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

Survey Methodology

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

by Melania Calinescu and Barry Schouten

Release date: December 17, 2015

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 toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

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.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

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, December 2015
Vol. 41, No. 2, pp. 403-425
Statistics Canada, Catalogue No. 12-001-X

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

Melania Calinescu and Barry Schouten[1]

Abstract

Assessing the impact of mode effects on survey estimates has become a crucial research objective due to the increasing use 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 be large relative to the precision of a survey. 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 an 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.

Key Words: Mode-specific selection bias; Mode-specific measurement bias; Survey costs; Survey quality.

1 Introduction

In this paper, we propose and demonstrate the minimization of mode effects through adaptive survey designs when a survey has a single statistic or indicator. We demonstrate this method for the Dutch Labour Force Survey (LFS), which has the unemployment rate as the key indicator.

The emergence of web as a survey mode has led to a renewed discussion about mixed-mode surveys. Market research companies quickly incorporated web in their designs, official statistics institutes are slower, but also these institutes are considering mixed-mode designs including web as one of the modes. Reasons for studying mixed-mode designs include increased costs in carrying out face-to-face surveys, decreasing coverage in telephone surveys and low participation in Web surveys (Fan and Yan 2010). As a consequence, survey organizations are gradually restructuring their single mode designs into mixed-mode designs. A large-scale project Data Collection for the Social Surveys (DCSS) was initiated within the EU statistical system in 2012 to investigate mixed-mode survey designs for the Labor Force Survey (LFS), see Blanke and Luiten (2012).

It is well-known that the survey mode impacts both non-observation survey errors (item-nonresponse, unit-nonresponse and undercoverage) as well as observation survey errors (measurement error and processing error). The overall difference between two modes is usually referred to as the mode effect. The difference between the measurement errors of two modes is termed the pure mode effect or measurement effect, while the difference in undercoverage and nonresponse is termed the selection effect, see, for example, de Leeuw (2005), Dillman, Phelps, Tortora, Swift, Kohrell, Berck and Messes (2009), Vannieuwenhuyze (2013) and Klausch, Hox and Schouten (2013b) for extensive discussions. There is evidence (Jäckle, Roberts and Lynn 2010, Schouten, van den Brakel, Buelens, van der Laan, Burger and Klausch 2013b, Dillman et al. 2009) that mode effects can be large. They may lead to incomparable statistics in time or incomparable statistics over population subgroups. Assessing, minimizing and stabilizing the impact of mode effects on survey estimates has become an important goal.

1. Melania Calinescu, Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081HV Amsterdam, Netherlands. E-mail: mcmelania@gmail.com; Barry Schouten, Statistics Netherlands, PO Box 24500, 2490HA Den Haag, Netherlands. E-mail: jg.schouten@cbs.nl.

There are four options to reduce the impact of mode effects in survey design and survey estimation. Thorough questionnaire design and data collection design should prevent them and survey estimation and calibration help accounting for mode effects by weighting. Careful questionnaire design reduces measurement differences between modes. This is possible by using a unified mode design for questionnaires, see Dillman et al. (2009), or by achieving an equivalent stimulus per mode, see de Leeuw (2005). Some measurement effects are, however, intrinsic to the survey mode administration process. For example, an oral versus visual presentation or the interview pace make it hard or impossible to completely remove such effects. Furthermore, questionnaire design cannot remove selection effects, although the length, layout and content may be a common cause to both measurement and selection effects. Also the history of the questions may prevent a questionnaire to be redesigned completely per mode as the survey users or stakeholders do not want to reduce the length of a questionnaire or change the wording of survey items. In summary, some mode effects will always remain, even after a thorough questionnaire redesign. If estimates of measurement effects and selection effects are available, then they can be used to design the data collection strategy of a survey, i.e., to avoid them, or to design the estimation strategy, i.e., to adjust them in future surveys.

The design option implies that some modes or sequences of modes are not applied because they are expected to lead to large mode effects with respect to some benchmark design, i.e., a survey design that is considered to be free of mode effects. The expectation of large mode effects is ideally based on pilot studies but may also lean on experience. When the choice of mode(s) is not made uniform over the whole sample but based on characteristics of persons or households, the survey design option amounts to an adaptive survey design, see Wagner (2008) and Schouten, Calinescu and Luiten (2013a). Such characteristics may be available before data collection starts or may become available during data collection in the form of paradata (i.e., data collection process data, see Kreuter 2013), leading to static and dynamic adaptive survey designs, respectively. The avoidance of mode effects by adaptive survey designs is the focus of this paper.

The adjustment option is especially interesting when there is a strong rationale or incentive to approximate true values of a statistic, i.e., when the focus is not just on comparability but also on accuracy of statistics. A drawback of the adjustment option is that it is more costly than the design option since precise estimates of mode effects are needed such that accuracy of resulting statistics is not affected. A benefit of the adjustment option is that it is more flexible. It allows for different adjustments to different survey variables, whereas the survey design option has to make an overall choice. We refer to Vannieuwenhuyze (2013), Klausch, Hox and Schouten (2013a) and Suzer-Gurtekin (2013) for a discussion of adjustment during estimation.

Another option is to stabilize mode effects, which is a useful last resort approach. Given that mode effects are conjectured to be present after questionnaire, data collection and estimation design, they can be stabilized over time by calibration of the distribution of modes in the response to some fixed distribution of modes. If the average proportion of a mode to response differs between months, the respondents to that mode get a larger weight and respondents to other modes get a smaller weight. For a discussion of this method, see Buelens and van den Brakel (2014).

In this paper, we minimize the adjusted method effect to a benchmark mode design by stratifying the population into relevant subgroups and assigning the different subgroups to different modes or sequences of modes. The adjusted method effect of a design is the difference of the nonresponse adjusted mean of

that design to the nonresponse adjusted mean of the benchmark design. The adjustment follows standard procedures, i.e., calibration of response to a population distribution. Hence, the adjusted method effect is the compound of the measurement effect between the two designs and the residual selection effect between the two designs that is not removed by the nonresponse adjustment.

Adaptive survey designs and the closely resembling responsive survey designs (Heeringa and Groves 2006, Kreuter 2013) are traditionally applied to reduce nonresponse error. As far as we know, to date, only Calinescu and Schouten (2013a) have attempted to focus adaptive survey designs on measurement error or the combination of nonresponse and measurement error. The main reasons are, first, that adaptive and responsive survey designs are still in their infancy and are not widely applied, and, second, that measurement error and measurement effects are inherently hard to measure. Many applications of adaptive survey designs involve a single survey mode in which it is plausible that measurement error is relatively stable for different design choices. When the survey mode is one of the survey design features, then it is no longer plausible to make this assumption. The survey mode is, however, the most interesting design feature in adaptive survey designs due to its large quality-cost differential.

A complication that arises when including the measurement error into adaptive survey designs is that, unlike nonresponse error, it is not the result of a simple yes-no decision. A sample unit provides a response or nonresponse whereas measurement error also has a magnitude. The magnitude of the measurement error may vary per item in the survey questionnaire. This implies that with multiple survey items or variables the choice of modes is a multidimensional decision. Calinescu and Schouten (2013a) attempt to reduce this multidimensionality by using response styles (or response latencies). When a survey has only one or a few key variables, which is in fact the case for the LFS, this complication does not exist and the focus can be directly on the main variables. This is the path that we follow in the current paper.

In this paper, we, therefore, bring two novel elements: we include method effects due to modes into adaptive survey designs and we focus on a single key variable. In our demonstration for the Dutch LFS, we consider three survey modes, namely, web, phone and face-to-face, and various sequences of these modes. In recent years, the Dutch LFS design underwent a series of changes in its transition from a full face-to-face survey to a mixed-mode survey. Extensive knowledge and historical survey data on the interaction between survey design features, the survey mode in particular, and the response process is available. We use this data to estimate the various parameters that are needed for the optimization model.

The outline of the paper is as follows. In Section 2, we formulate the multi-mode optimization problem. In Section 3, we describe an algorithm for the optimization of the mode effect problem. We present the optimization results in Section 4. In Section 5, we discuss the results of the paper. Appendix A and B provide extensions to be numerical results of Section 4.

2 The multi-mode optimization problem

In this section, we construct the multi-mode optimization problem that accounts for mode effects on a single key survey variable. Apart from the survey mode, we also consider caps on the number of calls in telephone and face-to-face as design features in the optimization. In the optimization model, we allow different design features to be assigned to different subpopulations. Hence, the optimization may lead to an adaptive survey design; it does so when the optimal allocation probabilities differ over the

subpopulations. In our case, the subpopulations are built on linked administrative data. Note that they could also be built based on paradata collected during the early stages of the survey. The last component to the optimization problem is given by a set of explicit quality and cost functions. In our case, the quality functions are derived from mode differences in selection and measurement bias and from requirements on the precision of statistics. As a cost function, we use the total variable costs of the survey design. In the following paragraphs, we discuss the components of the optimization problem.

We begin with the survey design features contained in the survey strategy set $\mathcal{S}$. We consider single mode and sequential mixed-mode strategies, i.e., a strategies where nonrespondents in a mode are followed-up in another mode. A single mode would be labelled as M and a sequential mixed-mode as $M_1 \rightarrow M_2$. We consider Web, telephone and face-to-face survey as the modes of interest and abbreviate them to Web, Tel and $F2F$. Examples of single mode and sequential mixed mode are Tel and $Web \rightarrow F2F$, respectively. For interview modes, we additionally consider a cap k on the number of calls, denoted as Mk. For example, $F2F3$ denotes a single mode survey strategy that uses face-to-face with a maximum of three visits. We let $Mk+$ denote the counterpart strategy where there is no explicit cap. We do not consider concurrent mixed-mode strategies (two or more modes are offered simultaneously to sample units) in this paper. This restriction is without loss of generality. It would be straightforward to apply the methodology to any set of multi-mode strategies, including hybrid forms of sequential and concurrent mixed-mode strategies. A wide or diffuse set of strategies will, however, come at the cost of a larger number of input parameters that need to be estimated. The survey strategy set $\mathcal{S}$ explicitly includes the empty strategy, denoted by Φ, which represents the case where a population unit is not sampled, i.e., no action is taken to get a response from the unit. We let $\mathcal{S}^R = \mathcal{S} \setminus \{\Phi\}$ denote the set of real, non-empty 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 or from paradata. Let $p(s,g)$ be the allocation probability of strategy s to group g, i.e., a proportion $p(s,g)$ from subpopulation g is sampled and approached through strategy s. In general, it may hold that multiple strategies have non-zero allocation probabilities, so that the subpopulation is divided over multiple strategies. Define the allocation probability $p(\Phi,g)$ as the probability that a unit from subpopulation g is not included in the sample. The ratio $p(s,g)/(1-p(\Phi,g))$ is the probability that a unit is assigned strategy s given that it has been sampled. For example, if only the allocation probabilities to the empty strategy $p(\Phi,g)$ vary and the allocation probabilities $p(s,g), \forall s \in \mathcal{S}^R$ are equal conditional on being sampled, then the design is stratified but non-adaptive. The probabilities must satisfy

$$\begin{aligned} \sum_{s \in \mathcal{S}^R} p(s,g) + p(\Phi,g) &= 1, \ \forall g \in \mathcal{G}, \\ 0 \leq p(s,g) &\leq 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G}. \end{aligned} \tag{2.1}$$

The allocation probabilities of survey strategies assigned to subpopulations $p(s,g)$ define the decision variables in the optimization model. More generally, and analogous to sampling designs, one could allow for dependencies between population units being sampled and/or being allocated to non-empty strategies $s \in \mathcal{S}^R$. We will not add that complexity here, but assume independence.

We now discuss the quality and cost functions. We assume that the interest lies in estimating the population means of a survey variable y. Given that we consider the survey mode as one of the design features, we view the nonresponse adjusted bias on y between the proposed design and a specified benchmark design BM as the main quality function. This bias may be viewed as the adjusted method effect with respect to BM, and it is a mix of mode-specific measurement biases and remaining mode-specific nonresponse biases after adjustment. If both the proposed design and the benchmark design are single mode, then the bias is a true (adjusted) mode effect. If one of the designs is multi-mode, then the bias represents a complex mixture of mode effects, see for instance Klausch, Hox and Schouten (2014).

Let N_g be the population size of group g, $w_g = N_g / N$ be the proportion of group g in the population of size N, and $\rho(s,g)$ be the response propensity for group g if strategy s is assigned. For a specific group, we define the adjusted method effect as the nonresponse adjusted difference between the survey estimate $\bar{y}_{s,g}$ and a benchmark estimate $\bar{y}_g^{\mathrm{BM}}$ of the population mean $\bar{Y}$, where the survey estimate $\bar{y}_{s,g}$ is obtained by allocating strategy $s \in \mathcal{S}^R$ to subpopulation $g \in \mathcal{G}$. Let $D(s,g)$ denote this difference. The adjusted method effect is expressed as

$$D(s,g) = \bar{y}_{s,g} - \bar{y}_g^{\mathrm{BM}}, \ \forall s \in \mathcal{S}^R, \ g \in \mathcal{G}. \tag{2.2}$$

For convenience, we omit the adjective "adjusted"' in the following and refer to $D(s,g)$ simply as the *method effect*.

In this paper, we seek to minimize the expected absolute overall method effect with respect to a given benchmark design BM, which is the weighted average of the method effects $D(s,g)$ per stratum and strategy to BM. The expected absolute overall method effect with respect to BM is equal to

$$\bar{D}^{\mathrm{BM}} = \left| \sum_{g \in \mathcal{G}} w_g \frac{\sum_{s \in \mathcal{S}^R} p(s,g)\rho(s,g)D(s,g)}{\sum_{s \in \mathcal{S}^R} p(s,g)\rho(s,g)} \right|. \tag{2.3}$$

This objective function represents the expected shift in the time series of the key survey statistic when a redesign is implemented from the benchmark design to the adaptive design using allocation probabilities $p(s,g)$. If a survey is new or if the benchmark design was never actually fielded, the objective function represents the bias of the adaptive survey design to the benchmark design. It is, therefore, a very useful objective function. 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. We implicitly 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 (2.4), while performing nonresponse adjustment within the optimization framework may lead to a very complex, perhaps even unsolvable, problem. We minimize the overall method effect $\bar{D}^{\mathrm{BM}}$ by optimally assigning strategies $s \in \mathcal{S}^R$ to the groups $g \in \mathcal{G}$, i.e.,

$$\underset{p(s,g)}{\text{minimize}} \ \bar{D}^{\mathrm{BM}}. \tag{2.4}$$

Ideally, $\bar{D}^{\mathrm{BM}} = 0$. However, achieving this situation may have serious practical issues such as requiring unlimited resources. Therefore, various practical aspects such as scarcity in resources are reflected through 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. The cost constraint is formulated as follows

$$\sum_{s,g} N_g p(s,g) c(s,g) \leq B. \tag{2.5}$$

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

$$\sum_{s \in \mathcal{S}^R} N_g p(s,g) \rho(s,g) \geq R_g, \ \forall g \in \mathcal{G}. \tag{2.6}$$

In addition to the objective function, the method effect between the proposed design and the benchmark design is also part of a constraint in the optimization problem: a constraint on comparability of population subgroups. The overall method effect as an objective function could lead to an unbalanced solution. For example, let a group g be assigned a strategy s such that the corresponding $D(s,g)$ is a large negative value and the other groups $h \in \mathcal{G} \setminus \{g\}$ receive strategies that yield positive $D(s,h)$ values. The large negative $D(s,g)$ is canceled out but group g will have a very different behavior compared to the other groups, and this complicates comparisons among groups. To prevent the occurrence of such designs, we limit the absolute difference in the method effect between two groups by the following constraint

$$\max_{g,h \in \mathcal{G}} \left\{ \frac{\sum_{s \in \mathcal{S}^R} p(s,g) \rho(s,g) D(s,g)}{\sum_{s \in \mathcal{S}^R} p(s,g) \rho(s,g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s,h) \rho(s,h) D(s,h)}{\sum_{s \in \mathcal{S}^R} p(s,h) \rho(s,h)} \right\} \leq M. \tag{2.7}$$

However, when

$$\frac{\sum_{s \in \mathcal{S}^R} p(s,g) \rho(s,g) D(s,g)}{\sum_{s \in \mathcal{S}^R} p(s,g) \rho(s,g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s,h) \rho(s,h) D(s,h)}{\sum_{s \in \mathcal{S}^R} p(s,h) \rho(s,h)} \leq M \tag{2.8}$$

is included in the optimization problem for each pair $(g,h) \in \mathcal{G}$, then (2.7) is automatically satisfied. For practical reasons, i.e., a depletion of the sampling frame, we also introduce a constraint on the maximum sample size $S_{\max}$, i.e.,

$$\sum_{s,g} N_g p(s,g) \leq S_{\max}. \tag{2.9}$$

Additionally, we require that at least one $p(s,g)$ be strictly positive,

$$\sum_{s \in \mathcal{S}^R} p(s,g) > 0, \forall g \in \mathcal{G}, \tag{2.10}$$

to avoid computational errors such as division by zero in (2.8).

Objective function (2.4) together with constraints (2.1), (2.5) – (2.10) form the multi-mode optimization problem to minimize method effects against a benchmark through adaptive survey designs. This problem is a nonconvex nonlinear problem.

3 An algorithm for solving the multi-mode optimization problem

In the previous section, we introduced the quality and cost functions and constructed a multi-mode optimization problem. The subpopulation comparability constraint, i.e., the upper limit to the maximum absolute difference between group method effects, makes the problem nonconvex and hard to solve. As a consequence, when trying to solve the multi-mode optimization problem, most general-purpose nonlinear solvers cannot do better than a local optimum. Therefore, the choice of starting points in the solvers plays an important role. As such, we propose a two-step approach. In the first step, we solve a linear programming problem (LP) that addresses the linear constraints (2.1), (2.5), (2.6) and (2.9) – (2.10). In the second step, we use the optimal solution obtained in step 1 as a starting point for a local search algorithm to solve the nonconvex nonlinear problem (NNLP).

We reformulate the optimization problem to make it computationally more tractable. Since $|f(x)| = \max\{f(x), -f(x)\}$, we can rewrite the objective function via an additional variable t and impose that $f(x) \leq t$ and $-f(x) \leq t$. Clearly, t has to be nonnegative. The constraints themselves do not change, they are simply replaced. The multi-mode optimization problem is given in (3.2).

We can derive the LP by removing the non-linear constraints on the comparability of method effects across subpopulations and by replacing the non-linear objective function by one of the linear constraints. We choose for minimization of costs as the LP objective. The resulting LP problem formulation is given by

$$\begin{aligned}
\underset{p(s,g)}{\text{minimize}} \quad & \sum_{s,g} N_g p(s,g) c(s,g) \\
\text{subject to} \quad & \sum_{s \in \mathcal{S}^R} N_g p(s,g) \rho(s,g) \geq R_g, \ \forall g \in \mathcal{G} \\
& \sum_{s,g} N_g p(s,g) \leq S_{\max} \\
& 0 \leq p(s,g) \leq 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}} p(s,g) = 1, \ \forall g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}^R} p(s,g) > 0, \ \forall g \in \mathcal{G}.
\end{aligned} \tag{3.1}$$

$$
\begin{aligned}
&\text{Minimize} \quad t \\
&\text{subject to} \quad \sum_{s,g} \frac{w_g\, p(s,g)\,\rho(s,g)\,D(s,g)}{\sum_{s' \in \mathcal{S}^R} p(s',g)\,\rho(s',g)} \le t \\
&\quad -\sum_{s,g} \frac{w_g\, p(s,g)\,\rho(s,g)\,D(s,g)}{\sum_{s' \in \mathcal{S}^R} p(s',g)\,\rho(s',g)} \le t \\
&\quad \sum_{s,g} N_g\, p(s,g)\, c(s,g) \le B \\
&\quad \sum_{s \in \mathcal{S}^R} N_g\, p(s,g)\,\rho(s,g) \ge R_g, \ \forall g \in \mathcal{G} \\
&\quad \frac{\sum_{s \in \mathcal{S}^R} p(s,g)\,\rho(s,g)\,D(s,g)}{\sum_{s \in \mathcal{S}^R} p(s,g)\,\rho(s,g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s,h)\,\rho(s,h)\,D(s,h)}{\sum_{s \in \mathcal{S}^R} p(s,h)\,\rho(s,h)} \le M \qquad (3.2) \\
&\quad \sum_{s,g} N_g\, p(s,g) \le S_{\max} \\
&\quad 0 \le p(s,g) \le 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G} \\
&\quad \sum_{s \in \mathcal{S}} p(s,g) = 1, \ \forall g \in \mathcal{G} \\
&\quad \sum_{s \in \mathcal{S}^R} p(s,g) > 0, \ \forall g \in \mathcal{G} \\
&\quad 0 \le t.
\end{aligned}
$$

To solve the linear problem, we use the simplex method available in R in package *boot*. Our proposed two-step algorithm thus handles (3.1) 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 (3.2). 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. The NNLP second step of the algorithm is performed only if the minimal required budget found in the LP first step is smaller than or equal to the available budget B. If the minimal budget is larger, then there is no feasible solution to the optimization problem.

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, Gould and Toint (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 strategy behind MMA is as follows. At each point $\mathbf{x_k}$, MMA forms a local approximation, that is both convex and separable, using the gradient of $f(\mathbf{x_k})$ 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 $\mathbf{x_{k+1}}$. If the constraints are met, then the process continues from the new point $\mathbf{x_{k+1}}$, 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. Since the optimization problem does not need to be solved during data collection, this will, however, not pose practical problems.

4 Case study: The Dutch Labor Force Survey

In this section, we discuss a case study linked to the Dutch Labor Force Survey (LFS) of the years 2010 – 2012. We briefly describe the design of the LFS first. We then proceed to a description of the selected design features and the selected population subgroups. Next, we explain how we have estimated the main input parameters to the optimization problem: response propensities, telephone registration propensities, variable costs and adjusted method effects with respect to two different benchmark designs. Following the estimation, we present the main optimization results. We end with a discussion of the sensitivity of optimal designs to inaccuracy of input parameters. For full details, we refer to Calinescu and Schouten (2013b).

4.1 The Dutch LFS design and redesign in 2010 – 2012

The Dutch LFS is a monthly household survey using a rotating panel with five waves at quarterly intervals. The LFS is based on an address sample using a two-stage design in which the first stage consists of municipalities and the second consists of addresses. A stratified simple random sample is drawn based

on the household age, ethnicity and registered unemployment composition. All households, to a maximum of eight, that are residents at the address are invited to participate. Within each household, all members of 15 years and older are eligible; they form the potential labor force population. The LFS contains a variety of topics, from employment status, profession and working hours to educational level, but the main survey statistic is the unemployment rate.

Up to 2010, the LFS consisted of a face-to-face first wave and telephone subsequent waves. For various reasons, costs being the most important, the first wave went through a major redesign. The other waves were left unchanged, except for a few relatively small changes to the questionnaires. The redesign consisted of two phases: First, telephone was added as a survey mode, and, second, also Web was added as a survey mode. In the first phase, the face-to-face first wave was replaced by a concurrent mode design where all households with at least one listed/registered phone number were assigned to telephone and all other households to face-to-face. The listed phone numbers consist of both landline and mobile phone numbers that can be bought from commercial vendors. In the second phase, the telephone and face-to-face concurrent design was preceded by a Web invitation, resulting in a mix of a sequential and a concurrent design. All households were sent an invitation to participate through an on-line questionnaire. Nonresponding households were approached by telephone if a listed number was available and otherwise by face-to-face. The first phase was performed during 2010 and the second phase during 2012. In both years large parallel samples were drawn in order to assess method effects between the designs on the unemployment rate. The 2010 parallel run compared the old design to the intermediate concurrent design and the 2012 parallel run compared the intermediate design to the final design with all three modes.

The redesign did not change the data collection strategy per mode. In all years, the face-to-face contact strategy for the LFS first wave consists of a maximum of six visits to the address and contacts are varied over days of the week and times during the day. If no contact is made at the sixth visit, then the address is processed as a noncontact. The telephone contact strategy consists of three series of three calls. The three series are termed contact attempts and represent three different interviewer shifts. In each shift the phone number is called three times with a time lag of roughly an hour. The Web strategy is an advance letter with a login code to a website and two reminder letters with time lags of one week.

We use the 2010 – 2012 first wave LFS data to estimate various input parameters for the optimization model. In order to keep the exposition simple, and since the subsequent waves were not redesigned, we restrict ourselves to methods effects on unemployment rate estimates based on the first wave only. However, the first wave redesign may clearly have influenced the recruitment and response to waves 2 to 5. In follow-up studies at Statistics Netherlands, recruitment propensities to subsequent waves were included in the optimization problem, but we do not discuss these here. The LFS data were augmented with data from two administrative registers: the POLIS register and the UWV register. The POLIS register contains information about employments, allowances, income from employment and social benefits. The UWV register contains persons that have registered themselves as unemployed and applied for an unemployment allowance. Both registers contain relevant variables for the LFS and will be used to stratify the population.

4.2 The strategy set

The parallel runs in the LFS allow us to consider a multi-mode optimization problem with various single mode and sequential mixed-mode strategies. In the following, we abbreviate the telephone and

face-to-face modes to Tel and $F2F$, respectively. Although, the sequential strategy $Web \rightarrow F2F$ is observed only for large households and for households without a registered phone, we do include this strategy in the optimization.

Since later face-to-face and telephone calls are relatively much more expensive than early calls, we also introduce a simple cap on calls. For Tel we set the cap after two calls and for $F2F$ after three calls. These values are motivated by historical survey data, e.g., after these numbers of calls the cost per call increases quickly. We let $Tel2$ and $F2F3$ denote the strategies where a cap is placed on the number of calls. $Tel2+$ and $F2F3+$ represent strategies where there is no cap and the regular contact strategy is applied. We do realize that placing a cap is not the same as restricting the number of calls in practice. This holds especially for face-to-face. With fewer calls, interviewers or interviewer staff may change behaviour and spread calls differently. At Statistics Netherlands the $Tel2$ and $F2F3$ strategies are viewed as censored strategies with shorter data collection periods, e.g., two weeks instead of four weeks. Hence, cases are removed from the interviewer workloads after the pre-specified data collection period. From this perspective, it is more reasonable to assume that the optimal contact strategy during the first two weeks of a $F2F3+$ strategy is not so different from the optimal contact strategy in $F2F3$. Still, we may expect that realized response propensities and costs in strategies with a cap are different from their simulated propensities and costs. The strategy set now becomes

$$\begin{aligned} \mathcal{S} = \{&Web, Tel2, Tel2+, F2F3, F2F3+, Web \rightarrow Tel2, \\ &Web \rightarrow Tel2+, Web \rightarrow F2F3, Web \rightarrow F2F3+, \Phi\}, \end{aligned} \tag{4.1}$$

where Φ denotes the nonsampling strategy.

The parallel runs for the LFS in 2010 and 2012 were large. In both years the LFS sample was doubled in size for six months. Still, estimated parameters are subject to sampling variation and in case of the $Web \rightarrow F2F$ strategies possibly also to bias. We return to this issue in Section 4.6.

4.3 Population groups

In order to stratify the population, the regular LFS weighting variables were used as a starting point: unemployment office registration, age, household size, ethnicity and registered employment. Crossing the five variables led to 48 population strata (yes or no registered unemployed in household times three age classes times two household size classes times two ethnicity classes times yes or no registered employment in household). These strata were collapsed to nine disjoint strata based on their response behavior and mode effects:

1. *Registered unemployed*: Households with at least one person registered to an unemployment office (7.5% of the population).
2. 65+ *households without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment and with at least one person of 65 years or older (19.8% of population)
3. *Young household members and no employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment,

with all persons younger than 65 years, and with at least one person between 15 and 26 years of age (2.4% of population).

4. *Non-western without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment, with all persons younger than 65 years and older than 26 years of age, and at least one person of non-western ethnicity (1.5% of population).
5. *Western without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment, with all persons younger than 65 years and older than 26 years of age and all persons of western ethnicity (11.0% of population).
6. *Young household member and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons younger than 65 years, and with at least one person between 15 and 26 years of age (15.6% of population).
7. *Non-western and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons older than 26 years of age, and at least one person of non-western ethnicity (3.9% of population).
8. *Western and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons older than 26 years of age and all persons of western ethnicity (33.5% of population).
9. *Large households*: Households with more than three persons of 15 years and older without a registration to an unemployment office (4.9% of population)

The nine population strata were given informal labels in order to aid interpretation. Note, however, that the strata 7, 8 and 9 may have household members that are 65+. Furthermore, some subgroups follow from collapsing certain strata. For instance, households with at least one employment are found by combining strata 6, 7 and 8, and households with no more than three members of 15 years and older by combining all strata from 1 to 8.

In the optimization model, the nine strata were allowed different strategies and with different strategy allocation probabilities. In addition, we added precision constraints following the regular LFS on another stratification. Minimum numbers of respondents were requested based on age, ethnicity and registered unemployment. We refer again to Calinescu and Schouten (2013b) for details about these strata and corresponding precision thresholds.

4.4 Estimation of input parameters

The input parameters to the multi-mode optimization problem are subpopulation response propensities per strategy, subgroup telephone registration propensities, subgroup costs per sample unit per strategy, and subgroup adjusted method effects per strategy. We sketch the estimation of each set of parameters in the following subsections. More details can be found in Appendix A.

There are three settings that may occur when estimating input parameters: 1) The strategy is directly observed in historical survey data, 2) the strategy is only partially observed in historical survey data, i.e., only for a subset of the sample, and 3) the strategy is not observed at all.

For the LFS case study, the first setting applies to strategies Web, $Tel2+$, $F2F3+$, $Web \rightarrow Tel2+$. The second setting applies to $Web \rightarrow F2F3+$ and the third setting applies to $Tel2$, $F2F3$, $Web \rightarrow Tel2$ and $Web \rightarrow F2F3$. Sequential mixed-mode designs with face-to-face as the follow-up mode are only observed for households without a listed phone number and fall under settings 2 or 3 depending on whether a cap is placed on the number of calls. We attempted to deal with setting 2 by modeling the input parameters based on the observed differences in parameters between $Tel2+$ and $F2F3+$. We assumed that the ratio in response propensity between $F2F3(+)$ and $Tel2+$ for households with a listed phone number can be applied to $Web \rightarrow F2F3(+)$ and $Web \rightarrow Tel2+$. Furthermore, in the estimation, we assumed that strategies involving caps on the number of calls are similar to simulated strategies, i.e., by artificially restricting strategies with the full number of calls to the specified cap. Hence, we attempted to deal with setting 3 by censoring strategies. Calinescu and Schouten (2013b) elaborate these modeling steps.

For the method effect $D(s,g)$, two benchmarks were selected $\mathrm{BM}_1 = \bar{y}_{F2F3+}$ and $\mathrm{BM}_2 = 1/3 * (\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+})$, where $\bar{y}_{\mathrm{mode}}$ represents the average unemployment rate estimated via the indicated survey mode. The first benchmark assumes that the average unemployment rate that is estimated via a single mode face-to-face design represents the target unemployment rate. The second benchmark assumes there is no preferred mode, hence, it assigns an equal weight to each of the three modes. The $F2F3+$ benchmark is chosen because it is the traditional mode for the LFS first wave and, hence, determines the LFS time series up to 2010. Furthermore, we believe it is the mode that provides the smallest nonresponse bias for many surveys, see, e.g., Klausch et al. (2013a). It is, however, unclear whether $F2F3+$ should also be considered the mode with the smallest measurement bias. Hence, we also introduced the second benchmark to investigate the importance of the benchmark choice.

Standard errors for the estimated input parameters were approximated using bootstrap resampling per sampling stratum, following the stratified sampling design.

4.5 Optimization results

In this section, we explore the optimal allocation and minimal method effect for various budget levels, between stratum method effect levels and sample size levels

$$B \in \{160{,}000;\ 170{,}000;\ 180{,}000\}$$

$$M \in \{1\%;\ 0.5\%;\ 0.25\%\}$$

$$S_{\max} \in \{9{,}500;\ 12{,}000;\ 15{,}000\}.$$

Appendix B presents the minimal method effects for the various levels and or the two benchmark designs, BM_1 and BM_2. For the sake of brevity, here, we highlight mostly the results for BM_1, which is the former LFS design. The actual values for the non-adaptive regular three mode LFS design are

$$B = 170{,}000 \quad M = 3.00\%$$

$$S_{\max} = 11{,}000 \quad \bar{D}^{\mathrm{BM}_1} = -0.15\%.$$

Two main conclusions can be drawn from the results. First, the adaptive design is able to decrease the absolute overall method effect with respect to both benchmarks while respecting a strict constraint on the maximal between stratum method effect and keeping the budget at the current level. The only constraint that need to be relaxed in order to reduce the overall method effect is the maximal sample size. Second, for benchmark BM_2, smaller minimal overall method effects are obtained than for BM_1, with the exception of $S_{\max} = 9{,}500$. This difference is the result of the generally smaller and more similar values of the stratum method effects $D(s, g)$. We can explore the impact of the sample size constraint by comparing the optimal allocations for $S_{\max} = 9{,}500$ and $S_{\max} = 15{,}000$. Assume thresholds are set at $B = 170{,}000$, $M = 1\%$ and BM_1. Figures 4.1 and 4.2 present the optimal allocation probabilities per stratum and strategy given that a unit is sampled. Each figure can be seen as a matrix where each row represents one of the strategies in $\mathcal{S}^R$ and each column one of the 9 strata described in Section 4.3, e.g., g_1 is the registered unemployed stratum. Each cell in the matrix, i.e., intersection of a row with a column, shows the probability of assigning the corresponding strategy to the corresponding stratum. The probabilities are depicted as bars; the larger a bar, the larger the proportion of the stratum that is allocated to the strategy. The probabilities sum up to one over the strategies, i.e., over the rows. The exact values are given in the bars in case they are 20% or larger. Figure 4.1 and 4.2 show a clear shift in allocation probabilities when the sample size is allowed to increase, e.g., stratum 6 (young household member and employment) is almost fully allocated to *Web* and stratum 8 (western and employment) and 9 (large households) change from sequential to face-to-face only strategies.

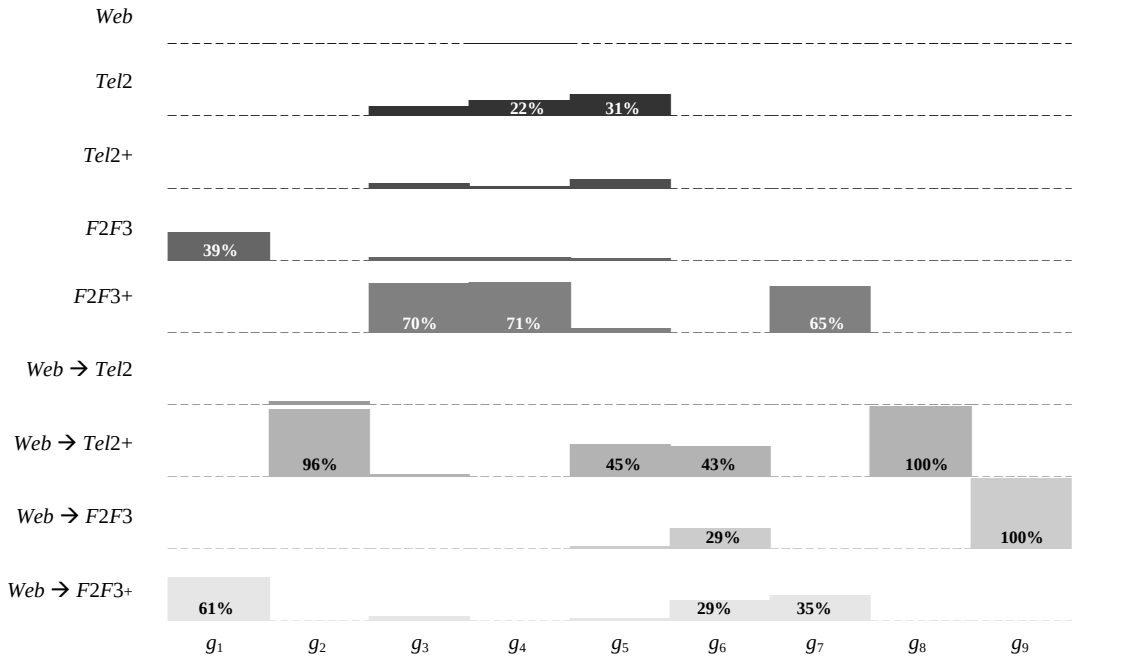

Figure 4.1 Strategy assignment given optimal solution for $S_{\max} = 9{,}500$, $B = 170{,}000$, $M = 1\%$, BM_1. The dotted line indicates that $p(s, g) = 0$.

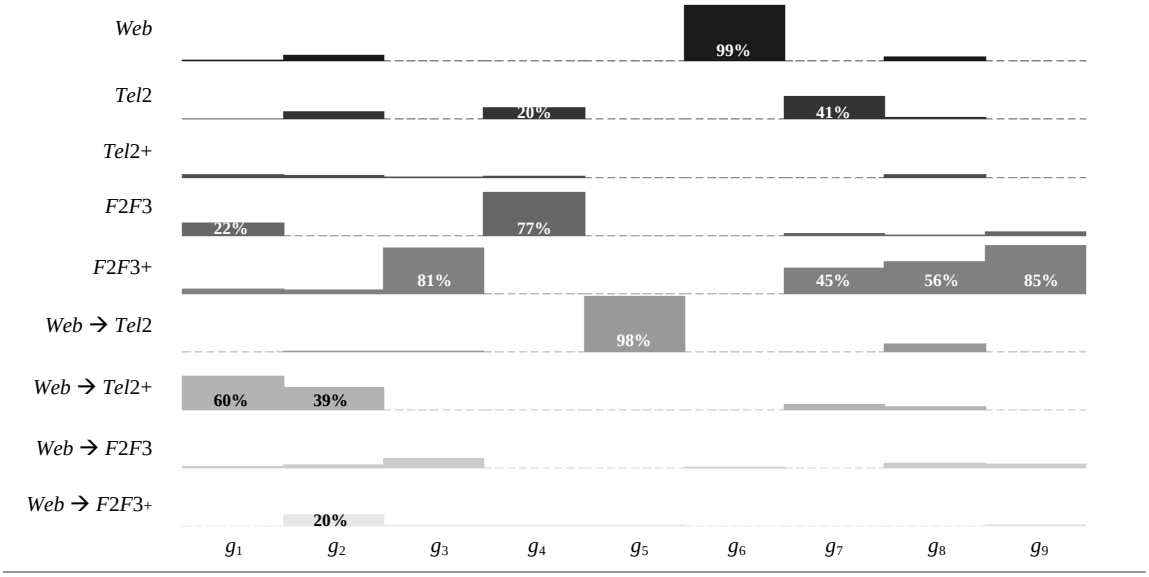

Figure 4.2 Strategy assignment given optimal solution for S_{max} = 15,000, B = 170,000, M = 1%, BM_1. The dotted line indicates that $p(s,g) = 0$.

The impact of the available budget can be seen very clearly for S_{max} = 12,000 and BM_1, where the minimal overall method effect drops from 0.10% for B = 160,000 to 0.01% for B = 180,000. The optimal allocation probabilities are shown in Figures 4.3 and 4.4. When increasing the budget, a shift takes place from telephone only strategies to a mix of face-to-face only strategies and, somewhat surprisingly, Web only strategies.

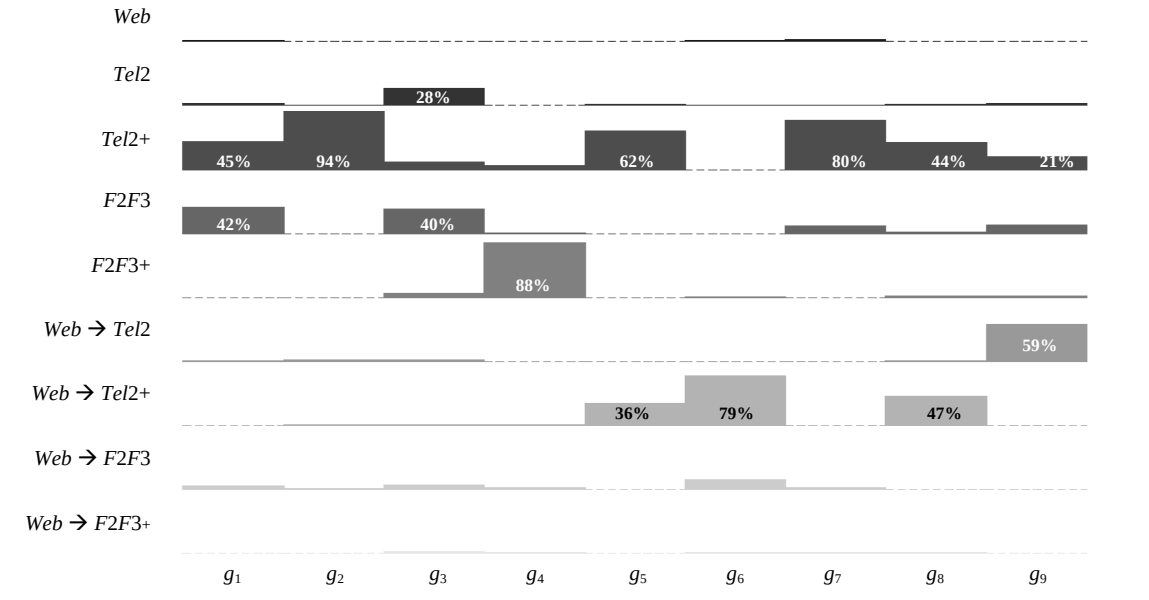

Figure 4.3 Strategy assignment given optimal solution for S_{max} = 12,000, B = 160,000, M = 1%, BM_1. The dotted line indicates that $p(s,g) = 0$.

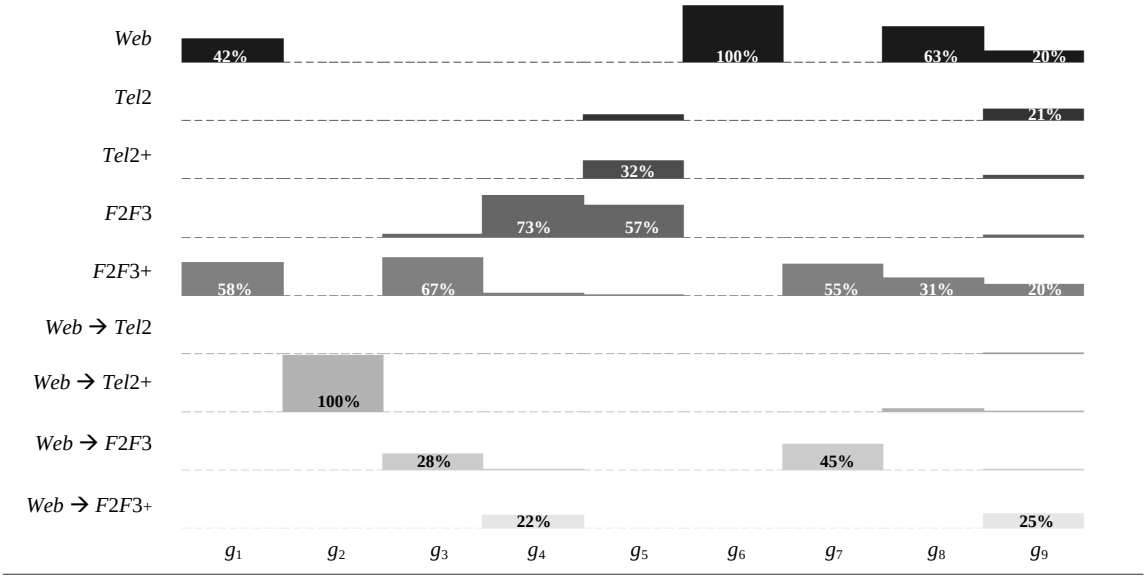

Figure 4.4 Strategy assignment given optimal solution for S_{max} = 12,000, B = 180,000, M = 1%, BM_1. The dotted line indicates that $p(s,g) = 0$.

A range of scenarios can be investigated using a wide range of threshold values, which we leave to other papers. We conclude by mentioning that optimal allocations with many small allocation probabilities lead to very intractable data collection processes. Lower thresholds to the allocation probabilities may be added to avoid strategies that get only small numbers of cases.

4.6 Robustness of optimal designs

In this section, we briefly discuss the robustness of the optimal designs. Sensitivity analyses are beyond the scope of this paper and are part of current research.

In the estimation of the response propensities, telephone registration propensities, costs per sample unit and adjusted methods effects, we make four main assumptions; apart from assumptions about the logistic link function between response – nonresponse, telephone registration – no registration and auxiliary variables. These are:

1. Model for $Web \rightarrow F2F3$ and $Web \rightarrow F2F+$: these two strategies have only been employed for households without a listed phone number.
2. Strategies with cap on calls estimated using censoring: The strategies with a cap on calls have not been conducted and we assume that their response propensities and costs can be approximated by censoring strategies with the full contact strategy.
3. Costs linear in size allocated to strategies: We assume that costs per sample unit do not depend on the size of the sample allocated to a strategy.
4. Time stability of methods effects during 2010 – 2012: Since the parallel runs were performed in two steps, the method effects for some strategies were estimated in two steps. We implicitly assume that the methods effects for these designs have not changed over 2010 – 2012.

Furthermore, all estimated input parameters are subject to sampling variation. Consequently, we expect that certain variations in the optimal designs might occur due to inaccuracy of parameters. In order to assess robustness of optimal designs we propose two types of sensitivity analysis:

- Repeated optimization for input parameters obtained from resampled data. In other words, all historical data are resampled multiple times and for each draw an optimization is performed. The resulting optimal values for quality and costs as well as the strategy composition of the optimal designs can thus be compared across the various draws.
- Performance evaluation of the optimal design on resampled data. In other words, given observed historical data, an optimization is performed. All historical data are then resampled and for each draw the optimization input parameters are recomputed. The optimal design is applied to each set of input parameters and the corresponding quality and cost values are computed. Finally, the statistical properties of quality and cost values are assessed across all draws of input parameters.

Exploratory sensitivity analyses show that there is relatively large variation in the strategy composition of the optimal designs, but that optimal method effects $\bar{D}^{\mathrm{BM}}$ are very stable. This implies that the method effect, as objective function, is a relatively smooth function.

5 Discussion

We constructed a multi-mode optimization problem that extends the framework of adaptive survey designs to mixed-mode survey (re)designs. This framework is especially useful when it is anticipated that method effects due to a change of mode design may impact the comparability and accuracy of statistics. To our best knowledge, this is the first research attempt of its kind and can be used as a basis for minimizing method effects subject to costs and other constraints.

In the optimization model, we included three quality criteria, one cost criterion and one logistical criterion. The quality criteria were the numbers of respondents in sampling strata, which acts as a surrogate for precision, the absolute adjusted overall method effect, which is the level shift caused by the design relative to the benchmark design and may be viewed as comparability in time, and the maximal absolute difference in method effects over important subpopulations, which may be viewed as comparability over population domains. The cost criterion is the total budget of the survey. The logistic criterion is the sample size, which needs to be limited in order to avoid a quick depletion of the sampling frame. The third quality criterion, the maximal absolute difference in subpopulation method effects, is nonlinear in the decision variables (the strategy allocation probabilities) and makes the optimization problem computationally complex. Although this criterion complicates the problem, it is a useful constraint that is often put forward by survey analysts and users. In regular redesigns, this criterion is often not considered and the Dutch LFS mixed-mode design leads to relatively large differences in method effects between subpopulations. Clearly, some of the criteria may be omitted and other quality, cost or logistical criteria may be added. In a follow-up on this research at Statistics Netherlands, various other criteria, mostly logistical, are considered.

In the optimization model, the focus was on maximizing quality, reflected by comparability in time, subject to cost constraints and other constraints on quality and logistics. The objective of the optimization may, however, be changed and each of the constraints could function as the objective. For instance, one

may minimize cost subject to quality and logistical constraints. One may also take a wider approach and perform several optimizations for different budget and quality levels in order to derive an informative multidimensional view on which a decision can be based.

Our attempt must be seen as merely a first step towards adaptive mixed-mode survey designs. There are various methodological and practical issues that need to be resolved. First, our approach is suited for surveys with only a few key statistics. For each of these statistics, an optimization can be performed and a weighted decision can be made. When a survey has a wide range of statistics, such an approach is not feasible. Second, the optimization leans heavily on the accuracy of its input parameters, i.e., estimated response probabilities, registered-telephone probabilities, cost parameters and mode effects in this case. It is important to assess the sensitivity of the optimization results to the accuracy of these parameters. It may be hypothesized that the objective function is relatively smooth with respect to these parameters, however, it is still important to perform sensitivity analyses. Third, it is essential to consider the sampling variation of the realized quality and costs of the optimized design when multiple waves of a survey are conducted. Such variation may be large and downsize the value of a precise optimization. Fourth, once nonlinear criteria are added to the problem, one has to rely on advanced solvers in statistical software. Even when using such solvers, convergence to global optimum is usually not assured and one has to be satisfied with local optima. For this reason, it is important to choose a useful set of starting points, including starting points that correspond to current designs. The practical issues concern the number of population strata, the number of strategies and the coordination to other surveys. Although survey administration systems and tools may support adaptive survey designs, such designs are harder to monitor and analyze. Furthermore, the tailoring of survey modes affects the size and form of interviewer workloads; interviewers may get only a specific range of subpopulations.

An important aspect of adaptive survey designs is the use of estimates for all kinds of input parameters such as response propensities, variable costs per sample unit and method effects between designs. Such estimates may not be readily available and there may only be weak historic survey data to support estimation. There are then four options: search for similar surveys that have historic support, be modest and restrictive in the choice of design features, perform a transitional period in which pilot studies and parallel runs are conducted, and develop a framework for learning and updating of parameters. In particular, designs with *Web* as one of the modes may still lack historic support for estimation in many countries, see, e.g., Mohorko, de Leeuw and Hox (2013). We also note that input parameters may gradually change in time, so that continuous updating will be needed. However, all of this is no different from a non-adaptive survey, except that now estimates are needed for relevant subpopulations instead of the overall population alone. Finally, we note that optimized adaptive designs, like optimized non-adaptive designs, provide an average, expected quality and costs. Due to sampling variation, the realized quality and costs will vary and unforeseen events may lead to deviations. Hence, monitoring and reacting to unforeseen events remain necessary.

Future research needs to address robustness of adaptive survey designs and should investigate other quality, cost and logistical criteria. It is also important that this study is replicated in order to evaluate whether the investment in terms of additional data collection and in terms of explicit optimization is worth the effort. The ultimate goal of this research is a data collection design strategy that allows for learning and updating optimization input parameters and that supports effective and efficient cost-benefit analyses in mixed-mode (re)designs. A Bayesian approach seems most promising for this purpose.

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.

Appendix A

Estimates of input parameters

In Section 4.4, we explain the estimation of input parameters for strategies that are observed only partially in the parallel runs. Here, we give the estimates for the response propensities, telephone registration propensities, variable costs per sample unit and adjusted method effects. Standard errors for all parameters were estimated using bootstrap resampling.

Table A2 presents the estimated response propensities $\rho(s,g)$ from available data and their corresponding standard errors. Table A1 shows the estimated propensity for a registered phone $\lambda(g)$.

Table A1
Estimated propensities for registered phone for group $g \in \mathcal{G}$ with the corresponding standard errors given in brackets

$\mathcal{G}$	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 A2
Estimated response propensities per strategy s and group g with the corresponding standard errors given in brackets

$\rho(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	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)
$Tel2$	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)
$Tel2+$	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)
$F2F3$	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)
$F2F3+$	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)
$Web \rightarrow Tel2$	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)
$Web \rightarrow Tel2+$	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)
$Web \rightarrow F2F3$	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)
$Web \rightarrow F2F3+$	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)

For the method effect $D(s,g)$, two benchmarks were selected after consultation with practitioners, i.e., $\mathrm{BM}_1 = \bar{y}_{F2F3+}$ and $\mathrm{BM}_2 = 1/3 * (\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+})$, where $\bar{y}_{\mathrm{mode}}$ represents the average unemployment rate estimated via the indicated survey mode. Tables A3 and A4 present the estimated method effects against the two benchmarks including their standard errors.

The estimates for the variable costs per sample unit plus estimated standard errors are given in Table A5. The costs are expressed relative to the $F2F3+$ strategy, which is set at one.

Table A3
Estimated method effects against benchmark $\mathrm{BM}_1 = \bar{y}_{F2F3+}$ with the corresponding standard errors given in brackets

$D^{\mathrm{BM}_1}(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	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)
$Tel2$	-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)
$Tel2+$	-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)
$F2F3$	-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)
$F2F3+$	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)
$Web \rightarrow Tel2$	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)
$Web \rightarrow Tel2+$	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)
$Web \rightarrow F2F3$	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)
$Web \rightarrow F2F3+$	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 A4
Estimated method effects against benchmark $\mathrm{BM}_2 = 1/3 * \left(\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+}\right)$ with the corresponding standard errors given in brackets

$D^{\mathrm{BM}_2}(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	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)
$Tel2$	-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)
$Tel2+$	-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)
$F2F3$	-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)
$F2F3+$	-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)
$Web \rightarrow Tel2$	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)
$Web \rightarrow Tel2+$	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)
$Web \rightarrow F2F3$	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)
$Web \rightarrow F2F3+$	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 A5
Estimated relative unit costs (in euros) per strategy s and group g with the corresponding standard errors given in brackets

$c(s,g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	0.03 (0.0)	0.04 (0.0)	0.04 (0.0)	0.03 (0.0)	0.04 (0.0)	0.03 (0.0)	0.03 (0.0)	0.03 (0.0)	0.03 (0.0)
$Tel2$	0.11 (0.1)	0.15 (0.1)	0.10 (0.1)	0.09 (0.1)	0.13 (0.1)	0.11 (0.1)	0.09 (0.1)	0.12 (0.0)	0.14 (0.1)
$Tel2+$	0.13 (0.1)	0.17 (0.1)	0.11 (0.1)	0.10 (0.1)	0.15 (0.1)	0.14 (0.1)	0.11 (0.1)	0.16 (0.1)	0.20 (0.2)
$F2F3$	0.84 (0.4)	0.89 (0.5)	0.83 (0.5)	0.82 (0.8)	0.86 (0.3)	0.84 (0.2)	0.81 (0.5)	0.84 (0.2)	0.89 (0.5)
$F2F3+$	1.00 (0.6)	1.00 (0.6)	1.00 (0.7)	1.00 (1.1)	1.00 (0.4)	1.00 (0.3)	1.00 (0.6)	1.00 (0.2)	1.00 (0.5)
$Web \rightarrow Tel2$	0.08 (0.0)	0.11 (0.1)	0.09 (0.1)	0.09 (0.1)	0.09 (0.0)	0.08 (0.0)	0.08 (0.0)	0.07 (0.0)	0.07 (0.0)
$Web \rightarrow Tel2+$	0.09 (0.1)	0.12 (0.1)	0.10 (0.1)	0.10 (0.1)	0.10 (0.1)	0.09 (0.0)	0.09 (0.1)	0.08 (0.0)	0.07 (0.0)
$Web \rightarrow F2F3$	0.60 (0.3)	0.66 (0.7)	0.64 (0.6)	0.70 (0.8)	0.59 (0.4)	0.56 (0.3)	0.65 (0.5)	0.51 (0.2)	0.61 (0.4)
$Web \rightarrow F2F3+$	0.71 (0.4)	0.71 (0.7)	0.80 (0.9)	0.84 (1.2)	0.73 (0.6)	0.68 (0.4)	0.81 (0.8)	0.62 (0.3)	0.71 (0.6)

Appendix B

Overview optimization results

In Section 4.5 we illustrate our approach to solve the multi-mode optimization problem for a range of input parameters. Tables B1 and B2 give a brief overview of the optimization results.

Table B1
Overview optimization results linear programming formulation - minimize costs

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

Table B2
Overview optimization results nonlinear problem - minimize average method effect in LFS

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%

References

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

Blanke, K., and Luiten, A. (2012). ESSnet project on data collection for social survey using multi modes (dcss). Paper for the UNECE Conference of European Statistics, Oct 31 - Nov 2, Geneva, Switzerland.

Buelens, B., and van den Brakel, J. (2014). Measurement error calibration in mixed-mode surveys. Forthcoming in *Sociological Methods and Research*.

Calinescu, M., and Schouten, B. (2013a). Adaptive survey designs that account for nonresponse and measurement error. Disscussion paper, Statistics Netherlands.

Calinescu, M., and Schouten, B. (2013b). Adaptive survey designs to minimize mode effects a case study on the dutch labour force survey. Disscussion paper, Statistics Netherlands.

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, 545-572.

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

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, 1-18.

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

Heeringa, S., and Groves, R. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3).

Jäckle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78, 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. (2013a). Assessing the mode-dependency of sample selectivity across the survey response process. Discussion paper, Statistics Netherlands.

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

Klausch, L., Hox, J. and Schouten, B. (2014). Evaluating bias in sequential mixed-mode surveys against single- and hybrid-mode benchmarks the case of the crime victimization survey. Discussion paper, Statistics Netherlands.

Kreuter, F. (2013). *Improving Surveys with Process and Paradata*. John Wiley & Sons, Inc.

Mohorko, A., de Leeuw, E.D. and Hox, J. (2013). Internet coverage and coverage bias in Europe: Developments across countries over time. *Journal of Offcial Statistics*, 29, 609-622.

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/.

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.

Suzer-Gurtekin, Z. (2013). *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys*. Ph.D. thesis, University of Michigan.

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

Vannieuwenhuyze, J. (2013). *Mixed-Mode Data Collection: Basic Concepts and Analysis of Mode Effects*. Ph.D. thesis, Katholieke Universiteit Leuven.

Wagner, J. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. Ph.D. thesis, University of Michigan.