Statistics Netherlands

Discussion Paper

Adaptive survey designs tot minimize survey mode effects

A case study on the Dutch Labor Force Survey

The views expressed in this paper are those of the author(s) and do not necesarily reflect the policies of Statistics Netherlands

2013 | 12

Melania Calinescu
Barry Schouten

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68
ISSN 1572-0314

60083 201312 X-10

Adaptive survey designs to minimize survey mode effects
A case study on the Dutch Labor Force Survey

Melania Calinescu, Barry Schouten

Summary: Assessing the impact of mode effects on survey estimates has become a crucial question due to the increasing appeal of mixed-mode designs. Despite the advantages of a mixed-mode design such as lower costs and increased coverage, there is sufficient evidence that mode effects may sometimes be large relative to the precision. They may lead to incomparable statistics in time or over population subgroups and they may increase bias. Adaptive survey designs offer a flexible mathematical framework to obtain the optimal balance between survey quality and costs. In this paper we employ adaptive designs in order to minimize mode effects. We illustrate our optimization model by means of a case-study on the Dutch Labor Force Survey. We focus on item-dependent mode effects and we evaluate the impact on survey quality by comparison to a "gold standard".

Keywords: Mode effects; Survey costs; Survey quality; Adaptive survey designs.

Acknowledgements: The authors would like to thank dr. Sandjai Bhulai (VU University Amsterdam) for his constructive comments on the mathematical framework presented in the current paper. The authors also thank Boukje Janssen (CBS) and Martijn Souren (CBS) for processing the raw field data for analysis and Joep Burger (CBS) for his comments that helped improve this paper.

1 Introduction

Survey mode effects appear when differences are noticed between estimates obtained from surveys using different survey modes, i.e., when the same question asked in different modes receives different answers. Assessing the impact of mode effects on survey estimates has become a crucial question due to the increasing appeal of mixed-mode designs. There are multiple reasons why survey practitioners prefer less and less unimode designs such as increased costs in carrying out face-to-face surveys, decreasing coverage in telephone surveys and low participation in online surveys (Fan and Yan 2010). As a consequence, survey organizations have been steadily restructuring their unimode designs into mixed-mode designs. However, there is significant theoretical and practical evidence (see Jäckle et al. 2010 and Dillman et al. 2009) that mode effects may sometimes be large relative to the precision. They may lead to incomparable statistics in time or over population subgroups and they may increase bias. Assessment of mode effects does not however follow a generally accepted technique. Literature and experimental studies note various viewpoints on what survey components suffer more from mode effects and how to test for manifestation of mode effects. The common ground among these perspectives is determining whether mode effects are item-specific or systematic phenomena (Klausch et al. 2013). Subsequently, when they are item-specific, changes to the survey design can be made in order to address the corresponding mode effects. If however, systematic errors are observed, then modes are incomparable and carrying out a mixed-mode design could result into misleading conclusions.

In the current paper, we investigate adaptive survey designs that address item-specific mode effects. Adaptive survey designs optimize quality of survey response given constraints on costs, or vice versa. For a general introduction to adaptive survey designs, see Schouten et al. (2013a). Survey mode effects stream from mode selection effect, i.e., different people have access to different modes, and mode measurement effects. However, as reported in Schouten et al. (2013b), for the Dutch Labor Force Survey (LFS), the mode selection effect for target variables can be adjusted for, due to strong auxiliary information available from registers. As a consequence, in the LFS case, the analysis may focus only on mode measurement effects. Moreover, it enables analysis of the adjusted mode effects, i.e., analysis of differences between modes after nonresponse adjustments.

In our analysis, we consider three survey modes, namely, web (CAWI), phone (CATI) and face-to-face (CAPI). As observed from historical data, web surveys are cheap to run but possibly more prone to measurement effects than CATI or CAPI. At the same time, CAPI surveys produce more reliable survey estimates but they are very expensive to run. A mixed-mode design balances costs but the mode measurement effects that may thus occur are harder to quantify and control. Therefore, in the current paper we investigate what mode combinations, from a specified list of combinations, should form the adaptive design such that overall mode measurement effects are minimized.

We focus our analysis on minimizing mode measurement effects that may impact the unemployment rate estimate, one of the key statistics produced in the LFS. In the recent years, the LFS design underwent a series of changes in its transition from a full face-to-face survey to a mixed-mode survey. Thus, extensive knowledge on the interaction between survey design features, survey mode in particular, and the response process is available. However, the approach we propose in quantifying mode measurement effects is survey item-dependent. For surveys with multiple survey variables, one would need to summarize mode effects across survey items. The method presented in Calinescu and Schouten (2013), where measurement profiles and response styles are used to summarize measurement effects across survey items, is designed for these settings.

The outline of the paper is as follows. Section 2 enumerates some of the difficulties that arise in the attempt to quantify mode effects and gives recommendations on possible techniques to deal with such issues. Section 3 presents the model formulation where we evaluate the mode measurement effect by comparison to a selected benchmark estimate and Section 4 describes the optimization algorithm. The optimization results are presented in Section 5 and Section 6 concludes the paper.

2 Survey mode effects: an introduction

Running mixed-mode surveys offers many advantages compared to unimode surveys such as lower costs and increased coverage. However, their implementation comes with a series of potential difficulties of which most troublesome is data comparability across modes. de Leeuw (2005) and Jäckle et al. (2010) note that before designing and implementing mixed-mode surveys survey practitioners should be able to understand and quantify the impact of mode effects on data quality. Therefore, extensive research by means of field studies and testing of mode differences is strongly recommended.

The most common framework to assess mode effects is given by the cognitive models of survey response process (see Tourangeau et al. 2000) which analyze the phases of the response process, i.e., interpretation and comprehension of the survey question, information retrieval, judgment and reporting of the answer, that are influenced by mode. As a result, manifestation of an answering behavior, e.g., social desirability, satistificing, is perceived as a mode effect that can lead to response bias. Another method for assessing mode effects is to test for differences in various quality indicators and response distributions (see Link and Mokdad 2005, Greenfield et al. 2000). If significant differences are displayed across modes then the analyzed survey items or indicators are subject to mode effects. A third approach is a model-based approach that analyzes the impact of various modes on the probability of providing the same answer under the different modes for two persons that possess the same true state on the question topic (see Millsap 2011).

Although sufficient tools are available to identify response differences across modes, the main challenge is how to decide that such differences translate into data quality difference between surveys (see Biemer 1988). Differences in mode coverage, sampling frames and nonresponse bias could easily perturb responses, making it hard for the researchers to disentangle the mode effect. Additionally, differences in questionnaire design across modes are an easy trap for overstatement of the mode effects existence (see Dillman 2000). The reason is that even if a questionnaire is designed specifically for a given survey mode, mode effects may still occur. Furthermore, mode effects may impact only certain survey estimates (see de Leeuw 1992). Dillman et al. (2009) have shown that different people are attracted by different modes which results in an inhomogeneous sample creating thus a selection effect next to a measurement effect.

As suggested by Biemer (1988), evaluating the mode effect impact on data quality could be done by comparing responses to a "gold standard" such as external records or prior knowledge on the direction of error. Following this recommendation we develop a mode effect indicator that aims at quantifying the deviation caused by mode on a survey item against a selected benchmark. Subsequently, we develop an optimization model that assigns optimally survey resources in order to minimize the mode effect impact on data quality given the mode effect indicator. We apply this evaluation method on the Dutch LFS in order to assess the mode effect on one of the survey items, i.e., the unemployment rate estimate. In the following we present the optimization model to develop an adaptive design that minimizes the mode effects given the selected benchmark and the mode effect indicator.

3 Problem formulation

The most influential survey design features are the mode and the number of visits/calls, which is due to their significant influence on survey costs and quality. Therefore, for our adaptive design framework, we focus on various combinations survey mode - number of attempts, further denoted as *survey strategies*, and we let the decision variables denote the allocation probability of a survey strategy to the survey units. For example, the set of survey strategies in the case study in Section 5 is given by

$$\begin{aligned}\mathcal{S} = \{&\text{CAWI, CATI2, CATI2+, CAPI3, CAPI3+,}\\ &\text{CAWI-CATI2, CAWI-CATI2+, CAWI-CAPI3, CAWI-CAPI3+}, \Phi\},\end{aligned}$$

where CAPI denotes face-to-face interviews, CATI telephone interviews and CAWI web survey. Note that the strategies alternate between no restriction of calls, i.e., CATI2+ and CAPI3+, and limitation to two calls for CATI, i.e., CATI2, and three for CAPI, i.e., CAPI3. Mixed-mode strategies, e.g., CAWI-CATI2+, are also considered in the case study, where the first mode (CAWI) is available for all sample units at the beginning of the survey fieldwork and the second mode (CATI2+) is employed later in the fieldwork to approach the nonrespondents from the first mode. Note that the sampling design is addressed explicitly in the model by including the nonsampling strategy Φ in the survey strategy set. In other words, allocating strategy Φ to population units denotes that such units are not sampled. The specified maximum number of attempts has been selected given the available historical data, where the greatest proportion of the overall response is obtained within the specified number of attempts. The selected mode combinations are a result of current practice, where web is part of most mixed-mode designs due to its reduced costs. Additionally, we are interested in discerning the mode measurement effect between unimode and mixed-mode strategies.

Population units are clustered into $\mathcal{G} = \{1, \ldots, G\}$ groups given a set of characteristics X such as age, ethnicity, that can be extracted from external sources of data. Let $p(s, g)$ be the allocation probability of strategy s to group g. Note that in the current paper, we implicity model the sampling probabilities. In other words, if $p(s, g) > 0$, then a proportion $p(s, g)$ from group g is sampled and approached through strategy s. Denote by $p(\Phi, g)$ the nonsampling probability. We then have that

$$\sum_{s\in\mathcal{S}} p(s, g) + p(\Phi, g) = 1, \ \forall\, g \in \mathcal{G}. \tag{1}$$

We define the mode effect measure as the nonresponse adjusted difference between the survey estimate $\bar{y}_{s,g}$ and a benchmark estimate $\bar{y}_{BM}$ of the population mean $\bar{Y}$, where the survey estimate $\bar{y}_{s,g}$ is obtained by allocating strategy $s \in \mathcal{S}$ to group $g \in \mathcal{G}$. Let $D(s, g)$ denote this difference. The mode effect measure is expressed as

$$D(s, g) = \bar{y}_{s,g} - \bar{y}_{BM}, \ \forall\, s \in \mathcal{S}, g \in \mathcal{G}, \tag{2}$$

with values in the same domain as $\bar{y}_{s,g}$ and $\bar{y}_{BM}$. We further refer to $D(s, g)$ as the *mode effect measure* or the *mode difference*. The mode effect measure considers the adjusted mode measurement effects, i.e., the survey estimates have been adjusted for nonresponse. For convenience, we refer in the following to mode effects but they are always adjusted for selection effects. The population average mode difference with respect to the indicated benchmark BM is given by

$$\bar{D}_{BM} = \left| \sum_{s,g} \frac{w_g p\,(s,g)\,\rho\,(s,g)\,D(s,g)}{\sum_{s' \in \mathcal{S}} p\,(s',g)\,\rho\,(s',g)} \right|,$$

where N_g is the population size of group g, $w_g = N_g/N$ represents the proportion of group g in the entire population N and $\rho\,(s,g)$ the response probability for group g if strategy s is assigned. In other words, the impact of any mode differences that arise when applying strategy s to group g are moderated by the contribution of the obtained group response to the total response, weighted by the group size. Since only response could trigger mode effects, the contribution of each group g to the overall mode effect measure needs to be proportional to the group's contribution to the overall response. $\bar{D}_{BM}$ takes values in the same domain as $D(s,g)$. Our goal is to minimize the overall mode effect $\bar{D}_{BM}$ by optimally assigning strategies $s \in \mathcal{S}$ to groups $g \in \mathcal{G}$, i.e.,

$$\underset{p(s,g)}{\text{minimize}}\ \bar{D}_{BM} = \left| \sum_{s,g} \frac{w_g p\,(s,g)\,\rho\,(s,g)\,D(s,g)}{\sum_{s' \in \mathcal{S}} p\,(s',g)\,\rho\,(s',g)} \right| \tag{3}$$

Note that $\bar{y}_{s,g}$ is a nonresponse adjusted estimate of $\bar{Y}$, while $\rho\,(s,g)$ is an unweighted estimate of the group g response probability in strategy s (see for details Section 5.2). We assume that the nonresponse adjustment does not influence the contribution of each group and strategy to the overall response. This allows us to write the objective function as in (3), while performing nonresponse adjustment within the optimization would be a very complex perhaps even irrealizable technique.

Scarcity in resources and other practical aspects impose a number of constraints in our model. A limited budget B is available to setup and run the survey. Let $c\,(s,g)$ be the unit cost of applying strategy s to one unit in group g (for estimation details, see Section 5.4). The cost constraint is formulated as follows

$$\sum_{s,g} N_g\, p\,(s,g)\, c\,(s,g) \leq B. \tag{4}$$

To ensure a minimal precision for the survey estimate of $\bar{Y}$, a minimum number R_g of respondents per group is required. This translates to the following constraint

$$\sum_{s} N_g\, p\,(s,g)\, \rho\,(s,g) \geq R_g,\ \forall g \in \mathcal{G}. \tag{5}$$

In addition to the objective function we address the mode effect also through a constraint. The structure of the objective function could lead to an unbalanced solution. For example, let a group g_i be assigned a strategy s such that the corresponding $D(s,g_i)$ is a large negative value and the other groups $g \in \mathcal{G} \backslash g_i$ receive strategies that yield positive $D(s,g)$ values. Thus, the large negative $D(s,g_i)$ is canceled out but group g_i

will have a very different behavior compared to the other groups, which renders mutual comparison among groups impossible. To prevent the occurrence of such solutions, we limit the absolute difference in the mode effect measured for any two groups by the following constraint

$$\left| \frac{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)D(s,g_i)}{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)} - \frac{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)D(s,g_j)}{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)} \right| \leq M, \ \forall\, g_i, g_j \in \mathcal{G}. \quad (6)$$

Given the definition of the absolute value, i.e., $|f(x)| \leq M \Leftrightarrow -M \leq f(x) \leq M$, reformulation of (6) such that the absolute value signs are discarded yields the following

$$-M \leq \frac{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)D(s,g_i)}{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)} - \frac{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)D(s,g_j)}{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)} \leq M, \ \forall\, g_i, g_j \in \mathcal{G}.$$

Furthermore, given that the two inequalities must hold for any $g_i, g_j \in \mathcal{G}$, it follows that (6) is equivalent to

$$\frac{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)D(s,g_i)}{\sum_{s\in\mathcal{S}} p\,(s,g_i)\rho\,(s,g_i)} - \frac{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)D(s,g_j)}{\sum_{s\in\mathcal{S}} p\,(s,g_j)\rho\,(s,g_j)} \leq M, \ \forall\, g_i, g_j \in \mathcal{G}. \quad (7)$$

For practical reasons we also introduce a constraint on the maximum sample size, i.e.,

$$\sum_{s,g} N_g p\,(s,g) \leq S_{max}. \quad (8)$$

The constraints on the decision variables $p\,(s,g)$ concern their definition as allocation probabilities, i.e.,

$$\begin{aligned} &0 \leq p\,(s,g) \leq 1, \ \forall\, s \in \mathcal{S},\, g \in \mathcal{G} \\ &\sum_{s\in\mathcal{S}} p\,(s,g) \leq 1, \ \forall\, g \in \mathcal{G}, \end{aligned} \quad (9)$$

where inequality is sufficient since every $p\,(s,g)$ implies assignment of strategy s after sampling from group g. Equality is necessary when taking into account the nonsampling probability $p\,(\Phi,g)$ as in (1). However, since mode effects cannot be defined in case of nonsampling, we have excluded this variable from the model and adjusted the constraints accordingly. Additionally, we require that at least one $p\,(s,g)$ be strictly positive,

$$\sum_{s\in\mathcal{S}} p\,(s,g) > 0, \ \forall g \in \mathcal{G}, \quad (10)$$

to avoid computational errors such as division by zero.

Objective function (3) together with constraints (4) – (10) form the optimization model to minimize overall mode effects in the context of adaptive survey designs, which leads to a nonconvex nonlinear problem (NNLP).

4 Algorithm

The previous section dealt with defining an adaptive design that optimally allocates survey resources in order to minimize mode effects. The model formulation however poses difficulties in terms of finding a suitable algorithm to solve the problem to optimality. The constraints on the maximum difference between group mode effects make the problem nonconvex and hard to solve. Therefore, most general-purpose nonlinear solvers cannot do better than a local optimum. In such cases, the choice for a starting point of search for an optimum plays an important role in trying to achieve the best local optimum. Given these considerations, we opt for a two-step approach where, in the first step, we solve a linear programming problem (LP) that addresses the linear constraints (4), (5) and (8) – (10) and use the optimal solution thus obtained as a starting point for a local search algorithm to solve the NNLP.

In the following we present a reformulation of the mode effect problem such that the absolute value signs are discarded.

$$
\begin{aligned}
\text{minimize} \quad & t \\
\text{subject to} \quad & \sum_{s,g} \frac{w_g\, p\,(s,g)\,\rho\,(s,g)\,D(s,g)}{\sum_{s'\in\mathcal{S}} p\,(s',g)\,\rho\,(s',g)} \le t \\
& -\sum_{s,g} \frac{w_g\, p\,(s,g)\,\rho\,(s,g)\,D(s,g)}{\sum_{s'\in\mathcal{S}} p\,(s',g)\,\rho\,(s',g)} \le t \\
& \sum_{s,g} N_g\, p\,(s,g)\, c\,(s,g) \le B \\
& \sum_{s} N_g\, p\,(s,g)\,\rho\,(s,g) \ge R_g, \;\; \forall g \in \mathcal{G} \\
& \frac{\displaystyle\sum_{s\in\mathcal{S}} p\,(s,g_i)\,\rho\,(s,g_i)\,D(s,g_i)}{\displaystyle\sum_{s\in\mathcal{S}} p\,(s,g_i)\,\rho\,(s,g_i)} - \frac{\displaystyle\sum_{s\in\mathcal{S}} p\,(s,g_j)\,\rho\,(s,g_j)\,D(s,g_j)}{\displaystyle\sum_{s\in\mathcal{S}} p\,(s,g_j)\,\rho\,(s,g_j)} \le M, \;\; \forall\, g_i, g_j \in \mathcal{G} \\
& \sum_{s,g} N_g\, p\,(s,g) \le S_{max} \\
& 0 \le p\,(s,g) \le 1, \;\; \forall\, s \in \mathcal{S},\, g \in \mathcal{G} \\
& \sum_{s\in\mathcal{S}} p\,(s,g) \le 1, \;\; \forall\, g \in \mathcal{G} \\
& 0 \le t.
\end{aligned}
\tag{11}
$$

Since $|f(x)| = \max\{f(x), -f(x)\}$, we can rewrite the objective function via an additional variable t and impose that $f(x) \le t$ and $-f(x) \le t$. Moreover, t has to be nonnegative. The constraints do not change from their initial formulation.

Before we sketch the LP, note that the dummy variable t in (11) appears only in two nonlinear constraints that would not be part of the LP. Hence, formulating the

LP with the same objective function as (11) is senseless. Consequently, one of the linear constraints should be reformulated as the objective function. We choose for minimization of costs as the LP objective. The resulting problem formulation is given in (12).

To solve the linear problem, we use the simplex method available in `R` in package *boot*. Our proposed two-step algorithm thus handles (12) in the first step. Denote by x^*_{LP} the optimal solution obtained in the LP. In the second step, x^*_{LP} is submitted to a nonlinear optimization algorithm as a starting point in order to solve (11). For this step, we use nonlinear algorithms available in NLOPT (see Johnson 2013), an open-source library for nonlinear optimization that can be called from `R` through the *nloptr* package.

$$
\begin{aligned}
\min \quad & \sum_{s,g} N_g\, p\,(s,g)\, c\,(s,g) \\
\text{s.t.} \quad & \sum_{s} N_g\, p\,(s,g)\, \rho\,(s,g) \geq R_g,\ \ \forall g \in \mathcal{G} \\
& \sum_{s,g} N_g\, p\,(s,g) \leq S_{max} \\
& 0 \leq p\,(s,g) \leq 1,\ \ \forall\, s \in \mathcal{S},\ g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}} p\,(s,g) \leq 1,\ \ \forall\, g \in \mathcal{G}.
\end{aligned}
\tag{12}
$$

Note that the choice for the LP objective function is also motivated by the intention to shorten the runtime in case of infeasibility due to limited budget. The algorithm does *not* perform the second optimization step if the LP objective value, i.e., minimum necessary budget to satisfy the survey design constraints, is larger than the available budget B.

Given that the performance of these algorithms is problem-dependent, we choose to combine two local search algorithms in order to increase the convergence speed. Global optimization algorithms are available in the NLOPT library but their performance for our problem was significantly worse than the selected local optimization algorithms. The two selected local search algorithms are COBYLA (Constrained Optimization by Linear Approximations), introduced by Powell (1998) (see Roy 2007 for an implementation in `C`) and the Augmented Lagrangian Algorithm (AUGLAG), described in Conn et al. (1991) and Birgin and Martinez (2008). The COBYLA method builds successive linear approximations of the objective function and constraints via a simplex of $n+1$ points (in n dimensions), and optimizes these approximations in a trust region at each step. The AUGLAG method combines the objective function and the nonlinear constraints into a single function, i.e., the objective plus a penalty for any violated constraint. The resulting function is then passed to another optimization algorithm as an unconstrained problem. If the constraints are violated by the solution of this sub-problem, then the size of the penalties is increased and the process is repeated. Eventually, the process must converge to the desired solution, if that exists.

As local optimizer for the AUGLAG method we choose MMA (Method of Moving Asymptotes, introduced in Svanberg 2002), based on its performance for our numerical experiments. The concept behind MMA is as follows. At each point $\mathbf{x}$, MMA forms a local approximation, that is both convex and separable, using the gradient of $f(\mathbf{x})$ and the constraint functions, plus a quadratic penalty term to make the approximations conservative, e.g., upper bounds for the exact functions. Optimizing the approximation leads to a new candidate point $\mathbf{x}$. If the constraints are met, then the process continues from the new point $\mathbf{x}$, otherwise, the penalty term is increased and the process is repeated.

The reason for using two local search algorithms is that AUGLAG performs better in finding the neighborhood of the global optimum but COBYLA provides a greater accuracy in locating the optimum. Therefore, the LP optimal solution is first submitted to AUGLAG and after a number of iterations, when the improvement in the objective value is below a specified threshold, the current solution of AUGLAG is submitted to COBYLA for increased accuracy. For our case study, given the precision requirements of the obtained statistics in the survey (0.5%), the results are considered accurate enough if the obtained objective value is within 10^{-4} away from the global optimum. Any further accuracy gains are completely blurred by the sampling variation and accuracy of the input parameters themselves. The computational times can run up to a few hours, which is not necessarily a practical issue since the optimization problem needs to be solved only once in several years.

5 Case study: the Dutch Labor Force Survey

The Dutch LFS is a monthly household survey using a rotating panel with five waves at quarterly intervals. The first wave was conducted using face-to-face interviews up to 2009. Over the years 2010-2012, the first wave was gradually redesigned to a mixed-mode survey employing web, telephone and face-to-face. In the four subsequent waves, data are collected by telephone. During these re-interviews, a condensed questionnaire is applied to establish changes in the labor market position of the respondents. The face-to-face contact strategy for the LFS consists of a maximum of six visits to the address. If no contact was made at the sixth visit, then the address is processed as a noncontact.

The key statistics produced based on the LFS data are estimates of the percentage of persons employed, unemployed and not in the labor force in the Netherlands and in various regional and socio-demographic subpopulations. The target population consists of persons aged 15 years and older (i.e., the potential labor force population). For all members of participating households, demographic variables are observed. For the target variables, only persons aged 15 years and older are interviewed. When a household member cannot be contacted, proxy interviewing is allowed by members of the same household. Households in which one or more members do not respond are treated as nonresponding households.

In order to keep the exposition simple, we restrict ourselves to the first wave. We use 2010–2012 LFS data to estimate various input parameters for the optimization model. Although the LFS sample is based on addresses, it is possible to zoom in on the

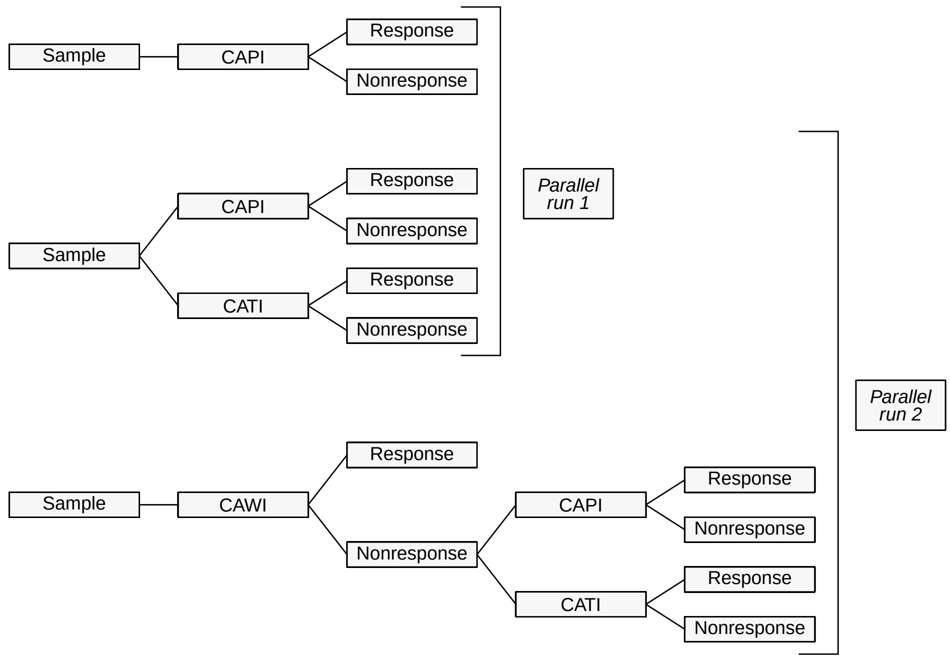

FIGURE 1: Parallel runs LFS redesign from unimode (CAPI) to mixed-mode.

individual level using the municipal registration of population data. Erroneous records such as in-existent addresses or empty house addresses are removed from the sample.

In order to investigate mode effects, we augmented the LFS with data from the POLIS and UWV registers. The POLIS register contains information about income from employment and social benefits; it does not however contain information on income from self-employment. From this register we can determine whether a person is employed and the number of jobs they have. The UWV assists unemployed people in finding a job. Unemployment benefits can be received only by those registered at UWV.

The redesign of the LFS from a unimode (CAPI) to a mixed-mode design involved two parallel runs of different designs (see Figure 1). As a consequence, adjustment across the parallel runs is necessary for certain input parameters (see Section 5.3). The first LFS mixed-mode run involved a parallel CATI-CAPI design where sample units were approached in CAPI only if no registered phone number was available or if the household size exceeded 3. The current mixed-mode design (second run) offers all sample units a web questionnaire. CAWI nonrespondents are subsequently approached in CATI or CAPI given availability of registered phone and the household size, following the same rules as in the first mixed-mode design. Thus, households with more than 3 members or without a publicly available phone number are approached in CAPI. Due to this structure, a CAWI-CAPI design, where all CAWI nonrespondents would be approached in a CAPI follow-up, is not observable. Sections 5.2 and 5.3 discuss an approximation method for producing suitable estimates of the optimization input parameters in the absence of such historical information.

We note here that accuracy in the estimators of the optimization input parameters (i.e., response probabilities $\rho(s,g)$, unit costs $c(s,g)$ and mode differences $D(s,g)$) is crucial for a successful implementation of the optimal design in practice. However, in this paper, we have not performed sensitivity analyses, which would be necessary to assess the robustness of the optimal design.

5.1 Population groups

The population units are clustered into $\mathcal{G} = \{g_1, g_2, \ldots, g_9\}$ homogeneous groups (see Table 1) given the following characteristics, X = (age, household size, UWV registration, POLIS registration of employment, ethnicity). The proportion w_g of each group in the total population is also provided. The characteristics enumerated in X were selected based on their close relationship to the survey target variables and the sampling frame variables. The list of characteristics may be extended, but the resulting groups should be big enough to ensure satisfactory precision of the optimization parameters.

Characteristic	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
UWV registration	1	0	0	0	0	0	0	0	0
Household size > 3		0	0	0	0	0	0	0	1
65+		1	0	0	0				
$15-26$			1	0	0	1	0	0	
POLIS employed		0	0	0	0	1	1	1	
Ethnicity non-western				1	0		1	0	
w_g (%)	7.46	19.77	2.38	1.53	10.97	15.59	3.91	33.50	4.89

TABLE 1: Population clustering given set of characteristics X

The listed characteristics in Table 1 are interpreted as follows. UWV registration: at least one person in the household is registered by the UWV; household size > 3: more than three members of the household are aged 15 or older; 65+: at least one person in the household is 65 years of age or older; $15-26$: at least one person in the household is between 15 and 26 years of age; POLIS employed: at least one person in the household is employed according to the POLIS register; ethnicity non-western: at least one person in the household is of non-western ethnicity. Presence of 0-1 marks whether the stated characteristic (1) or its complement (0) is active for the given group. Consequently, absence of 0-1 marks that the specified characteristic is not active for the given group. For example, group 1 clusters population units that are registered at UWV. For this group, none of the other characteristics are active. For group 4, a number of characteristics are used in defining the cluster, namely, UWV registration (households in this group do *not* have any member registered at the UWV), household size (households in this group have at most 3 members older than 15 years of age), age (households in this group do *not* have members older than 65 or younger than 26 years of age), POLIS registration of employment (according the POLIS records, at least one member of the household in this group is *not* employed) and ethnicity (population units in this group have non-western origins).

The LFS targets people with age between 15 and 64 years; therefore, undersampling occurs for addresses with households consisting only of persons of 65 years of age and higher. Moreover, households with persons of age between 15 and 26 or from non-western countries present more interest; therefore, such households are oversampled. Let $\mathcal{G}' = \{g'_1, g'_2, \ldots, g_5\}$ be the clustering of population units according to the set of characteristics $X' = (\text{age, UWV registration, ethnicity})'$ (see Table 2) that have unequal sampling probabilities. This clustering occurs before the random sampling procedure has been carried out and we further refer to $\mathcal{G}'$ as the *sampling strata*. Information in Table 2 should be interpreted analogously to Table 1.

Characteristic	g'_1	g'_2	g'_3	g'_4	g'_5
65+	1	0	0	0	0
UWV registration		1	0	0	0
$15-26$			1	0	0
Ethnicity non-western				1	0

TABLE 2: Sampling strata definition.

Similarly to constraint (5), we impose that a minimum number of households in each $g' \in \mathcal{G}'$ should be a respondent by the end of the survey. To formulate such a constraint, the response probabilities have to estimated at (g, g') level (see Section 5.2). The corresponding constraint is given by

$$\sum_{s,g} N_{g,g'}\, p\,(s,g)\, \rho\,(s,g,g') \geq R_{g'},\ \forall g' \in \mathcal{G}',$$

where $N_{g,g'}$ represents the population size for group (g, g').

5.2 Estimation of response probabilities

In every survey each population unit is assigned a non-zero probability of being sampled through a random selection procedure. Let d_i^D be the inverse of this probability for a population unit i which is most commonly known as the *design weight*. The sample estimate of a population mean $\bar{Y}$ is then computed as

$$\bar{y} = \frac{1}{N} \sum_{i \,\in\, sample} d_i^D Y_i,$$

where Y_i is the value of parameter Y for unit i. This yields an unbiased estimate of the population mean and it is also known as the *Horvitz-Thompson estimator*. To account for nonresponse, SÄRNDAL ET AL. (1992) modify the Horvitz-Thompson estimator as follows

$$\bar{y}^r = \frac{1}{N} \sum_{i \,\in\, resp} \frac{d_i^D Y_i}{\rho_i},$$

where ρ_i represents the unknown response probability of unit i and $resp$ the respondent sample. The unknown response probabilities can be replaced by their corresponding estimates based on auxiliary information (see SÄRNDAL 1981) or by the Horvitz-Thompson estimator for the mean response probability (see BETHLEHEM 1988) that also uses auxiliary information. Let d_i^A be the inverse of the Horvitz-Thompson estimator for the mean response probability. We then have

$$\tilde{y}^r = \frac{1}{N} \sum_{i \in resp} d_i^D d_i^A Y_i.$$

In order to prevent large capacity variations between subsequent modes in mixed-mode surveys, a subsampling of the remaining nonrespondents is carried out before the follow-up mode. In this case, the Horvitz-Thompson estimator for mixed-mode designs becomes

$$\tilde{y}^r_{\text{MM}} = \frac{1}{N} \sum_{i \in resp} d_i Y_i,$$

with

$$d_i = d_i^D d_i^A d_i^S,$$

the total adjusted weight and d_i^S is the subsampling rate. Note that $\tilde{y}^r_{\text{MM}}$ can be used for unimode surveys if d_i^S is set to 1 for all units i that did not respond in the first mode but were respondents in the follow-up mode.

Aggregating individual response probabilities to group g level yields the following

$$\rho\,(s,g) = \frac{\sum_{i \in g} d_i \frac{1}{d_i^A} R_i^s}{\sum_{i \in g} d_i R_i^s},$$

where $R_i^s \in \{0,1\}$ indicates whether unit i is a respondent through strategy s.

A final step in the estimation of the response probabilities is the sub-/over-sampling of rate of the sampling strata $\mathcal{G}'$. Let $z(g')$ be the sub-/over-sampling rate for stratum $g' \in \mathcal{G}'$, relative to a base stratum g'_{base}. Then the unadjusted design weight for group g', i.e., the design weight in the absence of sub-/over-sampling, is given by $d^D_{g',\text{UN}} = z(g')\, d^D_{g'_{base}}$. The sample size is computed as

$$n = \sum_{g' \in \mathcal{G}'} d^D_{g',\text{UN}}\, N_{g'},$$

with $N_{g'}$ the population size of stratum g'. Assuming all $N_{g'}$ are known, we can now derive $d^D_{g'_{base}}$ by replacing $d^D_{g',\text{UN}}$ accordingly. With this adjustment, the response probabilities estimates are given by

$$\rho\,(s,g,g') = \frac{\sum_{i \in (g,g')} d^D_{i,\text{UN}} d_i^S R_i^s}{\sum_{i \in (g,g')} d_i R_i^s}, \tag{13}$$

$\rho(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	23.2% (0.3)	23.6% (0.6)	15.5% (0.6)	10.8% (0.6)	27.9% (0.4)	27.7% (0.2)	17.5% (0.5)	36.7% (0.2)	22.4% (0.5)
CATI2	12.2% (0.5)	31.4% (1.1)	8.5% (0.8)	4.7% (0.8)	19.7% (0.6)	13.3% (0.4)	7.2% (0.5)	18.1% (0.4)	21.2% (0.8)
CATI2+	20.8% (0.6)	41.3% (1.1)	15.2% (1.0)	8.6% (1.0)	31.1% (0.7)	23.8% (0.5)	14.3% (0.7)	33.3% (0.5)	37.5% (0.9)
CAPI3	43.5% (1.5)	53.5% (1.7)	42.2% (2.4)	34.1% (2.4)	45.1% (1.1)	45.3% (0.9)	35.9% (1.5)	46.7% (0.7)	54.6% (1.4)
CAPI3+	52.4% (1.3)	58.3% (1.6)	51.0% (2.5)	41.2% (2.2)	51.2% (1.1)	54.9% (0.8)	46.0% (1.4)	56.8% (0.7)	61.4% (1.3)
CAWI-CATI2	28.3% (0.4)	41.0% (0.8)	20.2% (0.7)	13.9% (0.8)	36.3% (0.4)	34.0% (0.3)	20.8% (0.5)	44.5% (0.3)	23.1% (0.5)
CAWI-CATI2+	32.8% (0.4)	48.4% (0.7)	23.8% (0.8)	17.5% (0.9)	42.1% (0.5)	41.1% (0.3)	25.8% (0.6)	52.1% (0.3)	24.4% (0.5)
CAWI-CAPI3	46.3% (0.5)	57.7% (1.0)	38.6% (1.0)	32.7% (1.0)	50.0% (0.6)	51.0% (0.4)	39.3% (0.7)	58.9% (0.4)	50.0% (0.5)
CAWI-CAPI3+	49.8% (0.5)	58.3% (0.9)	43.4% (0.9)	36.6% (0.9)	52.6% (0.5)	54.7% (0.4)	44.3% (0.6)	62.0% (0.4)	54.2% (0.5)

TABLE 3: Estimated response probabilities per strategy s and group g.

where the summation is taken only over units in (g, g'). Aggregating over all sampling strata we obtain the response probability estimates for group g, i.e.,

$$\rho(s,g) = \sum_{g' \in \mathcal{G}'} \frac{N_{g'}}{N_g} \rho(s,g,g'), \ \forall\, s \in \mathcal{S},\ g \in \mathcal{G}. \tag{14}$$

Table 3 presents the estimated response probabilities $\rho(s,g)$ from available data. Note that given the complex definition of these estimates, it is not possible to compute directly their standard deviations. We perform a bootstrap analysis in order to assess the standard errors that are provided in brackets. Additionally, given the weighting technique necessary to estimate the response propensities, the weights d_i need to be adjusted in the bootstrap analysis in order to correctly scale up bootstrap sample estimates to the same population composition.

As expected, restricted strategies, i.e., strategies with a cap on the number of attempts, yield lower response probabilities than full strategies. The only reason why restricted strategies may be present in the optimal solution would be the incurred lower costs (see Section 5.4). Additionally, note that all mixed-mode strategies yield higher response probabilities than the CAWI-only strategy and the mixed-mode involving CAPI is more appealing in terms of response than mixed-mode involving CATI. However, in terms of costs, the situation is opposite, i.e., mixed-mode with CATI being significantly less expensive than mixed-mode with CAPI.

However, when the selected strategy s does not have historical support, i.e., there is no survey design in historical data that matches the strategy s specific combination of survey mode - number of attempts, additional modeling is necessary. The structure of the current LFS mixed-mode design assigns CAWI nonrespondents to CATI or CAPI given availability of phone numbers. Thus, CAWI nonrespondents with a publicly available phone number are approached in CATI in the follow-up. Correspondingly, CAWI nonrespondents without a publicly available phone number are approached in CAPI in the follow-up. For our analysis this translates to lack of historical information for strategies CAWI-CAPI3 and CAWI-CAPI3+ that assign all CAWI nonrespondents to CAPI. As a consequence, we must build approximations for response probabilities $\rho\left(s_8, g, g'\right)$ and $\rho\left(s_9, g, g'\right)$. This is done as follows

$$\begin{aligned}
\rho\left(s_9, g, g'\right) &= \frac{\sum_{i\in(g,g')} d_i^D d_i^S R_i^{s_9}}{\sum_{i\in(g,g')} d_i R_i^{s_9}} \\
&\quad + \frac{\rho\left(s_5, g, g'\right)\lambda(g,g')}{\rho\left(s_3, g, g'\right)}\left[\frac{\sum_{i\in(g,g')} d_i^D d_i^S R_i^{s_7}}{\sum_{i\in(g,g')} d_i R_i^{s_7}} - \frac{\sum_{i\in(g,g')} d_i^D d_i^S R_i^{s_1}}{\sum_{i\in(g,g')} d_i R_i^{s_1}}\right], \qquad (15) \\
\rho\left(s_8, g, g'\right) &= \rho\left(s_9, g, g'\right)\frac{\sum_{i\in(g,g')} d_i^D d_i^S R_i^{s_8}}{\sum_{i\in(g,g')} d_i^D d_i^S R_i^{s_9}}
\end{aligned}$$

where $\lambda(g, g')$ represents the probability for registered phone number in group (g, g'). More specifically, the response probability for strategy CAWI-CAPI3+, $\rho\left(s_9, g, g'\right)$, sums up two terms. The first term computes the response probability for strategy s_9 according to formula (13). The second term adjusts this result for the available historical data, where s_9 is not a stand-alone design. Thus, the second term represents an approximation of the response probability CATI respondents would have if approached in CAPI in the mixed-mode design. This probability is obtained by subtracting $\rho\left(s_1, g, g'\right)$ from $\rho\left(s_7, g, g'\right)$, i.e., removing CAWI respondents from the pool of respondents to strategy CAWI-CATI2+, and adjusting the result by the probability for registered phone for respondents in CAPI3+ and the response ratio between strategies CAPI3+ and CATI2+. Table 4 shows the estimated probability for registered phone $\lambda(g)$, where we aggregate from the (g, g') level similarly to the response probabilities, i.e.,

$$\lambda(g) = \sum_{g'\in\mathcal{G}'} \frac{N_{g'}}{N_g}\,\lambda(g, g'),\ \forall\, g. \qquad (16)$$

The response probability for s_8 is computed analogously to $\rho\left(s_9, g, g'\right)$, with the distinction that only units that respond within three CAPI visits in the follow-up are considered.

5.3 Estimation of the mode effect measure $D(s, g)$

For the mode effect measure $D(s, g)$, two benchmarks were selected after consultation with practitioners, i.e., $BM_1 = \bar{y}_{CAPI}$ and $BM_2 = 1/3 * \left(\bar{y}_{CAWI} + \bar{y}_{CATI} + \bar{y}_{CAPI}\right)$, where $\bar{y}_{mode}$ represents the average unemployment rate estimated via the indicated survey

$\mathcal{G}$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
$\lambda(g)$	38.1% (0.9)	76.4% (1.6)	30.2% (2.0)	22.4% (2.2)	60.0% (1.1)	38.9% (0.7)	32.0% (1.3)	53.4% (0.6)	62.4% (1.2)

TABLE 4: Estimated probabilities for registered phone for group $g \in \mathcal{G}$.

mode. The first benchmark assumes that the average unemployment rate estimated via CAPI only, i.e., when the LFS was carried out only as a face-to-face interview, represents the true average unemployment rate. The second benchmark assumes there is no preferred mode, hence, it assigns an equal weight to each of the three modes. Given a benchmark, the mode effect measure for group g under strategy s is computed as

$$D(s,g) = \frac{\sum_{i\in g} d_i R_i^s \, y_i^{unemployed}}{\sum_{i\in g} d_i \, R_i^s \, (y_i^{unemployed} + y_i^{employed} + y_i^{nonlabor})} - \bar{y}_{BM}, \tag{17}$$

where $\bar{y}_{BM}$ represents the benchmark estimate of the unemployment rate, $y_i^{unemployed}$ the number of unemployed household members, $y_i^{employed}$ the number of employed household members and $y_i^{nonlabor}$ the number of household members aged younger than 15. Note that the unemployment rate estimate is a quantity in $[0,1]$, therefore $D(s,g) \in [-1,1]$ which implies that $\bar{D}_{BM} \in [0,1]$.

As remarked in Section 5.2, due to the structure of the mixed-mode design, estimation of $D(s,g)$ for $s \in \{\text{CAWI-CAPI3,CAWI-CAPI3+}\}$ cannot be carried out directly. Moreover, adjustments are necessary to account for estimation of quantities across the parallel runs of the LFS. Let $D(s,s',g) = \bar{y}_{s,g} - \bar{y}_{s',g}$ be such an adjustment step between estimates of the unemployment rate obtained in strategies s and s'. Then,

$$D(s,g) = D(s',g) + D(s,s',g), \text{ for } s' \neq s. \tag{18}$$

Using (18), the adjusted mode differences are computed as

$$\begin{aligned}
D(s_1,g) &= D(s_3,g) + D(s_1,s_3,g)\\
D(s_6,g) &= D(s_6,s_2,g) + D(s_2,g)\\
D(s_7,g) &= D(s_7,s_3,g) + D(s_3,g)\\
D(s_8,g) &= p_{CAWI} D(s_1,g) + (1-p_{CAWI}) D(s_4,g)\\
D(s_9,g) &= p_{CAWI} D(s_1,g),
\end{aligned} \tag{19}$$

for all $g \in \mathcal{G}$, with p_{CAWI} the proportion of CAWI respondents in the total respondent sample in the mixed-mode design. Tables 5 and 6 present the estimated mode differences against the two benchmarks.

Generally, $D_{BM_1}(s,g) > D_{BM_2}(s,g)$ for strategies involving CAWI. This is understandable since BM_2, as a mode mix, is "closer" to CAWI than BM_1. Furthermore, for

$D_{BM_1}(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	1.5%	0.0%	−2.3%	−4.5%	0.9%	−0.4%	−2.2%	0.6%	−0.4%
	(1.0)	(0.5)	(1.5)	(3.1)	(0.7)	(0.4)	(1.5)	(0.5)	(0.6)
CATI2	−0.2%	−0.1%	−2.6%	−6.8%	−1.0%	−0.9%	−1.1%	0.2%	−1.3%
	(0.7)	(0.1)	(0.9)	(1.8)	(0.4)	(0.3)	(1.1)	(0.4)	(0.4)
CATI2+	−0.1%	−0.1%	−2.3%	−4.9%	−0.6%	−1.0%	−0.8%	−0.2%	−1.2%
	(0.7)	(0.1)	(0.8)	(1.7)	(0.4)	(0.3)	(1.0)	(0.3)	(0.4)
CAPI3	−0.5%	−0.1%	0.0%	0.7%	−0.1%	0.0%	0.5%	0.3%	0.1%
	(0.3)	(0.1)	(0.4)	(0.6)	(0.1)	(0.1)	(0.3)	(0.1)	(0.1)
CAPI3+	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
CAWI-CATI2	0.9%	0.0%	−2.4%	−3.4%	−0.1%	−0.7%	−4.4%	0.9%	−0.7%
	(1.0)	(0.4)	(1.5)	(3.7)	(0.6)	(0.5)	(1.9)	(0.5)	(0.6)
CAWI-CATI2+	0.9%	−0.1%	−3.7%	−1.7%	0.5%	−0.7%	−3.0%	0.6%	−0.4%
	(0.9)	(0.3)	(1.4)	(3.2)	(0.7)	(0.4)	(1.4)	(0.5)	(0.6)
CAWI-CAPI3	0.7%	0.0%	−1.2%	−1.6%	0.6%	−0.3%	−1.0%	0.5%	−0.2%
	(0.6)	(0.3)	(0.8)	(1.4)	(0.5)	(0.3)	(0.8)	(0.3)	(0.3)
CAWI-CAPI3+	0.9%	0.0%	−1.2%	−2.0%	0.6%	−0.3%	−1.2%	0.4%	−0.2%
	(0.6)	(0.3)	(0.8)	(1.4)	(0.5)	(0.3)	(0.8)	(0.3)	(0.3)

TABLE 5: Estimated mode differences against benchmark $BM_1 = \bar{y}_{CAPI}$.

BM_1, the mode differences $D(s_1, g)$ are higher than $D(s, g)$ for $s \neq s_1$. Looking at differences across groups, group 4 produces the highest $D(s, g)$ relative to the other groups for both benchmarks. Although units in this group do not have a UWV registration, i.e,. they are not looking for a job, they are registered as unemployed. Moreover, it has been observed in the past that unemployment rates for non-western ethnicities are generally higher than for other ethnicities. Additionally, group 3, that includes unemployed young people (15-26), displays slightly higher mode differences than the other groups. Population group 2 yields usually very low values, which can be explained by the fact that most population units aged 65 and higher fall into this group. Most often, such persons are either retired or employed, therefore producing a group unemployment rate very close to 0. Note that the standard deviations of $D_{BM_1}(\text{CAPI3+}, g)$ will always be 0 since its value is constant across bootstrap runs, i.e., it is always equal to zero given the definitions of the survey strategy and the benchmark, respectively.

5.4 Estimation of unit costs

The cost estimation process follows closely the actual cost computations from practice. This means that all major cost-incurring activities are taken into consideration such as average number of attempts until contact, interview time and travel time. Other costs such as questionnaire design or interviewer training, are considered one-time costs that occur before the start of the data collection, hence, they do not depend on the selected strategy or group. Consequently, it is not necessary to include overhead costs in the analysis. Furthermore, we assume that LFS workload for CAPI and CATI in-

$D_{BM_2}(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	1.0%	0.1%	−0.8%	−1.4%	0.8%	0.1%	−1.2%	0.5%	0.1%
	(0.5)	(0.3)	(0.9)	(1.8)	(0.4)	(0.2)	(0.8)	(0.2)	(0.3)
CATI2	−0.6%	−0.1%	−1.0%	−3.7%	−1.2%	−0.5%	−0.1%	0.1%	−0.8%
	(0.3)	(0.2)	(0.6)	(1.4)	(0.2)	(0.2)	(0.8)	(0.2)	(0.2)
CATI2+	−0.6%	−0.1%	−0.8%	−1.7%	−0.7%	−0.5%	0.2%	−0.3%	−0.6%
	(0.2)	(0.2)	(0.5)	(1.0)	(0.2)	(0.1)	(0.5)	(0.1)	(0.2)
CAPI3	−1.0%	−0.1%	1.6%	3.8%	−0.2%	0.5%	1.5%	0.2%	0.6%
	(0.7)	(0.2)	(0.8)	(1.6)	(0.4)	(0.2)	(0.8)	(0.3)	(0.3)
CAPI3+	−0.5%	0.0%	1.6%	3.1%	−0.1%	0.5%	1.0%	−0.1%	0.5%
	(0.5)	(0.2)	(0.7)	(1.4)	(0.4)	(0.2)	(0.7)	(0.3)	(0.3)
CAWI-CATI2	0.4%	0.0%	−0.9%	−0.3%	−0.2%	−0.2%	−3.4%	0.7%	−0.1%
	(0.5)	(0.3)	(1.0)	(2.9)	(0.4)	(0.3)	(1.5)	(0.3)	(0.4)
CAWI-CATI2+	0.5%	0.0%	−2.1%	1.5%	0.4%	−0.2%	−2.0%	0.5%	0.1%
	(0.4)	(0.2)	(0.8)	(2.0)	(0.4)	(0.2)	(0.8)	(0.2)	(0.3)
CAWI-CAPI3	0.3%	0.0%	0.4%	1.5%	0.5%	0.2%	0.0%	0.4%	0.3%
	(0.2)	(0.1)	(0.3)	(0.6)	(0.2)	(0.1)	(0.3)	(0.1)	(0.1)
CAWI-CAPI3+	0.4%	0.0%	0.4%	1.1%	0.5%	0.2%	−0.2%	0.3%	0.3%
	(0.1)	(0.1)	(0.3)	(0.5)	(0.2)	(0.1)	(0.3)	(0.1)	(0.1)

Table 6: Estimated mode differences against benchmark
$BM_2 = 1/3 * (\bar{y}_{CAWI} + \bar{y}_{CATI} + \bar{y}_{CAPI})$.

terviewers resulting from the optimization model gets subsumed in regular interviewer workloads, i.e., small allocated samples can be treated as larger ones since they are part of larger workloads. With these assumptions we do not have to account for clustering of addresses. Essentially, the $c\,(s,g)$ estimate represents the expected costs to address one population unit from group g using strategy s, i.e., it includes the corresponding response probability $\rho\,(s,g)$ such that the outcome of the survey attempt is considered.

Note that in the case of $c\,(s_1,g)$ the standard deviation will always be 0. The costs for the CAWI-only strategy do not depend on the response rate but only on the sample size, i.e., sending a web questionnaire to all sample units, which is constant across the bootstrap runs.

5.5 Optimization results

In our numerical experiments, we explore the solution structure for various values of the constraint thresholds, namely we let

$$B \in \{160{,}000; 170{,}000; 180{,}000\}$$
$$M \in \{1\%; 0.5\%; 0.25\%\}$$
$$S_{max} \in \{9{,}500; 12{,}000; 15{,}000\}.$$

$c(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	1.6 (0.0)	1.6 (0.0)	1.6 (0.0)	1.5 (0.0)	1.6 (0.0)	1.6 (0.0)	1.6 (0.0)	1.6 (0.0)	1.5 (0.0)
CATI2	4.9 (0.1)	5.8 (0.1)	4.3 (0.1)	3.9 (0.1)	5.3 (0.1)	5.3 (0.1)	4.4 (0.1)	5.6 (0.0)	7.2 (0.1)
CATI2+	6.0 (0.1)	6.6 (0.1)	5.0 (0.1)	4.3 (0.1)	6.3 (0.1)	6.8 (0.1)	5.3 (0.1)	7.4 (0.1)	10.0 (0.2)
CAPI3	38.8 (0.4)	34.5 (0.5)	36.4 (0.5)	35.7 (0.8)	35.9 (0.3)	39.9 (0.2)	38.7 (0.5)	39.7 (0.2)	45.4 (0.5)
CAPI3+	46.2 (0.6)	38.7 (0.6)	43.9 (0.7)	43.3 (1.1)	41.6 (0.4)	47.6 (0.3)	47.7 (0.6)	47.5 (0.2)	51.2 (0.5)
CAWI-CATI2	3.7 (0.0)	4.2 (0.1)	3.9 (0.1)	3.9 (0.1)	3.6 (0.0)	3.7 (0.0)	3.8 (0.0)	3.3 (0.0)	3.4 (0.0)
CAWI-CATI2+	4.1 (0.1)	4.6 (0.1)	4.2 (0.1)	4.2 (0.1)	4.0 (0.1)	4.3 (0.0)	4.3 (0.1)	3.9 (0.0)	3.6 (0.0)
CAWI-CAPI3	27.5 (0.3)	25.6 (0.7)	28.1 (0.6)	30.4 (0.8)	24.7 (0.4)	26.7 (0.3)	31.1 (0.5)	24.0 (0.2)	31.3 (0.4)
CAWI-CAPI3+	33.0 (0.4)	27.3 (0.7)	35.2 (0.9)	36.5 (1.2)	30.2 (0.6)	32.5 (0.4)	38.5 (0.8)	29.5 (0.3)	36.2 (0.6)

TABLE 7: Estimated unit costs (in euros) per strategy s and group g.

For the minimal precision constraints, we keep the constraints' right-hand side terms unchanged, with the following values

$$R_{g'} = (165.35, 533.50, 1359.34, 303.03, 2135.15)$$
$$R_g = (533.69, 162.07, 142.47, 84.77, 529.79, 933.66, 215.83, 1603.43, 290.63).$$

These values have been computed such that a 95% confidence interval is built for the population unemployment rate given the survey estimate. Since $\bar{D}_{BM} \in [0,1]$, it follows that the left hand side of (7) also takes values in the $[0,1]$ interval. The low values chosen for the threshold M are determined by the maximum absolute differences in mode effects observed among the groups when x^*_{LP}, the LP optimal solution, is applied (see Table 8). If for example, $M \geq 2.06\%$, the optimal solution for $S_{max} = 9{,}500$ and BM_1 would simply be the LP solution. Table 9 provides an overview of the optimization results for the original nonlinear problem in (11).

Two conclusions can be drawn. First, increasing the sample size and/or the budget brings the objective value down, reaching 0 for $S_{max} = 15{,}000$ and $B = 180{,}000$ for all levels of M. Second, using BM_2 as benchmark, yields lower objective values than BM_1 (except the case of $S_{max} = 9{,}500$), which is mainly due to the smaller values of $D(s,g)$. Additionally, there is an increased similarity among groups with respect to the deviation from the benchmark, i.e., $D_{BM_2}(s,g)$'s are close in absolute value. This allows feasibility even for $M = 0.01\%$ and for $B = 180{,}000$ and S_{max} the algorithm still yields an objective value very close to 0 (0.00001%).

A more counterintuitive effect is shown by the invariance of the objective value

Sample size (S_{max})	Objective value (min costs)	Benchmark	Mode effect ($\bar{D}_{BM}$)	Max difference in mode effects (M)	Response rate
9,500	123,748.50	BM_1	0.16%	2.06%	48.0%
		BM_2	0.29%	3.31%	
11,000	88,408.95	BM_1	0.05%	5.97%	39.9%
		BM_2	0.19%	2.98%	
12,500	82,270.72	BM_1	0.08%	5.97%	36.9%
		BM_2	0.21%	2.98%	
15,000	74,350.44	BM_1	0.12%	5.97%	29.4%
		BM_2	0.25%	2.39%	

Table 8: Overview optimization results linear programming formulation - minimize costs.

given decreasing values of M for BM_2. The reason is that the yielded solution is a point contained in all three feasible regions, i.e., a stricter bound on the maximum difference in group mode effects does not remove this point from the larger feasible region. The same holds for the invariance of $\bar{D}_{BM_2}$ with respect to increasing budget for $S_{max} = 9{,}500$, where larger feasible regions do not add points that could improve the objective value. For $\bar{D}_{BM_1}$ on the other hand, any change in the budget level or a lower value of M causes a change in the objective value. Note that the improvement step in the objective value decreases when tighter bounds are imposed on the group differences in mode effects.

We can analyze the impact of the sample size by comparing the optimal solutions for $S_{max} = 9{,}500$ and $S_{max} = 15{,}000$. Consider $B = 170{,}000$, $M = 1\%$ and BM_1 with the corresponding optimal solutions given in Tables 10 and 11. For a clearer exposition of the results, we do not provide the values of $p\,(s,g)$ directly, since they can be as small as 10^{-14}, but instead a derived quantity, namely the probability of being assigned a strategy after having been sampled, which is computed as

$$\mathbb{P}\{\text{assign}|\,\text{sample}\} = \frac{p\,(s,g)}{1 - p\,(\Phi, g)}.$$

Additionally, Table 12 presents the corresponding sampling probabilities, i.e., $1-p\,(\Phi,g)$, which allows the reader to derive the individual values of $p\,(s,g)$. Note that in Table 12, the sampling probabilities for groups g_3, g_4 and g_5 are only rounded to 0%, otherwise constraint (10) would be violated, and their actual value is equal to 0.0000001%. The very low sampling rates for these groups can be explained through their large deviations in mode effect from the benchmark that cannot be balanced by other groups when the maximum sample size is small. For larger sample sizes, e.g., $S_{max} = 15{,}000$, we see that the sampled proportion of these groups increases significantly. Consequently, the proportion of other population groups may go down in order to lower the average mode difference $\bar{D}_{BM}$, e.g., groups g_8 and g_9.

The impact of available budget can be most clearly seen for $S_{max} = 12{,}000$ and BM_1, when the objective value drops from 0.097% for $B = 160{,}000$ to 0.009% for $B = 180{,}000$. The corresponding solutions are provided in Tables 13 and 14 and the sampling probabilities are given in Table 15. Strategy CAPI3+ is often chosen in

S_{max}	B	BM	M	$\bar{D}_{BM}$	M	$\bar{D}_{BM}$	M	$\bar{D}_{BM}$
9,500	160,000	BM_1 BM_2	1%	0.155% 0.170%	0.5%	Infeasible	0.25%	Infeasible
	170,000	BM_1 BM_2	1%	0.131% 0.170%	0.5%	Infeasible	0.25%	Infeasible
	180,000	BM_1 BM_2	1%	0.100% 0.170%	0.5%	Infeasible	0.25%	Infeasible
12,000	160,000	BM_1 BM_2	1%	0.097% 0.046%	0.5%	0.119% 0.046%	0.25%	0.123% 0.046%
	170,000	BM_1 BM_2	1%	0.076% 0.036%	0.5%	0.093% 0.036%	0.25%	0.101% 0.036%
	180,000	BM_1 BM_2	1%	0.009% 0.014%	0.5%	0.058% 0.014%	0.25%	0.095% 0.014%
15,000	160,000	BM_1 BM_2	1%	0.051% 0.006%	0.5%	0.094% 0.006%	0.25%	0.112% 0.006%
	170,000	BM_1 BM_2	1%	0.020% 0.004%	0.5%	0.080% 0.004%	0.25%	0.097% 0.004%
	180,000	BM_1 BM_2	1%	0.005% 0.000%	0.5%	0.058% 0.000%	0.25%	0.095% 0.000%

Table 9: Overview optimization results nonlinear problem
- minimize overall mode effects in LFS.

large proportions when the available budget is sufficient, which leads to a low objective function. On the other hand, for smaller budget levels, the optimal solution presents a mix between telephone strategies and mixed-mode strategies. Although it may seem appropriate to assign CAWI strategies that are cheapest, the corresponding mode effect deviations from benchmark are significantly higher (see again the group mode differences in Table 5).

A more careful consideration of the optimal solution reveals that its implementation in practice may be difficult from a logistics point of view. Take for example, the solution from Table 10. Carrying out the survey design prescribed by this solution implies offering a CAWI-only survey to only 3 sample units due to the fact the group g_4 is small. Additionally, only 4% of group g_2, i.e., 73 sample units, would receive CAWI-CATI2, while the remaining 96% would receive CAWI-CATI2+. If we adjust this solution by sending the three units from CAWI to one of the interviewer-assisted modes and approach entire group g_2 in CAWI-CATI2+, then the costs will increase by 0.2% and the objective function will decrease by 0.22%. It follows that, if slightly more budget becomes available, then the optimization yields a better objective value and the corresponding optimal solution becomes more practical.

6 Discussion

Survey research has tried to improve designs of surveys such that the quality of the estimates is high. The development of technology and case management systems allows

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	0%	0%	0%	1%	0%	0%	0%	0%	0%
CATI2	0%	0%	14%	22%	31%	0%	0%	0%	0%
CATI2+	0%	0%	6%	2%	12%	0%	0%	0%	0%
CAPI3	39%	0%	3%	4%	2%	0%	0%	0%	0%
CAPI3+	0%	0%	70%	71%	5%	0%	65%	0%	0%
CAWI-CATI2	0%	4%	0%	0%	0%	0%	0%	0%	0%
CAWI-CATI2+	0%	96%	2%	0%	45%	43%	0%	100%	0%
CAWI-CAPI3	0%	0%	0%	0%	2%	29%	0%	0%	100%
CAWI-CAPI3+	61%	0%	5%	0%	2%	29%	35%	0%	0%

TABLE 10: Strategy assignment given optimal solution for $S_{max} = 9{,}500$, $B = 170{,}000$, $M = 1\%$, BM_1.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	2%	10%	0%	0%	0%	99%	0%	7%	0%
CATI2	1%	13%	0%	20%	0%	0%	41%	3%	0%
CATI2+	5%	4%	1%	2%	0%	0%	0%	5%	0%
CAPI3	22%	0%	0%	77%	0%	0%	3%	1%	6%
CAPI3+	8%	7%	81%	0%	0%	0%	45%	56%	85%
CAWI-CATI2	0%	1%	1%	0%	98%	0%	0%	14%	0%
CAWI-CATI2+	60%	39%	0%	0%	0%	0%	10%	6%	0%
CAWI-CAPI3	2%	5%	17%	0%	0%	1%	0%	8%	7%
CAWI-CAPI3+	0%	20%	1%	1%	1%	0%	0%	0%	2%

TABLE 11: Strategy assignment given optimal solution for $S_{max} = 15{,}000$, $B = 170{,}000$, $M = 1\%$, BM_1.

for detailed monitoring of the data collection process, which in turn gives more information about the response process. Additionally, introducing mixed-mode surveys helped getting a better grasp on the budget spendings. However, the added mode effects may have a strong impact on the accuracy of statistics. As JÄCKLE ET AL. (2010) point out, many mode effects are nonlinear in nature and appropriate adjustment methods are still not available. Therefore, additional research effort should be put into analyzing and reducing the occurrence and magnitude of mode effects in the context of developing new survey designs.

In the current paper, we discuss an optimization model that combines the mathematical framework of adaptive designs with mode effect assessment methods in an attempt to minimize mode effects for a given survey. To our best knowledge, this is the first research attempt of its kind and due to its flexibility, our methodology can be used as a basis for more complex settings that aim at addressing mode effects. However, our method requires that candidate strategies have been implemented and accurate estimates exist of mode differences in response and survey outcomes.

We use the adjusted mode effect for the comparison between the survey estimate and a "gold standard" as suggested by BIEMER (1988). We propose an optimization model

S_{max}	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
9500	0.06%	0.07%	0.00%	0.00%	0.00%	0.22%	0.23%	0.17%	0.09%
15000	0.08%	0.08%	0.21%	0.06%	0.55%	0.36%	0.36%	0.04%	0.01%

TABLE 12: Sampling probabilities for $S_{max} = 9{,}500$ and $S_{max} = 15{,}000$, when $B = 170{,}000$, $M = 1\%$, BM_1.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	2%	0%	0%	0%	0%	2%	3%	0%	0%
CATI2	3%	1%	28%	0%	2%	1%	1%	2%	3%
CATI2+	45%	94%	12%	7%	62%	0%	80%	44%	21%
CAPI3	42%	0%	40%	1%	0%	0%	12%	2%	14%
CAPI3+	0%	0%	7%	88%	0%	1%	0%	3%	3%
CAWI-CATI2	1%	2%	2%	0%	0%	0%	0%	1%	59%
CAWI-CATI2+	0%	1%	1%	1%	36%	79%	0%	47%	0%
CAWI-CAPI3	6%	1%	7%	3%	0%	16%	3%	0%	0%
CAWI-CAPI3+	0%	0%	2%	1%	0%	1%	1%	1%	0%

TABLE 13: Strategy assignment given optimal solution for $S_{max} = 12{,}000$, $B = 160{,}000$, $M = 1\%$, BM_1.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
CAWI	42%	0%	0%	0%	0%	100%	0%	63%	20%
CATI2	0%	0%	0%	0%	10%	0%	01%	0%	21%
CATI2+	0%	0%	0%	0%	32%	0%	0%	0%	7%
CAPI3	0%	0%	5%	73%	57%	0%	0%	0%	4%
CAPI3+	58%	0%	67%	3%	1%	0%	55%	31%	20%
CAWI-CATI2	0%	0%	0%	0%	0%	0%	0%	0%	1%
CAWI-CATI2+	0%	100%	0%	0%	0%	0%	0%	6%	1%
CAWI-CAPI3	0%	0%	28%	1%	0%	0%	45%	0%	1%
CAWI-CAPI3+	0%	0%	0%	22%	0%	0%	0%	0%	25%

TABLE 14: Strategy assignment given optimal solution for $S_{max} = 12{,}000$, $B = 180{,}000$, $M = 1\%$, BM_1.

B	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
160000	0.09%	0.09%	0.02%	0.48%	0.53%	0.27%	0.17%	0.06%	0.03%
180000	0.07%	0.07%	0.00%	0.00%	0.00%	0.42%	0.24%	0.20%	0.00%

TABLE 15: Sampling probabilities for $B = 160{,}000$ and $B = 180{,}000$, when $S_{max} = 12{,}000$, $M = 1\%$, BM_1.

that develops an adaptive survey design such that the overall population mode effect is minimized, subject to constraints on differences in mode effects between important population groups. If it is the designer's choice to focus on a different survey item, then the method is still applicable. Note that, in this case, an appropriate "gold standard" must be specified and the optimization input parameters must change accordingly. If it is the designer's choice to address multiple items simultaneously, then a composite indicator of the mode effects influencing these items must be developed. We find this a challenging task since the survey mode may have different effects on different survey items. In this case, an approach similar to the one suggested in Calinescu and Schouten (2013) may be more suitable, where an indicator that summarizes measurement effects across survey items, namely the measurement profile, is employed.

We illustrate our methodology on the unemployment rate, one of the key statistics of the Labor Force Survey (LFS). In our case study on LFS data, we are able to focus on mode measurement effects since mode selection effect can largely be adjusted for given auxiliary information, as concluded by Schouten et al. (2013b). We find that, for realistic values of the input parameters, the overall mode effect can be brought to zero. We also study the differences between applying two different "gold standards" in the definition of the mode effect measure. It follows that, if the population groups have similar behavior with respect to the benchmark, then it is easy to lower the overall mode effect even for low budget levels.

The accuracy of input parameters to the optimization model requires additional consideration. Section 5.2 dealt with the estimation of the optimization input parameters, i.e., response probabilities, unit costs and mode differences, for all considered population groups and survey strategies. This analysis step is of crucial importance for a successful implementation of the yielded optimal solution. Hence, sufficient historical data should be available to produce reliable optimization input parameters. However, since implementation of mixed-mode designs is rather new, it could happen that certain survey strategies are not backed up by historical data, as it was also the case for our numerical experiments. In such situations, approximation methods could be applied. As a consequence, sensitivity analysis should be performed to test the robustness of the optimal solution in case of small perturbations in the input parameters. Future research should develop a robust and effective model.

In this paper, we have chosen to optimize quality given constraints on the sample size, the budget and the precision of survey estimates. The dual problem, where cost is minimized given a constraint on quality, can be optimized almost analogously, leading, however, to different solutions in general. In fact, any constraint can be used as the objective function in the optimization. Practical implementations of our method may consist of a small number of repeating iterations alternating the optimization of quality and the optimization of costs in order to reach the right tradeoff.

This paper is part of a larger project at Statistics Netherlands, called Refining the Data Collection Strategy, (in Dutch: Verfijning Waarneemstrategie). This project aims at detailing the quality-cost tradeoff in the Statistics Netherlands mixed-mode data collection strategy for social surveys, introducing adaptive survey designs, restructuring monitoring and analysis of mixed-mode survey data and paradata, and increasing robustness of mixed-mode data collection for unexpected changes in quality. The results of this paper are input to a follow-up project in which a sensitivity analysis is performed and practical and logistical constraints are added to the optimization.

References

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, **4**(3), 251 – 260.

Biemer, P. (1988). *Telephone survey methodology*, chapter Measuring data quality, pages 273–282. Wiley & Sons. Edited by R.M. Groves and P.P. Biemer and L.E. Lyberg and J.T. Masswey and W.L. Nicholls II and J. Waksberg.

Birgin, E.G. and Martinez, J.M. (2008). Improving ultimate convergence of an augmented lagrangian method. *Optimization Methods and Software*, **23**(2), 177 – 195.

Calinescu, M. and Schouten, B. (2013). Adaptive survey designs that account for nonresponse and measurement error. In revision.

Conn, A.R., Gould, N.I.M., and Toint, P.L. (1991). A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, **28**(2), 545 – 572.

de Leeuw, E. (1992). *Data quality in mail, telephone and face-to-face surveys.* Amsterdam: TT Publications.

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, **21**(2), 233 – 255.

Dillman, D. (2000). *Mail and internet surveys: the tailored design method.* New York: John Wiley Co.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messes, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social Science Research*, **38**(1), 1 – 18.

Fan, W. and Yan, Z. (2010). Factors affecting response rates of the web survey: a systematic review. *Computers in Human Behavior*, **26**(2), 132 – 139.

Greenfield, T., Midanik, L., and Rogers, J. (2000). Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*, **95**(2), 277 – 284.

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, **78**(1), 3 – 20.

Johnson, S.G. (2013). The nlopt nonlinear-optimization package. Available online at http://ab-initio.mit.edu/nlopt.

Klausch, T., Hox, J., and Schouten, B. (2013). Measurement effects of survey mode - On the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, **42**(3), 227 – 263.

Link, M. and Mokdad, A. (2005). Alternative modes for health surveillance surveys: an experiment with web, mail and telephone. *Epidemiology*, **16**(5), 701 – 704.

Millsap, R. (2011). *Statistical approaches to measurement invariance*. Routledge New York.

Powell, M.J.D. (1998). Direct search algorithms for optimization calculations. *Acta numerica*, **7**, 287 – 336.

Roy, J.S. (2007). Stochastic optimization - scipy project. Available online at http://js2007.free.fr/.

Särndal, C. (1981). Frameworks for inference in survey sampling with application to small area estimation and ajustment for nonresponse. *Bulletin of the International Statistical Institute*, **49**, 494 – 513.

Särndal, C., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York.

Schouten, B., Calinescu, M., and Luiten, A. (2013a). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, **39**(1), 29 – 58.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., Burger, J., and Klausch, T. (2013b). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, **42**, 1555 – 1570.

Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization*, **12**(2), 555 – 573.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.