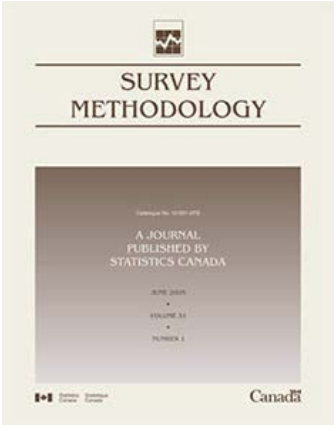
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Considering interviewer and design effects when planning sample sizes



by Stefan Zins and Jan Pablo Burgard

Release date: June 30, 2020

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 93-119
Statistics Canada, Catalogue No. 12-001-X

Considering interviewer and design effects when planning sample sizes

Stefan Zins and Jan Pablo Burgard[1]

Abstract

Selecting the right sample size is central to ensure the quality of a survey. The state of the art is to account for complex sampling designs by calculating effective sample sizes. These effective sample sizes are determined using the design effect of central variables of interest. However, in face-to-face surveys empirical estimates of design effects are often suspected to be conflated with the impact of the interviewers. This typically leads to an over-estimation of design effects and consequently risks misallocating resources towards a higher sample size instead of using more interviewers or improving measurement accuracy. Therefore, we propose a corrected design effect that separates the interviewer effect from the effects of the sampling design on the sampling variance. The ability to estimate the corrected design effect is tested using a simulation study. In this respect, we address disentangling cluster and interviewer variance. Corrected design effects are estimated for data from the European Social Survey (ESS) round 6 and compared with conventional design effect estimates. Furthermore, we show that for some countries in the ESS round 6 the estimates of conventional design effect are indeed strongly inflated by interviewer effects.

Key Words: Design effect; Interviewer effect; Multilevel model; Sample size; European Social Survey (ESS).

1 Introduction

Determining the sample size of a survey can be very demanding. The complexity of the task is often exacerbated by a lack of information and data on which to plan the survey. That is why survey planners seek to reduce the complexity of the problem using simplifications and statistical models. One such approach is to use the so-called *design effect* to select a sample size. The design effect is then defined as the ratio between the variance of an estimator under the sampling design of the planned survey and the variance of the same estimator under a simple random sample design. As such, the design effect is a property of an estimation strategy, i.e., a sampling design and an estimator (Chaudhuri and Stenger, 2005, page 4), not of the survey. The weighted sample mean of a single variable is usually used as a reference estimator. However, for reasons of simplification, if we speak in the following of the design effect of a sampling design, then we do this always with respect to the sampling variance of a weighted sample mean.

To plan the sample size, an effective sample size target can be set, meaning that the planned sample size divided by the planned design effect should be above a certain value. The effective sample size of a sampling design is the simple random sample equivalent of its sample size, in terms of efficiency, i.e., if a sampling design has an effective sample size of 1,000, then its sampling variance is equal to that of a simple random sample of size 1,000.

Ideally, a survey planner designs a survey with a specific analysis or hypotheses test in mind and formulates their opinion about tolerable sampling error levels or type II error probabilities. These opinions should be based on two things. First, some level of experience with the substantial research question, and

1. Stefan Zins, Institute for Employment Research (IAB) of the German Federal Employment Agency (BA), Regensburger Strasse 104, D-90478 Nürnberg. E-mail: st.zins@gmail.com; Jan Pablo Burgard, RIFOSS - Research Institute for Official and Survey Statistics, Trier University, D-54286 Trier.

second, on assumptions over target population parameters necessary for sampling error planning and power calculations. Assumptions about target population parameters can stem from previous rounds of a survey, or be based on data collected during the field test for the survey. Power calculations and sampling error planning are much less complex and require less information about the target population if done under the assumption of a simple random sampling design. That is why most methods addressing sample size planning found in textbooks are suited for determining an effective sample size. The effect of complex sampling is then factored in by multiplying the planned effective sample size with a planned design effect. Determining a design effect can thus be separated from selecting an effective sample size. For example, if a simple random sample of size 1,000 ensures the following: The sampling error of an estimator does not exceed a given value with a probability of 95%, or that the power of a statistical test is 80%, that is, the probability of rejecting a null hypothesis in case the alternative is true should be 80% (Ellis, 2010, Chapter 3). Then multiplying 1,000 by the assumed design effect of the a study will give the survey planner the required net sample size to achieve set precision targets.

The decision on an effective sample size also has to reflect a certain trade-off between the cost of the survey and the precision of survey estimates. Regarding this trade-off, the survey planner should, for example, consider what the consequences are if a type II error is committed, i.e., if a null hypothesis is not rejected even though the alternative hypothesis is true.

For surveys that are primarily intended for secondary analysis, i.e., they provide data to the research community with no single application in mind, like the European Social Survey (ESS) or the European Value Study (EVS), the decision on an effective sample size cannot be planned for a single research question or hypothesis test. For that reason, the ESS uses an average effective sample size. This means that ESS sample designs are planned such that the average design effect for a set of items from the ESS core questionnaire should have a certain value. The planned average design effect is multiplied by the required average effective sample size to calculate the planned net sample size. The net sample size is the sample size after unit-nonresponse, i.e., the number of completed interviews. To plan the gross sample size – that is, the sample size before unit-nonresponse – the net sample size is divided by the product of the assumed response rate and eligibility rate. The eligibility rate is the fraction of sampled persons that belong to the target population, which can be lower than 100% because of sampling frame imperfections.

However, design effects can still be difficult to quantify, given the complexity of the sampling design. Hence, to reduce complexity, statistical models for survey data are used to approximate the design effect. Such models commonly try to incorporate the effect of cluster sampling, which can have a large effect on the sampling variance of estimates. Clusters can be spatial areas like settlements, organizational units like municipalities, or institutions such as hospitals and schools. They are either used as so-called *Primary Sampling Units* (PSUs), which are selected first and then an additional sampling takes place within them, or they are surveyed in their entirety. For example, the German ESS round 6 (ESS6) sampling design has two sampling stages. The PSUs are municipalities, and the secondary sampling units are persons registered within the municipalities. Variables of interest can often not be considered as identically

distributed over all clusters in the population. In fact, it can be assumed that respondents within the same cluster are usually more similar to one another than those belonging to a different cluster. Kish (1965), page 162, gives the following formula for a design effect due to clustering:

$$\text{deff} = 1 + (b - 1)\ \rho. \tag{1.1}$$

This design effect deff consists of two parameters, b is typically an average cluster size in terms of realized respondents, and ρ, the intra-cluster correlation coefficient, which is a measure for the homogeneity of the measurements of a variable within the same cluster. ρ can be defined using variance decomposition as the between-cluster variance divided by the sum of the within-cluster and between-cluster variances. The higher the variance between the clusters the higher ρ will be.

To use deff when selecting a sample size, assumptions have to be made about the unknown parameter ρ. The cluster size b does not depend on the measured variable and can be influenced by the survey planner. For ρ, data from previous surveys can be used to formulate the necessary assumption. Especially for repeated cross-sectional surveys, their accumulated data is of great help in planning the sampling design for the next implementation of the survey.

Lynn, Häder, Gabler and Laaksonen (2007) describe how predicted design effects are used by the ESS to plan sample sizes that achieve a certain average effective sample size under a given sampling design. For recent rounds of the ESS, the prediction of the design effect and its components was informed by estimates of these statistics based on data from the preceding ESS rounds (The ESS Sampling Expert Panel, 2016).

An important factor that can also introduce homogeneity to measurements in face-to-face surveys is the interviewer. Embedded in the Total Survey Error (TSE) framework (Groves, 2009), different mechanisms have been described for how an interviewer can influence survey measurements. Similar to cluster sampling, interviewers have long been identified as a source of dependent measurements (Kish, 1965, page 522, Kish, 1962),with interviewers introducing homogeneity through measurement errors and selection effects, rather than the homogeneity of clusters that is intrinsic to the population. West and Blom (2017) give an overview of the research on interviewer effects. They detail how interviewer tasks like generating and/or applying sampling frames, making contact, and gaining cooperation and consent can have a selection effect on the recruitment of respondents. West and Blom (2017) also outline evidence that interviewers conducting measurements, making observations and finally recording the gathered information can introduce measurement and processing errors into the data that is used for analysis. For an overview of other sources of variance in surveys, we refer to the TSE framework as described, e.g., by Groves and Lyberg (2010) and Biemer (2010).

Analysis of interviewer effects using ESS data from different countries and years showed that this effect can be considerable (Beullens and Loosveldt, 2016). Such findings raise a question: To what extent ρ in equation (1.1) is driven by intra-cluster correlation, rather than intra-interviewer correlation? Schnell and Kreuter (2005) show that the interviewer effect can be higher than the cluster effect, even for

variables where a strong spatial correlation can be assumed. Consequently, the estimated design effect for face-to-face surveys is typically conflated with the interviewer effect. Hence, the design effect is systematically over-estimated in face-to-face surveys. This might pose a problem to surveys that predict design effects using historical data to plan sample sizes, as there is a risk of misallocating funds. A survey planner could try to offset an increase in the predicted design effect by increasing the sample size to hold the effective sample size constant. If the driving factor inflating the predicted design effect is the interviewer effect, funds could be more effectively allocated by hiring additional interviewers and/or training them better to improve measurement accuracy and reduce selection effects.

The novel part of the presented approach is that the proposed method allows for estimating a corrected design effect that is not conflated with the interviewer effects. With the proposed corrected design effect, the survey planner is able to make evidence-based decisions on changes in the sampling design, such as sample size and number of PSUs, and/or about the deployment of interviewers.

The article is structured as follows: Section 2 introduces the framework for describing the effects of the sampling design and the interviewer. The framework follows the model based justification of the design effect as outlined by Gabler, Häder and Lahiri (1999) and the introduction of an interviewer effect to this framework by Gabler and Lahiri (2009). The measurement models used to describe the observed data follow a multilevel structure. The influence of multi-stage or cluster sampling, and that of interviewers on the observed data, is modeled with the help of random effects that imply a certain variance-covariance structure. This approach allows for a factorization of the overall effect into separate sampling and interviewer effects. This separation is essential when addressing effects separately in order to control for them.

In Section 3, the sampling and interviewer effects described in Section 2 are estimated for ESS6 data with the help of multilevel models. First, we present the results from a simulation study conducted to assess the possibility of disentangling cluster and interviewer variances for the observed PSU-interviewer structure in the ESS6 data. Afterwards, we evaluate the applicability of the different measurement models for a selected set of ESS variables. The selected models are used to estimate the variances of different random effects in multilevel models, which are in turn used for estimating the intra-PSU and intra-interviewer correlation.

In Section 4, we present our conclusions and give recommendations for survey planners based on both our theoretical work in Section 2 and the empirical findings in Section 3. We then point to possible future research to adapt our relatively simplistic measurements models to better reflect complex sampling designs and the heterogeneity of interviewers.

2 Interviewer and design effects

We define a sample as a set of n distinct respondents, which we denote as $s = \{1, \ldots, n\}$, with $n \in \mathbb{N}$. For the k^{th} respondent our variable of interest y is a real valued variable, where y_k is the

observation of this variable for the k^{th} respondent in our sample s. The observed data is given by $\mathbf{y} = (y_1, \ldots, y_n)^\top$. We associate survey weights with every respondent in the sample, given by $\mathbf{w} = (w_1, \ldots, w_n)^\top$, where w_k is the weight of the k^{th} respondent and $w_k > 0$, for all $k \in s$.

We consider the weighted sample mean of $\mathbf{y}$ as our estimator, given by

$$\bar{y}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{y}}{\mathbf{w}^\top \mathbb{I}_n} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}, \tag{2.1}$$

where $\mathbb{I}_n$ is a column vector of ones of length n. We focus on one estimator of interest, $\bar{y}(\mathbf{w})$, as it is the most common choice for describing interviewer and design effects (Kish, 1965, Section 8.1, Kish, 1962; Särndal, Swensson and Wretman, 1992, page 53). This choice enables us to use an established framework (Gabler et al., 1999) and produce formulas that are recognizable to readers that are already somewhat familiar with the topic. However, design effects of other estimators have been studied, notably, Lohr (2014), derives design effects for estimators of regression coefficients and Fischer, West, Elliott and Kreuter (2018), describe the impact of interviewer effects on the estimation of regression coefficients.

In the following, the variance of $\bar{y}(\mathbf{w})$ is derived under different measurement models for y. The different models serve to distinguish between complex and simple sampling designs, as well as when there is and is not an interviewer effect. It should be noted that the model based variance of estimator $\bar{y}(\mathbf{w})$, which we use, is, in general, not the same as its design based variances, i.e., the variance of $\bar{y}(\mathbf{w})$ under a given sampling design (Särndal et al., 1992, page 492). Design based variances can be very complex and thus difficult to display in an accessible fashion, especially for multi-stage sampling. The model based approach reduces complexity while retaining the essential property of the complex sampling designs that we study, the cluster effect of multi-stage sampling. It also makes it possible to easily integrate cluster and interviewer effect into a common framework.

2.1 Simple random sampling without an interviewer effect

To model simple random sampling in the absence of an interviewer effect, i.e., without intra-PSU and intra-interviewer correlation, we assume the following measurement model (M_0)

$$y_k = \mu_k + \mathrm{e}_k, \tag{M_0}$$

where μ_k is the value of y for the k^{th} respondent and e_k is the measurement error. The measurement errors e_k for all $k \in s$ are independent and identically distributed (iid) random variables with a variance-covariance structure of

$$\mathrm{Cov}_{M_0}(e_k, e_l) = \begin{cases} \sigma^2, & \text{if } k = l \\ 0, & \text{else} \end{cases}, \tag{2.2}$$

where σ is a real value parameter greater than zero. Under model (M_0), the variance of $\bar{y}(\mathbb{I}_n)$ is given by $V_{M_0}(\bar{y}(\mathbb{I}_n)) = \sigma^2/n$. This variance can be interpreted as the variance of the unweighted sample mean under simple random sampling with replacement (Särndal et al., 1992, page 73). Simple random sampling with estimator $\bar{y}(\mathbb{I}_n)$ typically serves as a reference estimation strategy, which is compared with more complex sampling designs and estimators.

2.2 Simple random sampling with an interviewer effect

Next, we introduce interviewer variance into our measurement model for y. Each respondent is interviewed by one and only one interviewer. There are $R \in \mathbb{N}_{>0}$, interviewers that conduct the interviews of all n respondents. We denote $s_i \subset s$ as the set of all respondents that are interviewed by the i^{th} interviewer and $\mathcal{R} = \{1, \ldots, R\}$ as the set of all interviewers. The workload of the i^{th} interviewer is given by n_i, $\mathbf{n}_I = (n_1, \ldots, n_R)^{\top}$ is the vector of interviewer workloads and $\sum_{i=1}^{R} n_i = n$. Under measurement model (M_1), which follows the explanations of Särndal et al. (1992), page 623, the observed values of y for $k \in s_i$ are described as

$$y_{ik} = \mu_k + \mathfrak{I}_i + \mathfrak{e}_{ik}, \qquad (M_1)$$

with $\mathfrak{I}_i$ being the interviewer effect associated with all measurements conducted for respondents $k \in s_i$. $\mathfrak{e}_{ik}$ represents the random error due to sources other than the interviewer. All $\mathfrak{e}_{ik}$ for $i \in \mathcal{R}$ and $k \in s$ are iid random variables with zero mean and variance σ_e^2. $\mathfrak{I}_1, \ldots, \mathfrak{I}_R$ are iid random variables with zero mean and variance σ_I^2, which we call interviewer variance, and they are independent of $\mathfrak{e}_{ik}$ for all $i \in \mathcal{R}$ and $k \in s$. Särndal et al. (1992) interprets model (M_1) as a random assignment of interviewers to a pre-defined partition of the sample s into R disjoint subsets s_i, $i = 1, \ldots, R$. These subsets could correspond to different geographical areas where the survey is conducted and the interviewers are then randomly allocated to them. In practice, in many surveys fieldwork agencies assign interviewers to geographical areas based on experience and proximity. As this process is not necessarily observable by the researcher estimating the design effect, we assume a random allocation of interviewers to the PSUs. This can be seen as the recruitment of interviewers from an infinite, or very large, pool of possible interviewers.

If we define the random part in y_{ik} as $\varepsilon_{ik} = \mathfrak{I}_i + \mathfrak{e}_{ik}$, then the variance-covariance structure of y_{ik} under model (M_1) is given by

$$\operatorname{Cov}_{M_1}(\varepsilon_{ik}, \varepsilon_{jl}) = \begin{cases} \sigma^2, & \text{if } i = j,\ k = l \\ \rho_I \sigma^2, & \text{if } i = j,\ k \neq l, \\ 0, & \text{else} \end{cases} \qquad (2.3)$$

where $\sigma_I^2 + \sigma_e^2 = \sigma^2$ and $\rho_I = \frac{\sigma_I^2}{\sigma^2}$ is the correlation between two different observations of y made by the same interviewer. To derive the variance of $\bar{y}(\mathbf{w})$ under model (M_1), we first determine the variance

of $\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}y_{ik}$, where w_{ik} and y_{ik} are the survey weight and the observation for respondent $k \in s_i$, respectively. Thus we have

$$\begin{aligned}\mathrm{Var}_{M_1}\left(\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}y_{ik}\right) &= \sigma^2\left(\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2 \;+\; \rho_I \sum_{i\in\mathcal{R}}\sum_{k\in s_i}\sum_{\substack{l\in s_i\\ l\neq k}} w_{ik}w_{il}\right)\\ &= \sigma^2\left(\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2 \;+\; \rho_I\left[\sum_{i\in\mathcal{R}}\left(\sum_{k\in s_i} w_{ik}\right)^2 \;-\; \sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2\right]\right)\\ &= \sigma^2\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2\left(1 \;+\; \rho_I\left[\frac{\sum_{i\in\mathcal{R}}\left(\sum_{k\in s_i} w_{ik}\right)^2}{\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2} \;-\; 1\right]\right),\end{aligned}$$

from which follows

$$\mathrm{Var}_{M_1}\left(\bar{y}\left(\mathbf{w}\right)\right) \;=\; \frac{\sigma^2\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2}{\left(\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}\right)^2}\left(1 \;+\; \rho_I\left[\frac{\sum_{i\in\mathcal{R}}\left(\sum_{k\in s_i} w_{ik}\right)^2}{\sum_{i\in\mathcal{R}}\sum_{k\in s_i} w_{ik}^2} \;-\; 1\right]\right). \tag{2.4}$$

2.3 Multi-stage sampling with an interviewer effect

We consider a two-stage sampling design, where first PSUs are selected, and at the second stage respondents are selected from within the sampled PSUs. PSUs are the clustering units and we will treat the terms cluster and PSU as interchangeable. The sample of PSUs is denoted $\mathcal{K} = \{1, \ldots, K\}$, with $K > 1$. Each respondent belongs to one PSU and one PSU only. Let $s_q \subset s$ be the set of all respondents belonging to the q^{th} PSU, n_q be the number of respondents observed within the q^{th} PSU, $\mathbf{n}_C = (n_1, \ldots, n_K)^{\mathrm{T}}$ the vector of cluster sizes, and $\sum_{q\in\mathcal{K}} n_q = n$. Again, each respondent is interviewed by one interviewer and one interviewer only. Interviewers can work across PSUs and PSUs can be visited by multiple interviewers. Although interviewers might concentrate their work in a particular region, these regions are usually composed of multiple PSUs and interviewers do not work exclusively in one PSU only. This situation is frequently found in face-to-face surveys across Europe, e.g., in the ESS or EVS. Table 3.1 in Section 3.1 gives an overview on the level of interpenetration between PSUs and interviewer for countries that use a multi-stage sampling design in ESS6. Interpenetration between PSUs and interviewer can be observed across all ESS rounds for countries that use multi-stage sampling design.

We now introduce measurement model (M_2), which incorporates both cluster and interviewer variance into the observed values of y. For $k \in s_{qi} = s_q \cap s_i$ we model observations of y as

$$y_{qik} = \mu_k + \mathbb{C}_q + \mathfrak{I}_i + \mathrm{e}_{qik}, \tag{M_2}$$

with $\mathbb{C}_q$ defined as a random variable with mean zero and variance σ_C^2, which we call PSU variance, common to all respondents in PSU q. $\mathbb{C}_1, \ldots, \mathbb{C}_K$ are iid random variables and are independent of e_{qik}

and $\mathfrak{I}_i$ for all $i \in \mathcal{R}, q \in \mathcal{K},$ and $k \in s_{qi}$. $\mathfrak{C}_q$ introduces a certain degree of similarity between respondents from the same PSU. It allows for a permanent random effect of the PSU on the measurement of $y,$ for the k^{th} respondent, causing it to deviate from μ_k (Chambers and Skinner, 2003, page 201).

To establish the effect of sampling and interviewers on $\bar{y}(\mathbf{w}),$ we define the random part of y_{qik} as $\varepsilon_{qik} = \mathfrak{C}_q + \mathfrak{I}_i + \mathfrak{e}_{qik},$ which has the following variance-covariance structure

$$\text{Cov}_{M_2}\left(\varepsilon_{qik}, \varepsilon_{pjl}\right) = \begin{cases} \sigma^2, & \text{if } q = p,\ i = j,\ k = l \\ \rho_C\sigma^2, & \text{if } q = p,\ i \neq j,\ k \neq l \\ \rho_I\sigma^2, & \text{if } q \neq p,\ i = j,\ k \neq l \\ (\rho_I + \rho_C)\sigma^2, & \text{if } q = p,\ i = j,\ k \neq l \\ 0, & \text{else} \end{cases}, \tag{2.5}$$

where $\sigma_C^2 + \sigma_I^2 + \sigma_e^2 = \sigma^2$ and $\rho_C = \frac{\sigma_C^2}{\sigma^2}$ is the correlation between observation from the same PSU. The variance-covariance structure of ε_{qik} implies that the measurements of y are correlated if they are made within the same PSU or the same interviewer. Further, measurements of y are more homogeneous if they are made by the same interviewer within the same PSU. Model (M_2) represents a generalization of model M_4 of Gabler and Lahiri (2009), by removing the restriction that no interviewer works in more than one PSU.

The variance of $\bar{y}(\mathbf{w})$ under model (M_2) is given by

$$\text{Var}_{M_2}\left(\bar{y}(\mathbf{w})\right) = \frac{\sigma^2 \sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}^2}{\left(\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}\right)^2} \left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right]\right), \tag{2.6}$$

where w_{qik} and y_{qik} are the survey weight and the observation for respondent $k \in s_{qi},$ respectively, and

$$\bar{m}_I(\mathbf{w}) = \frac{\sum_{i\in\mathcal{R}}\left(\sum_{q\in\mathcal{K}}\sum_{k\in s_{qi}} w_{qik}\right)^2}{\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}^2} \quad \text{and} \quad \bar{m}_C(\mathbf{w}) = \frac{\sum_{q\in\mathcal{K}}\left(\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}\right)^2}{\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}^2}.$$

We can alter model (M_2) to allow for a PSU interviewer interaction effect, meaning that the covariance between the observations made by the same interviewer within the same PSU is not equal to the sum of the intra-PSU and intra-interviewer covariance. We call this measurement model (M_{2*}) and for $k \in s_{qi}$ the observation of y is modeled as

$$y_{qik} = \mu_k + \mathfrak{C}_q + \mathfrak{I}_i + \mathfrak{D}_{qi} + \mathfrak{e}_{qik}, \tag{M_{2*}}$$

with $\mathfrak{D}_{qi}$ as a random variable with mean zero and variance σ_{IC}^2 common to all respondents in PSU q that were interviewed by interviewer i. All $\mathfrak{D}_{qi}$ for $q \in \mathcal{K}$ and $i \in \mathcal{R}$ are iid random variables and are independent of $\mathfrak{e}_{qik}, \mathfrak{I}_i, \mathfrak{C}_q$ for all $q \in \mathcal{K}, i \in \mathcal{R},$ and $k \in s_{qi}$. Random effect $\mathfrak{D}_{qi}$ introduces some

additional correlation between observations made by the same interviewer within the same PSU, which cannot be explained by the separate PSU and interviewer variances.

For $k \neq l$ and $\varepsilon_{qik} = \mathfrak{C}_q + \mathfrak{I}_i + \mathfrak{D}_{qi} + \mathfrak{e}_{qik}$ we have under model (M_{2*}) $\text{Cov}_{M_{2*}}\left(\varepsilon_{qik}, \varepsilon_{qil}\right) = (\rho_I + \rho_C + \rho_{IC})\,\sigma^2$. Thus, we can write the variance of $\bar{y}(\mathbf{w})$ under model (M_{2*}) as

$$\text{Var}_{M_{2*}}\left(\bar{y}(\mathbf{w})\right) = \frac{\sigma^2 \sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}^2}{\left(\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}\right)^2} \left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right] + \rho_{IC}\left[\bar{m}_{IC}(\mathbf{w}) - 1\right]\right), \tag{2.7}$$

where

$$\bar{m}_{IC}(\mathbf{w}) = \frac{\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\left(\sum_{k\in s_{qi}} w_{qik}\right)^2}{\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} w_{qik}^2}.$$

2.4 Survey effect

After we establish the variance of $\bar{y}(\mathbf{w})$ under the different measurement models, we can define the effect associated with complex sampling and interviewers. We will refer to this effect as the *survey effect*, which we define as

$$\text{eff}_{ab}(\mathbf{w}) = \frac{\text{Var}_{M_a}\left(\bar{y}(\mathbf{w})\right)}{\text{Var}_{M_b}\left(\bar{y}(\mathbf{w})\right)}, \tag{2.8}$$

where M_a is the measurement model assumed for our survey of interest and M_b is the reference model. We use the term *survey effect* to distinguish $\text{eff}_{ab}(\mathbf{w})$ from design and interviewer effect, as $\text{eff}_{ab}(\mathbf{w})$ incorporates both effects. Other sources of variance, as described in the TSE framework, are not considered. Consequently, we will use the term *survey design* for the combination of a sampling design and interviewer workplan.

The survey effect associated with measurement model (M_2), is given by

$$\text{eff}_{20}(\mathbf{w}) = \frac{\text{Var}_{M_2}\left(\bar{y}_w\right)}{\text{Var}_{M_0}\left(\bar{y}\right)} = \text{eff}_w(\mathbf{w})\left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right]\right), \tag{2.9}$$

where

$$\text{eff}_w(\mathbf{w}) = \frac{n\sum_{k\in s} w_k^2}{\left(\sum_{k\in s} w_k\right)^2} \geq 1.$$

Factor $\text{eff}_w(\mathbf{w})$ does not depend on the measurement model and can be interpreted as a measure for the variance of the weights $\mathbf{w}$. If we write the variance of the weights as $\sigma^2_{\mathbf{w}} = 1/n \sum_{k \in s} w_k^2 - \bar{w}^2$, with $\bar{w} = 1/n \sum_{k \in s} w_k$, this relationship becomes more clear, as $\text{eff}_w(\mathbf{w}) = \text{CV}^2_{\mathbf{w}} + 1$, with $\text{CV}_{\mathbf{w}} = \sigma_{\mathbf{w}} / \bar{w}$ as the coefficient of variation of the survey weights. If the weights are all equal, then $\text{CV}_{\mathbf{w}} = 0$ and $\text{eff}_w(\mathbf{w})$ becomes 1. Terms $\bar{m}_I(\mathbf{w})$ and $\bar{m}_C(\mathbf{w})$ can be seen as measures for the average workload of the interviewers and the PSU size, respectively. If all weights are equal, $\bar{m}_I(\mathbf{w})$ has the value $\bar{m}_I(\mathbb{I}_n) = \sum_{i \in \mathcal{R}} n_i^2 / n$. Furthermore, if all interviewers have the exact same workload, i.e., $n_i = n/R$ for $i = 1, \ldots, R$, we have $\bar{m}_I(\mathbb{I}_n) = n/R$. $\bar{m}_C(\mathbf{w})$ has similar properties.

Following Gabler et al. (1999) and Gabler and Lahiri (2009) we can give the following upper bound for the survey effect.

Result 1.

$$\text{eff}^*_{20}(\mathbf{w}) \leq \text{eff}_w(\mathbf{w}) \text{eff}^*_{20}(\mathbb{I}_n),$$

where eff^*_{20} is the survey effect under the condition that $n_i = n/R$ for all $i \in \mathcal{R}$ and $n_q = n/K$ for all $q \in \mathcal{K}$. The upper bound of $\text{eff}^*_{20}(\mathbf{w})$, given in Result 1, follows from $\bar{m}_I(\mathbf{w}) \leq n/R$ if $n_i = n/R$ for all $i \in \mathcal{R}$ (Gabler et al., 1999). The proof is given in the Appendix. For $\bar{m}_C(\mathbf{w})$ an analogous result holds. It should be noted that, in general, we do not have

$$\text{eff}_{20}(\mathbf{w}) \geq \text{eff}_{20}(\mathbb{I}_n) = 1 + \rho_I \left[\sum_{i \in \mathcal{R}} \frac{n_i^2}{n} - 1 \right] + \rho_C \left[\sum_{q \in \mathcal{K}} \frac{n_q^2}{n} - 1 \right]. \tag{2.10}$$

That is, we cannot say that the survey effect is greater or equal to the survey effect of an equally weighted design. If the weights have the same relative frequency distribution across all sets s_{qi} inequality (2.10) holds (Gabler and Lahiri, 2009), i.e., if we have

$$n_{qig} = \frac{n_{qi}}{n} n_g, \quad g = 1, \ldots, G, \tag{2.11}$$

where G is the number of unique values in $\mathbf{w}$, n_g the frequency of the g^{th} weighting value, and n_{qig} the frequency of the g^{th} weighting value for respondents interviewed by the i^{th} interviewer in the q^{th} PSU.

We can, however, give a lower bound to $\text{eff}_{20}(\mathbf{w})$. Using the same argument that Gabler and Lahiri (2009) give in the proof of their Result 6, we get

$$\text{eff}_{20}(\mathbf{w}) \geq \left(1 + \rho_I \left[\frac{n}{R} - 1 \right] + \rho_C \left[\frac{n}{K} - 1 \right] \right). \tag{2.12}$$

With the right-hand side of inequality (2.12) an easy to calculate minimum of $\text{eff}_{20}(\mathbf{w})$ is given, which does not depend on the weights, the distribution of interviewer workloads, or the PSU sizes. This gives some valuable guidance at the planning stage of a survey design, as the planned survey effect of the survey should be at least as high as $\text{eff}^*_{20}(\mathbb{I}_n)$. The practical utility of the upper bound in Result 1 is somewhat limited by strong assumptions about $\mathbf{n}_I$ and $\mathbf{n}_C$. The further the values of $\mathbf{n}_I$ and $\mathbf{n}_C$ deviate

from the one point distribution of interviewer workloads and PSU sizes, the less this bound should serve as a guide. To give survey planners a less complex statistic to plan the value of $\bar{m}_I(\mathbf{w})$, Lynn and Gabler (2004) proposed using

$$\bar{m}'_I(\mathbf{w}) = \frac{H_{\mathbf{n}_I}}{H_{\mathbf{w}}}, \tag{2.13}$$

as a predictor for $\bar{m}_I(\mathbf{w})$, where $H_{\mathbf{n}_I} = \sum_{i\in\mathcal{R}}(n_i/n)^2$ is the Herfindahl index for the interviewer workload, a concentration measure, with $1/R \le H_{\mathbf{n}_I} \le 1$ (Fahrmeir, Heumann, Künstler, Pigeot and Tutz, 1997, page 83). $H_{\mathbf{n}_I} = 1$ corresponds to $R = 1$ and $H_{\mathbf{n}_I} = 1/R$ corresponds to $n_i = n/R$ for all $i \in \mathcal{R}$. $H_{\mathbf{w}} = \sum_{k\in s}\left(w_k / \sum_{k\in s} w_k\right)^2$ is the Herfindahl index for the weights. If equation (2.11) holds, we have $\bar{m}_I(\mathbf{w}) = \bar{m}'_I(\mathbf{w})$, but for most surveys this will not apply. For that reason, Lynn and Gabler (2004) suggested looking at $\mathrm{Cov}(w_{qik}, n_i)$, the covariance between the weights and interviewer workloads. The closer $\mathrm{Cov}(w_{qik}, n_i)$ is to zero the smaller the distance between $\bar{m}_I(\mathbf{w})$ and $\bar{m}'_I(\mathbf{w})$. Planning a survey with assumed values for $H_{\mathbf{n}_I}$ and $H_{\mathbf{w}}$ should be easier than with exact values of $\mathbf{n}_I$ and $\mathbf{w}$. Finding reasonable values for $H_{\mathbf{n}_I}$ and $H_{\mathbf{w}}$ could be guided by comparing these values from surveys with similar survey designs. Under equation (2.11) the findings are analogous for $\bar{m}_C(\mathbf{w})$.

It should be noted that we can also write $\mathrm{eff}_w(\mathbf{w})$ as

$$\mathrm{eff}_w(\mathbf{w}) = H_{\mathbf{w}} n. \tag{2.14}$$

The expression of $\mathrm{eff}_w(\mathbf{w})$ in equation (2.14) might also be useful at the planning stage of a survey, showing that it is possible to plan with a certain weight concentration, instead of specific values for $\mathbf{w}$.

Giving a general close upper bound for $\mathrm{eff}_{20}(\mathbf{w})$ is difficult if there are no restrictions on the values of $\mathbf{n}_I$, $\mathbf{n}_C$ and $\mathbf{w}$. However, survey weights are usually scaled to either the sample or the population size and it is not uncommon for them to be bounded. For example, the ESS provides weights to its users that are greater than zero and smaller or equal to 4 and scales them to the sample size (ESS, 2014c, 2014b). If $a \le w_k \le b$ for all $k \in s$ with $b < \infty$ and $a > 0$, then with a given value for $\mathbf{n}_I$ (or $\mathbf{n}_C$) upper limits of $\bar{m}_I(\mathbf{w})$, (or $\bar{m}_C(\mathbf{w})$) can be found, by solving a linear optimization problem. An upper limit for $\mathrm{eff}_w(\mathbf{w})$ can be deduced for given values of a and b, as shown in equation (A.5) in the Appendix.

The obtained upper bound of $\mathrm{eff}_{20}(\mathbf{w})$ will correspond to weight distributions with a very high concentration, i.e., a maximal number of the highest possible weights. However, adjusting the constraints of the linear optimization problem, based on the weight distribution of surveys with comparable sampling designs, can help to find bounds that are of higher practical relevance. (See Appendix for the formulation of this linear program.)

2.5 Corrected design effect

Now that we have established the survey effect of a survey design, we propose a new type of survey effect that we call *corrected design effect*. This statistic aims at quantifying the marginal effect of a complex survey design if an interviewer effect is present. We do this by defining the following effect

$$\begin{aligned}\mathrm{eff}_{21}(\mathbf{w}) &= \frac{\mathrm{Var}_{M_2}(\bar{y}_w)}{\mathrm{Var}_{M_1}(\bar{y})} \\ &= \mathrm{eff}_w(\mathbf{w})\,\mathrm{eff}_I\left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right]\right), \end{aligned} \tag{2.15}$$

where

$$\mathrm{eff}_I = \frac{n}{n + \rho_I\left(\sum_{i\in\mathcal{R}} n_i^2 - n\right)}.$$

The reference model (M_1) in $\mathrm{eff}_{21}(\mathbf{w})$ models a simple random sample with an interviewer effect. Factor eff_I, indicates how close the corrected design effect to the survey effect is. For $\mathrm{eff}_I = 1$ the corrected design and survey effect are equal and the closer eff_I is to zero the further apart are both effects. Hence, we can use eff_I to construct a measure for the contribution of the interviewer effect to the survey effect eff_{20}. For this, we first establish the following bounds for eff_I given in Result 2.

Result 2.

$$\frac{1}{n} \le \frac{n}{\rho_I(n-R)(n-R+1) + n} \le \mathrm{eff}_I \le \frac{R}{R + (n-R)\rho_I} \le 1.$$

The proof for Result 2 can be found in the Appendix.

Now we define a measure of the contribution of the interviewer effect to the survey effect inv_I as

$$\begin{aligned}\mathrm{inv}_I:&\quad \left[\frac{1}{n}, 1\right] \mapsto [0, 1], \\ \mathrm{inv}_I(a) &:= \frac{n(1-a)}{n-1} \quad \text{for} \quad \left[\frac{1}{n} \le a \le 1\right].\end{aligned} \tag{2.16}$$

For any given value of interviewer workloads $\mathbf{n}_I$, measure $\mathrm{inv}_I(\mathrm{eff}_I)$ is strictly increasing with decreasing eff_I. The maximum of $\mathrm{inv}_I(\mathrm{eff}_I)$ occurs at $R = 1$ and $\rho_I = 1$, which occurs when there is only one interviewer that always produces the same measurement. The minimum of $\mathrm{inv}_I(\mathrm{eff}_I)$ occurs at $\rho_I = 0$ for any given value of $\mathbf{n}_I$. If the concentration of the distribution of the workload over the interviewers increases and ρ_I stays fixed, $\mathrm{inv}_I(\mathrm{eff}_I)$ also increases. This relation becomes clearer if we write

$$\mathrm{eff}_I = \frac{1}{1 + \rho_I\left(H_{\mathbf{n}_I} n - 1\right)}. \tag{2.17}$$

Alternatively, the coefficient of variation for the interviewer workloads $\mathrm{CV}_{\mathbf{n}_I} = R\sigma_{\mathbf{n}_I}/n$, with $\sigma^2_{\mathbf{n}_I} = 1/R\sum_{i\in\mathcal{R}} n_i^2 - (n/R)^2$, could also be used to describe eff_I, since $H_{\mathbf{n}_I} = \left(1 + \mathrm{CV}^2_{\mathbf{n}_I}\right)/R$ (Lynn and Gabler, 2004). Note that for $\sigma^2_{\mathbf{n}_I} = 0$ we have $\mathrm{eff}_I = R/(R + (n-R)\rho_I)$.

Using Results 1 and 2, as well as inequality (2.12), we can give the following bounds for the corrected design effect.

Result 3.

$$\frac{n}{\rho_I(n-R)(n-R+1)+n}\,\mathrm{eff}_{20}^{*}(\mathbb{I}_n) \le \mathrm{eff}_{21}^{*}(\mathbf{w}) \le \mathrm{eff}_w(\mathbf{w})\,\frac{R}{R+(n-R)\rho_I}\,\mathrm{eff}_{20}^{*}(\mathbb{I}_n),$$

where eff_{21}^{*} is the corrected design effect when there are equal interviewer workloads and equal PSU sizes. The bounds of eff_{21}^{*}, given in Result 3 , do not depend on $\mathbf{n}_I$, but it should be noted that in the lower bound of $\mathrm{eff}_{21}^{*}(\mathbf{w})$, eff_I takes on its value for the maximum concentration in $\mathbf{n}_I$, whereas $\mathrm{eff}_{20}^{*}(\mathbb{I}_n)$ corresponds to the minimal concentration of $\mathbf{n}_I$. Since eff_I does not depend on $\mathbf{w}$, an upper (or lower) bound for $\mathrm{eff}_{21}(\mathbf{w})$ can be found by obtaining the upper (or lower) bounds of $\bar{m}_I(\mathbf{w})$, $\bar{m}_C(\mathbf{w})$ and $\mathrm{eff}_w(\mathbf{w})$ as described in the Appendix.

Finally, we introduce a corrected design effect that assumes the measurement model (M_{2^*}), given by

$$\begin{aligned}\mathrm{eff}_{2^*1}(\mathbf{w}) &= \frac{\mathrm{Var}_{M_{2^*}}(\bar{y}_w)}{\mathrm{Var}_{M_1}(\bar{y})}\\ &= \mathrm{eff}_w(\mathbf{w})\,\mathrm{eff}_I\left(1+\rho_I[\bar{m}_I(\mathbf{w})-1]+\rho_C[\bar{m}_C(\mathbf{w})-1]+\rho_{IC}[\bar{m}_{IC}(\mathbf{w})-1]\right). \quad (2.18)\end{aligned}$$

Similarly to Result 3 we can establish the following bounds for $\mathrm{eff}_{2^*1}(\mathbf{w})$.

Result 4.

$$\frac{n}{\rho_I(n-R)(n-R+1)+n}\,\mathrm{eff}_{2^*0}^{*}(\mathbb{I}_n) \le \mathrm{eff}_{2^*1}^{*}(\mathbf{w}) \le \mathrm{eff}_w(\mathbf{w})\,\frac{R}{R+(n-R)\rho_I}\,\mathrm{eff}_{2^*0}^{*}(\mathbb{I}_n).$$

Here $\mathrm{eff}_{2^*1}^{*}$ corresponds to the case where n_{qi}, the number of respondents that belong to the q^{th} PSU and are interviewed by the i^{th} interviewer, is a constant, i.e., $n_{qi} = n/(RK)$ for all $i \in \mathcal{R}$ and $q \in \mathcal{K}$. This also implies that for eff_{2^*1} we have $n_i = n/R$ and $n_q = n/K$. The proof of Result 4 can be found in the Appendix. Using model (M_{2^*}) instead of (M_2) gives some additional flexibility in fitting the measurement model to the observed data. Whether this is required is a part of Section 3.2, where the different measurement models are tested against each other for ESS6 data.

3 Empirical findings from the ESS

After we established the effects associated with interviewers and multi-stage or cluster sampling, we now estimate the survey effect and our proposed corrected design effect for ESS6 data (ESS, 2016).

There were 29 participating countries in ESS6 (ESS, 2018a), but not all have been considered in our analysis. We excluded all countries with a single-stage design (there were no single-stage cluster sampling designs in ESS6). In addition, we excluded those countries that had a multi-domain sampling design.

These countries employed different sampling designs in different regions of the country, but they all refer to a certain level of the Nomenclature of Territorial Units for Statistics (NUTS), as established by Eurostat (ESS, 2013, pages 21-22). For example, Norway used a single stage sample for its more densely populated regions, which, combined, contained almost 75 percent of the target population, and a two-stage sampling design for the rest of the country.

First, in Section 3.1 we assess whether the estimation of the measurement models described in Section 2 is generally feasible, given the PSU-interviewer structure found in ESS6. To this end, we use a model-based simulation study. In Section 3.2, we test the different measurement models against each other in order to use the most appropriate ones for the estimation of the survey effect and the corrected design effect. Afterwards, we compare our results with the design effect that was used by the ESS to plan the sample size.

The PSU and interviewer identification variables needed for our simulation study and the estimation of the effects were obtained from the so-called Sampling Design Data Files (SDDFs) and the Interviewer Questionnaire, respectively (ESS, 2014a). The SDDFs contain information on the sampling design, including a PSU identifier. For ESS6, the SDDFs have to be downloaded individually for each country (ESS, 2018b).

3.1 Simulation for the stability assessment of effect estimates

Interviewers and sampling have long been recognized as principal sources of survey error. The way interviewers are deployed during fieldwork makes it difficult to separate the interviewer variance from the PSU variance. To make data collection more efficient, interviewers are usually assigned to work exclusively in certain regions (Von Sanden, 2004, Section 1.3). Correspondingly, interviewers in ESS6 seldom work across regions. For ESS6, we observe the following situation: In general, interviewers work in a number of PSUs within a certain area, but never in all PSUs. PSUs might be visited by more than one interviewer, but never by all of them. For 25% of all considered countries, the mean number of regions (variable *region*, ESS (2013), pages 21-22) an interviewer visited was 1.017 or lower. For 75% of all countries, the mean number of regions per interviewer was 1.256 or lower.

The non-hierarchical structure of PSUs and interviewers can be considered typical of large scale social surveys like the ESS. A so-called *fully interpenetrated* survey design, where all interviewers work in all PSUs, is in general unfeasible for country-wide surveys. This makes it difficult to decide what amount of observed similarity between observations made by an interviewer is due to intra-interviewer correlation or instead due to intra-PSU correlation. This problem has been addressed in a number of studies. For instance, by using a fully nested survey design, where multiple interviewers work in the same PSU but not across them (Schnell and Kreuter, 2005). But also so-called *partially interpenetrated* surveys, where different interviewers work in multiple PSUs and PSUs are visited by multiple interviewers, have been analyzed, (Davis and Scott, 1995; O'Muircheartaigh and Campanelli, 1998). These partially interpenetrated surveys resemble more the situation we observe for ESS6.

To test our measurement models and to disentangle the different variance components, we fit a multilevel model with crossed random effects. In another context, Raudenbush (1993) proposes to allow for so-called *crossed effects* in the random effects structure. These crossed effects allow for the situation of partially interpenetrated factors, and are able to estimate all three variance components of measurement model (M_{2*}), σ_I^2, σ_C^2, and σ_{IC}^2.

Vassallo, Durrant and Smith (2017) show, using simulations on synthetic data, how well a multilevel model with crossed random effects for cluster and interviewer can estimate the variance-covariance structure of the data model under different patterns of interpenetration between cluster and interviewer. They identify the sample size, the number of interviewers and PSUs, and the level of interpenetration as the driving factor for the quality of the estimates of the variance components. The level of interpenetration plays a decisive role for the quality of the variance component estimates. Vassallo et al. (2017) found that already 2-3 interviewers per PSU lead to relatively stable estimates of the variance components. However, their survey designs were all balanced and symmetric, meaning that the interpenetration of PSUs by interviewers was constant for all PSUs and vice versa. This is not the case for countries in ESS6. Therefore, we perform a simulation to test whether under the partial interpenetrated survey designs of ESS6 the variance components of our measurement model (M_2) can be estimated or not.

For the simulation, we generate samples from a n-dimensional multi-variate normal distribution $\text{MVN}(\boldsymbol{\mu}, \Sigma)$. The vector of means $\boldsymbol{\mu}$ contains, for each dimension, the same value. The covariance matrix Σ follows the variance-covariance structure of measurement model (M_2) and was constructed for each country based on the observed PSU-interviewer structure. The variance components were set to $\sigma_I^2 = 0.2$, $\sigma_C^2 = 0.08$, $\sigma^2 = 2$. We generated 1,000 samples from the superpopulation model $\text{MVN}(\boldsymbol{\mu}, \Sigma)$ for each country and estimated measurement model (M_2) for each of these samples. The simulation was implemented in *R* (R Core Team, 2019). The samples for the simulation were generated with the help of the *mvtnorm* package (Genz, Bretz, Miwa, Mi and Hothorn, 2019) and the estimation of the model was done using the *lme4* package (Bates, Mächler, Bolker and Walker, 2015, 2019).

Table 3.1 depicts the relative Monte Carlo bias of the estimators for the variance components of model (M_2). For an estimator $\hat{\theta}$ of θ we define this measure as

$$\text{MC-RBias } \hat{\theta} = \frac{\bar{\hat{\theta}}_{\text{MC}}}{\theta} - 1,$$

where $\bar{\hat{\theta}}_{\text{MC}} = \sum_{d=1}^{D} \hat{\theta}_d / D$, θ is the true value, $\hat{\theta}_d$ the value of $\hat{\theta}$ for the d^{th} sample of the simulation and D is the total number of samples generated, i.e., $D = 1{,}000$, in our simulation. We see that σ_I^2 and σ_C^2 are estimated with a relative low bias for all considered countries in ESS6. In addition to the relative Monte Carlo bias, we have added the number of PSUs K, the number of interviewers R, the sample size n, the average number of PSUs that an interviewer works in $\bar{K}_I$, and the average number of interviewer that work in a PSU $\bar{R}_C$ to Table 3.1. $\bar{K}_I$ and $\bar{R}_C$ are used as measures for the level of interpenetration of PSUs by interviewers and interviewers by PSUs, respectively. For all countries, other than Germany, there

are more PSUs than interviewers and $\bar{K}_I$ is greater than $\bar{R}_C$. $\bar{K}_I$ reaches from 1.423 in Germany to 17.396 in Albania. The level of $\bar{K}_I$ observed for all countries seems to be high enough to disentangle the variance components of model (M_2). We can observe a negative relationship between $\bar{K}_I$ and MC-RBias $\hat{\sigma}_I^2$, which can be mediated by n and K. Higher n and K correspond to a higher accuracy of $\hat{\sigma}_I^2$. An analogous observation can be made for MC-RBias $\hat{\sigma}_C^2$. A higher $\bar{R}_C$ also improves the precision of the estimates and can compensate for a low $\bar{K}_I$. A high enough one-sided interpenetration, either of the PSUs by the interviewers or vice versa, is sufficient to accurately estimate σ_I^2 and σ_C^2 for model (M_2). For example, the Czech Republic, which has the lowest $\bar{R}_C$, but a $\bar{K}_I$ of round 1.848, enables relative precise estimates for the variance components.

It should be noted that for measurement model (M_{2*}), both $\bar{K}_I$ and $\bar{R}_C$ are of importance. For example, σ_C^2 and σ_{IC}^2 cannot be estimated with precision if $\bar{R}_C$ is too low. For example, in a similar simulation for model (M_{2*}), it was not possible to obtain accurate estimates of σ_C^2 and σ_{IC}^2 for the Czech Republic, although the relative bias of $\hat{\sigma}_I^2$ was around 1 percent.

For Bulgaria and Czech Republic $\bar{R}_C = 1$, that is, their PSUs are nested within the interviewers. In this case, we do not have crossed random effects, but nested random effects, as we never have the case where respondents are within the same PSU but not interviewed by the same interviewer. For this special case, strictly speaking, σ_C^2 should be labeled σ_{IC}^2. But, for simplicity, for both cases we use σ_C^2 as a label for the variances of the PSU random effect. This is not entirely unjustified, as σ_{IC}^2 defines the additional correlation between respondents that are in the same PSU, compared to those respondents that are interviewed by the same interviewer, but are in different PSUs.

Table 3.1
Relative bias of random effect variance estimates

	MC-RBias $\hat{\sigma}_I^2$	MC-Bias $\hat{\sigma}_C^2$	K	R	n	$\bar{K}_I$	$\bar{R}_C$
Albania	0.00	-0.02	264	53	1,201	17.40	3.49
Belgium	0.00	-0.02	363	155	1,869	3.00	1.28
Bulgaria	-0.01	0.04	400	247	2,260	1.63	1.00
Czech Republic	0.01	-0.01	426	231	2,009	1.85	1.00
France	0.01	0.01	267	165	1,968	1.99	1.23
Germany	0.01	-0.00	156	194	2,958	1.42	1.77
Ireland	-0.01	0.01	212	116	2,628	2.15	1.17
Israel	-0.00	0.01	190	114	2,508	3.00	1.80
Italy	-0.02	0.05	129	117	960	1.49	1.35
Kosovo	0.01	-0.02	160	72	1,295	2.29	1.03
Slovakia	-0.02	0.04	249	132	1,847	1.93	1.02
Slovenia	-0.01	0.00	150	50	1,257	3.30	1.10
Spain	-0.01	0.03	422	74	1,889	8.20	1.44
Ukraine	0.00	0.00	306	237	2,178	1.44	1.11
United Kingdom	-0.01	0.00	226	150	2,286	2.36	1.57

Our simulation study confirms and extends the findings of Vassallo et al. (2017) for the unbalanced situation of the ESS6. We also saw that the PSU-interviewer structure observed for ESS6 does not prohibit the disentanglement of σ_C^2 and σ_I^2 for measurement model (M_2).

3.2 Survey effects in ESS round 6

As seen in our simulation study, the estimation of the interviewer and cluster variance is feasible in ESS6. Now we test, for a set of selected variables from the ESS main questionnaire (ESS, 2013), each variance component of model (M_{2*}) on its significance. All used variables, except age and gender, have an ordinal scale, but are treated as metric variables for the purpose of this analysis. A list of all used variables can be found in the Appendix.

As a variance component has its minimum at zero, the test is performed on the boundary of the parameter space, which imposes classical problems from test theory. Scheipl, Greven, and Kuechenhoff (2008) proposed a restricted likelihood ratio test, designed to test for a zero random effects variance. We use their implementation of this test in the R-Package *RLRsim* and perform three test decisions.

First, we test on the significance of the interaction variance of interviewers and PSUs, when assuming relevant interviewer and PSU variances. Our null hypothesis is $H_0\colon \sigma_{IC}^2 = 0$ versus alternative hypothesis $H_A\colon \sigma_{IC}^2 > 0$. The per country average of rejected null hypothesis over the different variables is displayed in Table 3.2. The first two columns correspond to two different type I error levels for the test of $H_0\colon \sigma_{IC}^2 = 0$, indicated by $\alpha = 0.01$ and 0.05. Israel is the country that has the highest number for significant interaction variance σ_{IC}^2 on all type I error levels. For all other countries the null hypothesis is not rejected for all variables at a significance level of 1%. Although not displayed in Table 3.2 it can be noted that at a 10% significance level two-thirds of the countries have at least some variables with a significant interaction variance. Therefore, the possibility of an interaction effect should be considered when estimating survey effects.

In our second test decision an interviewer variance but no interaction variance is assumed. The null hypothesis is that the PSU variance is not relevant, that is $H_0\colon \sigma_C^2 = 0$ versus the alternative hypothesis $H_A\colon \sigma_C^2 > 0$. Average test results for the different type I error levels can be found in the columns 3 to 4 of Table 3.2. For some variables, the PSU variances are not significant as an addition to the interviewer variance. This result is especially strong for Belgium, where only 3% of the variables seem to have a PSU variance. However, also for France and Slovenia, the PSU variance is only significant at a level of 1% for a relative small number of the variables and for Albania for none of the variables. In contrast to that, Bulgaria, Ireland, Israel and Slovakia have significant PSU variance for the majority of variables. Overall, the PSU variance appears to be relevant in most countries and thus should be considered when estimating survey effects.

For the third test decision we perform, a PSU variance but no interaction variance is assumed. The null hypothesis is that the interviewer effect is not relevant $H_0\colon \sigma_I^2 = 0$ versus the alternative hypothesis $H_A\colon \sigma_I^2 > 0$. Average test results can be found in columns 5 to 6 of Table 3.2. The lowest rejection rates

are found in Germany and France, although 19% of the variables for Germany and 23% for France still have a significant interviewer variance at a 1% significance level. The other countries show a far higher proportion of variables with significant interviewer variance. On the 1% and 5% significance level, the interviewer variance has a higher rejection rate than the PSU variance for 13 out of the 15 countries. Thus, the interviewer variance appears to be of relevance for all countries in ESS6, indicating that possible interviewer effects should be taken into account when assessing the efficiency of survey designs.

Table 3.2
Rejection rates for existence of variance components

H_0:	$\sigma^2_{CI} = 0$		$\sigma^2_{C} = 0$		$\sigma^2_{I} = 0$	
α	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
Albania	0.00	0.03	0.00	0.16	0.55	0.77
Belgium	0.00	0.00	0.03	0.03	0.77	0.90
Bulgaria	0.00	0.00	0.81	0.90	0.90	1.00
Czech Republic	0.00	0.00	0.52	0.58	1.00	1.00
France	0.00	0.00	0.10	0.23	0.23	0.45
Germany	0.00	0.00	0.26	0.61	0.19	0.42
Ireland	0.00	0.06	0.77	0.81	0.94	0.97
Israel	0.13	0.32	0.94	1.00	0.84	0.94
Italy	0.00	0.03	0.10	0.32	0.42	0.65
Kosovo	0.00	0.00	0.45	0.58	0.94	0.97
Slovakia	0.00	0.00	0.77	0.90	0.97	0.97
Slovenia	0.00	0.00	0.03	0.16	0.74	0.84
Spain	0.00	0.00	0.13	0.23	0.74	0.84
Ukraine	0.00	0.00	0.55	0.74	0.90	0.94
United Kingdom	0.00	0.03	0.19	0.35	0.71	0.87

Based on the selected models for the different variables, survey effects defined in equation (2.8) are estimated. Table 3.3 shows the country specific average of estimated survey effects over all considered variables. In addition Table 3.3 also contains the average of design effect deff, as it is used by the ESS to plan sample sizes. In our notation this design effect has the form

$$\text{deff} = \text{eff}_w \left(1 + \rho_C \left(\bar{m}_C \left(\mathbf{w}\right) - 1\right)\right).$$

To estimate ρ_C in deff we used an ANOVA estimator (The ESS Sampling Expert Panel, 2016; Ganninger, 2010, page 45) and do not test for the significance of the PSU variance. Measurement model a used in eff_{a0} can include interviewer, PSU and interaction variance, if the model selection identifies it as significant at a level of 0.05. The same applies to measurement model a used in eff_{a1}, i.e., the corrected design effect. If interviewer variance is identified as not significant for a variable, then eff_{a1} becomes eff_{a0}. To measure the influence of the interviewer on the survey effect inv_I is also shown.

By comparing deff and eff_{a1} in Table 3.3 an interesting observation can be made: For Germany deff is clearly lower than for Ireland and the Czech Republic. From this we could deduce that Germany would

need a much lower sample size to achieve the same average effective sample size as Ireland and the Czech Republic. However, if we look at eff_{a1}, this relation switches. Table 3.3 shows that the cluster effect of the complex sampling design is higher in Germany than it is in Ireland or the Czech Republic. Meaning that, if we are interested in equal average effective sample across countries, Germany would need a higher sample size than in Ireland or the Czech Republic. For example, for the Czech Republic to achieve an effective sample size of 1,500 with the standard design effect deff from Table 3.3 we would plan with a net sample of round 3,925 and for Germany with one of 3,115. If instead we use the corrected design effect eff_{a1}, to base the planning of the net sample size solely on the effect of the sampling design, we would select a net sample size of round 1,707 and 2,598, for the Czech Republic and Germany, respectively. This finding is also reflected in the values of inv_I, which indicates that a large part of eff_{a0} for Ireland and the Czech Republic can be attributed to an interviewer effect, whereas for Germany, the interviewer effect is smaller and eff_{a0} seems to be dominated by the cluster effect. Apart from Israel, Slovakia, and Slovenia, all countries have different ranks for deff and eff_{a1}, indicating that the allocation of the sample size over all countries would be very different, if the corrected design was used to plan effective samples sizes, instead of the conventional design effect deff.

Table 3.3
Average effect sizes for ESS6

	deff	eff_{a0}	eff_{a1}	inv_I
Albania	2.07	2.87	1.68	0.35
Belgium	1.18	1.75	1.01	0.37
Bulgaria	2.32	3.88	1.21	0.65
Czech Republic	2.62	6.58	1.14	0.78
France	1.69	1.80	1.46	0.16
Germany	2.08	2.28	1.73	0.19
Ireland	3.32	5.42	1.26	0.73
Israel	2.41	4.67	1.42	0.61
Italy	1.76	2.20	1.32	0.34
Kosovo	4.01	10.97	1.51	0.80
Slovakia	5.02	20.28	2.27	0.85
Slovenia	1.59	3.03	1.06	0.55
Spain	1.16	2.01	1.05	0.42
Ukraine	2.97	5.61	1.18	0.73
United Kingdom	1.76	2.24	1.32	0.38

deff: average design effects as defined in equation (1.1).
eff_{a0} : average survey effect with measurement model of interest (M_a) and (M_0) as reference.
eff_{a1} : average corrected design effects with measurement model of interest (M_a) and (M_1) as reference.
inv_I : average contribution of interviewer effects to the design effect as defined in equation (2.16).

eff_{a1} is smaller than deff for all countries, and their distance, $|\text{deff} - \text{eff}_{a1}|$, has a positive but non-linear relationship with inv_I. The lowest values of $|\text{deff} - \text{eff}_{a1}|$ are observed for Spain, Belgium,

France, and Germany, which are all countries whose inv_I value is below the median of inv_I. The opposite is observed for Slovakia, Kosovo, Ireland, and Ukraine, the countries with the highest distance between deff and eff_{a1}. These countries all have a value of inv_I that is higher than the median value of inv_I. These patterns for countries with a relatively high distance between deff and eff_{a1} are consistent with what we would expect if there is a high interviewer effect present in the data. The opposite can be said for countries when a relatively small distance between deff and eff_{a1} is observed.

Interviewer effects depend on many different factors (West and Blom, 2017), including the type of the question asked and the used ESS6 data is mostly gathered from attitude questions. Hence, the presented results in this section cannot be extrapolated to other types of surveys in the same countries.

4 Conclusions

Using a design effect to select a sample size is a commonly used method to account for the loss of efficiency that a complex sampling design might entail. However, the design effect can be inflated by an interviewer effect in face-to-face surveys. This can lead to erroneous conclusions about the effect that complex sampling has on the efficiency of a sampling strategy. As a consequence, this could lead to misallocation of resources. The planned sample size might be too high, if it is based on an overestimated design effect. Therefore, we propose to consider both the design and the interviewer effect simultaneously when planning a sample size. The survey effect, which we develop in Section 2, accounts both for interviewer and PSU variance to assess the efficiency of a survey design. Based on the survey effect we introduce a corrected design effect, which uses as a reference design a simple random sample with an interviewer effect. As a result, the corrected design effect is no longer conflated with the interviewer effect and can be used to better base the decision on the samples size on the effect the sampling design has on the precision of survey estimates.

For ESS6, our empirical findings in Section 3.2 show that high design effects are related to high interviewer effects. The average corrected design effects that we observe suggest that the sampling design influences the variance of an estimator to a lesser degree than interviewers for many countries in the ESS6. The ability to estimate the corrected design effect, e.g., from historical data as guide for the survey planner, depends mainly on the PSU-interviewer structure and the allocation of interviewer workloads and cluster sizes. We find a partially interpenetrated survey design, i.e., on a regional level, can be sufficient to disentangle PSU and interviewer variance. In our simulation study an average number of 1.5 PSUs per interviewer or interviewers per PSU was enough to estimate the variance components of measurement model (M_2). For actual survey data, that is categorical, this level of interpenetration might not be high enough, but a high number of PSUs, interviewers, and a large sample size might off-set a low interpenetration. For practical applications, we recommend testing via simulation if the assumed measurement model can be estimated with the given PSU-interviewer structure, as we did in Section 3.1.

When using the survey effect and corrected design effect for the planning of a sample size it can be helpful to work with the upper and lower bounds of these statistics. In Section 2, we derive such bounds, but under somewhat unrealistic assumptions regarding the distribution of survey weights, interviewer workloads and PSU sizes. However, if realistic assumptions about the concentration of survey weights, interviewer workloads and PSU sizes can be made, then we propose to use a linear optimization, as shown in the Appendix, to derive bounds that are of much higher practical relevance and can serve as valuable guidance for survey planners. Generally, we recommend to have lowly concentrated distributions of interviewer workloads and PSU cluster sizes in order to increase the precision of survey estimates. Thus, interviewer workloads and PSU cluster sizes should be as equal as possible for any given number of interviewer and PSUs.

The measurement models we introduce in Section 2 are arguably simplistic. This makes the models applicable to most survey designs. The only information, besides the survey data, used to compute the estimates for Table 3.3 were the PSU and interviewer indicators. However, there are certain aspects of survey measurements that could be incorporated into a practical measurement model, such as stratification, which, in general, increases the efficiency of an estimation strategy (Särndal et al., 1992, Section 3.7). This was neglected in our analysis, despite the fact that many ESS6 countries used a stratified design for their PSU sample. Gabler, Häder and Lynn (2006) develop a design effect for estimation strategies that combine different sampling designs for sampling domains. This approach could possibly be adapted to add a stratification effect to the PSU variance. Furthermore, it might be plausible to assume that interviewers differ with regard to the degree of homogeneity that they add to their measurements. This interviewer heterogeneity could be incorporated into a measurement model by allowing groups of interviewers to have different distributions of $\mathfrak{I}_i$, i.e., values for σ_I^2 (West and Elliott, 2014). However, a procedure to classify interviewers would be needed. Preferably one that does mainly rely on the survey data and not so much on information available about the interviewers, which might differ from survey to survey.

A future application for the presented framework of the survey effect would be to find an optimal budget allocation with respect to the number of PSUs and interviewers, for a given effective sample size. Such an optimization requires a cost model for the deployment of interviewers to a possible set of PSUs. Fieldwork institutes could possibly provide the necessary information to calculate such a model for a particular country. Such a method could help survey planners to conduct face-to-face surveys more effectively, which is of increasing importance as surveys based on probability samples are under pressure from the comparably cheap alternative of recruiting respondents from online-access panels.

Further research could also focus on the development of survey effect for other estimators than the weighted sample mean. For estimators that can be described as functions of estimated totals, which includes the Ordinary Least Square Estimator for regression coefficients (Särndal et al., 1992, Section 5.10), it should be possible to derive survey effects, under the framework shown in Section 2, that allow for a similar factorization as the survey effect presented in this work.

Appendix

For the Appendix we will introduce a short notation of multiple sums, where, for example, $\sum_{qik} y_{qik}$ will be shorthand for

$$\sum_{q\in\mathcal{K}}\sum_{i\in\mathcal{R}}\sum_{k\in s_{qi}} y_{qik}.$$

Results 1

$$\mathrm{eff}^{*}_{20}(\mathbf{w}) \le \frac{n\sum_{qik} w_{qik}^2}{\left(\sum_{qik} w_{qik}\right)^2}\left(1+\rho_I\left[\frac{n}{R}-1\right]+\rho_C\left[\frac{n}{K}-1\right]\right).$$

Proof: We need to show that

$$\frac{\sum_i\left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w_{qik}^2} \le \frac{n}{R} \tag{A.1}$$

and

$$\frac{\sum_i\left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w_{qik}^2} \le \frac{n}{K} \tag{A.2}$$

hold, if $n_i = \frac{n}{R}$ and $n_q = \frac{n}{K}$, for all $i = 1, \ldots, R$ and $q = 1, \ldots, K$.

As shown in Gabler et al. (1999), if $a_{qik} = 1$ for all $q \in \mathcal{K}$, $i \in \mathcal{R}$, $k \in s_{qi}$, using the Cauchy-Schwarz inequality, we know that

$$\left(\sum_{qk} w_{qik}a_{qik}\right)^2 = \left(\sum_{qk} w_{qik}\right)^2 \le n_i\sum_{qk} w_{qik}^2 = \sum_{qk} a_{qik}^2\sum_{qk} w_{qik}^2$$

$$\frac{\sum_i\left(\sum_{qk} w_{qik}\right)^2}{\sum_i\sum_{qk} w_{qik}^2} \le \frac{\sum_i n_i\sum_{qk} w_{qik}^2}{\sum_i\sum_{qk} w_{qik}^2}.$$

If we have $n_i = \frac{n}{R}$ for all $i = 1, \ldots, R$, then it follows that

$$\frac{\sum_i\left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w_{qik}^2} \le \frac{n}{R}.$$

The proof for inequality (A.2) is analogous to the one above, which completes the proof of Result 1.

Upper bounds for $\bar{m}_I(\mathbf{w})$, $\bar{m}_C(\mathbf{w})$ and $\mathrm{eff}_w(\mathbf{w})$

For given $\mathbf{n}_I^{\top}$ and $\mathbf{n}_C^{\top}$ and $w_k \in [a, b]$ with $a, b \in \mathbb{R}_+$ for all $k \in s$, and $\sum_k w_k = n$ we can construct an upper bound for $\bar{m}_I(\mathbf{w})$ and $\bar{m}_C(\mathbf{w})$.

We know that

$$\frac{\sum_i\left(\sum_{qk} w_{qik}\right)^2}{\sum_i\sum_{qk} w_{qik}^2} \le \frac{\sum_i n_i\sum_{qk} w_{qik}^2}{\sum_i\sum_{qk} w_{qik}^2} \le \frac{\sum_i n_i\sum_{qk} w_{qik}^2}{n}. \tag{A.3}$$

Now we need to find a sufficiently high value for $\sum_i n_i \sum_{qk} w_{qik}^2$. For this we define $x_i = \sum_{qk} w_{qik}^2$ and $\mathbf{x} = (x_1, \ldots, x_I)^\top$. Thus we have to solve the following problem:

$$\begin{aligned} &\max_{\mathbf{x} \in \mathrm{R}^R} \mathbf{n}_I^\top \mathbf{x} \\ &\text{s.t.} \\ &x_i \geq a^2 n_i \ \forall i \in \mathcal{R} \\ &x_i \leq b^2 n_i \ \forall i \in \mathcal{R} \\ &\sum_i x_i \geq n \\ &\sum_i x_i \leq f_{sqm}(a, b, n), \end{aligned} \tag{A.4}$$

where

$$f_{sqm}(a, b, n) = b^2 \left\lfloor \frac{n - nb}{a - b} \right\rfloor + (n - nb) - \left\lfloor \frac{n - nb}{a - b} \right\rfloor (a - b) + b + a^2 \left(n - \left\lfloor \frac{n - nb}{a - b} \right\rfloor - 1 \right),$$

where $\lfloor \rfloor$ means rounded to the nearest lower integer. The problem formulated in equation (A.4) can be solved using a solver for linear programs, e.g., with the *solveLP* function from the *R* package Henningsen (2012). Function f_{sqm} gives a maximum of $\sum_k w_k^2$ given the upper and lower bounds of the weights a and b and the fact that the weights are scaled to n, i.e., $\sum_k w_k = n$. The sum of squares is maximized by giving as many weights their highest possible value b under the condition that each weight must have at least a value of a and that $\sum_k w_k = n$. The problem can then be solved using a simplex algorithm. An upper bound for $\bar{m}_C$ can be determined in the same fashion. Changing the problem to minimization and a lower bound for eff_{20} can be found. However, it is not guaranteed that separate optimization of $\bar{m}_C$ and $\bar{m}_I$ will yield values of $\mathbf{x}$ that allow for a value of $\mathbf{w}$ that jointly maximizes (or minimizes) $\bar{m}_C$ and $\bar{m}_I$. Although, if, $\mathbf{x}_C$ and $\mathbf{x}_I$ are the vectors that optimizes $\bar{m}_C$ and $\bar{m}_I$ respectively, it should be possible to find a possible value for $\mathbf{w}$, e.g., using iterative proportional fitting.

For $\mathrm{eff}_w(\mathbf{w})$ we have under the same assumptions as made above

$$1 \leq \mathrm{eff}_w(\mathbf{w}) = \frac{\sum_{k \in s} w_k^2}{n} \leq \frac{f_{sqx}(a, b, n)}{n}. \tag{A.5}$$

Result 2

$$\frac{n}{\rho_I (n - R)(n - R + 1) + n} \leq \mathrm{eff}_I \leq \frac{R}{R + (n - R)\rho_I}.$$

Proof: The upper bound in Result 2 can be shown by using the Cauchy-Schwarz inequality, which gives us

$$\begin{aligned} R \sum_i n_i^2 &\geq \left(\sum_i n_i \right)^2 \\ \sum_i n_i^2 &\geq \frac{n^2}{R}. \end{aligned} \tag{A.6}$$

With a some algebra we can formulate the upper bound of eff_I.

To prove the lower bound in Result 2 we solve the following problem:

$$\max_{\mathbf{n}_I \in \mathrm{N}_{>0}^R} \mathbf{n}_I^\top \mathbf{n}_I$$
$$\text{s.t.} \tag{A.7}$$
$$\sum_i n_i = n.$$

A solution to the problem formulated in (A.7) can be found by considering that if we have $n_i - 1 \geq 1$ and $n_i \leq n_j$ it follows that $(n_i - 1)^2 + (n_j + 1)^2 > n_i^2 + n_j^2$. Thus for $n_j = \max_{i \in \mathcal{R}} n_i$ we can increase $\sum_i^R n_i^2$ if we reduce any $n_i > 1$ $i \neq j$ by one and add one to n_j. Hence, if $n_i = 1$ for all $i \neq j \in \mathcal{R}$ and $n_j = n - R + 1$ then $\sum_i n_i^2$ is at its maximum, with $\sum_i n_i^2 = (R - 1) + (n - R + 1)^2$.

Result 4

Proof: Given Result 2, to prove the right-hand side of Result 4 we need to show that

$$\mathrm{eff}_{2*0}^*(\mathbf{w}) \leq \frac{n \sum_{qik} w_{qik}^2}{\left(\sum_{qik} w_{qik}\right)^2} \left(1 + \rho_I \left[\frac{n}{R} - 1\right] + \rho_C \left[\frac{n}{K} - 1\right] + \rho_{IC} \left[\frac{n}{RK} - 1\right]\right). \tag{A.8}$$

To prove inequality (A.8) we only need to show that

$$\frac{\sum_{qi} \left(\sum_k w_{qik}\right)^2}{\sum_{qik} w_{qik}^2} \leq \frac{n}{RK}.$$

The rest follows from the proofs of inequalities (A.1) and (A.2). Thus it is sufficient to show that

$$\left(\sum_k w_{qik}\right)^2 = \left(\sum_k w_{qik} a_{qik}\right)^2 \leq n_{qi} \sum_k w_{qik}^2 = \sum_k a_{qik}^2 \sum_k w_{qik}^2,$$

if $a_{qik} = 1$ for all $q \in \mathcal{K}$, $i \in \mathcal{R}$, $k \in s_{qi}$, which also follows from the Cauchy-Schwarz inequality. Inequality (A.8) then follows if $n_{qi} = \frac{n}{RK}$ for $i = 1, \ldots, R$ and $q = 1, \ldots, K$.

The left-hand side of Result 4 follows from the proof of Result 6 in Gabler and Lahiri (2009) and Result 2.

ESS6 variables used for empirical evaluation

Table A.1
ESS6 variables used for empirical evaluation

pplfair	trstprt	stfdem	imueclt	iorgact
pplhlp	trstep	stfedu	imwbcnt	agea
polintr	trstun	stfhlth	happy	gndr
trstprl	lrscale	gincdif	aesfdrk	
trstlgl	stflife	freehms	health	
trstplc	stfeco	euftf	rlgdgr	
trstplt	stfgov	imbgeco	wkdcorga	

The definition of these variables including question text can be found in ESS (2013).

References

Bates, D.M., Mächler, M., Bolker, B.M. and Walker, S.C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1, 1-48. https://doi.org/10.18637/jss.v067.i01.

Bates, D.M., Mächler, M., Bolker, B.M. and Walker, S.C. (2019). *Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4.

Beullens, K., and Loosveldt, G. (2016). Interviewer effects in the European social survey. *Survey Research Methods*, 10, 2, 103-118.

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, Oxford University Press, 74, 5, 817-848.

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Chaudhuri, A., and Stenger, H. (2005). *Survey Sampling: Theory and Methods*. CRC Press.

Davis, P., and Scott, A.(1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, 21, 2, 99-106. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1995002/article/14405-eng.pdf.

Ellis, P.D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

European Social Survey (ESS) (2013). *ESS6 Data Protocol. 1.4*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/download.html?r=6.

European Social Survey (ESS) (2014a). *European Social Survey Round 6 Interviewer Questionnaire*. Dataset edition: 2.1. London: ESS ERIC.

European Social Survey (ESS) (2014b). *Weighting European Social Survey Data*. London: ESS ERIC. www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf.

European Social Survey (ESS) (2014c). *ESS6 – 2012 Documentation Report*. Edition: 2.3. London: ESS ERIC. http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_data_documentation_report_e02_3.pdf.

European Social Survey (ESS) (2016). *European Social Survey Round 6 Data*. Dataset edition: 2.2. London: ESS ERIC.

European Social Survey (ESS) (2018a). *Countries by Round (Year)*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/country_index.html.

European Social Survey (ESS) (2018b). *Data and Documentation by Round European Social Survey (ESS)*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/download.html?r=6.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. and Tutz, G. (1997). *Statistik: Der Weg Zur Datenanalyse*. 1st ed. Berlin: Springer-Verlag.

Fischer, M., West, B.T., Elliott, M.R. and Kreuter, F. (2018). The impact of interviewer effects on regression coefficients. *Journal of Survey Statistics and Methodology*, May. https://doi.org/10.1093/jssam/smy007.

Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 1, 105-106. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4718-eng.pdf.

Gabler, S., Häder, S. and Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, 32, 1, 115-120. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9256-eng.pdf.

Gabler, S., and Lahiri, P. (2009). On the definition and interpretation of interviewer variability for a complex sampling design. *Survey Methodology*, 35, 1, 85-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10886-eng.pdf.

Ganninger, M. (2010). *Design Effects: Model-Based Versus Design-Based Approach*. Edited by GESIS - Leibniz-Institut für Sozialwissenschaften. Array 3.

Genz, A., Bretz, F., Miwa, T., Mi, X. and Hothorn, T. (2019). *Mvtnorm: Multivariate Normal and T Distributions*. https://CRAN.R-project.org/package=mvtnorm.

Groves, R.M. (2009). *Survey Methodology*. 2nd ed. Wiley Series in Survey Methodology. Hoboken, New York: John Wiley& Sons, Inc.

Groves, R.M., and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, Oxford University Press, 74, 5, 849-879.

Henningsen, A. (2012). *Linprog: Linear Programming/Optimization*. https://CRAN.R-project.org/package=linprog.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 297, 92-115. https://doi.org/10.1080/01621459.1962.10482153.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lohr, S.L. (2014). Design effects for a regression slope in a cluster sample. *Journal of Survey Statistics and Methodology*, 2, 2, 97-125. https://doi.org/10.1093/jssam/smu003.

Lynn, P., and Gabler, S. (2004). *Approximations to B* in the Prediction of Design Effects Due to Clustering*. ISER Working Paper Series.

Lynn, P., Häder, S., Gabler, S. and Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European social survey experience. *Journal of Official Statistics*, 23, 1, 107.

O'Muircheartaigh, C., and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161, 1, 63-77.

Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 4, 321-349. https://doi.org/10.2307/1165158.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

Scheipl, F., Greven, S. and Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52, 7, 3283-3299.

Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 3, 389-410.

The ESS Sampling Expert Panel (2016). *Sampling Guidelines: Principles and Implementation for the European Social Survey*. London: ESS ERIC Headquarters. http://www.europeansocialsurvey.org/docs/round8/methods/ESS8_sampling_guidelines.pdf.

Vassallo, R., Durrant, G. and Smith, P. (2017). Separating interviewer and area effects by using a cross-classified multilevel logistic model: Simulation findings and implications for survey designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 2, 531-550.

Von Sanden, N.D. (2004). *Interviewer Effects in Household Surveys: Estimation and Design*. Ph.d. Thesis, Wollongong: University of Wollongong. http://ro.uow.edu.au/theses/312.

West, B.T., and Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 2, 175-211. https://doi.org/10.1093/jssam/smw024.

West, B.T., and Elliott, M.R. (2014). Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Survey Methodology*, 40, 2, 163-188. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14092-eng.pdf.