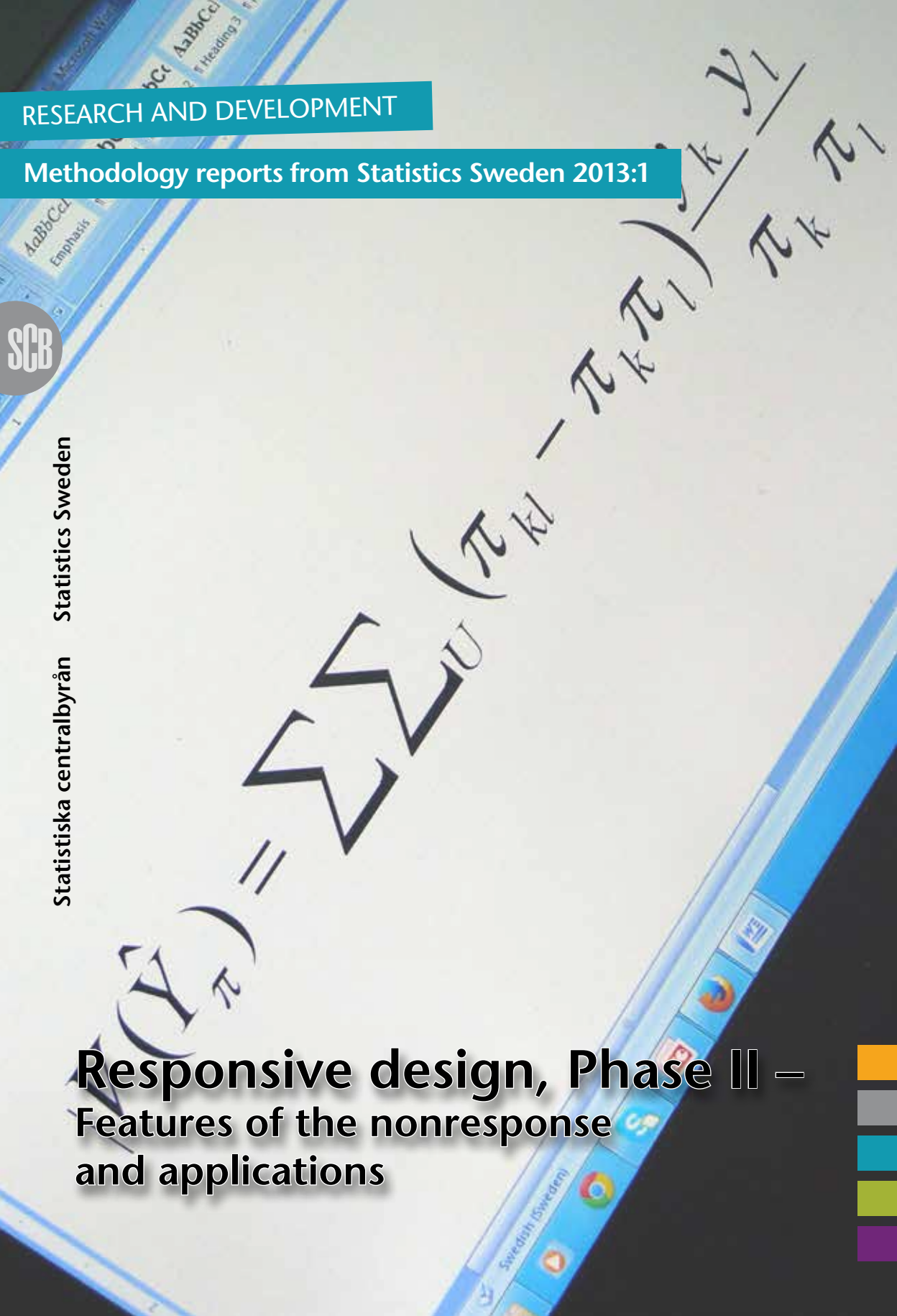
RESEARCH AND DEVELOPMENT

Methodology reports from Statistics Sweden 2013:1



Statistiska centralbyrån Statistics Sweden

Responsive design, Phase II – Features of the nonresponse and applications

Research and Development – Methodology reports from Statistics Sweden 2013:1

Responsive design, Phase II – Features of the nonresponse and applications

Peter Lundquist
Carl-Erik Särndal

Statistiska centralbyrån
2013

Research and Development – Methodology reports from Statistics Sweden 2013:1

Responsive design, Phase II – Features of the nonresponse and applications

Statistics Sweden
2013

Producer

Statistics Sweden, Research and Development Department
SE-701 89 ÖREBRO
+ 46 19 17 60 00

Inquiries

Peter Lundquist, +46 8 506 949 18

Cover Ateljén, SCB
Photo Jan-Aage Haaland

ISSN 1653-7149 (Online)
URN:NBN:SE:SCB-2013-X103BR1301_pdf

This publication is only published electronically on Statistics Sweden's website www.scb.se

Preface

This is the second report about Statistics Sweden's project on Responsive Design. It provides further theory and illustrations relative to the reduction of nonresponse bias and data collection cost for surveys on individuals and households at Statistics Sweden. The Living Conditions Survey is the main source for the illustrations. The project can be seen as part of the ongoing quality improvement efforts at Statistics Sweden. The first report appeared in the same series as R&D 2012/1.

Statistics Sweden, October 2013

Lilli Japec

Acknowledgment
The authors gratefully acknowledge the cooperation of colleagues at Statistics Sweden for several aspects of this work. Anders Ljungberg supported the idea to use the Swedish Living Conditions Survey for these studies. Tommy Blomqvist and Vanja Hultkrantz contributed in making available the necessary data files. Frida Danielsson and Birgitta Göransson played important roles in the planning and the execution of the embedded experiment.

Contents

Summary

A project on Responsive Design was carried out at Statistics Sweden in 2011 and 2012. This R&D report about Phase 2 of that project consists of four parts, dealing with different aspects of survey nonresponse. These parts share a common background of concepts and ideas.

The first part (Sections 1 to 4) reviews and extends theoretical and empirical findings in Phase 1 of the project, reported in Lundquist and Särndal (2012). In a data collection extending over a period of time, weeks or perhaps months, the set of responding units present at any given point in time is more or less well balanced. The concept of Imbalance of the set of respondents is a central one. Imbalance can be measured, and we can use this concept to influence the data collection in the direction of better balance in the final set of respondents. The empirical illustrations come from Statistics Sweden's 2009 Living Conditions Survey (abbreviated LCS 2009).

Section 1 reviews the survey background and some of the literature on Responsive Design. Section 2 presents the properties of Imbalance, which is a statistic based on a vector of selected auxiliary variables (monitoring variables). The concepts Balance (of the set of respondents), and Distance (between respondents and nonrespondents) are derived as functions of Imbalance. These are also discussed in Section 2.

As Section 3 notes, the auxiliary variables available in a survey can be employed to meet two different goals. Some find use at the data collection stage to monitor the data collection and to carry out suitable interventions. Others enter into consideration only at the estimation stage, where they serve to compute calibrated weights for the nonresponse adjusted estimates (calibration estimates).

Analysis of Imbalance (ANIMB) for the data collection is a tool proposed in Section 3.2. Important in that context are the notions of total, marginal and conditional imbalance.

Response propensity scoring is a tool for monitoring the data collection. The concept is presented in Section 4, and their use for interventions in the data collection is considered. The idea is that data collection attempts can be reduced or stopped entirely for units with high response propensity relative to less responsive units, so

that the data collection period ends with a well balanced response set. Section 4 presents the results of an ANIMB of the LCS 2009, and illustrates different types of intervention.

Part two of the report (Section 5) explores the use of Imbalance and related concepts in estimation for domains (subpopulations). To this end, we used as study objects several survey variables considered important in the LCS 2009.

Part three of the report (Section 6) applies the concepts of Balance and Distance to a totally different survey, namely, the Swedish part of the international PIAAC 2011. The essential concepts in the report, Imbalance, Balance and Distance, were conceived with a general purpose in mind; therefore, the objective with this section was to confirm that they can be integrated into a different survey setting. Dr. Sara Westling developed the empirical content for this section.

Part four of the report (Section 7) deals with an embedded experiment in the 2011 version of the LCS. The objective was to propose and test a new carefully elaborated calling strategy intended to achieve both higher overall response rate and better balanced response. The experimental design is described and the results analyzed.

The concluding discussion in Section 8 contains comments on the various parts of the report. A reader who does not seek a detailed account may still get important insights into the project by reading first Sections 1.1 and 1.2, and by then proceeding directly to Section 8.

1 Introduction and review of earlier results

1.1 Project plan

This report presents results from Phase 2 of a methodology project about responsive survey design at Statistics Sweden. The objective was more specifically to explore procedures for obtaining a balanced set of respondents, from a given probability sample. Such procedures rely on indicators of balance (or representativity) that can be monitored during data collection, so that a well balance response will be the end result.

Phase 1 of the project included a preliminary study of indicators. Using process data from Statistics Sweden's WinDATI data collection system, experiments were carried out on the data from the 2009 Swedish Living Conditions Survey, which we call LCS 2009. The behaviour of the indicators were studied.

The objectives for Phase 2 are as follows (Projektbeställning 2011-10-06): The balance indicators are to be further developed and studied. The concepts of conditional balance and conditional partial balance are to be investigated. The performance of the indicators is to be further illustrated empirically, by applying them to at least one more important survey at Statistics Sweden.

Finally, recommendations are to be given for the implementation of the indicators in future field work for important surveys at Statistics Sweden. A dynamic data collection design is aimed at, where interventions may be made during the data collection, at suitably chosen points. For example, data collection may be stopped in sample subgroups where a realistic response rate has already been achieved, so that the resources saved can be directed to improving other aspects of the survey and the data collection.

1.2 Quality of the survey response – what is it?

Survey quality, in its various dimensions, is very much in focus at Statistics Sweden. But "quality of the survey response" is seldom mentioned or measured. It is necessary to do so, at least for this project.

Probability sampling is the dominating mode of sampling at Statistics Sweden. The response set is the subset of the probability sample for which the study variable values (the y-values) are recorded and available for estimation.

The response rate receives much attention at Statistics Sweden as in other survey organizations. Its decreasing tendency, over more than 30 years, is much deplored. The nonresponse rate measures one aspect of the data collection. It has become increasingly clear that the nonresponse rate is not by itself suitable, or at least not sufficient, for an effective monitoring of the data collection. For example, it may be inefficient to continue according to an unchanging scenario driven primarily by the desire to obtain the highest possible response rate in the end, or to reach, by a costly and time consuming effort, a predefined rate of response, such as 70%.

All agree that high quality response is very important. The quality of the survey estimates depends directly on the quality of the response. What do we mean (or what should we mean) by the *quality of the response set*? There is no commonly accepted definition. Several concepts, not always clearly defined, figure in discussions of nonresponse:

Be representative (of the population, of the probability sample)

Be well balanced (with respect to observable (auxiliary) variables)

Be near (in terms of distance) the whole probability sample

Be a large portion of the probability sample (have a high response rate)

The properties *representative*, *well balanced* and *small distance* are based on *observable characteristics* of the sample units, and they are therefore much more descriptive and informative than the last mentioned property, the response rate, which needs only a count of how may responses were collected.

In an oversimplified example, "balanced" can mean that the response set and the entire sample have the same proportion of men and the same proportion of women. In practice, we can seek balance on many more characteristics than just "gender"; balance is then a multivariate statistic, defined in terms of multiple auxiliary variables.

Representativity, balance and distance are measurable aspects of the quality of the survey response. They should be regularly measured during the data collection period.

As mentioned, the *response rate* is very different from the first three items in the list. It does not refer to characteristics of the respondents. It only tells how many responding units were obtained, out of the selected sample. A response set representing 56% of the sample can, if unbalanced, have poor quality compared with a response set representing 49%. While this is common sense to experienced survey methodologists and to insightful interviewers, it is a message that is difficult to get across to more conservative survey administrators.

The overall survey response rate receives a disproportionate amount of attention. It does not address the critical issue. The critical issue is the quality of the response set realized at the end of the data collection.

The interest in the response rate is a remnant of the old classical probability sampling era of the 1950's: We want 100% response, in other words that all those selected for the sample respond, in order to guarantee unbiased estimation, and we want many responses to keep variance low. Certainly those are desirable objectives.

But recent decades show that realizing full response in socio-economic surveys is a far cry, a long since abandoned hope. Many statisticians find this fact difficult to accept. Whether the non-response is 35% or 45% is in itself of little interest to the experienced survey methodologist. The damage – usually a considerable non-response bias - is done. A necessity for the future is to deal rationally with an undesirable fact, if statistics of national importance are going to be produced at all. What counts is if the remaining 65% or 55% are a high quality response set. This aspect of quality can be measured, with methods discussed in this report.

1.3 Adaptive survey design and responsive design; a literature review

The terms *adaptive design* and *responsive design* are frequently used in recent literature. Bethlehem, Cobben and Schouten (2011) regard responsive design as a special case of adaptive design. Adaptive design seems to refer mainly to situations where treatments applied to sampled elements are identified prior to the start of the data collection, although they may also be revised or modified during the data collection. Responsive design is used mainly for situations where the data collection may involve two or more phases, with decisions taken underway about steps for the subsequent phases.

The general objectives of responsive design are formulated in Groves and Heeringa (2006). A number of applications of related approaches have subsequently appeared. Options for responsive design in a Canadian setting are discussed in Mohl and Laflamme (2007) and Laflamme (2009). Work on the development of adaptive designs has been presented for example in Wagner (2008).

Optimal scheduling of contact attempts is an example of adaptive design. Calienscu, Bhulai and Schouten (2011) formulate the scheduling as an optimization problem involving time slots, $t = 1,...,T$, groups of units, $g = 1,...,G$, and different survey modes, $m = 1,...,M$. An optimal solution is presented whereby one can decide at each time slot *t* to approach units in a given group *g* through a specific survey mode *m*.

1.4 The R-indicator

The recent book by Bethlehem, Cobben and Schouten (2011), titled *Handbook of Nonresponse in Household Surveys*, contains important material on different aspects of the nonresponse problem, some of it related to the issues in this report. Although the objectives are similar, the arguments in BCS (for Bethlehem, Cobben and Schouten) differ from those in this report.

The indicator for representativity of the survey response (R-indicator) proposed by BCS is based on the standard deviation of the response probabilities for the population units. Since these are unknown, estimated response probability must be used. The basic R-indicator is given by $\hat{R} = 1 - 2\hat{S}(\hat{\theta})$, where $\hat{S}(\hat{\theta})$ is the standard deviation of the estimated response probabilities $\hat{\theta}_k$ for the units *k* in the sample s, that is, the square root of $\hat{S}^2(\hat{\theta}) = \sum_s d_k(\hat{\theta}_k - \bar{\hat{\theta}}_s)^2 / \sum_s d_k$, where $d_k = 1/\pi_k$ is the sampling weight, π_k being the inclusion probability of unit k.

The concepts of conditional representativity and conditional partial representativity are discussed by BCS. Partial R-indicators can be of two types: Unconditional partial R-indicators and conditional partial R-indicators. BCS remark that "unconditional partial indicators are designed typically for comparisons of different surveys or surveys in time. Conditional partial indicators are especially suited for data selection monitoring." In this report we are interested mainly in the

latter perspective, that is, data selection monitoring in one and the same survey.

BCS present (in Example 7.7, page 193) an example of an analysis of unconditional and conditional partial R-indicators. They remark that in this example the partial R-indicators change little in the later stages of the data collection. The partial indicators identify the categories that are overrepresented or underrepresented in the course of the data collection.

1.5 The Swedish Living Conditions Survey (LCS), a brief description

As mentioned in Lundquist and Särndal (2012), from now on called LS (2012), the LCS is a sample survey designed to measure different aspects of social welfare in Sweden, in particular among different population subgroups. The LCS 2009 sample consists of a sample of individuals 16 years and older, drawn from the Swedish Register of Total Population. The data set used in the analysis in this report is a subsample of n = 8,220 individuals, taken from the entire LCS 2009 sample. This subsample can be regarded as a simple random sample.

In the LCS telephone interviews were conducted by a staff of interviewers using the Swedish CATI-system, WinDATI. All attempts by interviewers to establish contact with a sampled person are registered by WinDATI. For every sampled individual, the WinDATI system thus records a series of "call attempts", which play an important role in our analysis.

"WinDATI events" include events such as call without reply, busy line, contact with household member other than the sampled person, and appointment booking for later contact. When contact and data delivery has occurred, the data collection effort is complete for the sample member in question. Every registered WinDATI event is a "(call) attempt" in the following.

The LCS 2009 ordinary field work lasted five weeks, at the end of which the response rate was 60.4%. For some sampled persons, 30 or more call attempts had then been recorded. This was followed by a three week break during which characteristics of non-interviewed individuals were examined, in order to prepare for the three week follow-up period, which concluded the data collection. All individuals considered by the survey managers to be potential respondents were included in the follow-up effort, which brought

the response rate up to an ultimate 67.4%. However, there was no separate strategy or revised procedure for the follow-up. It followed the same routines as the ordinary field work. Hence, there were no attempts at responsive design such as for example a follow-up focusing on underrepresented groups, in an effort to thereby reduce the nonresponse bias.

1.6 A summary of Phase 1 of the project

Results of Phase 1 of the Responsive Design project are reported in LS (2012), containing a review of the recent literature on responsive design. Indicators of important characteristics of the data collection such as balance and distance were created. The empirical work focused on the data from LCS 2009, with its final response rate of 67.4%.

The LCS 2009 data file contains, in addition to study variables measured in the survey, a number of auxiliary variables, as well as process data from the WinDATI system. The measure of balance (of the set of respondents) and the measure of distance between respondents and nonrespondents) are computed on a vector of selected auxiliary variables. As shown in LS (2012) we can use these measures to follow the data collection as a function of the number of call attempts made to the sampled individuals. The indicators are of general scope and can be used for monitoring in other surveys as well.

For obvious reasons, one would like to see that the balance improves and that the distance decreases as the data collection progresses. This did not happen for the LCS 2009 data, as Tables 4.1 and 5.1 in LS (2012) show. Instead, the balance of the set of respondents deteriorates, and is less favouable at the end of the follow-up than at the end of the ordinary data collection. This casts doubt on essential aspects of the LCS data collection, as currently carried out. Furthermore, the estimates for important variables showed little change, or a change in the wrong direction, particularly in the later stages of the data collection.

In view of such disappointing results, it was decided to use the LCS 2009 data file for "experiments in retrospect". A set of important sample subgroups was defined, and data collection was deemed terminated in a group when the response rate had reached a specified threshold value. These interventions, placed at suitably chosen points in the data collection, bring considerable

improvement – better balance, reduced distance – compared with the traditional LCS data collection.

The following table taken from LS (2012) is based on LCS 2009. Benefits, Income and Employment are register variables (available for the full sample), used here as study variables. The table shows the relative deviation from the unbiased estimate, $RDF_{CAL} = (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}$ and $RDF_{EXP} = (\hat{Y}_{EXP} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}$, where $\hat{Y}_{CAL}$, $\hat{Y}_{EXP}$ and $\hat{Y}_{FUL}$ denote, respectively, the calibration estimator, the expansion estimator and the full response (Horvitz-Thompson) estimator. The line "Actual LCS 2009" is computed on the response set actually achieved, without any interventions or attempts at balancing. The experimental strategies 1, 2 and 3 represent response sets with increasingly better balance, obtained interventions in the LCS 2009 data set

	RDF_{CAL}			RDF_{EXP}		
	Sickness benefits	**Income**	**Employment**	**Sickness benefits**	**Income**	**Employment**
Actual LCS 2009	**-3.6**	**2.9**	**3.1**	**-9.4**	**6.7**	**4.8**
Strategy 1	**-1.6**	**2.7**	**3.0**	**-7.5**	**4.6**	**3.4**
Strategy 2	**-1.2**	**2.6**	**3.2**	**-7.6**	**3.1**	**2.2**
Strategy 3	**1.0**	**1.0**	**2.3**	**-6.5**	**1.0**	**0.3**

As expected, the three variables behave differently, but some general patterns are evident. For Actual LCS 2009, RDF_{EXP} is very high compared with RDF_{CAL}, showing that the calibration brings considerable improvement, as one would hope. Then, as the level of balance increases (Strategies 1 to 3), both RDF_{EXP} and RDF_{CAL} are steadily reduced; the reduction is especially pronounced for RDF_{EXP}, at least for Income and Employment.

For Strategy 3, the response set is very well balanced, given the 8-dimensional monitoring vector used here. Then we can expect that much of the difference between RDF_{EXP} and RDF_{CAL} will have disappeared. This is clearly seen for the variable Income, and as the bottom line shows, RDF_{EXP} is even smaller than RDF_{CAL} for the variable Employment.

In order to use the insight from such experiments in future practice, we need to anticipate a "reasonable expectations" response rate, say 60% or 50%, to be used as a stopping rule for the data collection in a group. Another important question is how to select the groups to be monitored in the data collection.

2 Review of theory

2.1 General theory background

We review in this section the basic theory in LS (2012), including the concepts of balance (of the set of respondents) and distance (between respondents and nonrespondents).

Denote by $U = \{1,...,k,...,N\}$ the finite population consisting of N units indexed $k = 1,2,...,N$. A probability sample s is drawn from U; unit k has the known inclusion probability $\pi_k = \Pr(k \in s) > 0$, and the known design weight $d_k = 1/\pi_k$. Usually, surveys of national importance involve many study variables. We denote by y_k the value for unit k of a study variable y for which we wish to estimate the population total $Y = \sum_U y_k$. (A sum $\sum_{k \in A}$ over a set of units $A \subseteq U$ will be written as $\sum_A$.) If the response were complete, this estimation would use the values y_k then available for all units $k \in s$. But the response is incomplete.

We follow the data collection as a function of the *call attempt number*. There is a series of successively larger response sets $r^{(a)}$, where a refers to the call attempt number, $a = 1, 2, \ldots$, and

$$r^{(1)} \subseteq r^{(2)} \subseteq \ldots \subseteq r^{(a)} \subseteq \ldots \tag{2.1.1}$$

Here $r^{(a)}$ is the set of units having delivered the value y_k at a certain point a (after a call attempts).

For a simplified notation, we let r refer to any one of the increasingly larger response sets. The values y_k for $k \in r$ are (together with auxiliary variable values) available for estimating the total $Y = \sum_U y_k$.

Data collection stops before r has reached the full probability sample s. We have $r \subset s \subset U$. For the response r the realized (design-weighted) response proportion is

$$P = \sum_r d_k / \sum_s d_k \tag{2.1.2}$$

We call P the response rate, but one should keep in mind that it is a proportion that increases during data collection, as the response set

r gets larger. When the data collection is finally stopped, the ending value P is the ultimate response rate for the survey. The response probability of unit *k*, denoted $\theta_k = \Pr(k \in r | k \in s)$, is unknown. It is a conceptually defined, non-random, non-observable number. The response rate *P* is an estimate of the (unknown) mean response probability in the population, $\bar{\theta}_U = \sum_U \theta_k / N$.

2.2 Auxiliary vectors and estimation by calibration

Several auxiliary variables are available to help the data collection and the estimation; the auxiliary variables values are assumed known for all $k \in s$. Auxiliary information plays an important role both at the data collection stage and at the estimation stage. We use $\mathbf{x}$ as the general notation for the auxiliary vector whose $\mathbf{x}_k$ is assumed available at least for all units $k \in s$, possibly for all $k \in U$. If $J \geq 1$ auxiliary variables are available and used, then $\mathbf{x}_k = (x_{1k},...,x_{jk},...,x_{Jk})'$, where x_{jk} is the value for unit *k* of the j^{th} auxiliary variable, x_j.

We use auxiliary vectors $\mathbf{x}_k$ of the following form: It is possible to state a constant vector $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all units *k*. This is not a major restriction. Vectors of importance in practice are usually of this kind. A simple illustration is when $\mathbf{x}_k = (1, x_k)$; then $\boldsymbol{\mu} = (1, 0)$ satisfies the requirement. When $\mathbf{x}_k = (0,...,1,...,0)'$ used to code a set of mutually exclusive and exhaustive categories, then $\boldsymbol{\mu} = (1, 1,...,1)'$ satisfies the requirement. The auxiliary variables can be categorical or continuous; in many applications at Statistics Sweden, they are categorical.

At the estimation stage, the response set *r* is finalized and fixed. The study variable values available for estimation are limited to y_k for $k \in r$. The estimation is done by calibrated weighting. The value y_k receives the weight $d_k m_k$, where $d_k = 1/\pi_k$ is the sampling weight, and the adjustment factor

$$m_k = (\sum_s d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$$

is constructed to realize a *calibration* to the estimate $\sum_s d_k \mathbf{x}_k$, computable for the full sample *s* and the given auxiliary vector $\mathbf{x}_k$. The calibration property is expressed by the equation

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$$

The importance of this calibration lies in the fact that the right hand side, $\sum_s d_k \mathbf{x}_k$, is unbiased for the population total $\sum_U \mathbf{x}_k$. The calibrated weights $d_k m_k$ will adjust the estimates in the direction of unbiasedness. The calibration estimator of $Y = \sum_U y_k$ based on the auxiliary vector $\mathbf{x}_k$ and the weights $d_k m_k$ is

$$\hat{Y}_{CAL} = \sum_r d_k m_k y_k \tag{2.2.1}$$

Calibration on a powerful vector $\mathbf{x}_k$ may substantially reduce the nonresponse bias, although without eliminating it completely. Some bias always remains. At Statistics Sweden, calibrated weighting is extensively used; many potential auxiliary variables are typically available for the estimation stage, and procedures are available for choosing the most efficient among them; see Särndal and Lundström (2008, 2010). In this report we are not specifically concerned with these selection methods.

2.3 Measuring the imbalance in the response set

At any given point in the data collection, the response set is more or less balanced. The imbalance is measured with respect to a given **x**-vector, whose computable (design weighted) mean is $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$ for the response set *r* and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$ for the whole probability sample s.

If $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ we say that the response set is *perfectly balanced* with respect to the vector $\mathbf{x}_k$. Normally we do not achieve this in practice, but we can strive to come close.

The vector of mean differences is denoted $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$. We have $\mathbf{D} = (D_1,...,D_j,...,D_J)'$, where $D_j = \bar{x}_{jr} - \bar{x}_{js}$ is the difference, for the j:th x-variable, between the respondent mean, $\bar{x}_{jr} = \sum_r d_k x_{jk} / \sum_r d_k$, and the full sample mean, $\bar{x}_{js} = \sum_s d_k x_{jk} / \sum_s d_k$. Large differences D_j signify that the response set is not well balanced.

The **x**-vector is multivariate; we need a univariate measure of the *imbalance* of the response set. It is defined as $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ $= (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$, a scalar quantity. For a response set that is perfectly balanced with respect to the vector $\mathbf{x}_k$, we have $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, $\mathbf{D} = \mathbf{0}$ (the zero vector) and $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = 0$. A reason for interposing the inverse of the weighting matrix $\boldsymbol{\Sigma}_s = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')/(\sum_s d_k)$ is that an upper bound can then be stated on the imbalance: Given s, we have $0 \le \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \le \frac{1}{P} - 1$, whatever *r* and $\mathbf{x}_k$. For example, for 1 – *P* = 20% nonresponse, $0 \le \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \le 0.25$; for 50% nonresponse, $0 \le \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \le 1$. Large nonresponse 1 – *P* and large mean differences D_j are factors contributing to large imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$.

A typical tendency during the data collection is that the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ decreases, because $\bar{\mathbf{x}}_r$ comes nearer the fixed sample mean $\bar{\mathbf{x}}_s$ (although this depends also on what particular units that happen to be in the set r.) If nearly complete response can realized, so that r is near s, then $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is near zero, because $\bar{\mathbf{x}}_r \approx \bar{\mathbf{x}}_s$. If $r = s$, then $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = 0$. Even if r is much smaller than s, we can have imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = 0$, namely if the response set is perfectly balanced with respect to the vector $\mathbf{x}_k$, that is, $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ and $\mathbf{D} = \mathbf{0}$.

2.4 The concepts of distance and balance

For a given $\mathbf{x}$-vector, the *distance between respondents and nonrespondents* is measured, at a given point in the data collection, by

$$dist_{r|nr} = [(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})]^{1/2}$$

where $nr = s - r$ is the nonresponse set and $\bar{\mathbf{x}}_{nr} = \sum_{s-r} d_k \mathbf{x}_k / \sum_{s-r} d_k$ Its relation to the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is expressed by

$$dist_{r|nr} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}/(1-P)$$

In a satisfactory data collection, the distance should decrease, or at least not get larger, as the response set gets larger. As pointed out earlier, it is normal that the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ decreases, because $\bar{\mathbf{x}}_r$

comes nearer the fixed sample mean $\bar{\mathbf{x}}_s$. The factor $1/(1-P)$ increases. Therefore, in order for the distance $dist_{r|nr}$ to decrease, the decrease in imbalance would have to be sufficiently pronounced so as to overcome the increase in $1/(1-P)$. An imbalance that stays roughly constant during data collection will definitely not give a decreasing distance.

From $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \leq 1/P - 1$ follows an upward bound on the distance, $dist_{r|nr} \leq 1/\sqrt{P(1-P)}$. For example, for 50% nonresponse, $dist_{r|nr} \leq 2$ for any r and auxiliary vector $\mathbf{x}$.

From the inequality $dist_{r|nr} \leq 1/\sqrt{P(1-P)}$ we get a measure of *balance of the response set*, placed in the unit interval, see LS (2012):

$$BI_1 = 1 - \sqrt{P(1-P)} \times dist_{r|nr} = 1 - \sqrt{\frac{P \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}}{1-P}}$$

Because $P(1-P) \leq 1/4$, an alternative indicator also contained in the unit interval is

$$BI_2 = 1 - 2P(1-P) \times dist_{r|nr} = 1 - 2P\sqrt{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}}$$

The case of mutually exclusive and exhaustive groups is of special interest because the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ takes a particularly simple form when the vector $\mathbf{x}_k$ points out membership in one out of J mutually exclusive and exhaustive categories. Then $\mathbf{x}_k$ has $J-1$ entries "0" and one single entry "1" pointing out the group to which unit k belongs. The imbalance statistic is then a sum of J nonnegative terms:

$$\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} C_j$$

where

$$C_j = W_j \times \left(\frac{P_j}{P} - 1\right)^2$$

is the imbalance attributable to category j, $W_j = \sum_{s_j} d_k / \sum_s d_k$ is the sample proportion and $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ the response rate in that

category, and P is the overall response rate. During the data collection we can follow the evolution of the contributions C_j of the different groups to the total imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, see LS (2012). An efficient data collection is signaled by a decreasing tendency in the terms C_j.

To use the terminology of partial imbalance, we can call C_j the unconditional partial imbalance for category j of the classification with J categories; see Sections 3.2 and 3.4 below, and the book by BCS, p. 190.

3 Theory development for phase 2 of the project

3.1 More on the concept of balance

We use the term *balance* to refer to the equality of respondent mean and full sample mean, for specified variables.

The important question examined in this section is: "Does balance on the auxiliary variables imply balance also on the study variables"?

In practice we can achieve good balance on the auxiliary variables, because these are available for responding units as well as for nonresponding units. By contrast, balance cannot in practice be ascertained for the study variables, which are observed only for the responding units.

We study the question in this section using *y*-variables that are register variables, hence available for all sample units, responding as well as nonresponding.

The response set *r* is *perfectly balanced on the study variable y* if

$$\bar{y}_r = \bar{y}_s \tag{3.1.1}$$

where

$$\bar{y}_r = \sum_r d_k y_k / \sum_r d_k \;\; ; \;\; \bar{y}_s = \sum_s d_k y_k / \sum_s d_k$$

It is a conceptual definition. Perfect balance on a real study variable, although desirable, is impossible under nonresponse. To say how close we get to the desired balance (3.1.1) is also impossible. If $\bar{y}_r \neq \bar{y}_s$, the response is unbalanced on the variable *y*.

Balance on the *y*-variable has implications for the estimation of the population *y*-total $Y = \sum_U y_k$. If the perfect *y*-variable balance (3.1.1) holds, then the estimator by simple **exp**ansion of the respondent mean,

$$\hat{Y}_{EXP} = (\sum_s d_k) \frac{\sum_r d_k y_k}{\sum_r d_k} = (\sum_s d_k)\, \bar{y}_r \tag{3.1.2}$$

is equal to the Horvitz-Thompson estimator

$$\hat{Y}_{FUL} = \sum_s d_k y_k \qquad (3.1.3)$$

The latter is unbiased under the full response where y_k is available for all $k \in s$. In other words, perfect balance on the *y*-variable would imply that even such a crude estimator as $\hat{Y}_{EXP}$ is without bias. Perfect or near-perfect balance is clearly a desirable (although unattainable) property.

In practice, the difference between (3.1.2) and (3.1.3) can be large. We should try at the data collection stage to obtain a well balanced ultimate response set r. This is impossible for the *y*-variable, but achieving good balance with respect to the auxiliary variables is possible.

The empirical studies in this report are of two types, depending on the nature of the *y*-variable in (3.1.2), (3.1.3) and in other estimators. We distinguish two situations: FULL *y*-data and INCOMPLETE *y*-data.

Case FULL: A register variable is used as a *y*-variable; the values y_k are available for $k \in s$. The register variable is a "pseudo *y*-variable." Although hypothetical, this case is important for methodological studies. It permits us to see how estimators computed under nonresponse behave relative to those that would be computed for the same *y*-variable under full response, where $\hat{Y}_{FUL}$ given in (3.1.3) can be computed and compared with the estimates made under nonresponse.

Case INCOMPLETE: The values y_k are available for the responding units $k \in r$ but missing for $k \in s - r$. The *y*-data are incomplete in that y_k is not available for all units in the selected sample s. It is the nonresponse situation that we are faced with in practice. The unbiased $\hat{Y}_{FUL}$ in (3.1.3) cannot be computed, but $\hat{Y}_{EXP}$ and other estimators computed on y_k for $k \in r$ are possible.

The auxiliary vector value $\mathbf{x}_k$ is available at least for $k \in s$. As pointed out in Section 2.3, if $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, the response set is *perfectly balanced* with respect to the auxiliary vector $\mathbf{x}_k$.

Why seek balance on the chosen **x**-vector? For both FULL and INCOMPLETE, we can strive for balance on the auxiliary vector **x** because available for $k \in s$. But balance on the *y*-variable can be

assessed only for FULL, because y_k is missing for $k \in s-r$. Now if *y* is well explained by the **x**-vector, then $y_k \approx \boldsymbol{\beta}'\mathbf{x}_k$ for $k \in s$ and some unknown vector $\boldsymbol{\beta}$. If the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ holds, then $(\sum_s d_k)\,\bar{\mathbf{x}}_r = \sum_s d_k \mathbf{x}_k$. We pre-multiply this equation by $\boldsymbol{\beta}'$ to conclude that $\hat{Y}_{EXP} \approx \hat{Y}_{FUL}$, where $\hat{Y}_{FUL}$ is the unbiased HT estimator (3.1.3) and $\hat{Y}_{EXP}$ is the primitive expansion estimator (3.1.2).

In words, if the response is balanced for a vector $\mathbf{x}_k$ highly related to the study variable *y*, then even the primitive expansion estimator is close to unbiased. Nonresponse bias would cease to be a problem. There is clear incentive to strive for balance on the **x**-vector.

We write the deviation $\bar{y}_r - \bar{y}_s$ as a two-term sum:

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s \qquad (3.1.4)$$

where $\mathbf{b}_r$ (for the response set) and $\mathbf{b}_s$ (for the whole probability sample) are the regression coefficient vectors derived by least squares fit of y_k on $\mathbf{x}_k$;

$$\mathbf{b}_r = (\sum_s d_k I_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k I_k \mathbf{x}_k y_k) \;\; ;$$

$$\mathbf{b}_s = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k \mathbf{x}_k y_k) \qquad (3.1.5)$$

where I is the response indicator with value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s-r$. $\mathbf{b}_r$ can be computed under INCOMPLETE, but $\mathbf{b}_s$ requires FULL. To verify equation (3.1.4) we need to note that $\mathbf{b}_r'\bar{\mathbf{x}}_r = \bar{y}_r$ and $\mathbf{b}_s'\bar{\mathbf{x}}_s = \bar{y}_s$. This follows from the form of the $\mathbf{x}$-vector specified in Section 2: There is a constant vector $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all units *k*.

The first term on the right hand side of (3.1.4), $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r$ can be computed in both cases, since $\bar{\mathbf{x}}_s$ is known. The d_k-weighted regression residuals $y_k - \mathbf{x}_k'\mathbf{b}_r$ sum to zero over *r* but generally not so over the whole sample *s*: We have $\sum_s d_k (y_k - \mathbf{x}_k'\mathbf{b}_r) = -(\sum_s d_k \mathbf{x}_k)'(\mathbf{b}_r - \mathbf{b}_s)$, which, with the opposite sign and divided by $\sum_s d_k$, equals the second term on the right

hand side of (3.1.4). This term cannot be computed under INCOMPLETE.

Equation (3.1.4) highlights two deviations, both undesirable:

(i) A usually nonzero difference $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ indicating imbalance with respect to the chosen $\mathbf{x}$-vector;

(ii) A usually non-zero difference $\mathbf{b}_r - \mathbf{b}_s$ indicating different regressions for respondents and for the whole sample.

A non-zero difference $\mathbf{b}_r - \mathbf{b}_s$ results from the problem referred to in the literature as selection bias in regression analysis. There is a large literature on this problem caused by a non-random selection of the units in the analysis. In our case, a regression model that may apply for the whole sample of units does not hold for the response set, which is not a random subset. As a result, the regression computed on those who happen to respond is systematically in error. Some early references are Heckman (1979) and Dubin and Rivers (1989).

From (3.1.4) we get an alternative representation of the deviation $\bar{y}_r - \bar{y}_s$:

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_s + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s + (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s)$$

Here the final term $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s)$ is interpretable as an interaction between the two vector differences $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ and $\mathbf{b}_r - \mathbf{b}_s$

It is desirable to build the response set *r* by a data collection such that in the end the deviation $\bar{y}_r - \bar{y}_s$ is small or zero. If realized, such a response set would be a perfect substitute for the unrealized full sample, if we disregard the comparatively speaking minor problem that r has fewer observations than *s*. Essential is that there is no bias. The increase in variance caused by a loss of observations can always be countered by drawing a larger sample in the first place.

In summary, equation (3.1.4) says that the response set r must satisfy two conditions to realize the *y*-variable balance $\bar{y}_r - \bar{y}_s = 0$:

(1) $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ (balance on the chosen **x**-vector)

(2) $\mathbf{b}_r = \mathbf{b}_s$ (consistent regression)

In the data collection, we can attempt to come close to realizing (1). But this does not imply that condition (2), $\mathbf{b}_r = \mathbf{b}_s$, is close to being fulfilled. The term $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ in (3.1.4) is not necessarily small.

Under nonresponse, a better alternative than $\hat{Y}_{EXP}$ is the calibration estimator

$$\hat{Y}_{CAL} = \sum_r d_k m_k y_k \tag{3.1.6}$$

with weights $d_k m_k = d_k (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ satisfying the calibration equation $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. It is computable both for INCOMPLETE and FULL. The calibrated weighting can make $\hat{Y}_{CAL}$ considerably less biased than $\hat{Y}_{EXP}$. The latter is the special case of $\hat{Y}_{CAL}$ for the uninformative vector $\mathbf{x}_k = 1$.

Equation (3.1.4) when multiplied by $\sum_s d_k$ expresses a relationship between the estimators (3.1.2), (3.1.3) and (3.1.6):

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = (\hat{Y}_{EXP} - \hat{Y}_{CAL}) + (\hat{Y}_{CAL} - \hat{Y}_{FUL}) \tag{3.1.7}$$

Here $\hat{Y}_{EXP} - \hat{Y}_{CAL}$ is a computable adjustment of the crude basic estimator $\hat{Y}_{EXP}$, and the other two terms are deviations from the unbiased estimate $\hat{Y}_{FUL}$. Dividing through by $\hat{Y}_{FUL}$, we get (3.1.7) expressed in terms of relative differences (in per cent),

$$RDF(\hat{Y}_{EXP}) = RADJ(\hat{Y}_{CAL}) + RDF(\hat{Y}_{CAL}) \tag{3.1.8}$$

where

$$RDF(\hat{Y}_{EXP}) = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}$$

$$RDF(\hat{Y}_{CAL}) = 100 \times (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{Y}_{FUL})$$

$$RADJ(\hat{Y}_{CAL}) = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{CAL}) / \hat{Y}_{FUL}$$

In words, (3.1.8) says:

Total relative deviation = Adjusted relative deviation + Residual relative deviation

In our empirical studies for case FULL, we compute the three components in (3.1.8). They can be positive or negative; those on the right hand side, although normally of same sign, can, rarely, be of opposite signs. (It depends on the choice of **x**-vector.) The sizes of the terms in (3.1.8) are of interest because they show how much of the relative difference $RDF(\hat{Y}_{EXP})$ has become "adjusted away" by the calibration (the term $RADJ(\hat{Y}_{CAL})$), and how much remains unadjusted (the term $RDF(\hat{Y}_{CAL})$).

In the case INCOMPLETE, only one term in (3.1.7) is computable, $\hat{Y}_{EXP} - \hat{Y}_{CAL}$.

In LS (2012) we compared four varieties of the LCS 2009 data: The data collection as it was really carried out (called Actual), and three experimental data collection strategies created in retrospect to produce a gradually decreasing imbalance and a gradually decreasing overall response rate, due to more and more stringent stopping rules. Tables 3.1.1 and 3.1.2 summarize results in Tables 6.1 to 6.7 in LS (2012). Table 3.1.1 shows the imbalance, $IMB = \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ $= (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$, the balance $BI_1 = 1 - \sqrt{IMB/(1/P - 1)}$ and the distance $distr_{r|nr} = \frac{1}{1-P}\sqrt{IMB}$. They are computed on the data as at the end of each of the four data collections, with the **x**-vector defined by the crossing of three dichotomous auxiliary variables: *Education level* (high, not high), *Property ownership* (owner, non-owner), *Country of origin* (Sweden, other). This gives eight mutually exclusive and exhaustive groups of units. The **x**-vector has dimension $J = 2^3 = 8$ and is of the form $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{jk},...,\gamma_{8k})'$, where $\gamma_{jk} = 1$ if *k* belongs to group *j* and $\gamma_{jk} = 0$ otherwise. Called experimental vector in LS (2012), we write it as

$$\mathbf{x} = (Educ \times Owner \times Origin) \qquad (3.1.9)$$

Table 3.1.1 shows a striking improvement in regard to both balance (it increases) and distance (it decreases) over the series of four data collections. It confirms empirically an improvement that was predictable, given the way in which the experimental strategies were set up. The entries depend only on the auxiliary vector, not on any *y*-variable.

Table 3.1.1
Comparison of four strategies examined in LS (2012), in regard to imbalance, balance, and distance. Computations based on the data at the end of each of the four data collections, with the x-vector (3.1.9)

	$100\times IMB$	BI_1	$dist_{r\|nr}$	$100\times P$
Actual	2.36	0.779	0.471	67.4
Strategy 1	1.39	0.843	0.326	63.9
Strategy 2	0.82	0.892	0.220	58.9
Strategy 3	0.20	0.955	0.089	50.3

Table 3.1.2
Comparison of four data sets examined in LS (2012), in regard to the terms of equation (3.1.8). Computations based on the data at the end of each of the four data collections, with the x-vector (3.1.9)

Pseudo y-variables and data set	$RDF(\hat{Y}_{EXP})$	$RDF(\hat{Y}_{CAL})$	$RADJ(\hat{Y}_{CAL})$
Benefits			
Actual	-9.41	-7.90	-1.51
Strategy 1	-7.52	-6.73	-0.79
Strategy 2	-7.61	-7.06	-0.55
Strategy 3	-6.47	-6.33	-0.14
Income			
Actual	6.75	2.91	3.84
Strategy 1	4.63	2.49	2.14
Strategy 2	3.06	2.04	1.03
Strategy 3	0.98	0.32	0.66
Employed			
Actual	4.76	3.12	1.64
Strategy 1	3.36	2.53	0.82
Strategy 2	2.17	1.77	0.40
Strategy 3	0.03	-0.22	0.25

Table 3.1.1 raises questions: Does the favorable trend in balance and distance also translate into favorable features for the estimates of the *y*-variables? To examine the question we use three pseudo *y*-variables (the case FULL), namely, the register variables Benefits, Income, and Employed. Not surprisingly they behave differently with respect to $RDF(\hat{Y}_{EXP})$ and $RDF(\hat{Y}_{CAL})$, but important to note is the persistent trend of improvement in both of them over the series of four data collections.

Note that $RDF(\hat{Y}_{EXP}) = (\bar{y}_r - \bar{y}_s)/\bar{y}_s$ expresses, apart from the constant $1/\bar{y}_s$, the imbalance $\bar{y}_r - \bar{y}_s$ for the study variable *y*. It drops substantially over the series of four data collections; in a very striking manner for Income and Employed.

Here $RDF(\hat{Y}_{CAL})$ measures the bias remaining (despite the balancing) in the nonresponse adjusted estimator $\hat{Y}_{CAL}$. The series of four data collections brings a significant bias reduction for all three *y*-variables.

By (3.1.8), the adjustment $RADJ(\hat{Y}_{CAL})$ is zero for the perfect x-vector balance, $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, something which is not far from being realized by Strategy 3. Therefore is comes as no surprise that $RDF(\hat{Y}_{CAL})$ and $RDF(\hat{Y}_{EXP})$ differ by little only for Strategy 3.

We recall that the comparisons in Tables 3.1.1 and 3.1.2 depend on the choice of x-vector. Therefore we recomputed the two tables with a new x-vector, but with the same four data sets, Actual and Strategies 1, 2, and 3, in order to illustrate what the consequences might be. This is done in Tables 3.1.3 and 3.1.4, with the "standard **x**-vector" explained in LS (2012). It contains more variables than (3.1.9).

The important message is that the favorable trends over the series of four data sets – increasing balance, decreasing distance, decreasing $RDF(\hat{Y}_{CAL})$ and $RDF(\hat{Y}_{EXP})$ – continue to apply in Tables 3.1.3 and 3.1.4.

The distance is greater, and balance is lower in Table 3.1.3 than in Tables 3.1.1, naturally, because the **x**-vector has more variables, making balance more difficult to achieve. As a result of the more extensive x-vector, $RDF(\hat{Y}_{CAL})$ for Benefits is much lower (and the adjustment $RADJ(\hat{Y}_{CAL})$ much higher) in Table 3.1.4 than in Table 3.1.2. For Income and Employed the differences are less striking. The effect of the auxiliary vector on the estimates is not always predictable.

Table 3.1.3
Comparison of four strategies examined in LS(2012), in regard to imbalance, balance, and distance. Computations based on the data at the end of each of the four data collections, with the "standard x-vector"

	$100\times IMB$	BI_1	$dist_{r\|nr}$	$100\times P$
Actual	3.86	0.717	0.603	67.4
Strategy 1	3.12	0.765	0.489	63.9
Strategy 2	3.17	0.787	0.433	58.9
Strategy 3	3.62	0.808	0.383	50.3

Table 3.1.4
Comparison of four strategies examined in LS(2012), in regard to the terms of equation (3.1.8). Computations based on the data at the end of each of the four data collections. The x-vector is the standard x-vector

Pseudo y-variables and Strategy	$RDF(\hat{Y}_{EXP})$	$RDF(\hat{Y}_{CAL})$	$RADJ(\hat{Y}_{CAL})$
Benefits			
Actual	-9.41	-3.58	-5.83
Strategy 1	-7.52	-1.56	-5.96
Strategy 2	-7.61	-1.21	-6.40
Strategy 3	-6.47	0.98	-7.49
Income			
Actual	6.75	2.90	3.84
Strategy 1	4.63	2.74	1.89
Strategy 2	3.06	2.57	0.49
Strategy 3	0.98	1.00	-0.01
Employed			
Actual	4.76	3.06	1.70
Strategy 1	3.36	3.01	0.34
Strategy 2	2.17	3.15	-0.98
Strategy 3	0.03	2.34	-2.31

3.2 Two kinds of auxiliary variables

To study the concepts conditional and unconditional imbalance we make an important distinction in regard to the auxiliary variables. They can be of two kinds, with different functions. Some are designated for monitoring the data collection, with an objective to obtain a final response set that is well balanced with respect to precisely those variables. They compose a vector denoted $\mathbf{x}_a$, of dimension J. Other available auxiliary variables stay neutral

during the data collection, but become important at the estimation stage. The vector of these variables is denoted $\mathbf{x}_b$, of dimension L. The variables in $\mathbf{x}_b$ are used, usually together with those in $\mathbf{x}_a$, to compute the calibrated weights for estimating the population total $Y = \sum_U y_k$. We call $\mathbf{x}_a$ the *monitoring (auxiliary) vector* and $\mathbf{x}_b$ the *supplement (auxiliary) vector*. Their values for unit k are denoted $\mathbf{x}_{ak}$ and $\mathbf{x}_{bk}$, respectively. Both can contain categorical as well as continuous variables; in essentially all our applications, they are categorical.

The distinction between monitoring vector and supplement vector helps to contrast the two stages of activity, the data collection stage and the estimation stage. The monitoring vector is an important instrument for the data collection; the supplement vector variables can be important at the estimation stage. Both categories of variables may be used in computing the calibrated weights for the estimation phase, with an objective to control bias and variance in the estimates.

For practical reasons, the monitoring vector $\mathbf{x}_a$ usually consists of a fairly limited number of selected *x*-variables. An important case is when the monitoring vector identifies a set of mutually exclusive and exhaustive sample subgroups. It is relatively easy for the survey manager to monitor the data collection with respect to a modest number of groups. How do we select these groups? They should be groups for which we expect considerable differences in response rate, because such differences are a sign of a large imbalance needing to be reduced.

In a regularly repeated survey, we usually know beforehand which those groups are. It is less obvious in a survey carried out for the first time. In that case, one possibility is to analyze the response at one or more points before the very end of the data collection, with an objective to identify groups with very different response rate. A tool for this is classification tree analysis, CHAID (see Kass 1980), which can for example be carried out at the end of the ordinary data collection and serve to identify the groups that should be pursued with particular emphasis in the follow-up. Tree analysis is described in BCS, p. 263-265, and is used in Section 6 for an application to the PIAAC survey. Once the groups have been identified, possibilities open up for the continuation of the data collection. For example, different groups can receive different number of call attempts.

3.3 Analysis of imbalance

The *analysis of imbalance* (ANIMB) that we now describe focuses on the imbalance statistic $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$, as defined in Section 2. We compute $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ for different specifications of the **x**-vector. Differences between computed values of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ become components in the ANIMB. The procedure is similar in some respects to a traditional analysis of variance. The concepts *conditional imbalance* and *partial conditional imbalance* arise naturally through the ANIMB. Although different both in derivation and in numeric aspects, they have some resemblance to *conditional representativity* and *conditional partial representativity*, as used by BCS and in the RISQ project.

We define the imbalance associated with all the auxiliary variables to be the value of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ when computed on the x-vector that combines the monitoring vector $\mathbf{x}_a$ and the supplement vector $\mathbf{x}_b$, so that $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{ak} \\ \mathbf{x}_{bk} \end{pmatrix}$, of dimension $J + L$. We denote that value of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ by IMB_{ab}. It is the *total imbalance*. The subscript ab indicates that all the auxiliary variables, both the *a* type and the *b* type, are included in the $\mathbf{x}$-vector. (It may happen that the $\mathbf{x}$-vector must be specified with dimension $J + L - 1$, namely if one category needs to be eliminated to make the matrix $\boldsymbol{\Sigma}_s$ invertible. It happens for example if both $\mathbf{x}_a$ and $\mathbf{x}_b$ are classification vectors, as in the analysis in Section 4.1.)

We are particularly interested in the impact on imbalance of the entire monitoring vector $\mathbf{x}_a$, as well as of the individual variables in $\mathbf{x}_a$. To this end we first compute $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ on $\mathbf{x}_k = \mathbf{x}_{bk}$ alone. In other words, the vector $\mathbf{x}_k$ used to compute $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is what remains of the complete vector $\begin{pmatrix} \mathbf{x}_{ak} \\ \mathbf{x}_{bk} \end{pmatrix}$ after excluding $\mathbf{x}_{ak}$. That value of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is denoted IMB_b, called the *marginal imbalance* attributed to $\mathbf{x}_b$, with its L variables.

The positive quantity $IMB_{ab} - IMB_b$ measures the increment in imbalance produced by adding $\mathbf{x}_a$ to the $\mathbf{x}$-vector defined by $\mathbf{x}_b$.

Hence $IMB_{ab} - IMB_b$ is the *conditional imbalance* due to $\mathbf{x}_a$, controlling for $\mathbf{x}_b$. The associated number of variables is $(J + L) - L = J$, which is the dimension of $\mathbf{x}_a$.

We also wish to assess the impact of each variable in $\mathbf{x}_{ak}$, which has the form $\mathbf{x}_{ak} = (x_{a1k},...,x_{ajk},...,x_{aJk})'$, where x_{ajk} is the value for the sampled unit *k* of the j^{th} (unidimensional) monitoring variable x_{aj}, $j = 1,...,J$.

To evaluate the impact of the single variable x_{aj}, we compute first $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ on the $\mathbf{x}$-vector composed of $\mathbf{x}_b$ and x_{aj}, $\mathbf{x}_k = \begin{pmatrix} x_{ajk} \\ \mathbf{x}_{bk} \end{pmatrix}$, containing now $1 + L$ variables. This new value of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ measures the increment in imbalance produced by adding the single variable x_{aj} to the $\mathbf{x}$-vector defined by $\mathbf{x}_b$. Hence $IMB_{a(j)b} - IMB_b$ is the *conditional partial imbalance* due to x_{aj}, the associated number of variables being $(1 + L) - L = 1$. The word "partial" means that x_{aj} is seen as a part of $\mathbf{x}_a$, conditionally on $\mathbf{x}_b$.

The sum of the imbalance increments obtained by taking the variables x_{aj} one at a time is $\sum_{j=1}^{J}(IMB_{a(j)b} - IMB_b)$. The increment by taking simultaneously all J variables in $\mathbf{x}_a$ (rather than one at a time) is $IMB_{ab} - IMB_b$. The difference, which can be of either sign, is

$$diff = (IMB_{ab} - IMB_b) - \sum_{j=1}^{J}(IMB_{a(j)b} - IMB_b)$$

$$= IMB_{ab} + (J-1) \times IMB_b - \sum_{j=1}^{J} IMB_{a(j)b}$$

The *diff* is often negative and small, it is an *interaction* component. The ANIMB is now complete and is summarized in Table 3.3.1.

Table 3.3.1
Analysis of imbalance (ANIMB) with monitoring vector $\mathbf{x}_a$ and supplement vector $\mathbf{x}_b$. #var is the number of variables associated with the component

Component of imbalance	Expression for computation	#var
Total (due jointly to $\mathbf{x}_a$ and $\mathbf{x}_b$)	IMB_{ab}	$J+L$
Marginal due to $\mathbf{x}_b$	IMB_b	L
Conditional due to $\mathbf{x}_a$ given $\mathbf{x}_b$	$IMB_{ab} - IMB_b$	J
Conditional partial due to x_{a1}	$IMB_{a(1)b} - IMB_b$	1
⋮	⋮	
Conditional partial due to x_{aj}	$IMB_{a(j)b} - IMB_b$	1
⋮	⋮	
Conditional partial due to x_{aJ}	$IMB_{a(J)b} - IMB_b$	1
Interaction	*diff*	---

An important special case is that of a categorical monitoring vector. We examine the form of the ANIMB components in Table 3.3.1 in this special case. Suppose $\mathbf{x}_a$ is categorical with J categories, $j = 1,...,J$. Further, let the supplement vector $\mathbf{x}_b$ be the simplest possible, with $L = 1$ and $\mathbf{x}_{bk} = 1$ for all *k*. It does not in any way distinguish the sample units, and $IMB_b = 0$. Then $IMB_{ab} - IMB_b = IMB_a$, where

$$IMB_a = \sum_{j=1}^{J} C_j \tag{3.3.1}$$

with $C_j = W_j\left(\frac{P_j}{P} - 1\right)^2$, where W_j is group *j*'s portion of the whole sample *s*, $W_j = \sum_{s_j} d_k / \sum_s d_k$, $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ is the response rate in group j, and P is the overall response rate, $P = \sum_r d_k / \sum_s d_k$.

The conditional partial imbalance due to the j^{th} category is $IMB_{a(j)b} - IMB_{ab} = IMB_{a(j)}$, where

$$IMB_{a(j)} = W_j(\frac{P_j}{P} - 1)^2 + W_{j_c}(\frac{P_{j_c}}{P} - 1)^2$$

where $W_j + W_{j_c} = 1$, $W_j P_j + W_{j_c} P_{j_c} = P$, and j_c denotes the complement category "not-j". (Hence j_c represents the union of all $J-1$ categories in the $\mathbf{x}_a$ classification except j itself.) This gives $W_{j_c}(\frac{P_{j_c}}{P} - 1)^2 = \frac{W_j^2}{1-W_j}(\frac{P_j}{P} - 1)^2$. Hence

$$IMB_{a(j)} = W_j(\frac{P_j}{P} - 1)^2 + \frac{W_j^2}{1-W_j}(\frac{P_j}{P} - 1)^2 = \frac{W_j}{1-W_j}(\frac{P_j}{P} - 1)^2$$

For small W_j, $IMB_{a(j)}$ close to (but not equal to) the *j*:th component $C_j = W_j(\frac{P_j}{P} - 1)^2$ in (3.3.1). The terms are not strictly additive. The (non-zero) interaction component is

$$diff = IMB_a - \sum_{j=1}^{J} IMB_{a(j)} =$$

$$\sum_{j=1}^{J} W_j(\frac{P_j}{P} - 1)^2 - \sum_{j=1}^{J} \frac{W_j}{1-W_j}(\frac{P_j}{P} - 1)^2 = -\sum_{j=1}^{J} \frac{W_j^2}{1-W_j}(\frac{P_j}{P} - 1)^2$$

Often *diff* is a small portion of the total IMB_a, especially when J is fairly large. If the J categories are of equal size, $W_j = 1/J$ for all j, then

$$\frac{diff}{IMB_{ab}} = -\frac{1}{J-1}$$

3.4 Imbalance analysis derived by propensity scores

The ANIMB components IMB_{ab}, IMB_b and $IMB_{a|b} = IMB_{ab} - IMB_b$ in Table 3.3.1 can be derived alternatively by computing first response propensity scores for the sample units. Then the ANIMB

components are obtained as coefficients of variation of the propensity scores.

This approach is important for guiding the data collection, and for the interventions or changes of emphasis that are called for. At suitable points in the data collection, we may wish to reduce the number of contact attempts for selected sample units, or to stop the attempts altogether for some units. The response propensity scores help identify the units that should be targeted in these interventions. The procedure to be presented is applicable for any monitoring vector $\mathbf{x}_a$ and any supplement vector $\mathbf{x}_b$.

As before, s denotes a probability sample from $U = \{1,...,k...,N\}$. The inclusion probability of unit k is π_k; the design weight is $d_k = 1/\pi_k$. The response set is r (the set of units for which the study variable value y_k has been recorded); $r \subseteq s \subseteq U$. The response indicator is $I_k = 1$ if $k \in r$; $I_k = 0$ if $k \in s - r$. The response rate is

$$P = \sum_s d_k I_k / \sum_s d_k = \sum_r d_k / \sum_s d_k$$

We view the data collection as a function of the number of call attempts made to the units in the selected sample *s*. In this process, the response set r is expanding, that is, larger (or the same) after *m* attempts than after *m* - 1 attempts. As explained earlier, the available auxiliary variables are of two kinds. Some are used to monitor the data collection, so as to obtain a well balanced final response set. These auxiliary variables form the vector $\mathbf{x}_a$ of dimension J. The other auxiliary variables are important at the estimation stage for computing calibrated weights. The vector of these latter variables is $\mathbf{x}_b$, of dimension L. We call $\mathbf{x}_a$ the "monitoring (auxiliary) vector" and $\mathbf{x}_b$ the "supplement (auxiliary) vector". Their respective values for unit k are $\mathbf{x}_{ak}$ and $\mathbf{x}_{bk}$. Both are assumed to be of the form mentioned in Section 2.2: It is possible to state a constant vector $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}'\mathbf{x}_{ak} = 1$ or $\boldsymbol{\mu}'\mathbf{x}_{bk} = 1$ for all *k*.

Suppose we are at a certain point in the data collection where the realized response is *r* and the realized response rate is $P = \sum_r d_k / \sum_s d_k$. By the response propensity of unit *k* at that point we mean the estimate of the response probability of *k*, estimated by least squares, taking into account both $\mathbf{x}_a$ and $\mathbf{x}_b$. The regression fit is done in two steps: Regress first I_k on $\mathbf{x}_{ak}$ to obtain predicted

values $\hat{I}_{ak}$. Then regress the residuals $I_k - \hat{I}_{ak}$ on $\mathbf{x}_{bk}$ to obtain the predicted differences $\hat{I}_{bk} - \hat{I}_{a:b,k}$ Combining the two steps, the prediction of I_k, computable for every unit $k \in s$, is the *response propensity score*

$$\hat{P}_k = \hat{I}_{bk} + \hat{I}_{ak} - \hat{I}_{a:b,k} \tag{3.4.1}$$

where

$\hat{I}_{ak} = (\sum_s d_k I_k \mathbf{x}_{ak})'(\sum_s d_k \mathbf{x}_{ak}\mathbf{x}'_{ak})^{-1}\mathbf{x}_{ak}$ (regression of I_k on $\mathbf{x}_{ak}$)

$\hat{I}_{bk} = (\sum_s d_k I_k \mathbf{x}_{bk})'(\sum_s d_k \mathbf{x}_{bk}\mathbf{x}'_{bk})^{-1}\mathbf{x}_{bk}$ (regression of I_k on $\mathbf{x}_{bk}$)

$\hat{I}_{a:b,k} = (\sum_s d_k \hat{I}_{ak} \mathbf{x}_{bk})'(\sum_s d_k \mathbf{x}_{bk}\mathbf{x}'_{bk})^{-1}\mathbf{x}_{bk}$ (regression of $\hat{I}_{ak}$ on $\mathbf{x}_{bk}$)

We have

$$\sum_s d_k \hat{P}_k = \sum_s d_k \hat{I}_{ak} = \sum_s d_k \hat{I}_{bk} = \sum_s d_k \hat{I}_{a:b,k} = (\sum_s d_k)\, P = \sum_r d_k$$

This shows that at any point in the data collection, all of $\hat{P}_k$, $\hat{I}_{bk}$, $\hat{I}_{ak}$ and $\hat{I}_{a:b,k}$ have (design weighted) mean equal to the overall response rate $P = \sum_r d_k / \sum_s d_k$ realized at that point.

The sum of squares $\sum_s d_k (\hat{P}_k - P)^2$ is decomposed into orthogonal non-negative components as follows:

$$\sum_s d_k (\hat{P}_k - P)^2 = \sum_s d_k (\hat{I}_{bk} - P)^2 + \sum_s d_k (\hat{I}_{ak} - \hat{I}_{a:b,k})^2 \tag{3.4.2}$$

This follows from $\sum_s d_k (\hat{I}_{ak} - \hat{I}_{a:b,k})\hat{I}_{bk} = 0$. Because $\hat{I}_{ak} - \hat{I}_{a:b,k} = \hat{P}_k - \hat{I}_{bk}$ by (3.4.1) we can also write (3.4.2) as

$$\sum_s d_k (\hat{P}_k - P)^2 = \sum_s d_k (\hat{I}_{bk} - P)^2 + \sum_s d_k (\hat{P}_k - \hat{I}_{bk})^2 \tag{3.4.3}$$

We divide through in (3.4.3) by $P^2\hat{N}$, where $\hat{N} = \sum_s d_k$ and write the resulting equation as

$$IMB_{ab} = IMB_b + IMB_{a|b} \tag{3.4.4}$$

where

$$IMB_{ab} = \frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{P}_k}{P}-1)^2 \; ; \; IMB_b = \frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{I}_{bk}}{P}-1)^2 \; ;$$

$$IMB_{a|b} = \sum_s d_k(\hat{P}_k - \hat{I}_{bk})^2 \tag{3.4.5}$$

These terms are equivalent to their counterparts in Table 3.3.1. The second term on the right hand side of (3.4.2) can be further developed to obtain

$$0 \le \sum_s d_k(\hat{I}_{ak} - \hat{I}_{a:b,k})^2 = \sum_s d_k(\hat{I}_{ak} - P)^2 - \sum_s d_k(\hat{I}_{a:b,k} - P)^2$$

Combining this equation with (3.4.2) and dividing by $P^2\hat{N}$, we get

$$\frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{P}_k}{P}-1)^2 =$$

$$\frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{I}_{bk}}{P}-1)^2 + \frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{I}_{ak}}{P}-1)^2 - \frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{I}_{a:b,k}}{P}-1)^2$$

The terms of this equation are the squared coefficients of variation of $\hat{P}_k$, $\hat{I}_{bk}$, $\hat{I}_{ak}$ and $\hat{I}_{a:b,k}$.

A special case is an *unconditional* analysis. In the expressions above, $\mathbf{x}_b$ is then the simplest possible (uninformative) vector, $\mathbf{x}_{bk} = 1$ for all *k*. Then $\hat{I}_{bk} = \hat{I}_{a:b,k} = P$ for all *k*. Equation (3.4.1) becomes $\hat{P}_k = \hat{I}_{ak}$, and (3.4.4) becomes $IMB_{ab} = IMB_a$, where

$$IMB_a = \frac{1}{\hat{N}}\sum_s d_k(\frac{\hat{I}_{ak}}{P}-1)^2$$

If in particular the $\mathbf{x}_a$-vector is categorical with J categories, then $\hat{I}_{ak}$ takes only J different values, $\hat{I}_{ak} = P_j$ for all k in group j, where $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ is the group j response rate. Then

$$IMB_a = \sum_{j=1}^{J} W_{js} (\frac{P_j}{P} - 1)^2$$

where $W_{js} = \sum_{s_j} d_k / \sum_s d_k$ is the sample share of group j. This formula played an important role in LS(2012), where it was used to generate several "experimental strategies", in which that data collection is stopped for a group j if its response rate P_j is high, compared with other groups.

3.5 Building a final response set

We use the results in Section 3.4 about the response propensity scores $\hat{P}_k$ to construct a desirable final response set. We compute the $\hat{P}_k$ for all $k \in s$ at several points during the data collection. They change continuously; they are *itinerant response propensities*.

Suppose that we have fixed a suitable set of intervention points. At each intervention point we stop call attempts for not yet responding units whose response propensity score $\hat{P}_k$ at that point is (comparatively) high. We let the continued data collection focus on the remaining nonrespondents, those with (comparatively) lower response propensity $\hat{P}_k$. The effect is to even out the variability in the $\hat{P}_k$, and, as we show, this reduces the imbalance in the response set.

At each point, the computed scores $\hat{P}_k$ for $k \in s$ are size ordered from smallest to largest. The mean of $\hat{P}_k$ is equal to the response proportion $P = \sum_r d_k / \sum_s d_k$ realized at that point. The units with the highest $\hat{P}_k$ are set aside; they are no longer pursued in the data collection. The justification is that their response propensity is "large enough" compared with other, less responsive units. Contact attempts continue for remaining units, those with lower $\hat{P}_k$. Some more units will respond, and P increases. At subsequent points, the values $\hat{P}_k$ are again computed for all $k \in s$, and so on.

There are variations of the procedure. One possibility is to set aside a fixed percentage of units at each specified intervention point, namely, those with the largest $\hat{P}_k$, as in the following illustration, called "the 25% dropping rule" and used later in the empirical Section 4.2.

Suppose three interventions points have been specified, that is, both their number (three) and their place in the data collection, identified by the call attempt number. The procedure is: At point 1, the set of units with the 25% highest $\hat{P}_k$ are set aside; let that set of *n*/4 units be s_1, where *n* is the size of the sample s. Data collection attempts continue for the nonrespondents in the rest of the sample, $s - s_1$. At point 2, the $\hat{P}_k$ are recomputed for all $k \in s$. The units already set aside have their $\hat{P}_k$ somewhat changed, but without consequence; they remain aside. At point 2, the set s_2 consisting of the $n/4$ sample units with the highest $\hat{P}_k$ in $s - s_1$ are set aside. The remaining set $s - s_1 - s_2$ has $n - (2 \times n/4) = n/2$ units. At point 3, the $\hat{P}_k$ are recomputed again for $k \in s$; those with the highest $\hat{P}_k$ in $s - s_1 - s_2$ are set aside; they form the set s_3 of size $n/4$. The remaining quarter of the sample continues until the end of the data collection period, resulting in a final response set r, and we compute the final values $\hat{P}_k$ for $k \in s$.

The number of sample units set aside is the same (25 % of the sample size *n*) at each of the three points, but the number of nonrespondents set aside is not the same, see Section 4.2.

More generally, if *H* is the specified number of sample subgroups, the procedure produces *H* sample subgroups denoted $s_1, s_2, ..., s_H$. For the finally computed $\hat{P}_k$ for $k \in s$, we compute the contribution of each group:

$$A_h = \frac{1}{\sum_s d_k} \sum_{s_h} d_k \left(\frac{\hat{P}_k}{P} - 1\right)^2 , \quad h = 1, 2, ..., H$$

Then $\sum_{h=1}^{H} A_h = IMB_{ab}$. That is, A_h measures the contribution of size group h to the total ultimate imbalance IMB_{ab}.

Alternatively, logistic regression fit can be used to obtain the scores $\hat{P}_k$, but there is no particular advantage to doing so. Then $\hat{P}_k$ has the form $\hat{P}_k = \exp(\mathbf{x}_a'\hat{\boldsymbol{\beta}})/[1+\exp(\mathbf{x}_a'\hat{\boldsymbol{\beta}})]$. The equations (3.4.1) to (3.4.3) would be modified accordingly.

4 New theory applied to the LCS 2009

4.1 Analysis of imbalance for the LCS 2009 response set

In this section we apply the techniques developed in Section 3 to the LCS 2009 data. Table 4.1.1 shows results of an ANIMB for the LCS 2009 data collection as actually carried out. It illustrates the kind of analysis possible by identifying a monitoring vector $\mathbf{x}_a$ and a supplementary vector $\mathbf{x}_b$, with values $\mathbf{x}_{ak}$ and $\mathbf{x}_{bk}$ known for all $k \in s$. In a survey where several auxiliary variables are available, the statistician must decide which of them to use in the monitoring vector $\mathbf{x}_a$ and which should be reserved for use at the estimation stage. This is a matter of professional judgment. As pointed out in Section 3.2, $\mathbf{x}_a$ may have to be limited to a rather small number of variables, in order to keep the monitoring task simple.

We choose here to define $\mathbf{x}_a$ by the crossing of three dichotomous auxiliary variables: *Educ* (which stands for Education level: high, not high), *Owner* (which stands for Property ownership: owner, non-owner), and *Origin (*which stands for Country of origin: Sweden, other). This gives eight mutually exclusive and exhaustive groups. We denote the vector as

$$\mathbf{x}_a = (Educ \times Owner \times Origin) \tag{4.1.1}$$

It has dimension $J = 2^3 = 8$ and value $\mathbf{x}_{ak} = (\gamma_{1k},...,\gamma_{8k})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ otherwise. This vector was used in LS (2012) and called experimental vector. Our vector $\mathbf{x}_b$ of dimension seven is composed of the variables *Phone* (for Phone access, equaling 1 for a person with phone number accessible at the start of the data collection; 0 otherwise), *Age* (for Age group; four zero/one coded age brackets: -24, 25-64, 65-74, 75+); *Civil* (for Civil status, equaling 1 if married or widowed; 0 otherwise) and *Gender* (equal to 1 if man, 0 otherwise). We denote this vector as

$$\mathbf{x}_b = (Phone + Age + Civil + Gender) \tag{4.1.2}$$

To place these vectors in a practical perspective, we pretend that $\mathbf{x}_a$ and $\mathbf{x}_b$ in (4.1.1) and (4.1.2) have been carefully chosen by the statistician, reflecting a desire to monitor the data collection with the aid of precisely the eight groups defined by $\mathbf{x}_a$, and that it is nevertheless important at the estimation stage to also benefit from the vector $\mathbf{x}_b$, when the calibrated weights are computed. In practice, the choice of $\mathbf{x}_a$ and $\mathbf{x}_b$ is a matter of professional judgment.

Total imbalance is measured here with respect to the vector that combines $\mathbf{x}_a$ and $\mathbf{x}_b$. Conditional imbalance and conditional partial imbalance are computed as specified by the ANIMB laid out in Table 3.3.1 . "Conditional" here means "given $\mathbf{x}_b$." "Partial" refers to the groups in $\mathbf{x}_a$. We measure the conditional imbalance of the monitoring vector $\mathbf{x}_a$ given $\mathbf{x}_b$, but also the conditional partial imbalance for each of the eight groups that make up $\mathbf{x}_a$. The total imbalance is IMB_{ab}, used here to compute the balance $BI_1 = 1 - \sqrt{IMB_{ab}/(1/P - 1)}$ and the distance $dist_{r|nr} = \frac{1}{1-P}\sqrt{IMB_{ab}}$.

Table 4.1.1 shows the ANIMB for the actual LCS 2009 data collection. Of particular interest is the trend in the different quantities (the tendency in the rows of the table) as the data collection progresses. The conditional partial imbalance $IMB_{a(j)|b}$ stays roughly constant from attempt five until the end for all but groups six and eight, for which some reduction occurs, but less than what one would like to see in a more satisfactory data collection. The sum of the eight conditional partial imbalances $IMB_{a(j)|b}$ and the term *diff* equals the conditional imbalance $IMB_{a|b}$, which stays roughly constant, not unexpectedly, because there was no attempt to direct the data collection in a more effective direction; it simply follows the preconceived original plan. Somewhat unexpectedly, the marginal imbalance IMB_b decreases. A clear sign of an unsatisfactory data collection is that the distance $dist_{r|nr}$ between respondents and nonrespondents increases from 0.44 to 0.62. The same unsatisfactory trend is seen in the balance BI_1, which drops from 0.85 to 0.71.

Table 4.1.1
The actual LCS 2009 data collection: Analysis of imbalance; monitoring vector $\mathbf{x}_a$ given by (4.1.1); supplement vector $\mathbf{x}_b$ given by (4.1.2)

Group characteristic			$100 \times IMB_{a(j)\|b}$ Conditional partial balance						
			Ord. field work attempt				Follow-up attempt		
Education	Property ownership	Origin	1	5	12	End	1	4	Final
Not high	Non-owner	Abroad	0.67	0.95	0.98	0.98	1.00	0.93	0.99
Not high	Non-owner	Sweden	0.11	0.20	0.25	0.26	0.21	0.20	0.19
Not high	Owner	Abroad	0.13	0.03	0.01	0.01	0.01	0.01	0.01
Not high	Owner	Sweden	0.32	0.03	0.02	0.02	0.02	0.02	0.02
High	Non-owner	Abroad	0.66	0.20	0.19	0.18	0.17	0.16	0.16
High	Non-owner	Sweden	0.31	0.36	0.28	0.25	0.22	0.20	0.22
High	Owner	Abroad	0.16	0.01	0.02	0.02	0.02	0.01	0.03
High	Owner	Sweden	0.31	0.49	0.48	0.49	0.46	0.39	0.30
		$100 \times diff$	-0.21	-0.07	0.10	0.18	0.21	0.24	0.32
		$100 \times IMB_{a\|b}$	2.5	2.2	2.3	2.4	2.3	2.2	2.2
		$100 \times IMB_b$	12.4	4.9	2.7	2.4	2.4	2.2	1.9
		$100 \times IMB_{ab}$	14.9	7.1	5.0	4.8	4.7	4.3	4.1
		$100 \times P$	12.8	44.3	57.7	60.4	61.4	64.6	67.4
		BI_1	0.85	0.76	0.74	0.73	0.73	0.72	0.71
		$dist_{r\|nr}$	0.44	0.48	0.53	0.55	0.56	0.59	0.62

We want to see if "better" data collections can change the undesirable trend in the distance and the balance in Table 4.1.1. We consider again the experimental strategies with "interventions in retrospect", as explained in LS (2012). Stopping rules are applied to the eight groups defined by $\mathbf{x}_a$ given by (4.1.1). The results are given in Tables 4.1.2, 4.1.3 and 4.1.4.

Experiment 1 (see Table 4.1.2) has two intervention points, attempt 12 of the ordinary data collection (point 1), and attempt 2 of the follow-up (point 2). The stopping rule says that data collection is terminated (so that no further y-values are taken into account) in a group having realized at least 65% response. Experiment 1 data

collection is terminated at point 1 for the line 6, 7 and 8 groups, and at point 2 for the line 4 group; for details see LS (2012). Remaining groups continue until the very end, at which point the line 1 and 5 groups still have a response rate far below 65%.

Experiments 2 and 3 use sharpened stopping rules. Experiment 2 is defined to declare data collection terminated for a group when its response has reached 60%. This gives the five intervention points shown in Table 4.1.3: Five groups terminate at four different points in the ordinary data collection, and one group terminates at follow-up attempt 3. The low-responding line 1 and line 5 groups continue to the end, but still do not come near 60% response. The still sharper stopping rule for Experiment 3 declares data collection terminated for a group whose response rate has reached 50%. The resulting six intervention points are shown in Table 4.1.4.

We note in Tables 4.1.2, 4.1.3 and 4.1.4 that the conditional partial imbalance now shows a clearly decreasing trend as the data collection progresses. This holds in particular for the critical groups in lines 1, 5, 6 and 8. The trend is, as one can expect, particularly pronounced for Strategy 3 in Table 4.1.4.

Table 4.1.2
Experimental strategy 1, LCS 2009: Analysis of imbalance; monitoring vector x_a (4.1.1) and x_b (4.1.2)

Group characteristic			$100 \times IMB_{a(j)\|b}$ Conditional partial balance		
Education	Property ownership	Origin	Attempt 12 ordinary	Attempt 2 follow-up	Final
Not high	Non-owner	Abroad	0.98	0.82	0.73
Not high	Non-owner	Sweden	0.25	0.11	0.00
Not high	Owner	Abroad	0.01	0.00	0.00
Not high	Owner	Sweden	0.02	0.05	0.00
High	Non-owner	Abroad	0.19	0.13	0.09
High	Non-owner	Sweden	0.28	0.08	0.02
High	Owner	Abroad	0.02	0.00	0.00
High	Owner	Sweden	0.48	0.19	0.11
		$100 \times diff$	0.10	0.30	0.38
		$100 \times IMB_{a\|b}$	2.3	1.7	1.3
		$100 \times IMB_b$	2.7	2.4	2.0
		$100 \times IMB_{ab}$	5.0	4.1	3.4
		$100 \times P$	57.7	61.5	63.9
		BI_1	0.74	0.74	0.76
		$dist_{r\|nr}$	0.53	0.52	0.51

Table 4.1.3
Experimental strategy 2, LCS 2009: Analysis of imbalance; monitoring vector x_a (4.1.1) and x_b (4.1.2)

Group characteristic			$100 \times IMB_{a(j)\|b}$ Conditional partial balance					
Education	Property ownership	Origin	Att 7 ord.	Att 8 ord.	Att. 9 ord.	Att. 15 ord.	Att. 3 fol.-up	Final
Not high	Non-owner	Abroad	0.99	1.02	0.94	0.70	0.52	0.38
Not high	Non-owner	Sweden	0.26	0.20	0.14	0.02	0.02	0.01
Not high	Owner	Abroad	0.01	0.01	0.01	0.00	0.00	0.00
Not high	Owner	Sweden	0.04	0.08	0.09	0.01	0.01	0.02
High	Non-owner	Abroad	0.21	0.18	0.18	0.11	0.06	0.03
High	Non-owner	Sweden	0.41	0.27	0.19	0.07	0.01	0.01
High	Owner	Abroad	0.01	0.02	0.01	0.01	0.00	0.00
High	Owner	Sweden	0.47	0.30	0.21	0.11	0.04	0.03
	$100 \times diff$		-0.06	0.00	0.06	0.22	0.29	0.34
	$100 \times IMB_{a\|b}$		2.3	2.1	1.8	1.2	1.0	0.8
	$100 \times IMB_b$		3.8	3.5	3.3	2.9	2.7	2.6
	$100 \times IMB_{ab}$		6.1	5.6	5.2	4.2	3.6	3.4
	$100 \times P$		50.9	52.5	53.8	56.0	58.6	58.9
	BI_1		0.75	0.75	0.76	0.77	0.77	0.78
	$dist_{r\|nr}$		0.50	0.50	0.49	0.46	0.46	0.45

Table 4.1.4
Experimental strategy 3, LCS 2009: Analysis of imbalance; monitoring vector $\mathbf{x}_a$ (4.1.1) and $\mathbf{x}_b$ (4.1.2)

Group characteristic			$100 \times IMB_{a(j)\|b}$ Conditional partial balance					
Education	Property ownership	Origin	Att 4 ord.	Att 5 ord.	Att. 6 ord.	Att. 7 ord.	Att. 8 ord.	Final
Not high	Non-owner	Abroad	0.93	0.88	0.79	0.69	0.67	0.02
Not high	Non-owner	Sweden	0.22	0.15	0.11	0.01	0.00	0.07
Not high	Owner	Abroad	0.05	0.02	0.00	0.00	0.00	0.00
Not high	Owner	Sweden	0.04	0.08	0.20	0.08	0.04	0.00
High	Non-owner	Abroad	0.32	0.18	0.18	0.12	0.09	0.01
High	Non-owner	Sweden	0.42	0.42	0.19	0.10	0.06	0.01
High	Owner	Abroad	0.00	0.01	0.01	0.01	0.01	0.00
High	Owner	Sweden	0.66	0.15	0.03	0.00	0.00	0.01
	$100 \times diff$		-0.15	-0.04	0.03	0.10	0.13	0.33
	$100 \times IMB_{a\|b}$		2.5	1.9	1.5	1.1	1.0	0.4
	$100 \times IMB_b$		5.9	4.9	4.4	4.1	3.9	3.4
	$100 \times IMB_{ab}$		8.4	6.8	6.0	5.2	4.9	3.8
	$100 \times P$		39.6	43.8	46.4	47.8	48.7	50.3
	BI_1		0.77	0.77	0.77	0.78	0.78	0.80
	$dist_{r\|nr}$		0.48	0.46	0.46	0.44	0.43	0.39

Table 4.1.5 compares the four data collections, the actual one and the three experimental ones. The ANIMB components in the table are computed at the end of the data collection. A striking feature is the reduction that occurs in the conditional imbalance $IMB_{a|b}$ over the series of four data collections. The final values of $IMB_{a|b}$ are 2.24 (Actual), 1.34 (Exp. 1), 0.82 (Exp. 2) and 0.45 (Exp. 3). By theory we can expect this downward trend, given how the experiments were set up, and the table confirms it empirically.

On the other hand, the marginal imbalance IMB_b goes up markedly from 1.90 for Actual to 3.36 for Experiment 3. This shows that a reduced conditional imbalance may happen at the expense of an increasing marginal imbalance. This can happen, because the experiments were not set up to control the marginal imbalance with respect to $\mathbf{x}_b$, which is a supplementary auxiliary vector here. As a consequence, the total imbalance IMB_{ab} stays about the same. The

distance $dist_{r|nr} = \sqrt{IMB_{ab}}/(1-P)$ is markedly reduced in going from Actual to Experiment 3, a desirable feature saying that respondents and nonrespondents are getting increasingly alike, and while this happens, the final response rate is reduced from 67.4% (Actual) to 50.3 (Exp. 3).

Table 4.1.5
Comparing the four data collections in Tables 4.1.1 to 4.1.4. The entries are the values at the end of each data collection

	$100 \times IMB_{a\|b}$	$100 \times IMB_b$	$100 \times IMB_{ab}$	$dist_{r\|nr}$	$100 \times P$
Actual	2.24	1.90	4.14	0.624	67.4
Exp. 1	1.34	2.03	3.37	0.508	63.9
Exp. 2	0.82	2.58	3.40	0.448	58.9
Exp. 3	0.45	3.36	3.81	0.393	50.3

It is useful to recall results in LS (2012), where the ANIMB was done unconditionally, with the $\mathbf{x}_a$-vector (4.1.1) but without any $\mathbf{x}_b$-vector. Then IMB_b=0 and $IMB_{ab} = IMB_{a|b} = IMB_a$, where $IMB_a = \sum_{j=1}^{J} C_j$, with $C_j = W_j \times (\frac{P_j}{P} - 1)^2$, as explained in Section 2. Not unexpectedly, the entries for IMB_a in Table 4.1.6, taken from Section 6 of LS (2012), are close to those for $IMB_{a|b}$ in Table 4.1.5, but the distances are greater in Table 4.1.5, because computed there on the more extensive x-vector that contains both $\mathbf{x}_a$ and $\mathbf{x}_b$, given respectively by (4.1.1) and (4.1.2).

Table 4.1.6
Comparing four data collections; monitoring vector $\mathbf{x_a}$ given by (4.1.1); no vector $\mathbf{x_b}$. The entries refer to the end of each data collection

	$100 \times IMB_a$	$dist_{r\|nr}$	$100 \times P$
Actual	2.36	0.471	67.4
Exp. 1	1.39	0.326	63.9
Exp. 2	0.82	0.220	58.9
Exp. 3	0.20	0.089	50.3

4.2 Interventions based on response propensities

This section illustrates the approach with response propensities, as outlined in Section 3.4. We compare the actual LCS 2009 data collection with three examples with response sets obtained by interventions in the LCS 2009 data file, so that data collection is stopped, at certain interventions points, for units having attained a high response propensity.

For a specified monitoring vector $\mathbf{x}_a$ and supplementary vector $\mathbf{x}_b$, we compute, for every unit $k \in s$, the *response propensity score*

$$\hat{P}_k = \hat{I}_{bk} + \hat{I}_{ak} - \hat{I}_{a:b,k} \tag{4.2.1}$$

where

$$\hat{I}_{ak} = (\sum_s d_k I_k \mathbf{x}_{ak})'(\sum_s d_k \mathbf{x}_{ak}\mathbf{x}'_{ak})^{-1}\mathbf{x}_{ak}$$

$$\hat{I}_{bk} = (\sum_s d_k I_k \mathbf{x}_{bk})'(\sum_s d_k \mathbf{x}_{bk}\mathbf{x}'_{bk})^{-1}\mathbf{x}_{bk}$$

$$\hat{I}_{a:b,k} = (\sum_s d_k \hat{I}_{ak} \mathbf{x}_{bk})'(\sum_s d_k \mathbf{x}_{bk}\mathbf{x}'_{bk})^{-1}\mathbf{x}_{bk}$$

We then compute the imbalance components shown in equations (3.4.4) and (3.4.5):

$$IMB_{ab} = IMB_b + IMB_{a|b}$$

where

$$IMB_{ab} = \frac{1}{\hat{N}}\sum_s d_k (\frac{\hat{P}_k}{P} - 1)^2 \; ; IMB_b = \frac{1}{\hat{N}}\sum_s d_k (\frac{\hat{I}_{bk}}{P} - 1)^2 \; ;$$

$$IMB_{a|b} = IMB_{ab} - IMB_b = \frac{1}{\sum_s d_k}\sum_s d_k (\frac{\hat{P}_k - \hat{I}_{bk}}{P})^2$$

Tables 4.2.2. to 4.2.5 describe the four scenarios (actual, and three examples of interventions) in terms of the intervention points, the vector pair ($\mathbf{x}_a$, $\mathbf{x}_b$), the imbalance, the distance $dist_{r|nr} = \sqrt{IMB_{ab}}/(1-P)$ and the balance $BI_1 = 1 - \sqrt{IMB_{ab}}/(1/P-1)$.

Table 4.2.2 is computed on the actual LCS 2009 data, as collected, without any interventions. We choose to place all the x-variables (see Section 4.1) in the monitoring vector, so that

$$\mathbf{x}_a = ((Educ \times Owner \times Origin) + Phone + Age + Civil + Gender) \tag{4.2.2}$$

of dimension 8+1+3+1+1 = 14. (One age category is suppressed in order to have an invertible matrix .) There is no $\mathbf{x}_b$ -vector, so $\hat{P}_k = \hat{I}_{ak}$, and $IMB_{ab} = IMB_a$ where

$$IMB_a = \frac{1}{\sum_s d_k} \sum_s d_k \left(\frac{\hat{I}_{ak}}{P} - 1\right)^2$$

Table 4.2.2 shows the distance $dist_{r|nr}$ and the balance BI_1 , computed at three intermediate points. These are: attempt 8 of the ordinary data collection, the end of the ordinary data collection, and attempt 3 of the follow-up. The row Final refers to the very end of the data collection. We contrast Table 4.2.2 with three examples involving interventions.

In Example 1 (Table 4.2.3) we again use $\mathbf{x}_a$ in (4.2.2) as monitoring vevtor, and no $\mathbf{x}_b$ -vector. The data collection is subject to "the 25% dropping rule". The values $\hat{P}_k$ are computed at each point and 25% of the units in the LCS 2009 sample, those with the largest $\hat{P}_k$, are set aside each time. Table 4.2.1 illustrates the procedure. For example, at point 2 we set aside 1/3 of the 3n/4 units still under consideration.

Table 4.2.1
The 25% dropping rule illustrated

Point	Set of units under consideration	Number set aside	Cumulative number set aside	Quantile for setting aside	Remaining number of units
1	n	$n/4$ largest	$n/4$	75%	$n - n/4 = 3n/4$
2	$3n/4$	$n/4$ largest	$n/4+(3n/4)\times(1/3) = n/2$	66.7%	$n - n/2 = n/2$
3	$n/2$	$n/4$ largest	$n/2+(n/2)\times(1/2) = 3n/4$	50%	$n - 3n/4 = n/4$
4	$n/4$				

In Example 2 (Table 4.2.4) we specify the monitoring vector as $\mathbf{x}_a = (Educ \times Owner \times Origin)$ of dimension 8, and a supplement vector, $\mathbf{x}_b = (Phone + Age + Civil + Gender)$ of dimension 7. The data set is obtained from the actual LCS 2009 by applying the 25% dropping rule at the points mentioned for Example 1.

Example 3 (Table 4.2.5) illustrates a variation of the approach with response propensities. Here the interventions consist in dropping, at

each intervention point, those units having attained a specified threshold value for response propensity $\hat{P}_k$, namely, in this case 0.60 or greater. This is a subjective choice; in practice, it would be justified by "a reasonable expectation" for the ultimate response rate in the survey. We have here chosen five intervention points: attempts 5, 8, 12 of the ordinary field work, the end of the ordinary field work, and attempt 3 of the follow-up (as identified by the WinDATI-events).

The end value (the row Final) of the distance $dist_{r|nr}$ is lower with the interventions in Examples 1, 2 and 3 (situated in the range 0.51 to 0.38), when compared with the Actual LCS (where it is as high as 0.62). With interventions we thus improve substantially over the actual LCS 2009 data collection; we expect it from theory, and the empirical results confirm it.

The best results (shortest distance, highest balance) happen for Example 3 (Table 4.2.5) with its five interventions and the 60% stopping rule. At "Final", the distance $dist_{r|nr}$ is then the shortest, 0.379, and the balance BI_1 is the highest, 0.814.

(By comparison, Table 6.8 of LS (2012) contrasted the actual LCS 2009 data collection with three experimental strategies, but derived differently. The best distance $dist_{r|nr}$ obtained in that Table 6.8 was 0.383; however, this was computed on the somewhat different "Standard x-vector," so the two situations are not completely comparable.)

We note that the row Final for Actual LCS 2009 in Table 4.2.2 (done with the response propensity approach) is confirmed by the column Final in Table 4.1.1 (done with the ANIMB approach in Table 3.1.1): $BI_1 = 0.71$, $dist_{r|nr} = 0.62$, $100 \times IMB_a = 4.1$.

Table 4.2.2
Actual LCS data collection (no interventions), vector x_a of dimension 14 given by (4.2.2); no vector x_b

Data collection point	$100 \times P$	BI_1	$dist_{r\|nr}$	IMB_a	Number of call attempts
Attempt 8 ordinary	53.0	0.743	0.515	0.0585	33807
End ordinary	60.4	0.730	0.552	0.0479	42652
Attempt 3 follow-up	63.8	0.721	0.581	0.0443	47711
Final	67.4	0.708	0.623	0.0414	53258

Table 4.2.3
Example 1. Vector x_a of dimension 14 given by (4.2.2); no vector x_b

Data collection point	$100 \times P$	BI_1	$dist_{r\|nr}$	IMB_a
Attempt 8 ordinary	53.0	0.743	0.515	0.0585
End ordinary	58.6	0.767	0.473	0.0385
Attempt 3 follow-up	60.0	0.782	0.446	0.0318
Final	60.5	0.796	0.418	0.0272
Final reduction of call attempts in %: 16.1				

Table 4.2.4
Example 2. Vector x_a of dimension 8 given by (4.1.1); x_b of dimension 7 given by (4.1.2)

Data collection point	$100 \times P$	BI_1	$dist_{r\|nr}$	IMB_{ab}	IMB_b	$IMB_{a\|b}$
Attempt 8 ordinary	53.0	0.725	0.551	0.0669	0.0390	0.0279
End ordinary	58.6	0.734	0.537	0.0494	0.0322	0.0171
Attempt 3 follow-up	60.5	0.742	0.528	0.0435	0.0309	0.0126
Final	60.8	0.750	0.513	0.0403	0.0304	0.0100
Final reduction of call attempts in %: 16.5						

Table 4.2.5
Example 3. Vector x_a of dimension 14 given by (4.2.2); no vector x_b . Stopping rule: response propensity 60% or greater

Data collection point	$100 \times P$	BI_1	$dist_{r\|nr}$	IMB_a
Attempt 5 ordinary	44.3	0.763	0.478	0.0709
Attempt 8 ordinary	52.5	0.760	0.481	0.0522
Attempt12 ordinary	56.0	0.777	0.449	0.0390
End ordinary	57.3	0.787	0.430	0.0337
Attempt 3 follow-up	58.7	0.801	0.404	0.0278
Final	60.1	0.814	0.379	0.0229
Final reduction of call attempts in %: 15.4				

5 Analysis of domains

5.1 Full *y*-data

As the Summary points out, this report consists of several parts. The present Section 5 of the report forms part 2, devoted to the estimation for domains in the presence of nonresponse.

As is well known, survey estimates are usually needed not only for the whole population but also for a variety of subpopulations (domains). In this section we illustrate domain estimation for the LCS 2009 data, for five selected domains. We are interested in seeing how the estimates for those domains evolve during data collection, including the changes in the estimates and their tendency to stabilize more or less as the data collection goes on. The LCS sample is considered a simple random sample from the whole population. The five domains are not identified in advance – they are "unplanned domains" – so the number of observations that fall in a given domain is random.

Persons in charge of a survey will typically maintain that accurate estimates require an ultimate overall response rate as high as possible. On the other hand, common sense tells us that it is useless to lengthen the data collection period unless it brings distinctly better features in the estimates, which can be computed and inspected continuously, along with their precision, as the data collection unfolds.

For those who manage and supervise the survey, two features of the data collection are usually of particular interest: (i) whether or not the estimates become "stable" relatively early in the data collection period, so that changes are small later in the period, and (ii) whether or not the estimates become sufficiently precise, as measured by the *coefficient of variation* (*cv*), if not early in the data collection period, so at least at the end. When the progression is closely examined – something seldom done in practice, for lack of time and other reasons – a typical pattern is that stability (small changes in the estimates) occurs quite early in the data collection and that the changes in the *cv* are minute, particularly in the later stages of the data collection. These patterns are confirmed by our analysis in this section.

In the terminology of Section 3.1, this Section 5.1 illustrates the FULL case, where register variables are used as study variables, with y_k-values available for all $k \in s$. They are called pseudo y-variables. Unbiased domain estimates can be computed, and we can study other aspects than just (i) and (ii). We use three pseudo y-variables, *Benefits*, *Income* and *Employed*. Related to important real y-variables in the LCS, they were explained in LS(2012) and were used in Section 3.1 of this report.

Section 5.2 illustrates the INCOMPLETE case, where the y_k-values are limited to $k \in r$. In that case we can compute the point estimates and their *cv* continuously during data collection, so aspects (i) and (ii) can still be studied.

The estimates for the three pseudo y-variables are examined in this section with the aid of four measures, or four aspects, all of them viewed as a function of the contact attempt number in the LCS 2009 data collection:

(i) the calibrated estimate of the domain mean

(ii) the precision of the domain estimate, measured by the coefficient of variation (*cv*),

(iii) the relative difference of the calibrated estimate and of the expansion estimate from the unbiased domain estimate (the *RDF*)

(iv) the variance proportion of the mean squared error, defined as the estimated variance divided with the estimated mean square error.

We choose here to examine estimates of y-variable means, for domains and for the whole population. One reason is that mean estimates are frequently used in the LCS production. Moreover, it is easier to compare domains through means than through totals. When the study variable is binary, the mean is a proportion, thus between zero and one.

The four measures are defined in detail below. Besides being simple to compute, they bring important insight into the properties of the estimates and the nature of the bias. Although their definitions are simple and intuitive, they have important differences and may lend themselves to conflicting interpretations. One must be aware of exactly which aspects of the estimates that matter the most.

With aspect (i) we can examine the stability of the estimates as the data collection proceeds; with aspect (ii) we can track the changes in the measure of precision (the *cv*). But neither (i) nor (ii) elucidate the main problem with nonresponse, namely the bias. But this is possible with (iii) and (iv), which require pseudo *y*-variables available for the whole sample. These aspects give important insight into the effect of the bias on the domain estimates. The choice of pseudo *y*-variables should be made in consultation with persons well familiar with the survey and its content; they should be linked to, or resemble, central real target variables.

The aspect (iv) is a proportion showing the contribution of the variance to the mean square error, which is the sum of variance and squared bias. A small value for this proportion indicates that the bias, not the variance, is the major cause of inaccuracy. This is an undesirable situation, because in practice the variance can usually be assessed (through the variance estimate) whereas the bias remains hidden. If we rely only on the variance estimate, we risk to get an unrealistically favorable impression of the accuracy.

In this Section 5.1 we examine the aspects (i)-(iv) for the three pseudo *y*-variables and for each of five domains, defined by age groups: 16-24 years, 25-44 years, 45-64 years, 65-74 years and 75+ years. Age groups are important in the presentation of LCS results.

We now specify the measures (i) to (iv). For the three pseudo *y*-variables we can compute the unbiased full sample (Horvitz-Thompson) estimate, and thereby obtain the *RDF* measures for aspect (iii). The calibration estimator, computed on the response set, is based on the combined **x**-vector of dimension 14 given by (4.2.2), that is,

$$\mathbf{x}_a = ((Educ \times Owner \times Origin) + Phone + Age + Civil + Gender) \qquad (5.1.1)$$

The computations rely on the following formulas: We denote by U_a a domain with size N_a of the population U, $U_a \subseteq U$, by s_a the part of the sample s falling in U_a and by r_a the set of respondents from s_a. There are five age group domains, $a = 1, \ldots, 5$. The domain mean $\bar{y}_a = \sum_{U_a} y_k / N_a$ is estimated under full response by the essentially unbiased estimator $\bar{y}_{a,s} = (\sum_{s_a} d_k y_k)/(\sum_{s_a} d_k)$, under nonresponse by $\bar{y}_{a,r} = (\sum_{r_a} d_k y_k)/(\sum_{r_a} d_k)$ (which uses design weights only and is likely to be considerably biased) or by $\bar{y}_{aCAL} =$

$(\sum_{r_a} d_k m_k y_k)/(\sum_{r_a} d_k m_k)$ (which uses calibrated weights and is likely to be less biased). The calibration factors m_k are common to all study variables and are given in Section 2.2.

Expressed as functions of domain total estimates we have $\bar{y}_{a,r} = \hat{Y}_{aEXP} / \hat{N}_{aEXP}$ with $\hat{Y}_{aEXP} = \sum_{r_a} d_k y_k$ and $\hat{N}_{aEXP} = \sum_{r_a} d_k$, and $\bar{y}_{aCAL} = \hat{Y}_{aCAL} / \hat{N}_{aCAL}$ with $\hat{Y}_{aCAL} = \sum_{r_a} d_k m_k y_k$ and $\hat{N}_{aCAL} = \sum_{r_a} d_k m_k$. When $U_a = U$, these formulas, with the index *a* suppressed, give the mean estimates for the entire population U.

We computed the mean estimates $\bar{y}_{aCAL}$ and their variance estimates $\hat{V}(\bar{y}_{aCAL})$ with Statistics Sweden's software ETOS, see Andersson (2012), Section 5.5. The coefficient of variation is computed as

$$c\hat{v}_{aCAL} = 100 \times \frac{\hat{V}(\bar{y}_{aCAL})^{1/2}}{\bar{y}_{aCAL}} \qquad (5.1.2)$$

The relative differences (in per cent) for domain *a* rely on the same principle, (3.1.8), as in the estimation of totals:

$$RDF(\bar{y}_{a,r}) = 100 \times (\bar{y}_{a,r} - \bar{y}_{a,s}) / \bar{y}_{a,s}$$

$$RDF(\bar{y}_{aCAL}) = 100 \times (\bar{y}_{aCAL} - \bar{y}_{a,s}) / \bar{y}_{a,s}$$

We note that $RDF(\bar{y}_{a,r}) = RDF(\hat{Y}_{aEXP})$. To obtain the measure (iv) for $\bar{y}_{aCAL}$ we need an estimation of its mean square error, the sum of the estimated variance and an estimate of the square of the bias. We note that $\bar{y}_{a,s}$ is constant for all call attempts (because based on the full sample), whereas $\bar{y}_{aCAL}$ changes, as new data enter with each additional call attempt. We use $(\bar{y}_{aCAL} - \bar{y}_{a,s})^2$ as an estimate of the squared bias. The estimated mean squared error is therefore

$$mse\hat{(}\bar{y}_{aCAL}) = \hat{V}(\bar{y}_{aCAL}) + (\bar{y}_{aCAL} - \bar{y}_{a,s})^2$$

It can be computed for each pseudo *y*-variable after each call attempt, on the data available at that point. The quantity

$$VarProp = 100 \times \hat{V}(\bar{y}_{aCAL}) / mse\hat{(}\bar{y}_{aCAL})$$

is a measure (in per cent) of the proportion contributed by the variance to the mean squared error. *VarProp* will be near 100% if

there is no bias. When *VarProp* is low, the interpretation is that the mean squared error is dominated by the bias, in other words a message about the importance of the nonresponse bias.

Because *VarProp* uses estimates of the two components of the mean squared error, it may be subject to considerable variability. Therefore *VarProp* should be interpreted with some care.

Tables 5.1.1-5.1.6 show the development of the measures (i) to (iv) as a function of the attempt number in the LCS 2009 data collection. Each table has three parts. From top to bottom, they refer to the three pseudo *y*-variables, *Benefits*, *Income*, and *Employed*.

Table 5.1.1 presents the results for the mean estimates $\bar{y}_{CAL}$ for the whole population. Tables 5.1.2 to 5.1.6 presents the results for the domain mean estimates $\bar{y}_{aCAL}$ for the five age domains, $a = 1,\ldots,5$.

In Table 5.1.1 we note:

- At Final, $\bar{y}_{CAL}$ has a distinctly smaller *RDF* than $\bar{y}_{EXP}$, which confirms the expectation that the auxiliary information used in $\bar{y}_{CAL}$ has a favourable effect.
- At a cursory look, the point estimate $\bar{y}_{CAL}$ looks quite stable from Attempt 12 to Final. But the table also shows that a different view of stability is obtained by looking instead at $RDF(\bar{y}_{CAL})$. For example, for *Benefits*, $RDF(\bar{y}_{CAL})$ ranges considerably, from -0.7 at Attempt 12 to as much as -4.6 at Final. The *RDF* can be positive or negative, so it specifies the position of $\bar{y}_{CAL}$ vis-à-vis the unbiased estimate $\bar{y}_s$. For all three variables in Table 5.1.1, $RDF(\bar{y}_{CAL})$ changes its sign during the data collection. That is, there is a point in the data collection where $\bar{y}_{CAL}$ and the unbiased estimate agree.
- As can be expected, the precision, as measured by the coefficient of variation *cv*, decreases as more and more data come in. But the drop is minute. The *cv* hardly changes at all from Attempt 12 to Final. There are virtually no gains in precision by pursuing the data collection beyond Attempt 12.
- *VarProp* decreases from Attempt 12 to Final, for all three variables. The principal reason is an increased squared bias component (because $RDF(\bar{y}_{CAL})$ increases). The decreasing trend in *VarProp* sends the message that the bias (and not the variance)

becomes increasingly harmful as the data collection goes on; it is an undesirable feature of LCS 2009 data collection.

For the five domains (Tables 5.1.2 to 5.1.6) we expect less stable patterns. One reason is that the sample size is not very large for some of them. Moreover, the variables have age related features, which can lead to unpredictable results. For example, a feature of the variable *Benefits* is that sickness and related benefits are quite rare among persons in the age domains 16-24, 65-74 and 75+. And for natural reasons, the variable *Employed* is at low levels in the domain 75+.

- An inspection of the domain estimates $\bar{y}_{aCAL}$ shows small changes over the course of the data collection. This gives a certain impression of stability. However if we inspect *RDF* instead, the variation is seen in a different light. In several cases, the *RDF* shifts sign during the data collection. That is, at some point in the data collection, the calibration estimator $\bar{y}_{aCAL}$ is very close to the unbiased estimate $\bar{y}_s$.
- The variable *Benefits* is at a very low level among the youngest (Table 5.1.2) and particularly among the oldest (Table 5.1.6). It is known that the benefits in question accrue predominantly to persons in the active age groups, 19-64 years. For the oldest group 75+, the values intended for Table 5.1.6 were unreliable and therefore suppressed.
- Comparing $RDF(\bar{y}_{a,r})$ and $RDF(\bar{y}_{aCAL})$ at the point Final we see that it can happen that $RDF(\bar{y}_{a,r})$ is the smaller of the two, stating that $\bar{y}_{a,r}$ is closer to the unbiased estimate than $\bar{y}_{aCAL}$. For example, in Table 5.1.5 (age group 65-74) the Final values for *Benefits* are 0.5 for $RDF(\bar{y}_{a,r})$ compared with 8.3 for $RDF(\bar{y}_{aCAL})$. This can of course happen in isolated cases. But a majority of the cases confirms the expectation that $\bar{y}_{aCAL}$ "works better", so that, at Final, $RDF(\bar{y}_{aCAL})$ is smaller in absolute value than $RDF(\bar{y}_{a,r})$.
- The low incidence of *Benefits* in the domains 16-24, 65-74 and 75+ gives high *cv*. Also not surprising, high *cv* for *Employment* occurs for the oldest domain, 75+ (Table 5.1.6).
- A near zero value of $RDF(\bar{y}_{aCAL})$ does not imply that *cv* is also low. They measure different, but important, aspects. The former indicates bias, the latter indicates precision (because it reflects the

variance). A comparison of the domains for the column Final reveals several different patterns in the relationship between $RDF(\bar{y}_{aCAL})$ and *cv*.

- At the point Final, *VarProp* is well below 50% in most of the 15 combinations of domain by study variable. It suggests that bias, not variance, is the principal cause of inaccuracy. Particularly large values of *VarProp* occur for the domain 64-75 (Table 5.1.5).

In summary, this Section 5.1 emphasizes that the measures (i) and (ii) can at best give an incomplete picture of the problem of inaccuracy caused by nonresponse. We have suggested that the measures (iii) and (iv) are valuable, not to say necessary, complements for an evaluation of the survey results.

Table 5.1.1
Total sample (all ages), size of sample: *n* = 8,220

	Attempt number			Follow-up			
	1	5	12	End ord.	1	4	Final
Pseudo variable ***Benefits***							
$RDF(\bar{y}_r)$	-10.0	-7.2	-7.9	-7.9	-8.0	-9.3	-9.4
$RDF(\bar{y}_{CAL})$	9.1	3.2	-0.7	-1.9	-2.1	-3.8	-4.6
$100 \times \bar{y}_{CAL}$	14.8	14.0	13.4	13.3	13.3	13.0	12.9
$c\hat{v}_{CAL}$	8.0	4.4	3.9	3.8	3.8	3.7	3.6
VarProp	47.7	66.9	96.7	79.4	76.4	46.3	36.6
Pseudo variable ***Income***							
$RDF(\bar{y}_r)$	0.3	3.6	6.8	7.4	7.1	6.7	6.7
$RDF(\bar{y}_{CAL})$	-0.2	0.8	3.3	3.7	3.7	3.5	3.3
$10^{-4} \times \bar{y}_{CAL}$	22.4	22.6	23.1	23.2	23.2	23.2	23.1
$c\hat{v}_{CAL}$	2.2	1.2	1.1	1.1	1.1	1.1	1.0
VarProp	99.3	67.4	11.3	8.7	8.8	9.0	9.7
Pseudo variable ***Employed***							
$RDF(\bar{y}_r)$	-9.0	-1.1	3.1	4.2	4.1	4.3	4.8
$RDF(\bar{y}_{CAL})$	-1.5	1.1	2.5	2.9	2.9	3.3	3.1
$100 \times \bar{y}_{CAL}$	66.0	67.6	68.6	68.9	68.9	69.1	69
$c\hat{v}_{CAL}$	2.1	1.1	0.9	0.9	0.9	0.9	0.9
VarProp	66.5	52.0	13.3	9.5	9.4	7.4	8.0

Table 5.1.2
Domain Age 16-24; size of sample in the domain: 1,118

	Attempt number				Follow-up		
	1	5	12	End ord.	1	4	Final
Pseudo variable ***Benefits***							
$RDF(\bar{y}_{a,r})$	-80.6	-21.7	-27.2	-22.4	-20.1	-17.4	-20.4
$RDF(\bar{y}_{aCAL})$	-83.4	-19.7	-26.7	-23.3	-21.4	-20.1	-23.0
$100 \times \bar{y}_{aCAL}$	0.6	3.0	2.8	2.9	3.0	3.0	2.9
$c\hat{v}_{aCAL}$	99.7	27.4	24.6	23.3	22.7	21.6	21.6
VarProp	3.8	55.5	31.3	37.0	40.9	42.2	34.2
Pseudo variable ***Income***							
$RDF(\bar{y}_{a,r})$	8.3	-4.9	-7.0	-5.6	-5.3	-4.4	-4.4
$RDF(\bar{y}_{aCAL})$	-1.3	-7.4	-8.7	-7.2	-6.6	-5.9	-6.4
$10^{-4} \times \bar{y}_{aCAL}$	5.6	5.2	5.2	5.3	5.3	5.3	5.3
$c\hat{v}_{aCAL}$	11.8	7.2	6.3	6.1	6.0	5.8	5.7
VarProp	98.7	44.4	30.7	38.2	42.4	46.1	40.4
Pseudo variable ***Employed***							
$RDF(\bar{y}_{a,r})$	5.0	-0.7	-0.7	0.4	0.5	2.0	2.3
$RDF(\bar{y}_{aCAL})$	3.4	-2.3	-2.3	-1.1	-1.0	0.6	0.8
$100 \times \bar{y}_{aCAL}$	70.5	66.6	66.6	67.4	67.5	68.6	68.7
$c\hat{v}_{aCAL}$	6.1	3.4	3.0	2.9	2.8	2.7	2.6
VarProp	77.5	67.7	60.9	87.8	88.7	94.9	92.3

Table 5.1.3
Domain Age 24-44; size of sample in the domain: 2,622

	Attempt number				Follow-up		
	1	5	12	End ord.	1	4	Final
Pseudo variable ***Benefits***							
$RDF(\overline{y}_{a,r})$	10.0	-0.7	-5.1	-5.6	-5.5	-6.6	-6.7
$RDF(\overline{y}_{aCAL})$	10.9	0.8	-3.0	-4.3	-4.2	-5.5	-4.7
$100 \times \overline{y}_{aCAL}$	13.0	11.8	11.4	11.2	11.2	11.1	11.2
$c\hat{v}_{aCAL}$	16.9	9.0	7.8	7.5	7.5	7.3	7.1
VarProp	74.6	99.2	86.2	73.9	74.0	61.7	67.6
Pseudo variable ***Income***							
$RDF(\overline{y}_{a,r})$	4.7	7.6	9.9	10.7	11.1	9.9	9.7
$RDF(\overline{y}_{aCAL})$	2.3	3.3	6.1	7.1	7.5	6.6	6.3
$10^{-4} \times \overline{y}_{aCAL}$	25.4	25.7	26.3	26.6	26.7	26.5	26.4
$c\hat{v}_{aCAL}$	4.5	2.1	1.8	1.8	1.8	1.8	1.7
VarProp	80.0	29.6	9.4	6.9	6.2	7.4	7.6
Pseudo variable ***Employed***							
$RDF(\overline{y}_{a,r})$	6.8	6.2	7.2	7.5	7.6	7.4	6.9
$RDF(\overline{y}_{aCAL})$	4.9	3.3	4.8	5.3	5.3	5.3	4.7
$100 \times \overline{y}_{aCAL}$	89.6	88.2	89.6	90.0	90.0	90.0	89.4
$c\hat{v}_{aCAL}$	2.4	1.3	1.0	1.0	1.0	0.9	0.9
VarProp	21.0	15.1	5.0	3.8	3.6	3.4	4.3

Table 5.1.4
Domain Age 45-64; size of sample in the domain: 2,626

	Attempt number				Follow-up		
	1	**5**	**12**	**End ord.**	**1**	**4**	**Final**
Pseudo variable ***Benefits***							
$RDF(\bar{y}_{a,r})$	5.0	-3.5	-8.1	-8.7	-8.6	-10.6	-11.6
$RDF(\bar{y}_{aCAL})$	10.6	2.6	-1.2	-2.2	-2.3	-4.8	-5.9
$100 \times \bar{y}_{aCAL}$	28.8	26.7	25.7	25.5	25.5	24.8	24.5
$c\hat{v}_{aCAL}$	9.2	5.2	4.6	4.5	4.5	4.4	4.4
VarProp	48.3	80.3	93.3	80.8	79.1	43.7	32.9
Pseudo variable ***Income***							
$RDF(\bar{y}_{a,r})$	0.4	2.8	6.4	6.2	5.8	5.9	5.7
$RDF(\bar{y}_{aCAL})$	-3.5	-1.5	2.0	2.0	1.6	2.1	1.9
$10^{-4} \times \bar{y}_{aCAL}$	28.0	28.6	29.6	29.6	29.5	29.7	29.6
$c\hat{v}_{aCAL}$	3.5	1.8	1.9	1.8	1.8	1.7	1.7
VarProp	47.9	58.7	46.6	45.8	56.4	41.3	45.8
Pseudo variable ***Employed***							
$RDF(\bar{y}_{a,r})$	-2.6	4.0	5.0	5.1	5.0	5.2	5.2
$RDF(\bar{y}_{aCAL})$	-6.3	0.3	1.4	1.5	1.5	1.9	1.9
$100 \times \bar{y}_{aCAL}$	74.9	80.2	81.0	81.1	81.1	81.4	81.4
$c\hat{v}_{aCAL}$	3.5	1.6	1.4	1.3	1.3	1.3	1.2
VarProp	21.9	96.0	47.8	43.1	44.9	30.4	29.4

Table 5.1.5
Domain Age 65-74; size of sample in the domain: 976

	Attempt number				Follow-up		
	1	5	12	End ord.	1	4	Final
Pseudo variable ***Benefits***							
$RDF(\bar{y}_{a,r})$	-22.0	-8.1	-0.8	-1.3	-2.4	0.6	0.5
$RDF(\bar{y}_{aCAL})$	-31.1	1.9	9.0	8.2	6.6	8.7	8.3
$100 \times \bar{y}_{aCAL}$	5.6	8.2	8.8	8.8	8.6	8.8	8.8
$c\hat{v}_{aCAL}$	28.7	16.1	13.8	13.6	13.6	13.1	12.9
VarProp	28.9	98.6	73.5	76.4	82.9	72.7	73.8
Pseudo variable ***Income***							
$RDF(\bar{y}_{a,r})$	2.3	3.1	4.2	4.9	4.6	4.5	4.3
$RDF(\bar{y}_{aCAL})$	-4.3	-1.6	0.2	1.0	0.8	0.8	0.8
$10^{-4} \times \bar{y}_{aCAL}$	21.8	22.4	22.8	22.9	22.9	22.9	22.9
$c\hat{v}_{aCAL}$	4.2	2.4	2.3	2.3	2.2	2.3	2.3
VarProp	46.3	68.5	99.4	85.1	88.8	88.9	87.9
Pseudo variable ***Employed***							
$RDF(\bar{y}_{a,r})$	12.3	5.2	10.1	10.0	8.8	7.4	7.9
$RDF(\bar{y}_{aCAL})$	-2.3	-1.4	3.5	3.7	2.5	1.6	2.3
$100 \times \bar{y}_{aCAL}$	33.4	33.7	35.4	35.5	35.1	34.8	35
$c\hat{v}_{aCAL}$	10.5	6.1	5.3	5.2	5.2	5.1	5.1
VarProp	95.0	95.0	70.9	68.6	81.8	91.7	83.0

Table 5.1.6
Domain Age 75+; size of sample in the domain: 878

	Attempt number				Follow-up		
	1	5	12	End ord.	1	4	Final
Pseudo variable ***Benefits***							
$RDF(\bar{y}_{a,r})$	-	-	-	-	-	-	-
$RDF(\bar{y}_{aCAL})$	-	-	-	-	-	-	-
$100 \times \bar{y}_{aCAL}$	-	-	-	-	-	-	-
$c\hat{v}_{aCAL}$	-	-	-	-	-	-	-
VarProp	-	-	-	-	-	-	-
Pseudo variable ***Income***							
$RDF(\bar{y}_{a,r})$	9.4	7.7	6.9	6.6	6.3	5.8	5.8
$RDF(\bar{y}_{aCAL})$	6.7	5.9	4.8	4.7	4.6	4.1	4.1
$10^{-4} \times \bar{y}_{aCAL}$	17.4	17.3	17.1	17.1	17.1	17.0	17.0
$c\hat{v}_{aCAL}$	3.1	2.0	1.9	1.9	1.9	1.8	1.8
VarProp	19.2	11.5	14.5	14.7	15.3	18.0	17.5
Pseudo variable ***Employed***							
$RDF(\bar{y}_{a,r})$	-25.1	26.0	25.9	24.7	29.4	27.0	27.7
$RDF(\bar{y}_{aCAL})$	-35.1	18.6	20.0	19.6	24.5	21.5	22.1
$100 \times \bar{y}_{aCAL}$	5.0	9.0	9.2	9.1	9.5	9.3	9.3
$c\hat{v}_{aCAL}$	33.3	14.9	13.8	13.8	13.4	13.1	13.0
VarProp	27.5	47.6	40.8	41.6	31.6	35.5	33.9

5.2 Incomplete *y*-data

We study the estimation of means for five real study variables for the five age-domains specified in Section 5.1. That is, we are dealing now with the INCOMPLETE case, in the terminology of Section 3.1: The y_k -values are available only for the responding units $k \in r$. The five study variables were selected in consultation with LCS subject matter specialists. They are dichotomous, with the following names and definitions:

- ***Smoke*** equals one for a person who smokes daily; zero otherwise.
- ***Dentist*** equals one for a person who has visited the dentist (in the last 12 months); zero otherwise.
- ***Trip*** equals one for a person who has gone on a holiday trip (in the last 12 months); zero otherwise.
- ***Threat*** equals one for a person victim of threat or violence (in the last 12 months); zero otherwise.
- ***No-space*** equals one for a person living in overcrowded conditions, according to a norm; zero otherwise.

The mean of each variable, in a given subpopulation (domain), is the proportion of persons with the characteristic in question. In this INCOMPLETE setting, we can compute the population mean estimate $\bar{y}_{CAL}$, the domain mean estimates $\bar{y}_{aCAL}$, and their coefficients of variation, *cv*, but not the measures *RDF* and *VarProp* used in Section 5.1. The auxiliary vector (5.1.1) is used to compute the calibration estimator $\bar{y}_{aCAL}$. The three pseudo *y*-variables considered in Section 5.1 are not present in this auxiliary vector. Also not present in (5.1.1) are a few other auxiliary variables used in producing the LCS estimates, namely, *Employed*, *Income*, *Social allowance*, *Geographical region* and *Immigrated after year 2000*. However, the conclusions in this section would not be markedly different for an auxiliary vector to some degree different from (5.1.1).

Figures 5.2.1 and 5.2.2 show graphs of the five population mean estimates $\bar{y}_{CAL}$ (in per cent) and their corresponding *cv*. Figures 5.2.3 to 5.2.12 show the five domain mean estimates $\bar{y}_{aCAL}$ (in per cent) and their corresponding *cv*. The variance estimates needed for

the *cv* were computed with the software ETOS, see Andersson (2012), section 5.5.

We note some features of the figures: Figure 5.2.1 shows smooth developments for most estimates of the whole population mean. Among the five inspected variables, the mean estimate for *Trip* changes the most, with some increase over the data collection. However, all five means in Figure 5.2.1 settle into a stable pattern quite early in the ordinary field work. The *cv*'s in Figure 5.2.2 show a declining trend, as was the case in Table 5.1.1 for the pseudo *y*-variables, and the rate of the decline of the *cv* is very small after attempt 5.

As Figures 5.2.3 to 5.2.12 show, the domain mean estimates $\bar{y}_{aCAL}$ are quite variable in the early stages the data collection, but the fluctuations subside after attempt 12, and from then on, then estimates change only negligibly until the end. A similarly stable pattern holds for the domain *cv*'s . Our conclusion is that if Figures 5.2.1 to 5.2.12 were the only basis for decisions about the LCS data collection, it would be hard to justify doing a follow-up in this survey.

Similarly as for the variable *Benefits* in Section 5.1, the estimated proportions are very small for some variables and groups, because of rarity of the characteristic in certain age brackets. Examples include the variable *Threat* in the two oldest age groups (Figures 5.2.9 and 5.2.11), and the variable *No-space* in the 65-74 group (Figure 5.2.9). The *cv*'s are large in these cases, even extremely large, as seen for example for *Threat* in Figures 5.2.10 and 5.2.12. Similar unreliability was the noted for the register variable *Benefits* in Section 5.1.

It is clear that one cannot base important conclusions about the domains on estimates such as those shown in figures 5.2.9 and 5.2.11. These estimates are quite unreliable, in part because of an unknown bias, particularly for small groups and rare characteristics.

In the figures we also see that the age domain estimates differ considerably for one and the same variable. For example, the final estimates for *Threat* and *No-space* are, comparatively speaking, very high among the youngest (Figure 5.2.3), whereas the age domain 25-44 shows a very low value on *Dentist* (Figure 5.2.5). The age domain 45-64 shows very high estimate for *Smoker* (Figure 5.2.7); and the

two oldest age domains show very low estimates for *No-space* and *Threat* (Figures 5.2.9 and 5.2.11).

Differences in the age domain estimates for a given y-variable are often predictable, given our general knowledge about society. For example, it may be expected that the *Threat* estimate is highest in the youngest age groups, but what we ignore about the estimates for these groups is the extent and direction of the bias; should the estimate be higher or lower?

The argument that the estimates have achieved stability at a certain point is often not a sufficient reason for stopping the data collection. Stability (a minute change) in the estimates does not imply that bias is low, or that a balanced response set has been achieved. However, a strong reason for actively striving to improve the balance during the data collection is that the risk of bias may be reduced.

When we inspect the evolution of the estimates during data collection, in the hope of bringing improvement (a reduction of bias), then we must have a good idea whether the estimate at a certain point is under- or overestimating. For some real *y*-variables, an examination of related register variables may help towards predicting the sign of the bias. For example, a perspective on the age domain estimates for *No-space* can be obtained by examining the related register variable *Owner*, considered in Section 5.1. But the ways in which register variables can help to assess the sign of the bias for real study variables needs to be studied in more depth; it is a topic of future research. In such work, one should inspect both register variables and real survey variables, with methods proposed in this Section 5.

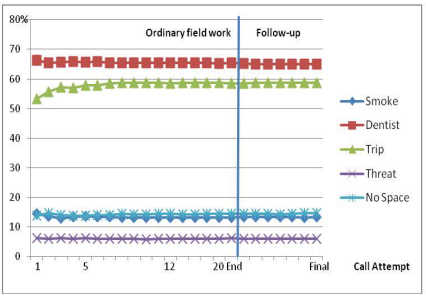


Figure 5.2.1 Point estimates in %, $\overline{y}_{CAL,j}$, for total response set

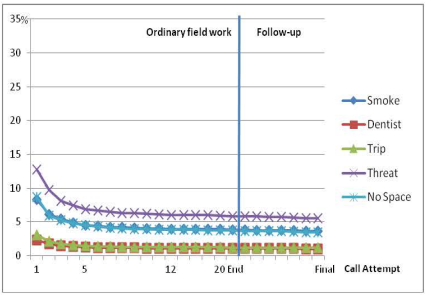


Figure 5.2.2 Estimates of $c\hat{v}_{\bar{y}_j}$ in %, for total response set

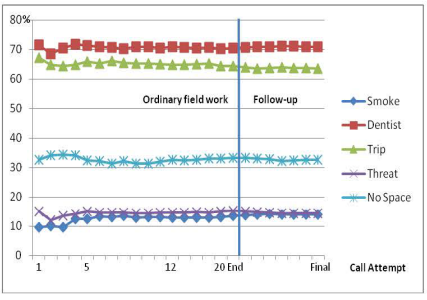


Figure 5.2.3 Point estimates in %, $\overline{y}_{CAL,j}$, for age domain 16-24

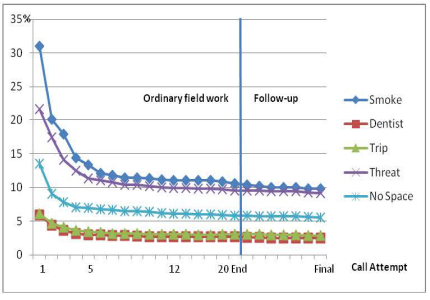


Figure 5.2.4 Estimates of, $c\hat{v}_{\bar{y}_j}$ in %, for age domain 16-24

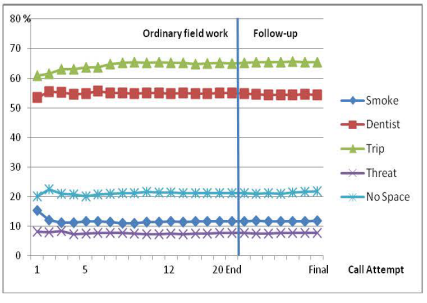


Figure 5.2.5 Point estimates in %, $\overline{y}_{CAL,j}$, for age domain 25-44

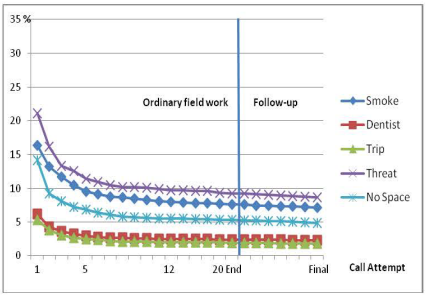


Figure 5.2.6 Estimates of, $c\hat{v}_{\bar{y}_j}$ in %, for age domain 25-44

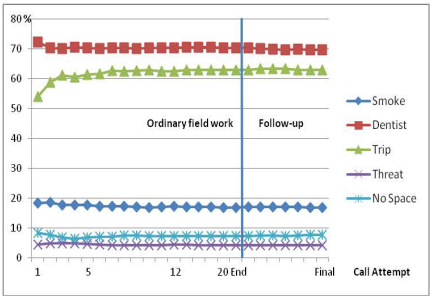


Figure 5.2.7 Point estimates in %, $\overline{y}_{CAL,j}$, for age domain 45-64

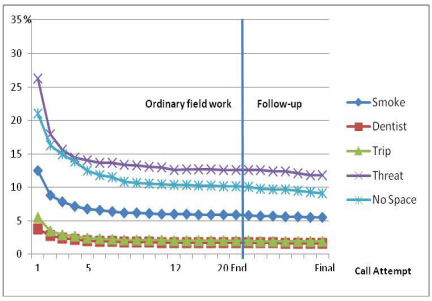


Figure 5.2.8 Estimates of $c\hat{v}_{\overline{y}_j}$ in %, for age domain 45-64

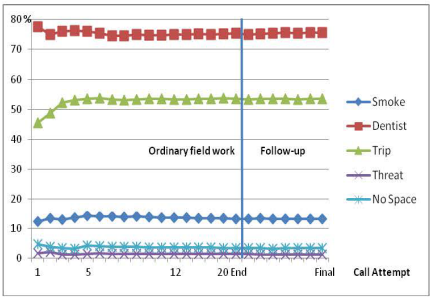


Figure 5.2.9 Point estimates in %, $\overline{y}_{CAL,j}$, for age domain 65-74

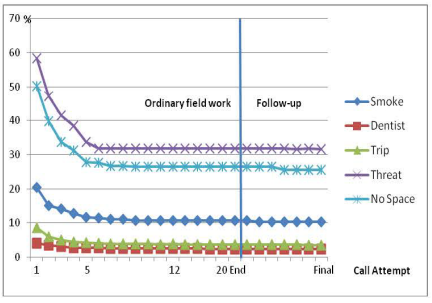


Figure 5.2.10 Estimates of, $c\hat{v}_{\overline{y}_j}$ in %, for age domain 65-74

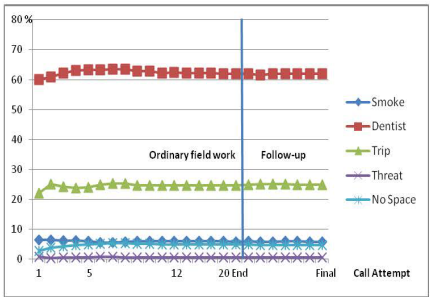


Figure 5.2.11 Point estimates in %, $\overline{y}_{CAL,j}$, for age domain 75+

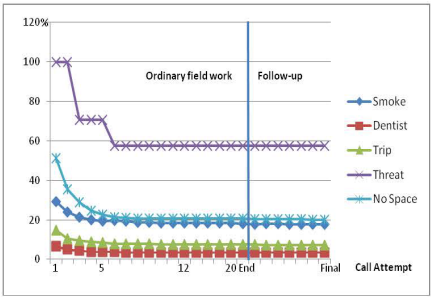


Figure 5.2.12 Estimates of, $c\hat{v}_{\overline{y}_j}$ in %, for age domain 75+

6 The Swedish PIAAC survey

6.1 Short description of PIAAC

The Programme for the International Assessment of Adult Competencies (PIAAC) is a multinational survey sponsored by the OECD. The PIAAC was carried out for the first time in 2011 and 2012. Its objective is to assess the level and the distribution of adult skills in a coherent and consistent way across countries. It focuses on the key cognitive and workplace skills that are required for successful participation in 21st century society and economy. PIAAC will also gather a range of other information including the antecedents and outcomes of skills, as well as information on usage of information technology and literacy and numeracy practices generally. Sweden is among the participating countries. The data collection is carried out by Statistics Sweden, while the development, estimation and analysis is carried out by a consortium of institutions under contract with OECD.

The background and the design are as follows: The target population is defined to consist of all non-institutionalized individuals aged 16-65 years. The Swedish Register of Total Population, which allows stratification by a number of variables, is used to draw the Swedish PIAAC sample. The stratification variables (with the number of categories in parenthesis) are: Gender (2), Age (5), Country of Birth (2), and Level of Education (3). The number of strata for the PIAAC sample S is thus $2\times5\times2\times3 = 60$. Simple random sampling with proportional allocation is used in each stratum; the sample is essentially self-weighting.

The survey requirements specify a minimum number of completed assessments (completed responses) of 5,000. The sample size used for the Swedish sample is 10,000, which, with an initially expected overall response rate of 51%, would satisfy the requirement of 5,000 completed assessments.

The PIAAC data collection was carried out as follows: Face-to-face interviews were conducted by a staff of interviewers using the Swedish CATI-system, WinDATI. An initial contact, by telephone, with the sampled person served to fix an appointment for the face-to-face interview.

All attempts by interviewers to establish contact with a sampled person are registered by the WinDATI system. For every sampled individual, WinDATI thus records a series of "call attempts", which are used in our analysis. The ordinary field work lasted 7 weeks, followed by a break during which the follow-up was prepared. The break and the follow-up period differ in length for different (random) sample groups. Field work was originally planned to be concluded by March 31, but was extended to May 16 because of the low response rate.

6.2 Applying balance and distance indicators to PIAAC

SAS routines for computing the balance and the distance measures were developed and used in LS(2012) to study the LCS 2009 data collection. One objective with this section is to illustrate that these programs can be applied to other surveys than LCS 2009. PIAAC uses stratified sampling, whereas simple random sampling (constant design weights for all units) was the case in LCS 2009. Some modification of the SAS routines was necessary; as expected, we found this modification feasible. Consequently, balance, R-indicators and distance can be computed for the PIAAC. Due to late incoming results from the follow-up, these measures are computed only up until the end of the ordinary data collection.

Another objective was to see if the unattractive patterns for the LCS 2009 data collection carry over to the PIAAC, notably, a decreasing balance and an increasing distance. For comparability, we used essentially the same auxiliary vector as in LCS 2009 to calculate the balance measure BI_1, the unadjusted and the adjusted representativity R, and the distance $dist_{r|nr}$, over the course of the PIAAC data collection.

The monitoring vector $\mathbf{x}_a$ used to analyze the PIAAC study is of dimension seven, defined by the following categorical auxiliary variables: *Phone* (equaling 1 for a person with accessible phone number; 0 otherwise); *Education* (1 if high; 0 otherwise). *Age group* (four zero/one coded groups; age brackets -25. 26-45. 46-55. 56-); *Region* (equaling 1 if resident in Stockholm; 0 otherwise); *Country of origin* (Sweden; other). We call it the *PIAAC standard* **x**-*vector*.

The value $\mathbf{x}_{ak}$ is known for all $k \in s$, that is, for respondents as well as for nonrespondents. The variable *Property ownership* used in the

standard **x**-vector for the LCS study was replaced here by *Region*, believed to be better for forming groups with substantial response rate differences.

The data set analyzed here covers the full PIAAC sample of 10,000 units (persons), but is restricted to field work and WinDATI events prior to April 13, 2012. Hence Table 6.2.1 refers only to the ordinary field work period.

Table 6.2.1 shows the development of the balance indicator BI_1, the unadjusted and the adjusted R, and the distance $dist_{r|nr}$, all viewed as functions of the call attempt number in the ordinary PIAAC data collection. As in LCS 2009, both the balance and the distance develop in "the wrong direction" during the data collection: The balance decreases and the distance increases.

As Table 6.2.1 also shows, the balance and the distance change very little from attempt eight (where the achieved response rate is 29.2%) until the end of the ordinary data collection (where the response rate is a disappointingly low 35.4%). Contributing to the minute changes in these measures is that the contact attempts in the later stages of the data collection bring very few additional responses.

The strategy for the follow-up in PIAAC differs from that of the LCS survey. After the termination of the ordinary field work, a classification tree analysis was performed in order to identify groups with low response rates and therefore likely to contribute considerably to nonresponse bias. These groups were given priority in the follow-up.

Classification trees is a technique that selects auxiliary variables and interactions among these variables. A well known method of this kind is CHAID (see Kass 1980). Here we used the *treedisc* macro developed and provided by the SAS institute. This method is similar to CHAID; for details, see the comments in the SAS code. The classification tree is constructed by partitioning the sample into subsets based on the categories of one of the predictor variables. The variable selected to form a partition is the one most significantly associated with the dependent variable, according to a chi-squared test of independence. The process is repeated up to a point where no further significant splits occur, or that a predefined stopping criterion is met.

In the tree analysis, the 0/1 response indicator is modeled using a number of register variables, available for the full sample, as explanatory variables. The register variables used in the analysis are

gender, *origin*, *marital status*, *employed*, *age*, *education level*, *income* and *region*. As already explained, the follow-up results came too late to be included in the analysis in Table 6.2.1.

Table 6.2.1
The PIAAC 2012 data collection: Progression of the response rate *P* (in per cent), the balance indicator BI_1, unadjusted *R* , adjusted *R*, and the distance $dist_{r|nr}$. The computations are based on the PIAAC standard x-vector explained in this section

Attempt number	100×*P*	BI_1	*R* unadj.	*R* adjusted	$dist_{r\|nr}$
2	5.0	0.940	0.965	0.970	0.276
3	11.7	0.916	0.929	0.934	0.262
4	17.8	0.893	0.894	0.898	0.281
5	22.0	0.879	0.871	0.875	0.293
6	25.1	0.872	0.856	0.860	0.295
7	27.7	0.859	0.844	0.847	0.315
8	29.4	0.856	0.837	0.841	0.316
9	30.7	0.854	0.834	0.838	0.316
10	31.8	0.852	0.830	0.834	0.318
11	32.5	0.852	0.830	0.834	0.316
12	33.1	0.853	0.828	0.832	0.313
13	33.6	0.851	0.827	0.831	0.315
14	33.9	0.854	0.829	0.833	0.309
15	34.3	0.852	0.827	0.831	0.311
⋮					
20	34.9	0.853	0.825	0.829	0.309
End ordinary field work	35.4	0.851	0.822	0.825	0.311
			Follow-up		
2					
3					
4		Not available			
5					
⋮					
10					
Final					

7 Embedded experiment with LCS 2011

7.1 Background

Concurrently with the work reported in LS (2012), a project was launched with a mission to improve the data collection routines in the LCS. The objective was to improve both resource allocation and survey quality. The results would also be useful for Statistics Sweden's data collection department, because an improved contact strategy would facilitate the control of costs and interviewer resources.

A part of the 2011 LCS sample (called the experiment sample) was reserved for testing a new contact strategy. The idea was to compare this sample, through an embedded experiment, with the rest of the sample (the control sample), which obeyed the ordinary LCS contact strategy. The experiment sample data collection was defined in terms of different call intensities for different persons, using paradata from the WinDATI system ("contact protocol data") and auxiliary data. For example, younger persons would receive more call attempts than older persons. An objective was to obtain in the end a well balanced final response, or at least one that is better balanced than with the ordinary data collection.

SCB's project group for the experiment included persons from the data collection department, the department responsible for the LCS, and one survey methodologist from the process department. Working together, this group designed the experiment in the spring of 2011.

7.2 Departures from the original plan

The original plan was to carry out the data collection differently in various subgroups of the experiment sample. For example, the number of contact attempts would depend on the group. For the follow-up in the experiment sample, the idea was initially to combine predefined call strategies with modified advance letters and incentives. Due to a disappointingly low data inflow in the ordinary field work for the control sample, the interventions planned for the follow-up in the experiment sample were

subsequently abandoned. The management decided to abandon the planned modified advance letters, and to administer instead an incentive for all follow-up individuals. Thus the old fixation on achieving the highest possible overall response became the ruling principle for the experiment sample also. Therefore, the empirical results reported below for the experiment sample are presented in two parts: (i) based on the actually observed data and (ii) based on "interventions in retrospect", by deleting some respondents and their *y*-data in the follow-up portion.

The LCS data collection in the experiment uses three starting weeks: Week 33, Week 35 and Week 37. During the first of these, the survey manager noted that the interviewers for the experiment sample felt uncomfortable with the idea to attempt just one single call to a sampled person in a selected time slot. They instead tended to place additional calls to persons they believed "would answer next time." The management worked hard trying to convince the interviewers of the importance of adhering to the experiment call strategy. In the second and particularly in the third starting week, the interviewers did follow the strategy. Section 7.3 deals with the ordinary data collection; Section 7.3.1 for the experiment sample and Section 7.3.2 for the control sample. For the follow-up, a cross-over experiment was designed with an idea to test if the experiment interviewers produced a higher response rate than the control sample interviewers. Although the cross-over was carried out, an analysis of its results is not presented in this report.

7.3 Ordinary data collection

7.3.1 Experimental design and data collection strategy in experiment sample

In this section we describe the design for the experiment, which started in September 2011, lasted until December 2011, and was conducted in Statistics Sweden's CATI-environment.

The LCS 2011 sample is considered a simple random sample from the Swedish population. Profiting from the experience with LCS 2009, we distributed ("post-stratified") the autumn portion of the LCS 2011 sample, denoted s of size n = 2,108 persons, into five sample groups denoted s_1 to s_5. They were formed with the aid of auxiliary variables, as described in the following. Then each of the five groups was further divided, randomly, into two equal-sized

parts, one experiment part and one control part. We obtain thereby the "experiment sample", consisting of five "call groups", s_{CG1} to s_{CG5}. The call groups are treated differently in the data collection. We can describe the experiment sample as a two-phase sample with stratification of the first-phase sample. The "control sample" also consists of five parts, but these play no role in the data collection; these five parts are put together and are subjected to the standard LCS data collection routine. But the units in the control sample groups can be identified and described in terms of auxiliary characteristics.

Once identified, the call groups s_{CG1} to s_{CG5} were made to differ in regard to features of the data collection, such as the number of call attempts and the time of day at which calls were to be placed. To manage the call schedules, the queue system in WinDATI was overruled. Instead, the call schedule was defined by a paper-based contact form for each unit in the experiment sample. This form specified the time points for the calls to be made. The interviewers were instructed to mark on the contact form the time and the outcome of each call attempt. Consequently, the survey manager had to spend time in coordinating the interviewing, and to make sure that just one call attempt would really happen in the specified time slots. The experiment was limited to five groups, so as to limit the survey manager's task.

The group s_1 was formed by a use of register variables known to be good indicators of over-coverage. Based on earlier experience (see SCB (2010), page 44), it is safe to conclude that a number of sampled persons are not (any longer) residing in the country during the data collection period in question. On good grounds, s_1 is assumed to contain a high proportion of the sample over-coverage, and by singling out this group, the number of unproductive call attempts can be reduced. The remaining four groups, s_2 to s_5, were identified with the aid of the following dichotomous auxiliary variables: *Age* (65 and older; under 65), *Origin* (non-Swedish born; Swedish born), and *Property ownership* (owner; non-owner). The groups were formed sequentially: First s_1 was identified. Then s_2 was formed as those in $s - s_1$ who are non-Swedish born and non-owners. Then s_3 was formed as those in $s - s_1 - s_2$ who are Swedish born and aged 65 and older, or who are non-Swedish born aged 65 and older and

owners. Then s_4 was formed as those in $s - s_1 - s_2 - s_3$ and who are Swedish born younger than 65 and owners. The remainder, of mixed composition, is the group s_5. Table 7.3.1 summarizes the procedure and shows the percentage proportion of each group out of the total sample s.

Table 7.3.1.
Principal characteristics of the five groups, s_1 to s_5, and their proportion (in per cent) out of the whole sample s

s_1	Probable over-coverage [2%]
s_2	All ages, non-Swedish born non-owners [14%]
s_3	Aged 65 and over, Swedish born and non-Swedish born owners [21%]
s_4	Aged up to 64, Swedish born owners [23%]
s_5	Aged up to 64, remaining mixed group [39%]

The breakdown of the entire sample s on the five groups s_1 to s_5, and the breakdown of each group on its control and experiment part, are shown in Table 7.3.2. The call groups, s_{CG1} to s_{CG5}, of the experiment sample are essential for the call scheduling; based on an analysis of the LCS 2009 data, these groups are made to differ in regard to the number of call attempts and the time-slots for making calls.

Earlier experience had suggested that s_{CG3} and s_{CG4} are groups "easy to contact". More particularly, persons in s_{CG3} are usually easy to locate, with a listed number for a fixed landline telephone. However, for age related reasons, some persons in s_{CG3} experience difficulty in participating in a telephone interview. Persons in s_{CG3} and s_{CG4} generally have a positive attitude to the survey and its topic; these are high responding groups, provided contact can be made at a suitable time of day. The group s_{CG5} is considered moderately difficult to enter in contact with; both s_{CG2} and s_{CG5} require particular attention in planning the contact routines.

Table 7.3.2
Distribution of the LCS 2011 autumn sample s on the five groups, and subdivided into control sample and experiment sample

Group	s_1	s_2	s_3	s_4	s_5	Entire s
Control	24	152	225	243	411	1055
Experiment	24	151	225	242	411	1053
Total	48	303	450	485	822	2108

Different call attempts schedules were created for the five call groups s_{CG1} to s_{CG5} that make up the experiment sample. This task was considerably facilitated by the co-operation of the survey manager, whose thorough knowledge of the available interviewer resources was essential for the time slot scheduling. It was decided to apply higher call attempt intensity during the first three data collection weeks, and that units who had shown four non-contact attempts would be inspected and if necessary traced again. The lowest call intensity was applied in s_{CG1} (the suspected over-coverage) and the highest in s_{CG2} (born abroad) and in s_{CG5} (the mixed although moderately difficult group).

It was also considered to start making calls in the evening, between 5 and 9 PM, primarily for s_{CG2} to s_{CG5}. For s_{CG3} it was deemed inappropriate to call after 7 PM, but for practical reasons it was not possible to adhere to this rule for persons 65 and older in s_{CG2}. At present, the LCS does not have a policy to abstain from calling older persons after 7 PM, although it is well known that it is better to avoid late calls for those persons.

Table 7.3.3 shows the call strategies planned for the five call groups s_{CG1} to s_{CG5}. The shaded areas indicate the number of call attempts during the first three field work weeks. For s_{CG2}, nine call attempts are scheduled during the first three weeks, as contrasted with only three attempts for s_{CG1}. For s_{CG2}, s_{CG4} and s_{CG5}, most of the calls are scheduled for evenings and weekends. There are minor differences between the three groups; fewer calls are planned for s_{CG4}. The group s_{CG3} containing persons 65 years or older has lower call intensity during the first three weeks, and the calls are planned for day time or early evening. Note that the groups have different caps

for the number of call attempts; the cap is lowest for s_{CG1} with four attempts, and highest for s_{CG2} and s_{CG5} with 12.

Table 7.3.3.
The experiment sample LCS2011, planned contact strategy in ordinary field work. The columns represent the five call groups: week in ordinary field work, day in week and time for call

	s_{CG1}			s_{CG2}			s_{CG3}			s_{CG4}			s_{CG5}		
Attempt	Week	Day	Time	Week	Day	Time	Week	Day	Time	Week	Day	Time	Week	Day	Time
1	w1	Mo-Th	12-17	w1	Mo-Th	19-21	w1	Mo-Th	17-19	w1	Mo-Th	19-21	w1	Mo-Th	17-19
2	w1	Fri	9-12	w1	Mo-Th	17-19	w1	Fri	9-12	w1	Mo-Th	19-21	w1	Mo-Th	19-21
3	w3	Sat	10-14	w1	Sun	12-17	w2	Mo-Th	17-19	w1	Sun	19-21	w1	Sun	17-19
4	w6	Sun	19-21	w2	Mo-Th	17-19	w3	Sat	11-14	w2	Mo-Th	17-19	w2	Mo-Th	19-21
				Control of unit											
5				w2	Fri	9-12	w3	Sun	17-19	w2	Sun	19-21	w2	Mo-Th	19-21
6				w2	Sun	17-19	w5	Mo-Th	17-19	w3	Mo-Th	17-19	w2	Fri	9-12
7				w3	Mo-Th	19-21	w6	Fri	9-12	w3	Sat	10-14	w3	Mo-Th	19-21
8				w3	Sat	10-14	w7	Sun	12-17	w5	Sun	12-17	w3	Sat	10-14
9				w3	Sun	19-21				w6	Mo-Th	19-21	w3	Sun	19-21
10				w5	Mo-Th	19-21				w7	Sun	17-19	w5	Mo-Th	19-21
11				w6	Sun	19-21							w6	Sun	12-17
12				w7	Mo-Th	19-21							w7	Mo-Th	19-21

7.3.2. Data collection strategy in the control sample.

The ordinary eight week data collection routines, as used earlier in LCS, were followed for the control sample. For that time period, twelve call attempts were planned; the interviewers should in principle spread the calls over suitable time points during week days and weekends. This is however hard to administer, because all interviews are done by the central CATI-group, with interviewers working shifts and under instruction to follow the call schedule device built into the central CATI-system, WinDATI. The ability to spread the twelve calls over time depends on the interviewers' working hours and on the CATI-system. The high pressure on interviewer resources at Statistics Sweden could lead to less than optimal time shifts for the control sample. After four unanswered

calls, it would be desirable to check if the phone number in use is valid or not, but there was no procedure for implementing this.

7.4. The follow-up data collection

As mentioned in Section 7.2, the ordinary data collection resulted in very low data inflow, especially for the control sample. As a consequence the management decided that all follow-up action should focus on improving the overall response rate for the control and experiment samples together. All nonresponding units, those in both samples, were inspected at the end of the ordinary data collection by the survey manager, and only those deemed likely to deliver a response were selected for the follow-up.

All persons in the follow-up portion of the experiment sample were subjected to the same call strategy, consisting in ten calls distributed over predetermined time slots in the three week follow-up period.

The follow-up portion of groups S_2 and S_5 in the control sample was randomly split into two halves. One half was to follow the experiment data collection strategy and the other half was to follow the ordinary LCS follow-up data collection. This set-up, a cross-over experimental design, made it possible to investigate if the experiment interviewers produced higher response than the control interviewers. The reason that only groups S_2 and S_5 were chosen for the cross-over was a particularly low response in the control sample for these groups, see Table 7.6.1. But as mentioned in Section 7.2, the analysis of the cross-over was not carried out because of lack of time.

Incentives were used in the follow-up, for both samples, in an attempt to encourage a response.

7.5 Interviewer allocation for experiment sample and control sample

The data collection was carried out, for both experiment and control sample, by interviewers in the central CATI-group, but they were not the same for the two samples. Before the start, the CATI interviewers were informed that an experiment was planned for the fall interviewing, and those interested in working as experiment interviewers were encouraged to sign up, although at no extra benefits. Thirteen interviewers signed up.

The experiment interviewers were briefly instructed about the purpose of the experiment and about the use of the call schedules specified on the contact form. During the field work, the survey manager was continuously in touch with these interviewers. When an interviewer had no more units to call in the work shift, he or she would switch to other surveys and/or to work with the LCS control sample. The call attempts for the experiment sample were done by the experiment interviewers. At the planning stage it was expected that the experiment interviewers might produce a higher response rate, because they would feel particularly motivated by participating in an experiment. Another hypothesis is that those who signed up were also more skilled or persuasive.

The control sample interviewers worked under the traditional data collection routines (see also Section 7.3.2). They were not allowed to call units in the experiment sample.

7.6 Results based on the actually observed data

Table 7.6.1 shows results of the ordinary field work. Response rates and mean number of call attempts are given for the five control sample groups and for the five experiment sample groups (the call groups) s_{CG1} to s_{CG5}. All groups show higher response rate in the experiment sample. It should be noted that while the control sample interviewers are not allowed to interview call group units, they can obtain data such as name, gender, age and geographic location for all sampled persons. This may influence an interviewer's contact attempt frequency. The tables show that older persons got fewer call attempts than planned, both in the control sample and in the experiment sample. This may reflect a "common sense attitude" on the part of the interviewers, namely that, despite the instructions they receive, they place fewer calls to older persons than what the contact strategy specifies.

The strategy for the experiment sample appears more successful in that call attempts become transferred from relatively high responding groups to more difficult groups, such as s_2, *Non-Swedish born non-owner*, and s_5, *Mixed group*. These have higher average number of call attempts in the experiment sample than in the control sample. Also, the use of register information had the positive effect of reducing the number of calls in the group s_1, *Suspected over-coverage*. The number of units in this group is however very small.

A logistic regression was carried out with the response indicator as the dependent variable and a number of explanatory variables: Paradata such as call group membership, number of calls, and starting week and auxiliary variables such as property ownership, education, country of origin, age, civil status and gender. This analysis showed that the experiment sample produced a significantly higher response rate at the one per cent level. The mean number of call attempts was practically the same in the experiment sample and in the control sample.

It is important to note that the results suggest possibility to realize a higher response rate in the LCS without increasing the number of calls.

Table 7.6.1
Ordinary field work: Response rate in per cent and average number of call attempts for the five groups of control and experiment sample. The groups s_1 to s_5 are explained in Table 7.3.1

Group		Control	Experiment
S_1	Response	29%	33%
	Call attempts	6.5	3.7
S_2	Response	27%	40%
	Call attempts	6.4	6.8
S_3	Response	51%	56%
	Call attempts	4.4	4.0
S_4	Response	48%	60%
	Call attempts	6.6	5.7
S_5	Response	42%	47%
	Call attempts	6.5	6.9
Total	Response	43%	50%
	Call attempts	6.1	5.9

Table 7.6.2
Total field work (ordinary and follow-up): Response rate in per cent and average number of call attempts for the five groups of control and experiment sample. The groups s_1 to s_5 are explained in Table 7.3.1

Group		Control	Experiment
S_1	Response	33%	38%
	Call attempts	7.8	5.4
S_2	Response	43%	50%
	Call attempts	9.7	9.0
S_3	Response	64%	62%
	Call attempts	5.5	5.1
S_4	Response	56%	72%
	Call attempts	8.2	8.2
S_5	Response	54%	60%
	Call attempts	8.8	9.9
Total	Response	55%	61%
	Call attempts	8.1	8.2

The total field work (ordinary and follow-up) is portrayed in Table 7.6.2. A comparison of the experiment sample and the control sample is more difficult to interpret than for the ordinary field work alone, as in Table 7.6.1. Initially, the idea with the experiment sample was to intervene in the data collection to boost the response in groups with particularly low response, but this idea was abandoned. As Section 7.4 describes, the follow-up cases for the control sample were divided into two parts, where the smaller random was given the same treatment as the experiment sample. Nevertheless, the bulk of the response is collected during the ordinary field work, where the experimental plan was in effect.

The overall response rate is significantly higher in the experiment sample, and there is no significant difference for the average number of call attempts.

For all groups but S_3, Table 7.6.2 shows a higher response rate for the experiment sample. Although essentially the same number of calls was placed in the control sample and in the experiment sample, the latter yielded a six per cent higher overall response rate, 61% as compared with 55%. A future task would be to analyze the results in Tables 7.6.1 and 7.6.2 in more depth.

We now compare the two samples in regard to (i) the relative difference RDF (for selected pseudo y-variables), (ii) the balance BI_1 of the response set, and (iii) the distance $dist_{r|nr}$ between respondents and non-respondents. Do those indicators show more favourable values for the experiment sample than for the control sample?

The **x**-vector used in the computations for Tables 7.6.3 and 7.7.1 is

$$\mathbf{x}_a = (S_1 + \ldots + S_5 + Educ + Civil + Gender) \qquad (7.6.1)$$

of dimension 5+1+1+1 = 8, where S_i, for i = 1,…,5, is the zero/one indicator of membership in group i, and *Educ*, *Civil* and *Gender* are the dichotomous variables defined in Section 3.1. The vector (7.6.1) does not include *Phone*, which was unavailable at the time of the analysis, but includes instead the group indicators.

First we examine how the measures change during the field work, for each of the two samples. The relative differences (in percent) $RDF(\hat{Y}_{CAL})$ and $RDF(\hat{Y}_{EXP})$ are shown in Table 7.6.3 for two pseudo y-variables (register variables): *Income* (continuous) and *Employment* (dichotomous), see Section 3.1. The register variable *Benefits*, used earlier for LCS 2009, was unavailable at the time of the analysis. Table 7.6.3 also shows the balance measure BI_1 and the distance measure $dist_{r|nr}$.

As Table 7.6.3 shows, these measures progress, over the series of call attempts, in a manner that is, on the whole, more satisfactory for the control sample. At the end of the ordinary field work, both balance and distance show slightly better values in the control sample, and the contrast with the experiment sample is even more pronounced for the *RDF*'s; the control sample shows $\hat{Y}_{CAL}$ and $\hat{Y}_{EXP}$ as clearly closer to the unbiased $\hat{Y}_{FUL}$ for both *Income* and *Employment*. In the follow-up, balance and distance were slightly improved for the control sample, but this did not happen in the experiment sample, where instead they deteriorated. $RDF(\hat{Y}_{CAL})$ progressed in the wrong direction (getting further away from zero) for both samples and for both pseudo y-variables. This raises serious questions about the efficiency of the follow-up.

It is difficult to explain why the experiment sample performs rather poorly in comparison with the control sample, when the opposite was expected. In the ordinary data collection, the experiment sample

interviewers followed a regimented contact strategy, while those for the control sample were not allowed to prioritize specific sampled units; instead their calling order was determined by the queuing system in WinDATI. (This does not apply to the follow-up, where the data collection routines differed negligibly between the two samples.) Our hypothesis is that in the ordinary data collection, the control sample procedure gives a "more random" set of respondents than the experiment sample procedure, something which could explain why the control sample shows more favorable values on most of the computed measures.

Table 7.6.3
The LCS 2011 data collection: progression of *RDF* (for Income and Employment) BI_1 and $dist_{r|nr}$. Computations based on the 8-dimensional vector (7.6.1)

Experiment sample

		Income		*Employment*			
Attempt #	$100\times P$	$RDF(\hat{Y}_{CAL})$	$RDF(\hat{Y}_{EXP})$	$RDF(\hat{Y}_{CAL})$	$RDF(\hat{Y}_{EXP})$	BI_1	$dist_{r\|nr}$
1	6.0	1.3	6.3	3.5	-9.2	0.900	0.422
2	18.6	0.5	15.0	1.8	5.1	0.815	0.475
3	25.5	-0.9	13.7	2.6	5.3	0.792	0.478
8	44.6	0.7	11.2	5.1	8.4	0.785	0.432
15	49.8	1.4	11.4	7.3	10.7	0.778	0.443
End Ordinary	50.3	1.3	11.4	8.0	11.2	0.773	0.453
Follow-up							
1	52.0	1.5	11.2	8.0	11.4	0.775	0.450
5	58.2	3.0	10.4	7.8	11.1	0.788	0.430
10	60.2	2.7	9.9	9.1	12.8	0.777	0.455
Final	61.4	2.2	9.3	9.9	13.8	0.771	0.470

Control sample

		Income		*Employment*			
Attempt #	$100\times P$	$RDF(\hat{Y}_{CAL})$	$RDF(\hat{Y}_{EXP})$	$RDF(\hat{Y}_{CAL})$	$RDF(\hat{Y}_{EXP})$	BI_1	$dist_{r\|nr}$
1	7.4	-9.9	-2.7	-12.9	-18.6	0.862	0.526
2	16.2	-5.4	-1.7	-2.7	-6.1	0.850	0.406
3	24.0	-3.4	3.9	0.0	0.3	0.811	0.442
8	38.2	-1.2	7.4	0.3	3.3	0.787	0.438
15	42.3	-2.1	5.4	-0.3	3.2	0.779	0.448
End Ordinary	42.7	-0.7	6.0	0.1	3.3	0.788	0.428
Follow-up							
1	44.9	1.8	9.0	-0.6	2.5	0.785	0.433
5	52.9	3.3	9.2	3.3	4.5	0.787	0.427
10	54.1	3.2	8.4	4.4	4.9	0.801	0.399
Final	54.5	3.2	8.3	4.7	5.2	0.803	0.396

7.7 Results based on interventions in retrospect

The analysis in this section serves to demonstrate, by an "analysis in retrospect", what could have happened if we had stopped data collection attempts for units having attained a high response propensity at the end of the ordinary field work. This method for "constructing" a final response set is explained in Section 3, and illustrations are given in Section 4.2. When applied after the end the ordinary field work, it is easy to build this technique into the LCS data collection. It should be noted that the sample size is not very large here, something which may affect the response propensity scores and the subsequent analysis.

We present in Table 7.7.1 (the last two lines) two illustrations of interventions in retrospect for the experiment sample. These are compared in Table 7.7.1 (the first two lines) with the results, given in Section 7.6 for the two samples, at the end of the actual data collection (which is without any interventions).

Both cases of intervention start from the LCS 2009 data, as they are at the end of the ordinary field work. We pretend that contact attempts are stopped at the end of the ordinary data collection for units with high response propensity at that point, and that the follow-up focuses only on units with low response propensity, calculated with the auxiliary vector (7.6.1),

$$\hat{P}_{ak} = (\sum_s d_k I_k \mathbf{x}_{ak})'(\sum_s d_k \mathbf{x}_{ak}\mathbf{x}'_{ak})^{-1}\mathbf{x}_{ak}$$

Hence there is one and only one intervention point, namely at the end of the ordinary data collection. Note that the procedure, as used here, will necessarily result in a lower final overall response rate. Given that the survey outcome is fixed, we cannot attempt to redirect the resources saved to more intensive contact attempts for units with low response propensity; this is what might happen in a real application. Our purpose is only to see if it is possible through this simple intervention to get better balance, reduced distance, and a lower risk of bias.

The first intervention strategy, labeled "Stopping rule 60%", consists in dropping, at the end of the ordinary field work, units with response propensity $\hat{P}_{ak} > 0.60$. The second intervention strategy, labeled "Stopping rule 48.7% (median)", uses the median response propensity, 0.487, computed at the end of the ordinary field work. That is, units with $\hat{P}_{ak} > 0.487$ are dropped for the follow-up. One

can describe the first illustration as more conservative, the second as more risky. Both are subjective choices, serving here only the purpose of illustration.

The primary interest in Table 7.7.1 lies in comparing the two actual data collections (the first two lines; taken from Table 7.6.3) with the two collections with intervention (lines three and four). The measures shown in the table are the relative difference $RDF(\hat{Y}_{CAL})$ (for the pseudo *y*-variables *Income* and *Employment*), the balance BI_1 of the final response set, and the final value of the distance $dist_{r|nr}$ between respondents and non-respondents.

Table 7.7.1
Comparison of actual data collection (Control and Experiment sample) with intervention at the end of ordinary field work for Experiment sample. Computations based on auxiliary vector $\mathbf{x}_a$ given by (7.6.1)

		Income	*Employment*			
Field work	$100\times P$	$RDF(\hat{Y}_{CAL})$	$RDF(\hat{Y}_{CAL})$	BI_1	$dist_{r\|nr}$	**Number of call attempts**
Actual data collection; from Table 7.6.3						
Control sample	54.5	3.2	4.7	0.803	0.393	8531
Experiment sample	61.4	2.2	9.9	0.771	0.470	8667
Examples of intervention after ordinary field work, Experiment sample						
Stopping rule: 60% response propensity	59.5	2.5	10.1	0.818	0.371	8380
Stopping rule: 48.7% [median] response propensity	56.7	2.5	10.1	0.850	0.302	7792

Table 7.7.1 shows that the intervention considerably improves both balance and distance for the experiment sample. The balance is higher and the distance is lower, despite lower response rates, compared with the actual data collection. This is noted already for the conservative 60% stopping rule (line three); the riskier stopping rule (line four) yields even better balance and even lower distance.

Similar interventions computed for the control sample also showed that they lead to higher balance and lower distance.

Turning to $RDF(\hat{Y}_{CAL})$ we see in Table 7.7.1 that the intervention does not give clear improvement. This can of course happen; it depends on the particular pseudo *y*-variables considered. Compared with the actual experiment sample, both *RDF*'s are slightly higher after intervention. That they are higher in comparison with the control sample is not surprising, in view of the discussion in Section 7.6.

The results in Section 7.6 indicated that the experiment call strategy produces a markedly different response set than the control sample data collection. For example, the data show that the control sample achieves high response for the group s_3 of older persons, whose income and rate of employment is usually lower than for people in the 34 to 64 age bracket. This has an impact on the results. In the auxiliary vector (7.6.1) used here for computation, the variable *Age* is confounded in the groups s_1 to s_5. A different choice of auxiliary vector might produce substantially better results for the experiment sample *RDF*'s.

The results also suggest that a sole intervention at the end of the ordinary field work may be too little and too late. We believe that interventions during the ordinary field work would have shown a different evolution, with better final values in the measures. Another factor is that they are computed here on samples of an approximate size of only 1,000. In the LCS 2009, the computations rest on a sample eight times greater, something which surely affects the estimates and thereby the *RDF*'s. It is not clear if for example the value 9.9 of $RDF(\hat{Y}_{CAL})$ for the experiment sample is significantly different from the corresponding value 4.7 for the control sample.

7.8 Conclusions.

In conclusion we remark the following:

- The experiment sample data collection gives mixed results. The higher response rate in the experiment sample is not accompanied by clearly better values on the indicators, balance, distance and *RDF*'s; see Table 7.6.3.
- Test statistics need to be developed for balance, distance and *RDF*, to make it possible to test if a data collection other than "a standard one" is significantly better. A test statistic for the *RDF* may be relatively easy to develop if the full sample estimate $\hat{Y}_{FUL}$ is regarded as a fixed constant. For the balance and distance indicators, the construction of test statistics is more demanding. One solution is to use the *R*-indicator see *The Program Cockpit* found at http://www.risq-project.eu/tools.html.
- As the results in Section 7.7 suggest, it may be "too little, too late" to wait until the follow-up to apply an intervention. The possibility of interventions during the ordinary field work needs to be investigated.
- For the particular x-vector (7.6.1), the calibration estimator gave better estimates than the expansion estimator, both for the experiment and the control sample, see Table 7.6.3. But alternative x-vectors should be tried, because the $\hat{Y}_{CAL}$ estimates depend on the vector specification, and so do the corresponding *RDF*'s.
- The data collection proposed for the experiment sample is probably similar to how the field interviewers approach their work, whereas the central CATI-group (which handled the control sample) works in ways which seem to give a more random response set and therefore more favorable values for the indicators ; this holds at least for the ordinary data collection (before the follow-up).
- The contact strategy for the experiment outlined in Table 7.3.3 is impractical for data collection in a typical survey at Statistics Sweden. However, a modified version might be implemented in the CATI-system. To prioritize cases by assigning them to different points in time is important in Statistics Sweden's continuing efforts to improve the survey response rates. Results in this study indicated that time of call was important for increasing response, and conversely that inconvenient timing increased the nonresponse.

- We conclude that more experimentation is needed to find ways to reduce the risk of bias.
- The interviewer effects, if any, should be studied. However, our preliminary conclusion is that the time of call has a stronger impact on the response rate than the interviewer differences.

8 Discussion

In this R&D report we have wanted to present the main features and results of the second phase of a project on Responsive Design at Statistics Sweden. Results of the first phase were reported in LS (2012).

A key concept in the report is *indicators*. They mirror different aspects of the nonresponse problem and are useful in guiding the data collection. They are complements to the response rate, which is in itself insufficient for describing the quality of the survey response.

Our indicators are general and designed for use in probability sampling surveys, where the inclusion probabilities are known, and several auxiliary variables are available. The report shows how to use the indicators in monitoring the data collection in already established surveys and also shows how they might be used in embedded experiments, where new types of data collection are tried.

Relevant recent articles dealing with the nonresponse problem, with an emphasis on the data collection phase, are for example Schouten et al. (2012) and Wagner (2012). The former gives a broad description of uses of indicators in connection with nonresponse error, including an overview of the *R*-indicator and the work accomplished in the RISQ-project. Wagner (2012) sets up a typology of data sources with three levels; our approach fits best in the level "the response indicator and frame data/paradata."

Beginning in Section 3, we use the important term *balance* to refer to the equality of respondent mean and full sample mean, for specified variables. In practice we can measure the balance on the chosen auxiliary vector (called **x**-vector), because the values of all variables in this vector are known for responding units as well as for nonresponding units.

By contrast, balance cannot in practice be ascertained for study variables (real *y*-variables). As equation (3.1.1) specifies, the response set *r* is perfectly balanced on the study variable *y* if response and full sample means agree, so that $\bar{y}_r = \bar{y}_s$. This is a conceptual definition. Nonresponse makes perfect balance on a real study variable impossible, because *y* is observed only for the responding units.

But in methodological study, the question of balance on *y*-variables can be examined for selected register variables (pseudo *y*-variables), available for all units in the selected sample s. This approach is also used in the report.

It is important is to seek balance on a suitable auxiliary vector. As Section 3 points out, if the response is balanced for a vector $\mathbf{x}_k$ highly related to the study variable *y*, then even the primitive expansion estimator is close to unbiased. More specifically, the response set r must satisfy two conditions to realize the perfect *y*-variable balance $\bar{y}_r - \bar{y}_s = 0$: (*i*) $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ (balance on the chosen **x**-vector) and (*ii*) $\mathbf{b}_r = \mathbf{b}_s$ (equality of the regression vectors for response set and full sample). In the data collection we can attempt to come close to realizing (*i*). But this does not imply that condition (*ii*) is close to being fulfilled, so *y*-variable balance may at best be partially realized. This reflects the well known fact that nonresponse bias is never completely eliminated, but at best to a degree. Further theoretical and empirical work around these questions - the importance of (*i*) in relation to (*ii*) - is being planned.

The traditional view of auxiliary variables holds that they are instruments for nonresponse adjustment at the estimation stage, after a completed data collection. To use the auxiliary variables to inform and modify the data collection is a relatively recent development, inspired by the literature on responsive design. Today, not much auxiliary information finds use in the data collection, regrettably, since such information may be abundant, as it is at least in Scandinavia.

A motivation for an increased and systematic use of auxiliary information in the data collection phase is that the production can be streamlined, and a better control of the field work becomes possible. The data inflow can be continuously studied, and its emphasis can be altered, for example by focusing on low-responding subgroups, identified with the aid of selected auxiliary variables. These variables form what we refer to as the *monitoring* vector.

The choice of auxiliary variables depends on the purpose. For example, if the objective is to improve the balance with respect to designated sample groups, then one selection of auxiliary variables may be preferred, whereas other auxiliary variables might be preferred if the objective is instead to reduce the standard errors of subpopulation (domain) estimates.

All available auxiliary variables need not or should perhaps not be used for the monitoring function. Auxiliary variables not chosen for the monitoring are included instead in a vector that we refer to as the *supplementary* vector.

The monitoring vector is an important instrument for the data collection, and the supplement vector becomes important at the estimation stage, where variables from both vectors may be used in computing the calibrated weights for the estimates, with an objective to control both bias and variance.

Two different ways of monitoring the field work are outlined and investigated in Section 3. We call the first approach *analysis of imbalance* (ANIMB), a name that suggests similarities with a traditional analysis of variance. The concepts *conditional imbalance* and *partial conditional imbalance* arise naturally through the ANIMB. Although different both in derivation and in empirical aspects, these concepts have some resemblance to *conditional representativity* and *conditional partial representativity*, as used in Bethlehem et al. (2011) and in the RISQ project.

The second approach uses *response propensities* for interventions during the data collection period. The *response propensity method* can be applied in the field work with little extra effort. It can be based on some or all of the available auxiliary variables.

Both methods, the ANIMB and the response propensity method, are illustrated empirically in Section 4, on LCS 2009 data and with different specifications of the auxiliary vector. Although they are computationally simple, their implementation in the production at Statistics Sweden requires further work.

In Section 5 we study the use of the indicators in estimation for domains. Five age group domains were used for this illustration. We find that the indicators are less stable (have more variability) for the domains than for the entire sample, not entirely surprising because of fewer data in the domains. We examine the development of the domain mean estimates as the data collection proceeds, in particular how and when the means stabilize, as more and more respondent data enter. With the coefficient of variation (the *cv*) we track the changes in the measure of precision. But although both are desirable, neither stabilized means nor low *cv*'s will give a clear message about the crucial problem with nonresponse, that is, the bias. The estimates can be stable and their *cv*'s satisfactorily low, but the bias nevertheless high.

Section 5 emphasizes that an examination of register variables (and their estimates) can contribute much to an understanding of inaccuracy caused by nonresponse. But the degree to which register variables can help in forecasting the size and the direction of the bias for real study variables is a topic that deserves more in-depth study.

Section 6 shows a small scale application of the existing SAS-program for computing balance and associated measures in a different survey setting, namely, the Swedish part of the multinational PIAAC survey. This exercise shows that the program code is sufficiently general to allow application in other survey conditions. The indicators and measures are fairly simple to code, making possible a SAS tool of general scope.

Most of this paper focuses on methods for an already established data collection strategy. But when we use our methods to intervene in the data collection (in an established survey or in a new survey), good statistical practice requires significance testing to see if an improvement in balance or other characteristics is really accomplished. Such significance tests need to be further developed.

Section 7 reports results of a carefully designed embedded experiment, in which a new contact strategy is contrasted with the traditional contact strategy in the LCS. The use of our measures facilitate and enrich the evaluation of the new contact strategy; despite a six per cent higher overall response rate with the new strategy, which is in itself very encouraging, the indicators show however that there is no clear improvement in balance, distance or *RDF*. The experiment also shows how difficult it can be to complete a carefully planned alternative data collection in harmony with traditional objectives firmly entrenched in a large statistical organization.

Based on the findings in Section 7, a small experiment was carried out on the 2012 version of the LCS. In particular, the response propensity method was used in the follow-up part of the data collection. The results of this test will be analyzed in Spring 2013.

References

Andersson, C. (2012). ETOS User's guide. Internal Document, Statistics Sweden.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley Series in Survey Methodology.

Calienscu, M., Bhulai, S. and Schouten, B. (2011). Optimal scheduling of contact attempts in mixed-mode surveys. Discussion paper 201131, Statistics Netherlands, The Hauge, 2011.

Dubin, J.A. and Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods and Research*, **18**, 360-390.

Groves. R.M. and Heeringa. S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A*. **169**, 439-457..

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153-162.

Kass, G.V. (1980). An explanatory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, 119-127.

Laflamme. F. (2009) Experinces in assessing. monitoring and controlling survey productivity and costs at Statistics Canada. Proceedings of the 57th Session of the International Statistical Institute, South Africa.

Lundquist, P. and Särndal, C.E. (2012), *Aspects of Responsive Design for the Swedish Living Conditions Survey*. R&D report 2012/1, Statistics Sweden. http://www.scb.se/statistik/_publikationer/OV9999_2012A01_BR_X103BR1201.pdf

Mohl, C. and Laflamme, F. (2007). Research and responsive design options for survey data collection at Statistics Canada. *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

SCB (2010). The overcoverage in the Total Population Register. Background Facts, *Population and Welfare Statistics 2010:5*, Statistics Sweden [In Swedish]

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N. and Skinner, C. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, **80**, 382-399.

Särndal. C.E. and Lundström. S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics*. **24**. 251-260.

Särndal. C.E. and Lundström. S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*. **36**. 131-144.

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

Wagner, J. (2012). Research synthesis: A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly*, **76**, 555-575.

ISSN 1653-7149 (Online)

All official statistics can be found at: www.scb.se
Statistics Service, phone +46 8 506 948 01