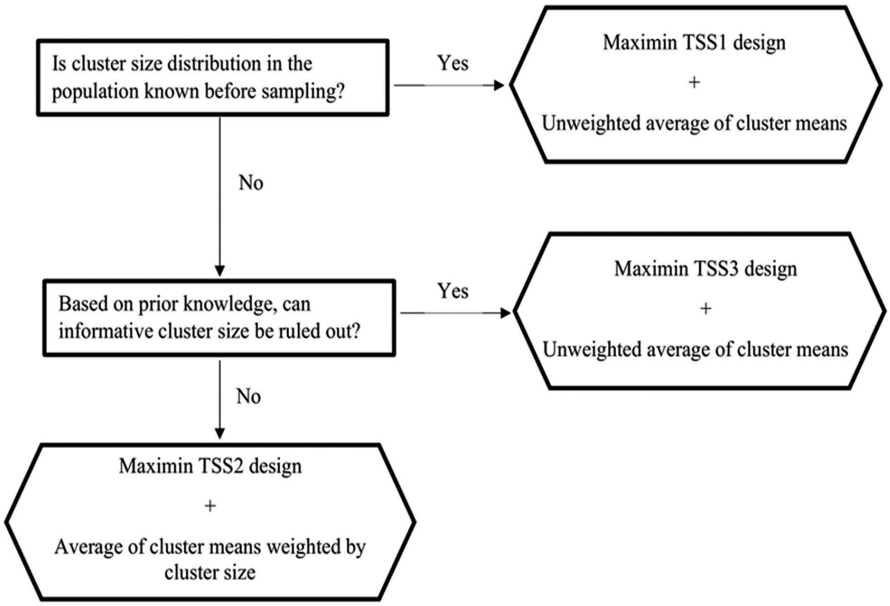


Figure 6.1. Decision tree on the best strategy (i.e. most efficient sampling scheme and unbiased mean estimator) to estimate the average of all individual outcomes in a two-level population with two-stage sampling.

Guidelines for designing a normative study

1. To decide upon the analysis technique for deriving norms: How many tests are to be normed?
    a. One, then use univariate regression (see chapter 4).
    b. More than one, then use multivariate regression (see chapter 5), because it allows to take into account the correlation between test scores of the same individual in testing hypotheses on the regression coefficients (Johnson & Wichern, 2007), and allows to compare individuals with the reference population based on the Mahalanobis distance.
2. In the choice of the predictors and of the regression model, the following results should be kept in mind:
    a. A model without interactions and a model including all possible interactions have the same optimal design (for two predictors, see chapters 4 and 5, for an arbitrary number of predictors, see Schwabe (1996)). However, the required sample size for the optimal design is an increasing function of the number of regression weights in

the model, and that number depends on whether interactions are included into the model (see chapters 4 and 5).

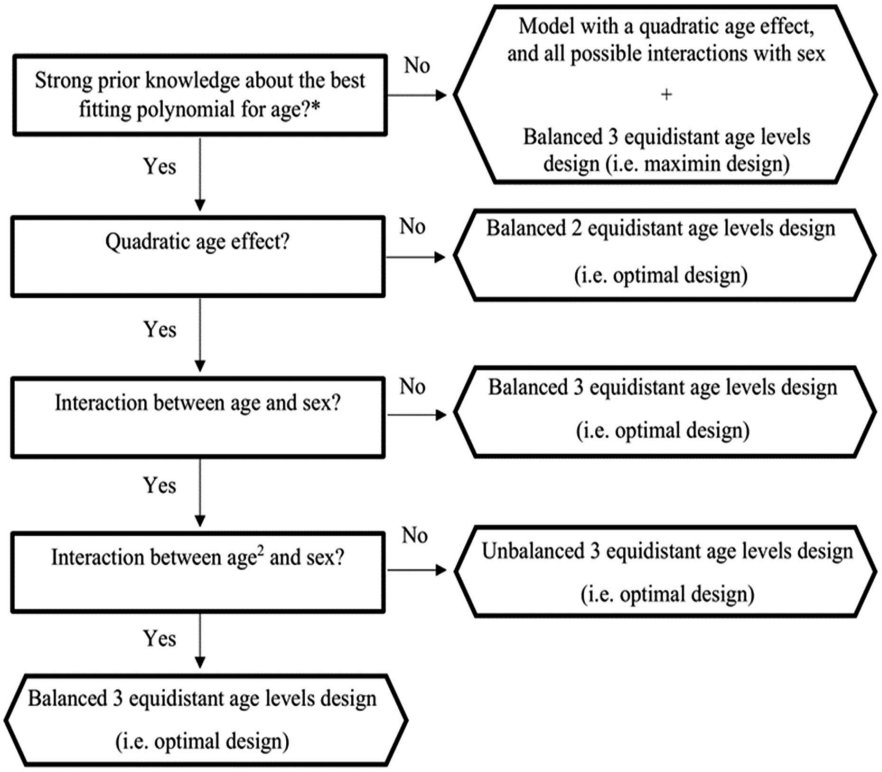
b. For quantitative predictors (e.g. age, number of years of education): The optimal design for a polynomial effect of order h (e.g. $h = 2$ for a quadratic effect) consists of (not necessarily equidistant) $h + 1$ levels only (Berger & Wong, 2009, p. 67). Including additional levels of the predictor into the normative sample yields a loss of efficiency, under the constraint of a fixed total sample size (see chapter 4).

c. For categorical predictors (e.g. sex, educational level): Increasing the number of categories increases the required sample size for a given optimal design (see chapter 4).

3. If there is uncertainty about the regression model for norming (e.g. about the best fitting polynomial for a quantitative predictor, or about how many interactions between predictors are nonzero), use the maximin design instead of the optimal design.

In Figure 6.2, these guidelines are applied to the univariate and multivariate regression models considered in chapters 4 and 5, which have only a quantitative predictor (e.g. age) and a categorical predictor (e.g. sex).

In the next section, ideas for future research are discussed.



**Note that in the 65 regression-based normative studies reviewed in supplement chapter 4, either a linear or a quadratic age effect was assumed. Hence, this decision tree is limited to at most a quadratic age effect.*

Figure 6.2. Decision tree on the choice of the regression model for norming and the corresponding best design (for details, see chapters 4 and 5). This decision tree is valid for univariate and multivariate regression models with only a quantitative predictor (e.g. age) and a categorical predictor (e.g. sex), as in chapters 4 and 5.

6.2 Ideas for future research

The work presented in this thesis can be extended in several ways. The results in chapters 2 and 3 could be extended to three-level populations (e.g. students nested within classes nested within schools), which raises two issues: (i) cluster size can be informative not only at the first level (i.e. number of students per class) but also at the second level (i.e. number of classes per school), and (ii) several three-stage sampling schemes need to be compared in terms of efficiency. Furthermore, the results of chapters 2 and 3 were restricted to a linear effect of cluster size on the outcome variable of interest, so the present results could be extended to account for higher order polynomial effects of cluster size. Another possible extension could be to consider without-replacement sampling, that is, to relax the assumption that the number of clusters in the population is very large relative to the number of sampled clusters. However, the results of chapters 2 and 3 could still be relevant, though as a conservative approach because with-replacement variances are bigger than without-replacement variances (Lohr, 2010). Furthermore, future research could deal with two-stage sampling schemes for estimating multiple means in a multilevel population, because surveys often target several variables (e.g. body mass index, cholesterol level, blood pressure). Optimal designs for such multipurpose surveys could be derived with a multivariate criterion, such as the volume of the 95% confidence ellipsoid for the population means of interest (i.e. D-optimality). Another interesting extension of chapters 2 and 3 could be the derivation of optimal designs for estimating a prevalence (i.e. the mean of a binary outcome) with the three considered TSS schemes. However, previous research on optimal designs for multilevel logistic models has shown that the optimal design for these models depends on the unknown regression coefficients (Abebe et al., 2015; Moerbeek et al., 2001; Tekle et al., 2008), so one can expect that the local optimality problem in optimal TSS designs for prevalence estimation will be more complicated than that it was for continuous outcomes.

Chapters 4 and 5 focused on test norming in one-level populations, so future research could extend the results of this thesis to multilevel populations. This extension is particularly relevant for normative studies of educational measures (e.g. arithmetic or language tests) in children, because they are nested within schools. Investigating the effect of informative cluster size on test norming could also be interesting. Furthermore, normality and homoscedasticity of the regression residuals were assumed in chapters 4 and 5, but these assumptions are not always realistic. Promising approaches to norming in such situations seem to be the use of generalized

additive models for location, scale, and shape (GAMLSS), which allow to model a wide range of test score distributions (see Voncken et al., 2019a, 2019b), and quantile regression that allows to directly model conditional percentiles of the test score of interest (see Sherwood et al., 2016). Optimal designs for model-based norming with GAMLSS models or quantile regression could be an interesting extension. Furthermore, future research could derive optimal designs for the models in chapters 4 and 5 using the I-optimality criterion, which minimizes the average of the standardized prediction variance over the design region (Goos & Jones, 2011), instead of the G-optimality criterion, which minimizes the maximum of the standardized prediction variance over the design region (see chapters 4 and 5). However, if the norms are expected to be mainly used for individuals at the extreme (or close to the extreme) of the quantitative predictor (e.g. neuropsychological tests to detect dementia for persons with or close to the highest age), minimizing the maximum of the standardized prediction variance is more approriate. Finally, in chapter 5 the Mahalanobis distance-based approach was compared with two classification rules (i.e. the conjunctive and the disjunctive rules) to combine Z-scores of the same individual under Van der Elst et al. (2017)'s multivariate approach. The comparison has shown that the Mahalanobis distance-based approach defines "abnormality" as unlikely performance (e.g. a high score on one test and a low score on the other test, when the two tests are positively correlated), whereas the conjunctive and the disjunctive rules identify "abnormality" with extreme performance (e.g. a high score on one or both tests). To detect this latter type of "abnormality" with the Mahalanobis distance-based approach one can assume zero correlation or use a lower percentile than the 95^{th} percentile of the cumulative distribution of the Mahalanobis distance as cut-off for decision making (for details, see chapter 5). However, this comparison was limited to two tests, so it could be interesting to extend it to more than two tests. As the number of tests P increases, the conjunctive rule could become too conservative as definition of "abnormality" (e.g. a subject with $P-1$ extreme Z-scores is classified as "normal" under this rule), while the disjunctive rule could become too liberal (e.g. a subject with only one extreme Z-score out of P is classified as "abnormal" under this rule). Consequently, the Mahalanobis distance-based approach, which combines the P test scores into a single norm score and takes into account the pairwise correlations between Z-scores, could become more appealing as P increases. Furthermore, the derivation of sample size calculation formulas for the conjunctive and disjunctive rules, which are affected by multiple testing issues, is an interesting topic for future research.

Chapter 7

Summary

This thesis deals with sample size planning and optimal design for two types of observational studies: (1) surveys for mean estimation in multilevel populations, such as school-based surveys for estimating mean alcohol consumption among high school students (see, for instance, ESPAD Group, 2016), and (2) normative studies to derive reference values for tests and questionnaires, such as neuropsychological tests to assess information processing speed (see, for instance, Van der Elst et al., 2006a), and clinical questionnaires to measure patients' orientation toward chronic pain (see, for instance, Van Breukelen & Vlaeyen, 2005).

Chapter 1 provides an introduction to these two types of studies. Specifically, the practical importance of these studies is explained, real-life examples are provided, the main results available in the literature are summarized, and the statistical models for analysing data obtained with these studies are introduced. Furthermore, a definition of optimal design for each type of study is given, and strategies to find robust designs are presented. Chapter 1 ends with an outline of the thesis.

Chapter 2 is on unbiased and efficient estimation of the average of all individual outcomes in two-level populations, with either simple random sampling (SRS) of individuals (i.e. individuals are drawn directly from the population) or two-stage sampling (TSS) (i.e. first clusters are sampled, and then individuals are sampled from the selected clusters). Cluster sizes are allowed to vary and to be related to the outcome variable of interest (i.e. cluster size is informative). Three TSS schemes are considered: sampling clusters with probability proportional to cluster size and then taking a SRS of the same number of individuals within each sampled cluster (TSS1); drawing a SRS of clusters and then sampling the same percentage of individuals per cluster (TSS2); taking a SRS of clusters and then the same number of individuals per sampled clusters (TSS3). In this chapter, it is shown that the average of all

individual outcomes and the average of all cluster-specific means (i.e. the two definitions of population means in a two-level population) coincide only if cluster sizes are either equal or non-informative. Unbiased estimation of the average of all individual outcomes is discussed under each sampling scheme. Furthermore, the three TSS schemes are compared in terms of efficiency with each other and with SRS of individuals, under the constraint of a fixed total sample size. The relative efficiency of the sampling schemes is shown to vary across different cluster size distributions. However, TSS1 is the most efficient TSS scheme for many cluster size distributions. Model-based and design-based inference are compared and are shown to give similar results, at least if the model assumptions are met. The results of this chapter are applied to two real-life cluster size distributions, that is, the distribution of high school size in Italy, and the distribution of patient list size in England.

Chapter 3 deals with optimal TSS schemes for mean estimation in two-level populations, when cluster size is informative. A simulation study is performed to assess the bias in the sampling variance formulas derived for TSS2 and TSS3 in chapter 2, because these variances are based on approximations. Optimal sample size equations are derived for each TSS scheme considered in chapter 2. The optimal design is the number of clusters and number of individuals per cluster that minimizes the sampling variance of the population mean estimator, subject to a cost constraint. The consequences for the optimal design of ignoring informative cluster size are investigated. It turns out that the optimal TSS designs are quite robust against misspecification of the degree of informativeness of cluster size, but assuming non-informative cluster size can lead to serious underestimation of the required research budget for the desired power level. Furthermore, the three optimal TSS schemes are compared, in terms of efficiency, with each other and with SRS of individuals, under the constraint of a fixed budget for sampling and measuring. The optimal TSS1 is shown to be the most efficient sampling scheme for many cluster size distributions. To overcome the dependency of the optimal designs on the prior knowledge of some model parameters, maximin designs are derived for each TSS scheme. Finally, a procedure is proposed to derive maximin sample sizes and a maximin budget split between two surveys to estimate and compare the means of two populations with TSS1. This procedure is illustrated when planning a hypothetical survey to compare adolescent alcohol consumption between France and Italy, using the real distributions of high school size in these two countries.

Chapter 4 focuses on normative studies to derive reference values or norms for tests and questionnaires. In this chapter, the regression-based approach to norming is adopted, because it

has three advantages over the traditional approach of norming per subgroup defined in terms of age and sex (or other demographic variables): (i) it is more efficient, that is, it requires a smaller sample size, (ii) it allows to identify the predictors that affect the test score of interest, and (iii) it allows to derive the optimal design, which is defined as the joint distribution of scores on the predictors that minimizes the sampling variance of the norm statistic under the assumed norming model. In chapter 4, sampling variance formulas are derived for commonly used norm statistics, that is, Z-scores and percentile rank scores. Since these variance formulas are based on approximations, two simulation studies are performed to assess the bias induced by these approximations. These sampling variance formulas are used to derive optimal designs for five regression models including a quantitative and a qualitative predictor, differing in whether they allow for interaction and nonlinearity. Efficient designs that are robust against misspecification of the norming model are derived using the maximin strategy with efficiency and relative efficiency as criteria. It is shown that, for the five considered regression models, the most robust designs obtained under these two criteria are the same. Furthermore, for the optimal design formulas are proposed to determine the required size of the normative sample for each norm statistic (i.e. Z-score and percentile rank score). These formulas can be used to ensure either the desired power level for hypothesis testing or the desired margin of error for interval estimation. The results of this chapter are illustrated using Van der Elst et al. (2006b)'s normative study of the Profession Naming verbal fluency test.

Chapter 5 extends chapter 4 to a scenario of several tests to be normed with the same sample. To take into account the correlation between test scores of the same individual, a multivariate regression model is used (Van der Elst et al., 2017), instead of estimating a univariate regression model for each test. However, in the multivariate regression-based approach of Van der Elst et al. (2017), each test is normed separately, thus ignoring the correlation between norm statistic values of the same individual. In chapter 5, a new multivariate regression-based approach is proposed that combines all separate scores for an individual in the Mahalanobis distance (i.e. between the multivariate test score for an individual and the multivariate average in the reference population), thus providing an indicator of the individual's overall performance across all tests. Furthermore, sampling variance and covariance formulas are derived for the Z-score estimator, as well as a sampling variance formula for the Mahalanobis distance estimator. Since all these formulas are based on approximations, two simulation studies are performed to assess the bias induced by these approximations, thus extending the results of the simulation studies in chapter 4 to the case of

two tests to be normed. For both multivariate regression-based approaches, optimal designs are derived for the multivariate version of the five regression models as considered in chapter 4, and efficient designs that are robust against model misspecification are obtained using the maximin strategy with efficiency and relative efficiency as criteria. It is shown that the most robust designs obtained under the two criteria coincide. Formulas to derive the required size for the optimal design of the normative sample are proposed for the Mahalanobis distance-based approach only, because Van der Elst et al. (2017)'s approach is hampered by multiple testing issues. The results of this chapter are illustrated using Van der Elst et al. (2006a)'s normative study of the oral and written versions of the Letter Digit Substitution Test.

Chapter 6 has provided some considerations about common challenges that are encountered in planning surveys and normative studies. Furthermore, a few practical guidelines on how to design each type of study have been given, and ideas for future research have been discussed.

In the next chapter, a reflection is given on the scientific and social impact of this thesis.

Chapter 8

Scientific and social impact of the thesis

A crucial step in the research process is the choice of the design of the study, because a poorly designed study can have serious consequences for science (e.g. biased or unreliable results) and society (e.g. a waste of resources or bad decisions in health and education based on invalid research conclusions). This thesis deals with the design of two types of studies, that is, surveys for mean estimation in multilevel populations (e.g. students grouped in schools, patients clustered in hospitals), and normative studies for estimating reference values (or norms) for tests (e.g. IQ test) and questionnaires (e.g. to measure patients' symptoms). Both types of studies are of practical importance. By allowing comparisons between different populations with respect to their means (e.g. comparing countries in terms of average length of stay for discharges from hospitals), surveys for mean estimation can be useful, for instance, for the implementation of new governmental policies (e.g. new interventions to reduce length of stay in public hospitals). Normative studies, instead, provide reference values that clinicians and educators need in order to compare individuals' performance on a test with the reference population (e.g. individuals with the same age, sex, and education), and to make decisions about individuals (e.g. assignment of a patient to a treatment or of a student to remedial teaching). Thus, it is important that population means and reference values are estimated with the highest possible precision, and without wasting resources (i.e. time and money). This goal can be attained by a careful design of the study. Hence, the main objective of this thesis is to provide guidelines for planning both types of studies in order to achieve precise estimates using minimum resources.

This thesis addresses three design issues of surveys for mean estimation in multilevel populations. First, it identifies the best strategy to draw a sample from the population (i.e. the

most efficient sampling scheme) and to analyse the data (i.e. the unbiased mean estimation). Specifically, data should be collected by sampling clusters (e.g. schools, hospitals) with probability proportional to their size (i.e. number of individuals belonging to a cluster) first, and then by sampling the same number of individuals per selected cluster. Furthermore, the average of all individual outcomes in the population (i.e. the population mean) should be estimated by computing the average of the means of the sampled clusters. Second, this thesis provides formulas to compute optimal sample sizes (i.e. number of clusters and number of individuals per cluster) that allow to either maximize precision of mean estimation for the available budget for sampling and measuring, or to minimize the budget for the required precision of estimation, thus avoiding a waste of the limited resources. Third, a strategy is proposed to overcome the dependency of the optimal sample sizes on some unknown model parameters. This strategy consists of defining a range of plausible values for each unknown parameter first, and then deriving the sample sizes (i.e. number of clusters and number of individuals per cluster) that maximize precision of estimation under the worst-case scenario that can occur within these parameter ranges. These results have a direct impact on science by extending the results available in survey sampling literature (Chambers & Clark, 2012; Cochran, 1977; Lohr, 2010; Särndal et al, 1992; Sukhatme, 1954; Valliant et al, 2000) to a scenario where cluster size affects individuals' outcome variables (e.g. when the length of stay of a patient in a hospital depends on the number of patients admitted to the hospital), and have an indirect impact on society by helping researchers in planning surveys for monitoring important social issues, such as alcohol consumption among adolescents or government expenditure on health, without wasting resources (i.e. time and money).

Three design issues of normative studies are addressed in this thesis. First, it provides the sample composition (e.g. which age groups to include) that maximizes precision of the estimation of reference values (i.e. the optimal design) under five regression models for the reference population. Second, since a design that is optimal under one regression model can be very inefficient under another regression model, and at the design phase of a study there is uncertainty about the "true" model, highly (but not maximally) efficient designs that are robust against the choice of the wrong model in the design stage are presented. Third, this thesis provides formulas to determine how many individuals must be sampled under the optimal design in order to achieve a desired statistical power for testing hypotheses on an individual's performance, or to achieve a desired margin of error in estimating an individual's performance. These results have a direct impact on science because they extend previous research on sample

size requirements for normative studies (Oosterhuis et al., 2016), by providing formulas to compute the required sample size (instead of plots based on simulation studies), and by providing efficient robust designs, not only for studies that derive norms for a single test but also for studies norming multiple tests with the same sample. These results have an indirect impact on society because these guidelines can be used by researchers to carefully plan normative studies, thus preventing mistakes in the assessment of individuals (e.g. not recommending remedial teaching because imprecise norms based on a poorly designed normative study lead to an overestimation of a child's vocabulary size and arithmetic skills).

The results of this thesis are relevant for scientists (not necessarily statisticians) carrying out these two types of studies, because they can find in this work helpful guidelines for planning such studies. Since surveys are often performed by governmental institutions, such as national statistical institutes, chapters 2 and 3 might be more interesting for researchers based at these organizations. Normative studies, instead, are mainly conducted in educational, clinical, and neuropsychological research, so chapters 4 and 5 might be more relevant for researchers from these fields. Furthermore, chapters 2-5 of this thesis can also be used as teaching materials for mathematical statistics students, whereas the application sections of these chapters could be used to introduce students without a mathematical statistics background to the design of research studies.

In order to make the results of this thesis available to these target groups, chapters 2, 3, and 4 have been published in three international scientific journals, and chapter 5 has been submitted for publication. Furthermore, chapters 2 and 4 have been presented at two international scientific conferences. To compute the required size of the sample with the sample size formulas in chapters 3-5, R codes have been developed that will be made available when the chapter to which a code belongs has been published. However, more user-friendly software for applying the results of this thesis could be developed and made freely available to researchers. Such software could be accompanied by a non-technical summary or tutorial paper for a psychological, social, biomedical or health science journal, where the results of this thesis are further illustrated.

Samenvatting

Efficiënte ontwerpen voor gemiddelde schatting in multilevel populaties en testnormering

Dit proefschrift behandelt de planning van de steekproefgrootte en het optimale ontwerp (design) voor twee soorten observationele studies: (1) surveys voor het schatten van een gemiddelde in multilevel populaties, zoals de gemiddelde alcoholconsumptie onder middelbare scholieren (zie, bijvoorbeeld, ESPAD Group, 2016) en (2) normatieve studies om referentiewaarden af te leiden voor tests en vragenlijsten, zoals neuropsychologische tests om de snelheid van de informatieverwerking te beoordelen (zie, bijvoorbeeld, Van der Elst et al., 2006a), en klinische vragenlijsten om de oriëntatie van patiënten op chronische pijn te meten (zie, bijvoorbeeld, Van Breukelen & Vlaeyen, 2005).

Hoofdstuk 1 geeft een inleiding tot deze twee soorten studies. Het praktische belang van deze studies wordt uitgelegd, er worden praktijkvoorbeelden gegeven, de belangrijkste resultaten in de literatuur worden samengevat en de statistische modellen voor het analyseren van gegevens die met deze studies zijn verkregen worden geïntroduceerd. Verder wordt een definitie gegeven van het optimale ontwerp (optimal design) voor elk soort studie en worden strategieën gepresenteerd om robuuste ontwerpen te vinden. Hoofdstuk 1 eindigt met een overzicht van het proefschrift.

Hoofdstuk 2 gaat over zuivere en efficiënte schatting van het gemiddelde van alle individuele uitkomsten in twee-niveau populaties, met ofwel simple random sampling (SRS) van individuen (d.w.z. individuen worden rechtstreeks uit de populatie getrokken) of two-stage sampling (TSS) (d.w.z. eerst worden clusters getrokken, en vervolgens worden individuen uit de geselecteerde clusters getrokken). Clustergroottes mogen variëren en gerelateerd zijn aan de uitkomstvariabele van belang (d.w.z. de clustergrootte is informatief). Er worden drie TSS methoden onderzocht: Het trekken van clusters met een waarschijnlijkheid evenredig aan de clustergrootte, en het vervolgens trekken van een SRS van hetzelfde aantal individuen binnen elk getrokken cluster (TSS1); Het trekken van een SRS van clusters en het vervolgens trekken van hetzelfde percentage individuen per getrokken cluster (TSS2); Het trekken van een SRS

van clusters en het vervolgens trekken van hetzelfde aantal individuen per getrokken cluster (TSS3). In dit hoofdstuk wordt aangetoond dat het gemiddelde van alle individuele uitkomsten en het gemiddelde van alle clusterspecifieke gemiddelden (d.w.z. de twee definities van populatiegemiddeldes in een twee-niveau populatie) alleen samenvallen als de clustergrootten gelijk of niet-informatief zijn. Zuivere schatting van het gemiddelde van alle individuele uitkomsten wordt voor elke steekproefmethode (sampling scheme) besproken. Bovendien worden de drie TSS-methoden in termen van efficiëntie met elkaar en met SRS van individuen vergeleken, uitgaande van een vastgestelde totale steekproefomvang. Aangetoond wordt dat de relatieve efficiëntie van de sampling schemes afhangt van de verdeling van de clustergrootte. TSS1 is echter de meest efficiënte TSS methode voor veel clustergrootte verdelingen. Model-based en design-based inference worden met elkaar vergeleken en er wordt aangetoond dat zij vergelijkbare resultaten opleveren indien aan de modelaannames wordt voldaan. De resultaten van dit hoofdstuk worden toegepast op twee echte clustergrootte verdelingen: de verdeling van de middelbare schoolgrootte in Italië, en de verdeling van de grootte van huisartspraktijken in Engeland.

Hoofdstuk 3 gaat over optimale TSS designs voor het schatten van het gemiddelde in twee-niveau populaties wanneer de clustergrootte informatief is. Er wordt een simulatiestudie uitgevoerd om de bias te beoordelen in de steekproefvariantie (sampling variance) formules die zijn afgeleid voor TSS2 en TSS3 in hoofdstuk 2, omdat deze varianties gebaseerd zijn op benaderingen. Voor elke TSS methode die in hoofdstuk 2 wordt beschouwd, worden vergelijkingen voor de optimale steekproefomvang afgeleid. Het optimale ontwerp (optimal design) is het aantal clusters en het aantal personen per cluster dat de steekproefvariantie (sampling variance) van de schatter van het populatiegemiddelde minimaliseert, onder een kostenbeperking. De gevolgen voor het optimale ontwerp (optimal design) van het negeren van informatieve clustergrootte worden onderzocht. Het blijkt dat de optimale TSS-ontwerpen heel robuust zijn tegen misspecificatie van de mate van informativiteit van clustergrootte, maar aannemen dat clustergrootte niet informatief is kan leiden tot ernstige onderschatting van het vereiste onderzoeksbudget voor het gewenste power niveau. Bovendien worden de drie optimale TSS methoden in termen van efficiëntie met elkaar en met SRS van individuen vergeleken uitgaande van een vastgesteld budget voor steekproeftrekking en meting. Aangetoond wordt dat de optimale TSS1 het meest efficiënte design is voor veel clustergrootte verdelingen. Om de afhankelijkheid van de optimale ontwerpen (optimal designs) van voorkennis over sommige modelparameters te overwinnen, worden maximin ontwerpen

(maximin designs) afgeleid voor elke TSS methode. Ten slotte wordt een procedure voorgesteld om maximin steekproefomvangen van, en maximin budgetverdeling (budget split) tussen, twee surveys af te leiden om de gemiddelden van twee populaties te schatten en te vergelijken met TSS1. Deze procedure wordt geïllustreerd met de planning van een hypothetische survey om het alcoholgebruik van adolescenten te vergelijken tussen Frankrijk en Italië, gebruikmakend van de echte verdelingen van de middelbare schoolgrootte in deze twee landen.

Hoofdstuk 4 richt zich op normatieve studies om referentiewaarden of normen voor tests en vragenlijsten af te leiden. In dit hoofdstuk wordt de regression-based benadering van normering gevolgd, omdat die drie voordelen heeft ten opzichte van de traditionele benadering van normering per subgroep gedefinieerd in termen van leeftijd en geslacht (of andere demografische variabelen): (i) de regression-based benadering is efficiënter, dat wil zeggen, deze vereist een kleinere steekproefomvang, (ii) en maakt het mogelijk om de voorspellers te identificeren die de testscore beïnvloeden, en (iii) en maakt het mogelijk om het optimale ontwerp (optimal design) af te leiden, dat wordt gedefinieerd als de gezamenlijke verdeling van scores op de voorspellers die de steekproefvariantie (sampling variance) van de referentiewaarde of norm onder het veronderstelde regressiemodel minimaliseert. In hoofdstuk 4 worden steekproefvariantie (sampling variance) formules afgeleid voor veelgebruikte normen nl. Z-scores en percentile rank scores. Aangezien deze variantieformules gebaseerd zijn op benaderingen, worden twee simulatiestudies uitgevoerd om de bias te beoordelen die door deze benaderingen wordt veroorzaakt. Deze steekproefvariantie (sampling variance) formules worden gebruikt om optimale ontwerpen (optimal designs) af te leiden voor vijf regressiemodellen met daarin een kwantitatieve en kwalitatieve voorspeller, die verschillen in of ze interactie en niet-lineariteit toelaten. Efficiënte ontwerpen (efficient designs) die robuust zijn tegen misspecificaties van het regressiemodel worden afgeleid met behulp van de maximin strategie (maximin strategy), met efficiëntie en relatieve efficiëntie als criteria. Aangetoond wordt dat voor de vijf onderzochte regressiemodellen de meest robuuste ontwerpen (robust designs) die voor deze twee criteria verkregen zijn, dezelfde zijn. Bovendien worden voor het optimale ontwerp (optimal design) formules voorgesteld om de vereiste omvang van de normatieve steekproef voor elke norm (d.w.z. Z-score en percentile rank score) te bepalen. Deze formules kunnen worden gebruikt om het gewenste power niveau voor hypothese toetsing of de gewenste foutmarge (margin of error) voor intervalschatting te garanderen. De resultaten van dit hoofdstuk worden geïllustreerd aan de hand van een normatieve studie van de Profession Naming Verbal Fluency Test (zie Van der Elst et al., 2006b).

Hoofdstuk 5 breidt hoofdstuk 4 uit tot een scenario van verschillende tests die met dezelfde steekproef moeten worden genormeerd. Om rekening te houden met de correlatie tussen de testscores van hetzelfde individu, wordt een multivariaat regressiemodel gebruikt (Van der Elst et al., 2017), in plaats van een univariaat regressiemodel voor elke test te schatten. In de multivariate regression-based benadering van Van der Elst et al. (2017) wordt elke test echter afzonderlijk genormeerd, waardoor de correlatie tussen normwaarden van hetzelfde individu wordt genegeerd. In hoofdstuk 5 wordt een nieuwe multivariate regression-based benadering voorgesteld die alle afzonderlijke scores voor een individu in de Mahalanobis-distance (d.w.z. de afstand tussen de multivariate testscore voor een individu en het multivariate gemiddelde in de referentiepopulatie) combineert, waardoor een indicator van de algehele prestatie van het individu op alle tests wordt verkregen. Bovendien worden steekproefvariantie (sampling variance) en covariantie formules afgeleid voor de Z-score schatter, evenals een steekproefvariantie (sampling variance) formule voor de Mahalanobis-distance schatter. Aangezien al deze formules gebaseerd zijn op benaderingen, worden twee simulatiestudies uitgevoerd om de door deze benaderingen veroorzaakte bias te beoordelen, waarbij de resultaten van de simulatiestudies in hoofdstuk 4 worden uitgebreid tot het geval van twee te normeren tests. Voor beide multivariate regression-based benaderingen worden optimale ontwerpen (optimal designs) afgeleid voor de multivariate versie van de vijf regressiemodellen zoals beschouwd in hoofdstuk 4, en worden efficiënte ontwerpen (efficient designs) die robuust zijn tegen modelmisspecificatie verkregen met behulp van de maximin strategie (maximin strategy), met efficiëntie en relatieve efficiëntie als criteria. Aangetoond wordt dat de meest robuuste ontwerpen (robust designs) die op grond van de twee criteria zijn verkregen, samenvallen. Formules om de vereiste omvang af te leiden voor het optimale ontwerp (optimal design) van de normatieve steekproef worden alleen gepresenteerd voor de Mahalanobis distance-based benadering, omdat de benadering van Van der Elst et al. (2017) gehinderd wordt door het probleem van multiple testing. De resultaten van dit hoofdstuk worden geïllustreerd aan de hand van de normatieve studie van de mondelinge en schriftelijke versies van de Letter Digit Substitution Test (zie Van der Elst et al., 2006a).

Hoofdstuk 6 geeft een aantal overwegingen over gemeenschappelijke uitdagingen die worden ondervonden bij het ontwerpen van enquêtes (surveys) en normatieve studies. Verder worden er enkele praktische richtlijnen gegeven voor het ontwerpen van elk soort studie en worden ideeën voor toekomstig onderzoek besproken.

References

Abebe, H. T., Tan, F. E. S., Van Breukelen, G. J. P., & Berger, M. P. F. (2015). Bayesian design for dichotomous repeated measurements with autocorrelation. *Statistical Methods in Medical Research*, *24*(5), 594-611. https://doi.org/10.1177/0962280213508850

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intracluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57,* 785-794. https://doi.org/10.1016/j.jclinepi.2003.12.013

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods, 35*(3), 439-460. https://doi.org/10.1080/03610920500476598

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS.* Oxford, UK: Oxford University Press.

Bayard, S., Gély-Nargeot, M-C., Raffard, S., Guerdoux-Ninot, E., Kamara, E., Gros-Balthazard, F., Jacus, J.-P., & Moroni, C. (2017). French version of the Hayling Sentence Completion test, part I: normative data and guidelines for error scoring. *Archives of Clinical Neuropsychology*, *32*, 585-591. https://doi.org/10.1093/arclin/acx010

Berger, M. P. F, King, C. Y. J., & Wong, W. K. (2000). Minimax D-Optimal designs for item response theory models. *Psychometrika*, *65*, 377-390.

Berger, M. P. F., & Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*. Chichester, UK: John Wiley & Sons.

Candel, M. J. J. M, & Van Breukelen, G. J. P. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research*, *24*, 557-573. https://doi.org/10.1177/0962280214563100

Casella, G., & Berger, R.L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chaloner, K. (1984). Optimal Bayesian experimental designs for linear models. *Annals of Statistics*, *12*, 283-300.

Chaloner, K. & Verdinelli, I. (1995). Bayesian experimental design. A review. *Statistical Science*, *10*, 273–304.

Chambers, R. L., & Clark, R. G. (2012). *An introduction to model-based survey sampling with applications*. Oxford, UK: Oxford University Press.

Chang, S. I. (1994). Some properties of multiresponse D-optimal designs. *Journal of Mathematical Analysis and Applications*, *184*, 256-262.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley & Sons.

DeFrances, C. J., Lucas, C. A., Buie, V. C., & Golosinskiy, A. (2008). 2006 National Hospital Discharge Survey. *National health statistics reports,* no. 5. Hyattsville, MD: National Center for Health Statistics.

Dette, H. (1996). A note on Bayesian C- and D-optimal designs in nonlinear regression models. *Annals of Statistics*, *24*, 1225-1234.

Dette, H., & Biedermann, S. (2003). Robust and efficient designs for the Michaelis-Menten model. *Journal of the American Statistical Association*, *98*, 679-686. https://doi.org/10.1198/016214503000000585

Dette, H., Lopez, I. M., Ortiz Rodriguez, I. M., & Pepelyshev, A. (2006). Maximin efficient design of experiment for exponential regression models. *Journal of Statistical Planning and Inference*, *136*, 4397-4418. https://doi.org/10.1016/j.jspi.2005.06.006

Direzione generale per i contratti, gli acquisti e per i sistemi informativi e la statistica (DGCASIS). (2018). Studenti per anno di corso e fascia di etá: Scuola statale. http://dati.istruzione.it/opendata/opendata/catalogo/elements1/?area=Studenti Accessed May 1, 2018.

Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association*, *62*(318), 607-625.

Eldridge, S. M., Ashby, D., Feder, G. S., Rudnicka, A. R., & Ukoumunne, O. C. (2004). Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials, 1,* 80-90. https://doi.org/10.1191/1740774504cn006rr

Eldridge, S. M., Costelloe, C. E., Kahan, B. C., Lancaster, G. A., & Kerry, S. M. (2016). How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research*, *25*, 1039-1056. https://doi.org/10.1177/0962280215588242

ESPAD Group. (2016). *ESPAD Report 2015: Results from the European School Survey Project on Alcohol and Other Drugs.* Luxembourg: Publications Office of the European Union. https://doi.org/10.2810/564360

Eurostat. (2018). Government expenditure by function. http://ec.europa.eu/eurostat/statistics-explained/index.php/Government_expenditure_by_function Accessed July 1, 2018.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York, NY: Academic Press.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Fox, J. (1991). *Quantitative applications in the social sciences, No. 79. Regression diagnostics.* Newbury Park, CA: Sage Publications.

Fujikoshi, Y, Ulyanov, V. V., & Shimizu, R. (2010). *Multivariate statistics. High-dimensional and large-sample approximations.* Hoboken, NJ: John Wiley & Sons.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester, UK: John Wiley & Sons.

Goos, P., & Jones, B. (2011). *Optimal design of experiments. A case study approach*. Chichester, UK: John Wiley & Sons.

Goos, P., Vermeulen, B., & Vandebroek, M. (2010). D-optimal conjoint choice designs with no-choice options for a nested logit model. *Journal of Statistical Planning and Inference*, *140*, 851-861. https://doi.org/10.1016/j.jspi.2009.09.006

Goretti, B., Niccolai, C., Hakiki, B., Sturchio, A., Falautano, M., Minacapelli, E., Martinelli, V., Incerti, C., Nocentini, U., Murgia, M., Fenu, G., Cocco, E., Marrosu, M. G., Garofalo, E., Ambra, F. I., Maddestra, M., Consalvo, M., Viterbo, R. G., Trojano, M.,

… Amato, M. P. (2014). The brief international cognitive assessment for multiple sclerosis (BICAMS): Normative values with gender, age and education corrections in the Italian population. *BMC Neurology*, *14*, 171-176. https://doi.org/10.1186/s12883-014-0171-6

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative designs. *Survey Methodology*, *30*(1), 93-103.

Gurland, J. (1967). An inequality satisfied by the expectation of the reciprocal of a random variable. *The American Statistician, 21*(2), 24-25.

Han, C., & Chaloner, K. (2004). Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biostatistics*, *60*, 25-33. https://doi.org/10.1111/j.0006-341X.2004.00148.x

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*(384), 776-793.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87. https://doi.org/10.3102/0162373707299706

Hoogenhout, E. M., Van der Elst, W., de Groot, R. H. M., Van Boxtel, M. P. J., & Jolles, J. (2010). The neurovegetative complaints questionnaire in the Maastricht aging study: Psychometric properties and normative data. *Aging & Mental Health*, *14*, 613-623. https://doi.org/10.1080/13607861003587297

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S., & Van Breukelen, G. J. P. (2019). Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine*, *38*, 1817-1834. https://doi.org/10.1002/sim.8070

Innocenti, F., Tan, F. E. S., Candel, M. J. J. M., & Van Breukelen, G. J. P. (2021). Sample size calculation and optimal design for regression-based norming of tests and questionnaires. *Psychological Methods*. (In press). https://doi.org/10.1037/met0000394

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*, 183-189.

Johnson, R. A., *&* Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Johnson, R. A, & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Jones, A. (2002). The National Nursing Home Survey: 1999 summary. *Vital and Health Statistics Series 13*, *152*, 1-116.

Kelly, E., & Stoye, G. (2014). *Does GP practice size matter? GP practice size and the quality of primary care.* Report no. R101. London, UK: Institute for Fiscal Studies. https://doi.org/10.1920/re.ifs.2014.0101

King, C. Y. J., & Wong, W. K. (2000). Minimax D-Optimal designs for the logistic model. *Biometrics*, *56*, 1263-1267.

Koziol, N. A., Bovaird, J. A., & Suarez, S. (2017). A comparison of population-averaged and cluster-specific approaches in the context of unequal probabilities of selection. *Multivariate Behavioral Research, 52*(3), 325-349. https://doi.org/10.1080/00273171.2017.1292115

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546-556.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston, MA: Brooks/Cole.

Magnus, J. R., & Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Makela, S., Si, Y., & Gelman, A. (2018). Bayesian inference under cluster sampling with probability proportional to size. *Statistics in Medicine*, *37*(26), 3849-3868. https://doi.org/10.1002/sim.7892

Manju, M. A., Candel, M. J. J. M., & Berger, M. P. F. (2014). Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Statistics in Medicine*, *33*(15), 2538-2553. https://doi.org/10.1002/sim.6112

Manju, M. A., Candel, M. J. J. M., & Berger, M. P. F. (2015). Optimal and maximin sample sizes for multicentre cost-effectiveness trials. *Statistical Methods in Medical Research*, *24*(5), 513-539. https://doi.org/10.1177/0962280215569293

Maus, B., Van Breukelen, G. J. P., Goebel, R., & Berger, M. P. F. (2010). Robustness of optimal design of fMRI experiments with application of a genetic algorithm. *Neuroimage*, *49*, 2433-2443. https://doi.org/10.1016/j.neuroimage.2009.10.004

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537-563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McNeely, C. A., Nonnemaker, J. M., & Blum, R. W. (2002). Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *Journal of School Health*, *72*, 138-146. https://doi.org/10.1111/j.1746-1561.2002.tb06533.x

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York, NY: Oxford University Press.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25*, 271-284. https://doi.org/10.3102/10769986025003271

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(1), 17-30. https://doi.org/10.1111/1467-9884.00257

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). Tokyo: McGraw-Hill.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York, NY: John Wiley & Sons.

Nevalainen, J., Datta, S., & Oja, H. (2014). Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers*, *55*(1), 71-92. https://doi.org/10.1007/s00362-013-0504-3

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191-202. https://doi.org/10.1177/1073191115580638

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2017). Standard errors and confidence intervals of norm statistics for educational and psychological tests. *Psychometrika*, *82*(3), 559-588. https://doi.org/10.1007/s11336-016-9535-8

Ouwens, J. N. M., Tan, F. E. S., & Berger, M. P. F. (2002). Maximin D-optimal design for longitudinal mixed effects models. *Biometrics*, 58, 735-741.

Panageas, K. S., Schrag, D., Localio, A. R., Venkatraman, E. S., & Begg, C.B. (2007). Property of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine*, *26*(9), 2017-2035. https://doi.org/10.1002/sim.2657

Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. B. (2010). The utility of regression-based norms in interpreting the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society*, *16*, 6-16. https://doi.org/10.1017/S1355617709990750

Patton, G. C., Hibbert, M., Rosier, M. J., Carlin, J. B., Caust, J., & Bowes, G. (1995). Patterns of common drug use in teenagers. *Australian Journal of Public Health*, *19*, 393-399. https://doi.org/10.1111/j.1753-6405.1995.tb00392.x

Pearson, K. (1929). Editorial note to 'Inequalities for moments of frequency functions and for various statistical constants'. *Biometrika*, *21,* 361–375.

Peña-Casanova, J., Gramunt-Fombuena, N., Quiñones-Ubeda, S., Sanchez-Benavides, G., Aguilar, M., Badenes, D., Molinuevo, J. L., Robles, A., Barquero, M. S., Payno, M., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Alfonso, V., Sol, J. M., & Blesa, R. (2009a). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for Boston Naming Test and Token Test. *Archives of Clinical Neuropsychology*, *24*, 343-354. https://doi.org/10.1093/arclin/acp039

Peña-Casanova, J., Quiñones-Ubeda, S., Gramunt-Fombuena, N., Aguilar, M., Casas, L., Molinuevo, J. L., Robles, A., Rodriguez, D., Barquero, M. S., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Molano, A., Alfonso, V., Sol, J. M., & Blesa, R. (2009b). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for the Rey–Osterrieth Complex Figure (Copy and Memory), and Free and Cued Selective Reminding Test. *Archives of Clinical Neuropsychology*, *24*, 371-393. https://doi.org/10.1093/arclin/acp041

Peña-Casanova, J., Quiñones-Ubeda, S., Gramunt-Fombuena, N., Quintana, M., Aguilar, M., Molinuevo, J. L., Serradell, M., Robles, A., Barquero, M. S., Payno, M., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Alfonso, V., Sol, J. M., & Blesa, R. (2009c). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for the Stroop Color-Word Interference Test and the Tower of London-Drexel. *Archives of Clinical Neuropsychology*, *24*, 413-429. https://doi.org/10.1093/arclin/acp043

Petersen, K. B., & Pedersen, M. S. (2012). *The Matrix Cookbook.* Technical University of Denmark.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, *61*(2), 317-337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60*(1), 23-40.

Pronzato, L., & Walter, E. (1988). Robust experiment design via maximin optimization. *Mathematical Biosciences*, *89*, 161-176. https://doi.org/10.1016/0025-5564(88)90097-1

R Core Team. (2020). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805-827.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., & Udry, J. R. (1997). Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, *278*(10), 823-832. https://doi.org/10.1001/jama.278.10.823

Rigueiro Neves, M., Sousa, C., Passos, A. M., Ferreira, A. I., & Sà, J. M. (2018). Verbal Selective Reminding Test (six-trial administration): Regression-based norms for a Portuguese version. *Applied Neuropsychology: Adult*, *25*, 523-531. https://doi.org/10.1080/23279095.2017.1336712

Safarkhani, M., Jolani, S., & Moerbeek, M. (2014). Optimal number of accrual groups and accrual group sizes in longitudinal trials with discrete-time survival endpoints. *Statistica Neerlandica*, *68*(1), 43-60. https://doi.org/10.1111/stan.12022

Salt, K. (2017). Patients registered at a GP practice. October 2017; Special Topic – Practice list size comparison 2017. Health and Social Care Information Centre. http://digital.nhs.uk/catalogue/PUB30111 Accessed November 8, 2017.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer.

Schwabe, R. (1996). *Optimum designs for multi-factor models.* New York, NY: Springer-Verlag.

Seaman, S., Pavlou, M., & Copas, A. (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine*, *33*(30), 5371-5387. https://doi.org/10.1002/sim.6277

Searle, S., & Pukelsheim, F. (1986). Effect of intraclass correlation on weighted averages. *The American Statistician, 40*, 103-105.

Shackleton, N., Hale, D., Bonell, C., & Viner, R. M. (2016). Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM Population Health*, *2*, 217-225. https://doi.org/10.1016/j.ssmph.2016.03.005

Sherwood, B., Zhou, A. X., Weintraub, S., & Wang, L. (2016). Using quantile regression to create baseline norms for neuropsychological tests. *Alzheimer's & Dementia:*

Diagnosis, Assessment & Disease Monitoring, *2*, 12-18. https://doi.org/10.1016/j.dadm.2015.11.005

Skinner, C., & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165-175. http://dx.doi.org/10.1214/17-STS614

Smith, T. M. F. (1994). Sample Surveys 1975-1990; An Age of Reconciliation? *International Statistical Review, 62*(1), 28-32.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage Publishers.

Sudgen, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, *71*(3), 495-506. https://doi.org/10.1093/biomet/71.3.495

Sukhatme, P. V. (1954). *Sampling theory of surveys with applications*. Ames, IA: Iowa State College Press.

Tan, F. E. S. (2010). Conditions for D_A-Maximin marginal designs for generalized linear mixed models to be uniform. *Communications in Statistics - Theory and Methods*, *40*, 255-266. https://doi.org/10.1080/03610920903411218

Tekle, F. B., Tan, F. E. S., & Berger, M. P. F. (2008). Maximin D-Optimal designs for binary longitudinal responses. *Computational Statistics & Data Analysis*, *52*, 5253-5262. https://doi.org/10.1016/j.csda.2008.04.037

Thompson, D. R., Iachan, R., Overpeck, M., Ross, J.G., & Gross, L. A. (2006). School connectedness in the health behavior in school-aged children study: The role of student, school, and school neighborhood connectedness. *Journal of School Health*, *76*(7), 379-386. https://doi.org/10.1111/j.1746-1561.2006.00129.x

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite population sampling and inference: a prediction approach.* New York, NY: John Wiley & Sons; 2000.

Van Breukelen, G. J. P., & Candel, M. J. J. M. (2012). Efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests. *Statistics in Medicine*, *31*(4), 397-400. https://doi.org/10.1002/sim.4449

Van Breukelen, G. J. P., & Candel, M. J. J. M. (2018). Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Statistics in Medicine*, *37*(21), 3027-3046. https://doi.org/10.1002/sim.7824

Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The pain cognition list. *Psychological Assessment*, *17*(3), 336-344. https://doi.org/10.1037/1040-3590.17.3.336

Van Breukelen, G. J. P., Candel, M. J. J. M., & Berger, M. P. F. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, *26*(13), 2589-2603. https://doi.org/10.1002/sim.2740

Van der Elst, W., Hurks, P., Wassenberg, R., Meijs, C., & Jolles, J. (2011). Animal Verbal Fluency and Design Fluency in school-aged children: Effects of age, sex, and mean level of parental education, and regression-based normative data. *Journal of Clinical and Experimental Neuropsychology*, *33*(9), 1005-1015. http://dx.doi.org/10.1080/13803395.2011.589509

Van der Elst, W., Molenberghs, G., Van Boxtel, M. P. J., & Jolles, J. (2013). Establishing normative data for repeated cognitive assessment: A comparison of different statistical methods. *Behavior Research Methods*, *45*, 1073-1086. https://doi.org/10.3758/s13428-012-0305-y

Van der Elst, W., Molenberghs, G., Van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: The multivariate regression-based approach. *The Clinical Neuropsychologist*, *31*, 1173-1187. https://doi.org/10.1080/13854046.2017.1294202

Van der Elst, W., Van Boxtel, M. P. J, Van Breukelen, G. J. P., & Jolles, J. (2006a). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24–81 from the Maastricht Aging Study (MAAS): Influence of Age, Education, and Sex. *Journal of Clinical and Experimental Neuropsychology*, *28*, 998-1009. https://doi.org/10.1080/13803390591004428

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006b). Normative data for the animal, profession and letter *M* naming verbal fluency tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the*

International Neuropsychological Society, *12*, 80-89. https://doi.org/10.1017/S1355617706060115

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019a). Model selection in continuous test norming with GAMLSS. *Assessment*, *26*(7), 1329–1346. https://doi.org/10.1177/1073191117715113

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019b). Improving confidence intervals for normed test scores: Include uncertainty due to sampling variability. *Behavior Research Methods*, *51*(2), 826–839. https://doi.org/10.3758/s13428-018-1122-8

Warren, C. W., Riley, L., Asma, S., Eriksen, M. P., Green, L., Blanton, C., Loo, C., Batchelor, S., & Yach, D. (2000). Tobacco use by youth: a surveillance report from the Global Youth Tobacco Survey project. *Bulletin of the World Health Organization*, *78*, 868- 876.

Winkens, B., Schouten, H. J., Van Breukelen, G. J. P., & Berger, M. P. (2005). Optimal time-points in clinical trials with linearly divergent treatment effects. *Statistics in Medicine*, *24*(24), 3743-3756. https://doi.org/10.1002/sim.2385

Winkens, B., Schouten, H. J., Van Breukelen, G. J. P., & Berger, M. P. (2007). Optimal designs for clinical trials with second-order polynomial treatment effects. *Statistical Methods in Medical Research*, *16*, 523-537. https://doi.org/10.1177/0962280206071847

Wong, W. K. (1992). A unified approach to the construction of minimax designs. *Biometrika*, *79*, 611-619. https://doi.org/10.1093/biomet/79.3.611

Wong, W. K. (1995). On the equivalence of D and G-optimal designs in heteroscedastic models. *Statistics & Probability Letters*, *25*, 317-321. https://doi.org/10.1016/0167-7152(94)00236-1

Wu, S., Wong, W. K., & Crespi, C. M. (2017). Maximin optimal designs for cluster randomized trials. *Biometrics*, *73*, 916-926. https://doi.org/10.1111/biom.12659

Yu, J., Goos, P., & Vandebroek, M. (2008). Model-robust design of conjoint choice of experiments. *Communications in Statistics – Simulation and Computation*, *37*(8), 1603-1621. https://doi.org/10.1080/03610910802244638

Zhang, L. (2007). Sample mean and sample variance: their covariance and their (in)dependence. *The American Statistician, 61*(2), 159-160. https://doi.org/10.1198/000313007X188379

Zheng, H., & Little, R. J. A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, *30*(2), 209-218.

Acknowledgments

This thesis would not have been possible without the support and the encouragement I received from many wonderful persons.

First of all, I would like to mention my supervisors Prof. dr. Gerard van Breukelen, Dr. Math Candel, and Dr. Frans Tan. I am grateful to all of you for your valuable guidance and support. I consider myself very lucky to have had you as supervisors! Gerard, thank you for your detailed feedback on my work from which I learned a lot on how to structure my thoughts, and for your continuous encouragement. Math, you were my daily supervisor in the first (and most difficult) years of the project, thank you for your helpful support, and for your time, patience, and enthusiasm. Frans, thank you for your time, and for your calm and balanced supervision in the second part of the project.

I would also like to thank all the members of the M&S department for making it a warm working environment. Many thanks to Edith van Eijsden and Marga Doyle for their administrative support. Also, thanks to all the PhD candidates of the department, or who visited the department in these four years, with whom not only I have shared the challenging experience of doing a PhD, but also of living abroad: Etienne, Gavin, Marco, Matthias, Mutamba, and Zaheer. Finally, I would like to thank Dr. Alberto Cassese and Dr. Jan Schepers, who have been my "guardian angels" in these four years. Our lunch breaks together have been very precious for me, thank you for your support and friendship! Furthermore, I am particularly grateful to Alberto because it is thanks to him that I learned about this position.

I would also like to thank Prof. dr. Antonio Villanacci and Prof. dr. Leonardo Grilli, my Bsc and Msc thesis supervisors, respectively, because working with them helped me preparing for the PhD experience. Furthermore, they both helped shift my interest from economics to statistics, for which I am extremely grateful!

I would also like to mention my lifelong friends: Alessandro, Cosimo, Francesco, Jacopo, Nicola, and Vieri. Your sarcastic jokes (worthy of the best Florentine tradition) about my person have tormented my ego as much as amused me! I am very lucky to have friends like you who have taught me not to take myself too seriously. I would also like to thank Alan,

Claudio, François, Pietro, Roberto, and Veneziano for our enlightening talks and your extensive support!

Finally, I would like to thank my beloved family: My little Lilli, for her devotion; my parents, Anna Maria and Andrea, who taught me, through the example of their lives, that "*fatti non* [fummo] *a viver come bruti, ma per seguir virtute e canoscenza*"[*]; and my dear Valentina, for her unlimited patience and for always reminding me that there is a whole world to discover outside that window!

[*] Dante Alighieri (1320), *Divina Commedia*, *Inferno*, Canto XXVI. Translation by H.W. Longfellow (1867): "Ye were not made to live like unto brutes, but for pursuit of virtue and of knowledge".

About the author

Francesco Innocenti was born on March 13, 1991, in Firenze, Italy. He attended liceo scientifico (scientific high school) “Guido Castelnuovo” in Firenze from 2005 to 2010. In 2013, he obtained his BSc degree in Economics and Trade (cum laude), with economics as a major, at Università degli Studi di Firenze. At the same university, he obtained his MSc degree in Statistics, Actuarial and Financial Science (cum laude), with statistics as a major, in 2016. In September 2016, he moved to Maastricht where he started his PhD project at the department of Methodology and Statistics, Maastricht University. In January 2021, he will start working as an assistant professor at the same department.