Research and Development – Methodology reports from Statistics Sweden 2012:1

Statistiska centralbyrån Statistics Sweden

Aspects of Responsive Design for the Swedish Living Conditions Survey

Peter Lundquist
Carl-Erik Särndal

The series entitled "**Research and Development – Methodology Reports from Statistics Sweden**" presents results from research activities within Statistics Sweden. The focus of the series is on development of methods and techniques for statistics production. Contributions from all departments of Statistics Sweden are published and papers can deal with a wide variety of methodological issues.

Previous publication:

2006:1 Quantifying the quality of macroeconomic variables

2006:2 Stochastic population projections for Sweden

2007:1 Jämförelse av röjanderiskmått för tabeller

2007:2 Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator.

2007:3 Kartläggning av felkällor för bättre aktualitet

2008:1 Optimalt antal kontaktförsök i en telefonundersökning

2009:1 Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

2009:2 Demographic methods for the statistical office

2011:1 Three Factors to Signal Nonresponse Bias – With applications to Categorical Auxiliary Variables

2011:2 Quality assessment of administrative data

Research and Development – Methodology reports from Statistics Sweden 2012:1

Aspects of Responsive Design for the Swedish Living Conditions Survey

Peter Lundquist
Carl-Erik Särndal

Statistiska centralbyrån
2012

Research and Development – Methodology reports from Statistics Sweden 2012:1

Aspects of Responsive Design for the Swedish Living Conditions Survey

Statistics Sweden
2012

Producer

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

Enquiries

Peter Lundquist, +46 8 506 949 18
peter.lundquist@scb.se

Cover Ateljén, SCB

ISSN 1653-7149 (online)
URN:NBN:SE:SCB-2012-X103BR1201_pdf (pdf)

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

Preface

Responsive design is a newly emerged view focusing on the possibilities to reduce the effects of nonresponse by monitoring the data collection process. This paper contributes with a study on the potentials of reducing bias and costs in the Swedish LCS by applying responsive design techniques. The paper adds to the quality improvement efforts made by Statistics Sweden.

Statistics Sweden, January 2012

Lilli Japec

Acknowledgment
The authors gratefully acknowledge the cooperation of colleagues at Statistics Sweden in preparing the data files used in this research.

Contents

Preface....3
Contents....5

Summary....7

1 Introduction and background....9

2 Earlier experiences at Statistics Sweden....11

3 The Swedish Living Conditions Survey....13

4 Analysis of the LCS 2009 data....15

5 Additional tools: indicators of balance, distance and representativity, computed on LCS 2009 data....21

6 Experimental data collection strategies derived from LCS 2009 data....29
6.1 Experimental strategies....29
6.2 The LCS 2009 data analyzed with the experimental vector....32
6.3 Experimental strategy 1 and its results....36
6.4 Experimental strategy 2 and its results....39
6.5 Experimental strategy 3 and its results....41
6.6 The experimental strategies compared with the actual LCS 2009 data collection....43

7 Discussion and implications for the future....45

References....49

Summary

High nonresponse is forcing Statistics Sweden to find new ways of designing and controlling the data collection in its surveys. The nonresponse rate is by itself an insufficient guide. More informative measures of the progress of the data collection have recently been proposed. These include balance indicators and *R*-indicators (where *R* stands for representativity). We use these and other indicators in conjunction with process data from the Swedish CATI system to study and monitor the progress of the data collection in the Swedish Living Conditions Survey (LCS). Results from earlier studies at Statistics Sweden are confirmed, namely, that the traditional way of LCS data collection is not as efficient as it should be; revision is needed to realize better efficiency and reduced cost. The report also uses the 2009 LCS data to formulate several "experiments in retrospect". The results show that it is possible to improve the data collection, through interventions in the data inflow at suitably chosen points. As a result, survey cost can be reduced, the ultimate set of respondents will be better balanced, and the estimates become less biased. The proposed indicators are general, making them suitable for inspecting the data collection in other surveys as well.

1 Introduction and background

Large nonresponse is typical of many sample surveys today. This can be a serious detriment to survey quality. Nonresponse causes systematic error, usually called *bias*, in the survey estimates. The purpose of this paper is to define and apply new tools, in the spirit of responsive design, to the Swedish Survey of Living Conditions (LCS), so as to improve the data collection for this important survey that has become affected by high nonresponse in recent years.

An extensive literature is devoted to nonresponse and its consequences. In dealing with survey nonresponse, statisticians need to consider (a) measures to be taken at the data collection stage, and (b) measures to be taken at the estimation stage.

The nonresponse rate measures one aspect of the data collection. However, it has become increasingly clear that the nonresponse rate is not by itself a suitable, or not a sufficient, measure for monitoring the data collection. For example, it may be wasteful to continue a data collection according to an unchanging scenario driven primarily by the desire to obtain the highest possible response rate in the end, or to reach, by an expensive and unrelenting effort, a preset rate of response, such as for example 70%.

Instead, at data collection stage, different forms of *responsive design* may be used. The general objectives of responsive design are formulated in Groves and Heeringa (2006). They use the term "phase capacity" for "the stable condition of an estimate in a specific design phase". When phase capacity has been reached in a given phase, it is no longer effective to continue data collection in the same mode or phase; there is an incentive to modify the design, if data collection is to be at all continued. Options for responsive design in a Canadian setting are discussed in Mohl and Laflamme (2007) and Laflamme (2009).

Responsive design may take different forms. Recent literature suggests that rather than narrowly fixing attention on the response rate, one should strive to obtain an ultimate set of respondents with favourable and measurable characteristics. For example, one may, especially in the later stages of the data collection, intervene and try to achieve an ultimate response set that is "better balanced" or "more representative" than if no special effort is made. Those

concepts must first be defined, and then quantified thorough indicators. Such indicators have recently been proposed, and they can now be used to monitor the data collection and to implement changes deemed advantageous.

The 7th EU Framework Programme funded a project called RISQ, which stands for Representativity Indicators for Survey Quality (see for example Schouten and Bethlehem 2009). An objective of that project was to develop and study indicators for the *representativity* of survey response. An important use of the *R*-indicator (for Representativity Indicator) proposed by Schouten, Cobben, and Bethlehem (2009) is to help comparing different surveys - the same survey in different countries, or different surveys within the same country - with respect to the representativity of the final set of respondents. The statistical concept behind the *R*-indicator is the variance of the response probabilities, estimated with the aid of auxiliary variables. The underlying motivation is that a small variability of such estimates would suggest a "representative set of respondents."

Indicators based on instead the concept of a *balanced response set* were developed in Särndal (2011a). The response set is said to be balanced if the means for a number of important auxiliary variables are the same or almost the same for the respondents as for all those selected in the probability sample. In other words, on average, the respondents are like all those sampled. These balance indicators are computable from the auxiliary variable values available for responding as well as for nonresponding units.

When the data collection is terminated, the estimation stage begins. The statistician's task is then to provide estimates that are adjusted for the nonresponse bias that affects the survey estimates, despite the efforts made at the data collection stage. The objective is to achieve the best possible reduction of a nonresponse bias that can never be completely eliminated. This is done by adjustment weighting, based on selected auxiliary variables. Nonresponse weighting adjustment techniques have been studied, theoretically and empirically, in a number of publications from Statistics Sweden, including Särndal and Lundström (2005, 2008, 2010), Särndal (2011b). There is no reason here to go into these techniques; they lie outside the primary scope of this report, which is devoted primarily to the data collection stage.

2 Earlier experiences at Statistics Sweden

It has become more and more clear that efforts motivated principally by a desire to get the best possible ultimate rate of response are inefficient, or even wasteful. A number of studies at Statistics Sweden illustrate this. They show that Statistics Sweden is spending resources on efforts that have little or no effect on the estimates or on the representativity of the final set of respondents. Unfortunately, little or no information is available to show how much this unproductive work is costing the agency. We give a brief review of these studies which involve telephone interviewing of individuals drawn (by probability sampling) from the Swedish Register of Total Population.

For the November 2002 edition of the Swedish Labour Force Survey (LFS), Japec and Hörngren (2005) compute the estimates after each contact attempt. An administrative data file variable plays the role of a study variable (y-variable); for this variable, the relative nonresponse bias can be computed. These authors find that the estimates change very little after the fifth contact attempt. They conclude "we found that a less elaborate fieldwork strategy, with four call attempts instead of twelve, could reduce the monthly cost in the LFS for call attempts by about 16,000 to 42,000 Euros."

The work by Japec and Hörngren (2005) was followed up in a new project with focus on contact strategies. The report by Lundquist et al. (2007) examines the data collection over time in the Swedish LFS. Within the same project, Westling (2008) presents a study along the same lines for the Household Finances (HF) survey. Estimates were computed successively over the data collection period, as a function of the number of call attempts identified by "WD-events," which are events registered by the data collection instrument WinDATI. In both studies, the study variables are register variables, that is, their values are available for respondents as well as for nonrespondents. Both studies show that the simple expansion estimator (the mean of units having responded up until that moment) stabilizes at an early stage in the data collection: After about 10 call attempts, the estimates change very little. Since the total number of call attempts for a sampled person may exceed 20, there is strong indication that

resources are wasted. For the HF survey, Westling (2008) also examines calibration estimates, where the calibration is carried out on selected auxiliary variables. These estimates stabilize even sooner, at around five call attempts.

Lundquist (2008) examines the nonresponse in the Living Conditions Survey (LCS). With the aid of two different indicators constructed as functions of an auxiliary vector, the representativity of the response set is examined as the data collection procedure unfolds. These indicators are found to change very little after a relatively early point in the collection, suggesting that the response set fails to become more similar to the selected sample. In addition, the follow-up (the field work after the ordinary data collection) appears to have little effect on the estimates.

The effect of a follow-up strategy for the HF survey is studied in Petric (2009). Low response rates had been observed in the primary data collection for several groups expected to have high impact on the nonresponse error. However, Petric (2009) finds that the ensuing follow-up has little effect on the estimates and that follow-up respondents are not the ones that influence the nonresponse error the most. At the end of the follow-up, the response rate remains disappointingly low for groups that were underrepresented already in the ordinary data collection.

3 The Swedish Living Conditions Survey

The Swedish Living Conditions Survey (LCS) is a sample survey designed to measure different aspects of social welfare in Sweden, in particular among different groups in the population. 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 completed for the sampled person 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 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 of the kind where for example the follow-up would focus on underrepresented groups, in an objective to reduce nonresponse bias.

4 Analysis of the LCS 2009 data

The results reported in this section reinforce the impression from the studies described in Section 2 for other surveys at Statistics Sweden, namely, that a data collection (including a follow-up) which proceeds according to an essentially unchanging format will produce very little change in the estimates, beyond a certain "stability point" reached quite early in the data collection.

Table 4.1 shows the progression of the population total estimates for three variables. These are register variables, used here as study variables. Their values are known for all sampled units, not only for responding units. In Table 4.1 we can follow the progress of the three estimates as a function of the number of the call attempt at which an interviewer made contact with a sampled person and data delivery occurred.

We need some notation. The finite population $U = \{1,...,k,...,N\}$ consists of N units indexed $k = 1,2,...,N$. A probability sample s is drawn from U; in this sampling procedure, unit k has the known inclusion probability $\pi_k = \Pr(k \in s) > 0$, and the known design weight $d_k = 1/\pi_k$. We denote by y_k the value of the study variable y. In the hypothetical case of full response, we would know the value y_k for all units $k \in s$. For the study variable y, the target parameter for estimation is 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 $\sum_A$.) Normally, the survey involves many study variables.

The set of units having responded at a certain point in the data collection is denoted r. The study value y_k is recorded for the units $k \in r$; these y-values (and auxiliary variable values) are material for estimating $Y = \sum_U y_k$. Here we follow the data collection as a function of the call attempt number. There is a series of successively larger response sets. For a completely rigorous notation, we could denote these increasingly large response sets as $r^{(a)}$, where a refers to "call attempt number", $a = 1, 2, \ldots$, and

$$r^{(1)} \subseteq r^{(2)} \subseteq \ldots \subseteq r^{(a)} \subseteq \ldots \tag{4.1}$$

But in order to not burden the notation, it is sufficiently clear to let the notation r refer to any one of the increasingly larger response sets. Data collection stops before r has reached the full probability sample *s*.

The (sample design-weighted) survey response rate is

$$P = \sum_r d_k / \sum_s d_k \tag{4.2}$$

The unknown response probability of unit *k* is denoted $\theta_k = \Pr(k \in r | k \in s)$. 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, $\overline{\theta}_U = \sum_U \theta_k / N$.

Auxiliary information plays an important role. We denote by $\mathbf{x}_k$ the auxiliary vector value for unit *k*, assumed available at least for all units $k \in s$, possibly for all $k \in U$. If $J \geq 1$ auxiliary variables are 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 consider auxiliary vectors $\mathbf{x}_k$ of the form such that for some constant vector $\boldsymbol{\mu}$ we have $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all *k*. It is not a major restriction. Vectors of importance in practice are usually of this kind, as when $\mathbf{x}_k = (1, x_k)$ and $\boldsymbol{\mu} = (1, 0)$.

For a given response set *r*, we compute estimation weights calibrated on the specified auxiliary information. The weight given to the value y_k observed for $k \in r$ is $d_k m_k$, the product of the sampling weight $d_k = 1/\pi_k$ and the adjustment factor

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

Hence the calibration estimator is

$$\hat{Y}_{CAL} = \sum_r d_k m_k y_k \tag{4.3}$$

The weights $d_k m_k$ conform to unbiased estimation expressed on the right hand side of the calibration equation

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

Hence the weights $d_k m_k$ give unbiased estimates for the variables present in the auxiliary vector. Here we need not go into the theory behind calibration; it is presented in for example Särndal and Lundström (2005). Calibration will generally reduce the

nonresponse bias, perhaps considerably if the auxiliary vector is powerful, but without eliminating it entirely. Some bias always remains. At Statistics Sweden, many potential auxiliary variables are typically available at the estimation stage. The question then arises as to how one should choose the best among those. Indicators to guide the construction of a powerful auxiliary vector are presented Särndal and Lundström (2008, 2010).

The adjustment factor m_k in $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$ is based on a fixed auxiliary vector $\mathbf{x}_k$ considered suitable for monitoring the estimates over the course of the data collection. We use a vector of dimension eight composed of the following categorical auxiliary variables: *Phone access* (equaling 1 for a person with accessible phone number; 0 otherwise), *Education level* (equaling 1 if high; 0 otherwise), *Age group* (four zero/one coded groups; age brackets -24, 25-64, 65-74, 75+ years); *Property ownership* (equaling 1 for a property owner; 0 otherwise); *Country of origin* (equaling 1 if born in Sweden; 0 otherwise). We refer to this vector as the *standard* **x**-*vector* (to distinguish it from the *experimental* **x**-*vector* needed in Section 6). The variables are a subset of those used to produce the calibration estimates in the LCS 2009.

(About the variable *Property ownership*: This variable is equal to one for a person identified in the property tax register as someone having paid taxes on real estate property owned. About the variable *Phone access*: This variable is equal to 1 if the phone number for a sampled person is made available and ready to be used at the very beginning of the data collection period.)

In Tables 4.1, 5.1 and 5.2, the entries for Attempt *a* (where *a* = 1, 2, 3 …) are computed on the union of the sets of persons having responded at attempts 1, 2, …, *a*, as expressed by (4.1). Not all call attempts are shown in the tables, but changes for deleted rows are minor. The entries for "End ordinary field work" are computed on the respondents at the end of the five week ordinary data collection period; "Final" is based on the total response recorded at the end of the follow-up period.

The three register variables used here as study variables are: *Sickness benefits* (a categorical variable equaling 1 for a recipient of benefits; 0 otherwise), *Income* (a continuous variable including employment as well as retirement income), and *Employed* (a categorical variable equaling 1 for an employed person; 0 otherwise). We chose these three

register variables because they are representative for (although not identical to) important real study variables observed in the LCS.

Since the three study variables are register variables, the value y_k is available for every $k \in s$, and we can compute the unbiased full sample (Horvitz-Thompson) estimate for these variables,

$$\hat{Y}_{FUL} = \sum_s d_k y_k \tag{4.4}$$

The computable percentage relative difference between $\hat{Y}_{FUL}$ and $\hat{Y}_{CAL}$ computed according to (4.3) is

$$RDF_{CAL} = 100 \cdot (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{Y}_{FUL} \tag{4.5}$$

The calibration estimator corresponding to the most primitive auxiliary vector, $\mathbf{x}_k = 1$ for all units *k*, serves as a benchmark in the study. Known as the expansion estimator, it is given by

$$\hat{Y}_{EXP} = (\sum_s d_k)(\sum_r d_k y_k)/(\sum_r d_k)$$

Its percentage relative difference from $\hat{Y}_{FUL}$ is

$$RDF_{EXP} = 100 \cdot (\hat{Y}_{EXP} - \hat{Y}_{FUL}) / \hat{Y}_{FUL} \tag{4.6}$$

Table 4.1 shows RDF_{CAL} and RDF_{EXP} computed for the three variables and for a number of steps in the series of call attempts. Table 4.1 leads to the following conclusions:

- RDF_{CAL} and RDF_{EXP}, defined by (4.5) and (4.6), change little in the later phases of the data collection.
- At the end of the data collection (the row "Final"), the value of RDF_{CAL} is (unacceptably) large, -3.6%, 2.9%, and 3.1%, respectively. To pursue the data collection according to an unchanging original plan does not result in small error in the estimates.
- For all three study variables, RDF_{CAL} is small, in fact near zero, at earlier stages of the data collection. For example, for Sickness benefits, RDF_{CAL} hovers around zero in the range of 9 to 14 call attempts. For the other two variables, RDF_{CAL} is near zero even earlier in the data collection.

- The numerically important changes in RDF_{CAL} and RDF_{EXP} tend to occur early in the series of attempts. Already after attempt number 3, both are in quite a stable pattern; later changes are small and move in a smooth continuous fashion.
- At almost all points in Table 4.1, RDF_{EXP} is greater than RDF_{CAL}. The high values of RDF_{EXP} indicate that the LCS 2009 data collection results in a skewed response, causing great departures from the unbiased estimate $\hat{Y}_{FUL}$. The auxiliary information used in RDF_{CAL} has the desired effect of reducing the departure from the unbiased estimate. But a noticeable feature for the variable Employed is that RDF_{EXP} is closer to RDF_{CAL} than for the other two variables. Thus, for Employed, the variables in the auxiliary vector used here are apparently not sufficiently effective to significantly reduce the estimation error.

The three selected study variables are representative for important real study variables in the LCS; we conclude that the data collection in the LCS 2009 is not as efficient as it ought to be.

Table 4.1
The LCS 2009 data collection: Progression of the response rate *P* (in per cent) and of *RDF* for three selected register variables. The calibration estimator is based on the *standard x-vector* explained in this section

		Sickness benefits		Income		Employed	
Attempt number	**100×*P***	***RDF*$_{CAL}$**	***RDF*$_{EXP}$**	***RDF*$_{CAL}$**	***RDF*$_{EXP}$**	***RDF*$_{CAL}$**	***RDF*$_{EXP}$**
1	12.8	10.5	-10.0	-0.05	0.3	-1.3	-9.0
2	24.6	3.3	-13.9	-1.1	0.4	-2.0	-8.1
3	32.8	1.6	-12.1	-0.4	1.6	0.2	-4.7
4	39.6	2.7	-10.1	0.2	2.9	0.4	-2.4
5	44.3	3.7	-7.2	0.7	3.6	1.1	-1.1
6	47.8	2.7	-7.0	1.2	4.5	1.7	0.4
7	50.9	1.6	-7.3	2.1	5.5	2.5	1.6
8	53.0	1.0	-7.4	2.4	6.2	2.4	2.3
9	54.6	0.2	-8.0	2.8	6.4	2.6	2.5
10	55.7	0.2	-8.0	2.8	6.6	2.6	2.8
11	56.8	-0.5	-8.5	2.7	6.5	2.6	3.0
12	57.7	0.1	-7.9	3.0	6.8	2.5	3.1
13	58.3	-0.3	-8.0	3.0	6.9	2.7	3.4
14	58.7	-0.1	-7.7	3.0	6.9	2.7	3.6
15	59.1	-0.5	-8.0	3.1	7.1	2.8	3.8
⋮							
20	60.1	-0.5	-7.7	3.4	7.5	3.0	4.1
End ord. fieldwork	60.4	-0.9	-7.9	3.3	7.4	2.9	4.2
				Follow-up			
1	61.4	-1.0	-8.0	3.3	7.1	2.9	4.1
2	62.6	-1.6	-8.2	3.1	6.7	3.0	3.9
3	63.8	-2.5	-9.2	3.0	6.7	3.2	4.2
4	64.6	-2.8	-9.3	3.1	6.7	3.3	4.3
5	65.3	-2.7	-9.0	3.1	6.8	3.1	4.3
⋮							
10	66.8	-2.9	-8.9	2.9	6.7	3.0	4.5
Final	67.4	-3.6	-9.4	2.9	6.7	3.1	4.8

5 Additional tools: indicators of balance, distance and representativity, computed on LCS 2009 data

In the theoretical first part of this section we define and explain several indicators designed to throw light on the progression of the data collection. In the latter part of the section we illustrate these indicators numerically, by computing them on the LCS 2009 data.

In the next section (Section 6), those indicators are used in an experiment with the LCS 2009 data, where we intervene "in retrospect" in the data collection process, with an objective to get a better balanced, or more representative, ultimate response set than if no action is taken. The indicators can be computed from the auxiliary variable values, known both for respondents and for all those sampled. Well known statistical concepts are reflected in the indicators.

One important concept is *balance*. The response set is balanced if it agrees on average with the whole set of sampled units, for one or more measurable variables. Another useful concept is *distance*, namely, between respondents and nonrespondents; that distance should be small. High balance and low distance are desirable.

A third useful concept is *variance of the response probabilities*. It is a good sign if that variability is small. Since the response probabilities are unknown, it is the variance of the estimated probabilities that will have to be computed.

We consider an auxiliary vector of dimension $J \geq 1$, $\mathbf{x}_k = (x_{1k},...,x_{jk},...,x_{Jk})'$, with known value for every $k \in s$ (or possibly for every $k \in U$). Here x_{jk} is the value for unit *k* of the *j*:th auxiliary variable x_j, which may be a continuous, or categorical equal to 1 or 0 to code the presence or the absence of a given trait of unit *k*. For the *j*:th auxiliary variable, we can compute the difference $D_j = \bar{x}_{jr} - \bar{x}_{js}$ between the respondent mean, $\bar{x}_{jr} = \sum_r d_k x_{jk} / \sum_r d_k$,

and the full sample mean, $\bar{x}_{js} = \sum_s d_k x_{jk} / \sum_s d_k$. If $D_j = 0$ for all J auxiliary variables, then r is called a *perfectly balanced* response set. Then the respondents are on average equal to the members of the whole sample, for every variable in the auxiliary vector $\mathbf{x}_k$.

Matrix language is needed because of the multivariate nature of $\mathbf{x}_k$. Let $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (D_1,...,D_j,...,D_J)'$. Under perfect balance, $\mathbf{D} = \mathbf{0}$, the zero vector. But normally, $\mathbf{D} \neq \mathbf{0}$, suggesting a departure from balance. We need to transform the multivariate $\mathbf{D}$ into a univariate measure of *lack of balance*, for the given survey outcome (*s*, *r*) and the given composition of $\mathbf{x}_k$. This purpose is filled by the quadratic form

$$\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)$$

where $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$ and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$ and the weighting matrix is $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$. Increased mean differences D_j tend to increase $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$.

It can be shown (Särndal, 2011a) that $0 \le \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \le Q - 1$ where $Q = 1/P$. Hence, $(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})/(Q-1)$ measures lack of balance on a unit interval scale.

We examine several balance indicators measured on the unit interval scale and such that the value "1" implies perfect balance. The first is

$$BI_1 = 1 - \sqrt{\frac{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}}{Q-1}} \tag{5.1}$$

It follows from $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \le Q - 1$ that $0 \le BI_1 \le 1$ for every survey outcome (*s*, *r*) and any composition of $\mathbf{x}_k$. Because $P(1-P) \le 1/4$, an alternative indicator also contained in the unit interval is

$$BI_2 = 1 - 2P\sqrt{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}} \tag{5.2}$$

The square root of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ can be viewed as a measure of distance, denoted $dist_{r|s}$, between the respondent set r and the whole sample s . Hence

$$dist_{r|s} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2} = [(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)]^{1/2} \quad (5.3)$$

Another distance of interest is the one between respondents and nonrespondents,

$$dist_{r|nr} = [(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})]^{1/2} \quad (5.4)$$

where $nr = s - r$ is the nonresponse set and $\bar{\mathbf{x}}_{nr} = \sum_{s-r} d_k \mathbf{x}_k / \sum_{s-r} d_k$. Intuitively, a desirable objective for the data collection is to achieve a low distance. The two distances are related by

$$dist_{r|s} = (1 - P) \times dist_{r|nr}$$

Both $dist_{r|s}$ and $dist_{r|nr}$ are varieties of the *Mahalanobis distance* between means of two sets of units. The values taken by $dist_{r|s}$ and $dist_{r|nr}$ depend closely on the choice of the x-variables entered into the auxiliary vector $\mathbf{x}_k$; for data encountered in practice, $dist_{r|s}$ seldom exceeds 0.5. Simple relationships exist between $dist_{r|nr}$ and the balance indicators:

$$BI_1 = 1 - \sqrt{P(1-P)} \times dist_{r|nr} \quad ; \quad BI_2 = 1 - 2P(1-P) \times dist_{r|nr}$$

Being a function of the response set r, the distance $dist_{r|nr}$ changes during the course of the data collection. We would like $dist_{r|nr}$ to decrease as r gets larger. But the opposite can occur. If the response rate in the data collection has not yet reached 50%, and $dist_{r|nr}$ is increasing, then the balance measured by BI_1 or by BI_2 is necessarily decreasing.

Consider now the concept of variability of estimated response probabilities. Let $\hat{\theta}_k$ be the estimated response probability for unit *k*. It may be obtained by one of several possible methods. Their variance is

$$S^2_{\hat{\theta}s} = \sum_s d_k (\hat{\theta}_k - \bar{\hat{\theta}}_s)^2 / \sum_s d_k \quad (5.5)$$

where $\bar{\hat{\theta}}_s = \sum_s d_k \hat{\theta}_k / \sum_s d_k$.

One method is to derive the estimates $\hat{\theta}_k$ by ordinary linear least squares, for the given specification of $\mathbf{x}_k$: Find **b** to minimize $\sum_s d_k(I_k - \mathbf{x}_k'\mathbf{b})^2$, where I_k is the response indicator, $I_k = 1$ for $k \in s$ and $I_k = 0$ for $k \in s - r$. This gives the estimate $\hat{\theta}_k = t_k$ for $k \in s$, where $t_k = \mathbf{x}_k'\hat{\mathbf{b}}$ with $\hat{\mathbf{b}} = \left(\sum_r d_k \mathbf{x}_k\right)'\left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$. The variance (5.5) computed with $\hat{\theta}_k = t_k$ is denoted S_{ts}^2. It can be shown that $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ and S_{ts}^2 are related according to $S_{ts}^2 = P^2 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, so the relation to the balance indicators is

$$BI_1 = 1 - \frac{S_{ts}}{\sqrt{P(1-P)}} \quad ; \quad BI_2 = 1 - 2S_{ts}$$

Low variability in the estimated response probabilities implies a high measure of balance. The *representativity indicator* proposed by Schouten, Cobben and Bethlehem (2009) is also based on the concept of estimated response probabilities. These authors use a logistic regression fit to obtain first $\hat{\boldsymbol{\beta}}$, then $\hat{\theta}_{k\log} = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})/[1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})]$ for $k \in s$. Their variance denoted $S_{\hat{\theta}\log,s}^2$ is computed by (5.5) with $\hat{\theta}_k = \hat{\theta}_{k,\log}$. The unadjusted *R-indicator* (where *R* stands for representativity) is given by

$$R = 1 - 2S_{\hat{\theta}\log,s} \tag{5.6}$$

These authors also suggest an *adjusted R*-indicator. Its objective is to reduce a bias that (5.6) may have when viewed as an estimate of a corresponding population quantity.

The indicators BI_1 and in particular BI_2 are usually numerically close to the *R*-indicators (unadjusted and adjusted). All four indicators behave similarly, as illustrated later in this section.

The distance $dist_{r|s}$ also has an interpretation as the coefficient of variation of the response probability estimates $\hat{\theta}_k = t_k$ for $k \in s$. It can be shown that

$$cv_{ts} = S_{ts}/\bar{t}_s = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2} = dist_{r|s}$$

where $\bar{t}_s = \sum_s d_k t_k / \sum_s d_k$. Often close in value to cv_{ts} is another coefficient of variation, namely, that of the adjustment factors m_k used in the estimator $\hat{Y} = \sum_r d_k m_k y_k$. It can be shown that

$$cv_{mr} = S_{mr} / \bar{m}_r = (\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D})^{1/2}$$

where $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k$; $S_{mr} = \left[\sum_r d_k (m_k - \bar{m}_r)^2 / \sum_r d_k\right]^{1/2}$ and

$\bar{m}_r = \sum_r d_k m_k / \sum_r d_k$. The only difference between these two coefficients of variation lies in the inverted weighting matrix, $\boldsymbol{\Sigma}_s^{-1}$ in one case, $\boldsymbol{\Sigma}_r^{-1}$ in the other.

Table 5.1 shows the progression of the balance indicators BI_1, BI_2, unadjusted *R* and adjusted *R* as functions of the call attempt number in the LCS 2009 data collection. The table shows (i) that the four measures are numerically close, and (ii) that the balance, as measured by any one of the four quantities, deteriorates as the data collection proceeds. The four indicators evolve very similarly over the series of attempts. Since they tell essentially the same story, any one of them serves the purpose to measure balance. We shall use BI_1, sometimes also BI_2.

Table 5.2 shows the development of the balance BI_1, the distance $dist_{r|nr}$, and the coefficients of variation $cv_{ts} = dist_{r|s} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}$ and $cv_{mr} = (\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D})^{1/2}$. Ideally, the balance BI_1 should increase and the distance $dist_{r|nr}$ decrease with an increasing number of attempts. But contrary to what we like to see, Table 5.2 shows that BI_1 and $dist_{r|nr}$ "go the wrong way": The balance is decreasing, and the distance between respondents and non-respondents is increasing. Thus Tables 5.1 and 5.2 reinforces the impression from Table 4.1, namely, that there is no strong motivation for keeping the present procedure for the LCS data collection. For example, it is questionable whether the ordinary field work should proceed as long as is currently the case, rather than to end altogether after say 12 or 15 attempts.

Table 5.1
The LCS 2009 data collection: Progression of the response rate *P* (in per cent), the balance indicators BI_1, BI_2, *R* unadjusted and *R* adjusted. As in Table 4.1, computations are based on the standard x-vector explained in Section 4

Attempt number	100×*P*	BI_1	BI_2	*R* unadj.	*R* adjusted
1	12.8	0.855	0.904	0.902	0.905
2	24.6	0.802	0.829	0.829	0.831
3	32.8	0.779	0.793	0.794	0.796
4	39.6	0.770	0.775	0.780	0.782
5	44.3	0.767	0.769	0.775	0.777
6	47.8	0.763	0.763	0.770	0.772
7	50.9	0.756	0.756	0.763	0.765
8	53.0	0.751	0.752	0.758	0.760
9	54.6	0.750	0.752	0.757	0.759
10	55.7	0.748	0.749	0.756	0.758
11	56.8	0.746	0.749	0.754	0.756
12	57.7	0.747	0.750	0.756	0.758
13	58.3	0.744	0.748	0.754	0.756
14	58.7	0.742	0.746	0.753	0.754
15	59.1	0.741	0.745	0.752	0.754
⋮					
20	60.1	0.737	0.743	0.751	0.753
End ordinary field work	60.4	0.738	0.744	0.752	0.754
			Follow-up		
1	61.4	0.736	0.743	0.751	0.752
2	62.6	0.734	0.742	0.750	0.752
3	63.8	0.730	0.741	0.748	0.750
4	64.6	0.728	0.740	0.747	0.749
5	65.3	0.727	0.740	0.748	0.749
⋮					
10	66.8	0.719	0.736	0.742	0.744
Final	67.4	0.717	0.735	0.742	0.743

Table 5.2
The LCS 2009 data collection: Progression of the response rate *P* (in per cent), BI_1, cv_{mr}, cv_{ts} and $dist_{r|nr}$. As in Table 4.1, computations are based on the standard x-vector explained in Section 4

Attempt number	100×*P*	BI_1	cv_{mr}	cv_{ts}	$dist_{r\|nr}$
1	12.8	0.855	0.437	0.378	0.433
2	24.6	0.802	0.399	0.347	0.460
3	32.8	0.779	0.379	0.316	0.470
4	39.6	0.770	0.348	0.285	0.471
5	44.3	0.767	0.320	0.261	0.469
6	47.8	0.763	0.306	0.248	0.475
7	50.9	0.756	0.294	0.240	0.488
8	53.0	0.751	0.291	0.234	0.499
9	54.6	0.750	0.283	0.227	0.501
10	55.7	0.748	0.280	0.225	0.508
11	56.8	0.746	0.277	0.221	0.512
12	57.7	0.747	0.273	0.217	0.513
13	58.3	0.744	0.274	0.217	0.519
14	58.7	0.742	0.275	0.216	0.523
15	59.1	0.741	0.274	0.215	0.527
⋮					
20	60.1	0.737	0.273	0.214	0.536
End ordinary field work	60.4	0.738	0.271	0.212	0.536
			Follow-up		
1	61.4	0.736	0.269	0.210	0.542
2	62.6	0.734	0.265	0.206	0.550
3	63.8	0.730	0.262	0.203	0.561
4	64.6	0.728	0.258	0.201	0.569
5	65.3	0.727	0.255	0.199	0.573
⋮					
10	66.8	0.719	0.255	0.198	0.596
Final	67.4	0.717	0.255	0.197	0.603

6 Experimental data collection strategies derived from LCS 2009 data

6.1 Experimental strategies

There is strong evidence that the best objective for the future LCS is not to try to achieve a predefined "respectable" overall response rate. It is hard to motivate a costly effort for a possible five per cent greater ultimate response rate if this is not accompanied by improved features of the respondents, such as better balance, and closeness to nonrespondents. Such objectives should be the guiding light for a data collection strategy. The LCS 2009 data show that an increasing response rate does not improve the balance and other important characteristics of the response set. Improvements on these regards call for changes in the data collection procedure itself. A goal is to try to reduce the cost of the field work, for example by restrictions on the total number of call attempts. Savings realized by fewer calls may instead be might be used to improve the balance of the response set.

This section presents the results of three "experiments in retrospect" carried out on the existing LCS 2009 data file. We cannot add any more data to that file, but we can use data from that file to illustrate the effects of different interventions in the data collection. We want to see the trend in the balance indicators, particularly BI_1 and BI_2, and in the distance $dist_{r|nr}$, as the data collection progresses; signs of a good procedure would be increasing balance and decreasing distance. The experiments consist in declaring data collection terminated for designated sample sub-groups at suitably chosen points in the data inflow. For example, it might stop in some groups after a given number of call attempts because "realistic expectations" for the response have already been met, whereas for other groups the data collection would continue for yet some time before stopping, and for remaining groups it would continue until the very end of the data collection period. In this example there are two intermediate *intervention points*, that is, points at which changes occur in the data collection procedure. In our experiments, the

interventions are of a simple kind, namely, to stop data collection for specified groups at suitably chosen points.

We thus delete data in the existing LCS 2009 data file in an organized fashion, pretending that data collection has terminated at specified points for specified sample sub-groups. For those groups, we deliberately sacrifice some LCS 2009 data values y_k that were in reality observed beyond the specified intervention points.

The quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ plays a key role in illustrating the experiments. It determines the balance measures $BI_1 = 1-\sqrt{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}/(Q-1)}$ and $BI_2 = 1-2P\sqrt{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}}$ and the distance measure $dist_{r|nr} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}/(1-P)$. We should direct the data collection so as to reduce the differences D_j that define the vector $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (D_1,...,D_j,...,D_J)'$; a decrease in $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ implies improved balance.

The quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ has a particularly useful expression when the vector $\mathbf{x}_k$ (which determines $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$) is defined in terms of *J* mutually exclusive and exhaustive traits or characteristics. A simple example is when "Age" is defined by, say, $J = 3$ traits, Young, Middle-aged and Elderly. In practice, several categorical variables are often crossed to define a set of mutually exclusive and exhaustive groups. The trait of unit *k* is then uniquely coded by the J-vector $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{jk},...,\gamma_{Jk})' = (0,...,1,...,0)'$ (with a single entry "1"), or equivalently by the J-vector $\mathbf{x}_k = (1,\gamma_{1k},...,,\gamma_{J-1,k})'$, where $\gamma_{jk} = 1$ if k has the trait *j* and $\gamma_{jk} = 0$ otherwise. Denote by s_j the (non-empty) subset of the sample s consisting of the units *k* with the trait *j*, and let r_j be the corresponding responding subset of *r*; $r_j \subseteq s_j$. For trait *j*, let $W_{js} = \sum_{s_j} d_k / \sum_s d_k$ be that trait's share of the whole sample s, and $W_{jr} = \sum_{r_j} d_k / \sum_r d_k$ its share of the whole response set r. Then the quadratic form is a sum of non-negative terms expressed as

$$\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} C_j \qquad (6.1)$$

with

$$C_j = \frac{(W_{jr} - W_{js})^2}{W_{js}} = W_{js} \times \left(\frac{P_j - P}{P}\right)^2 \qquad (6.2)$$

where $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ is the response rate for the j^{th} group and P is the overall response rate given by (4.2). The terms C_j reflect the impact on each group of the data collection strategy. We call $(P_j - P)/P$ the *response rate differential* for the j^{th} group. These differentials, some positive, some negative, reflect the differences in response between the J groups; they can be substantially different, although seldom greater in absolute value than 0.3. Their weighted average is zero: $\sum_{j=1}^{J} W_{js} \times \frac{P_j - P}{P} = 0$. If the maximum $\left|P_j - P\right|/P$ over the J groups is equal to let us say 0.5, it follows that $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \leq 0.25$ and that $dist_{r|s} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2} = cv_{ts} \leq 0.5$. In the applications in Sections 5 and 6, $dist_{r|s} = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}$ is of the order of 0.4 or less.

If all group response rates P_j are equal, there is perfect balance with respect to the chosen set of groups; both BI_1 and BI_2 are then equal to 1.

The results in Sections 4 and 5 were based on the *standard* $\mathbf{x}$*-vector*, an auxiliary vector close to the one used to produce the estimates (by calibration) for the LCS 2009. For the experimental strategies in this section, we choose a more appropriate vector. We report the results of three *experiments in retrospect*, each based on an *experimental data collection strategy* consisting of these features:

a suitably chosen vector $\mathbf{x}_k$, called *experimental* $\mathbf{x}$*-vector*, with known value for every sampled unit $k \in s$;

one or more specified *intervention points*, with a *stopping rule* for each intervention point.

For every unit $k \in s$, the experimental $\mathbf{x}$-vector is of the form $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{Jk})'$, pointing out membership in one of the *J* mutually exclusive and exhaustive sample groups. Every intervention point marks a change in the data collection procedure. The stopping rule is formulated in terms of a predefined response rate for each group; data collection will stop at a certain intervention point for groups having at that point reached the specified response rate.

The experimental $\mathbf{x}$-vector used throughout this section is defined by the complete crossing of three dichotomous auxiliary variables: *Education level* (high, not high), *Property ownership* (owner, non-owner), *Country of origin* (Sweden, other). This defines eight mutually exclusive and exhaustive groups. The experimental **x**-vector is of dimension $J = 2^3 = 8$ and of the form $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{8k})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ otherwise. The vector value $\mathbf{x}_k$ is known for all $k \in s$.

We chose those three dichotomous variables because the eight sample groups that they define are especially important to follow. The group response rates are expected to differ considerably. Strikingly low response can be expected in some groups.

Although important, the auxiliary variable Phone access is not used in the experimental **x**-vector. This variable, as defined in Section 4, equals one for a person with a phone number known at the beginning of the data collection, and zero otherwise. By tracing during the course of the data collection, telephone numbers are however found for around one third of those with value zero on the Phone access variable, that is, one third of those without number recorded at the start of the data collection.

6.2 The LCS 2009 data analyzed with the experimental vector

This section illustrates again that the current LCS 2009 data collection is not as efficient as it should be. Tables 6.1, 6.2 and 6.3 are computed on the entire LCS 2009 data set, although based on the experimental **x**-vector defined in Section 6.1.

Table 6.1 shows (as did Table 5.2 for the standard **x**-vector) that the balance, as measured by BI_1 and BI_2, decreases as the data collection proceeds, and that the distance $dist_{r|nr}$ increases. These unfavorable features point again to the inefficiency in the 2009 data collection with its predefined unchangeable format.

Table 6.1
The LCS 2009 data collection. Values of RFD_{CAL} for three study variables; values of several indicators. Experimental x-vector defined by crossing of Education (high, not high), Property ownership (owner, not owner) and Country of origin (Sweden, other)

		RDF_{CAL}							
Attempt number	**100×*P***	**Sickness benefits**	**Income**	**Employed**	BI_1	BI_2	cv_{ts}	cv_{mr}	$dist_{r\|nr}$
1	12.8	-8.4	-2.7	-10.2	0.922	0.948	0.203	0.276	0.233
2	24.6	-13.2	-3.2	-9.7	0.887	0.902	0.198	0.262	0.263
3	32.8	-11.5	-2.3	-6.3	0.867	0.875	0.190	0.232	0.283
4	39.6	-8.5	-1.5	-4.4	0.850	0.854	0.185	0.215	0.306
5	44.3	-5.8	-0.5	-3.0	0.846	0.847	0.173	0.198	0.310
⋮									
8	53.0	-5.7	1.7	0.2	0.812	0.812	0.177	0.199	0.377
⋮									
12	57.7	-6.1	2.5	1.2	0.805	0.807	0.167	0.185	0.394
⋮									
20	60.1	-6.0	3.1	2.2	0.795	0.799	0.167	0.183	0.418
⋮									
End ordinary field work	60.4	-6.2	3.1	2.3	0.796	0.801	0.165	0.182	0.417
				Follow-up					
1	61.4	-6.2	3.0	2.3	0.796	0.802	0.162	0.179	0.418
⋮									
4	64.6	-7.9	2.8	2.6	0.792	0.801	0.154	0.171	0.435
⋮									
Final	67.4	-7.9	2.9	3.1	0.779	0.793	0.154	0.171	0.471

Table 6.2 shows the development over the data collection of the eight terms C_j (defined by (6.2)) whose total equals $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ in the bottom line. The table shows the nearly constant level of C_j for the highly influential lines 1, 4, 5 and 8, in going from attempt 5 (where the data collection has attained some stability) and on to the very end. The problematic line 1 group, *education not high, non-owner, foreign origin*, has a distinctly negative response differential $(P_j - P)/P$, and $100 \times C_j$ decreases somewhat from 1.44 at attempt 5

to a final value of 1.18. We would have liked to see a much more pronounced decrease toward zero, but it does not happen with the unchanging format in LCS 2009. The line 5 group is another with a distinctly negative response differential. The next highest C_j, from attempt 5 and on to the end, occurs for the line 8 group, *high education, owner, Swedish origin*. This group, with a distinctly positive response differential $(P_j - P)/P$, shows a value of $100 \times C_j$ that decreases somewhat, from 0.58 at attempt 5 to end at 0.44, but not as much as one would like to see.

Table 6.2

The LCS 2009 data collection; values of the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100). Experimental x-vector defined by crossing of Education (high, not high), Property ownership (owner, non-owner) and Country of origin (Sweden, other)

			$100 \times C_j$						
Group characteristic			Ordinary fieldwork attempt				Follow-up attempt		
Education	**Property ownership**	**Origin**	**1**	**5**	**12**	**End**	**1**	**4**	**Final**
Not high	**Non-owner**	**Abroad**	1.49	1.44	1.26	1.23	1.25	1.16	1.18
Not high	**Non-owner**	**Sweden**	0.00	0.06	0.11	0.11	0.08	0.07	0.07
Not high	**Owner**	**Abroad**	0.06	0.01	0.00	0.00	0.00	0.00	0.00
Not high	**Owner**	**Sweden**	0.72	0.24	0.21	0.19	0.17	0.17	0.18
High	**Non-owner**	**Abroad**	1.28	0.39	0.29	0.26	0.25	0.23	0.22
High	**Non-owner**	**Sweden**	0.11	0.26	0.25	0.24	0.21	0.20	0.23
High	**Owner**	**Abroad**	0.18	0.01	0.03	0.03	0.03	0.02	0.04
High	**Owner**	**Sweden**	0.29	0.58	0.64	0.66	0.62	0.53	0.44
$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$			4.13	2.99	2.78	2.72	2.61	2.37	2.36

Table 6.3
The LCS 2009 data collection; column-wise percentages of values in Table 6.2

Group characteristic			Ordinary fieldwork attempt				Follow-up attempt		
Education	**Property ownership**	**Origin**	**1**	**5**	**12**	**End**	**1**	**4**	**Final**
Not high	**Non-owner**	**Abroad**	36.1	48.2	45.2	45.1	47.6	48.8	49.8
Not high	**Non-owner**	**Sweden**	0.1	2.0	3.9	4.1	3.1	2.8	2.9
Not high	**Owner**	**Abroad**	1.6	0.2	0.01	0.00	0.00	0.00	0.01
Not high	**Owner**	**Sweden**	17.3	8.1	7.6	7.0	6.7	7.0	7.7
High	**Non-owner**	**Abroad**	31.0	13.1	10.4	9.4	9.6	9.8	9.4
High	**Non-owner**	**Sweden**	2.6	8.7	8.8	8.9	8.2	8.4	9.6
High	**Owner**	**Abroad**	4.4	0.3	1.0	1.2	1.1	1.0	1.7
High	**Owner**	**Sweden**	6.9	19.5	23.1	24.2	23.8	22.2	18.8
		Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 6.3, shows the entries in Table 6.2 converted to percentages of each column total. They illustrate the rigidity of the data collection: The main contributors to the column total, line 1 and line 8, remain essentially constant over the seven points, at around 45% for the first and around 23% for the last.

The purpose with Table 6.4 is to examine the variable Phone access, as defined in Section 4, for the eight groups. The groups in lines 1 and 5 have distinctly lower Phone access. Persons in these groups of foreign origin are likely to use cell phones with pay cards, which amounts in principle to an unlisted phone number. On the other hand, groups with high education and/or property ownership are likely to be more firmly rooted in society and more likely to participate in the LCS. Although known to be problematic, the line 1 and line 5 groups do not seem to have received special attention in the LCS 2009 field work: Their mean number of call attempts is not any higher than for other, "easier" groups. Non-contact rates (not shown here) are high for the two problem groups, suggesting that their members are hard to find, some of them perhaps living abroad. A special investigation would be needed for these groups.

Table 6.4
Initially found telephone number (Phone access = 1) and mean number of call attempts, for the eight groups of the experimental x-vector, computed on the LCS 2009 data

Group characteristic Education, Property ownership, Origin	Phone access in %	Mean call attempts	Individuals in sample
Not high, Non-owner, Abroad	72.9	6.7	847
Not high, Non-owner, Sweden	90.4	6.6	3210
Not high, Owner, Abroad	94.2	6.1	171
Not high, Owner, Sweden	97.7	6.4	2036
High, Non-owner, Abroad	78.0	6.7	236
High, Non-owner, Sweden	96.0	6.5	816
High, Owner, Abroad	95.8	7.2	72
High, Owner, Sweden	98.4	6.0	832

In the following sections 6.3 to 6.5 we present the results of three experimental strategies. All three use the experimental **x**-vector of dimension $J = 2^3 = 8$ as defined in Section 6.1. The three strategies differ in other aspects: The points of intervention, and the stopping rule at the points of intervention.

6.3 Experimental strategy 1 and its results

Strategy 1 uses two intervention points: Attempt 12 of the ordinary data collection (point 1), and Attempt 2 of the follow-up (point 2). The stopping rule for data collection in a group is a realized response rate of at least 65%. Computed on the entire LCS 2009 data, Table 6.5 shows how the experimental Strategy 1 data collection is determined. Marked in grey are the groups for which data collection is deemed terminated at each point. Thus the Strategy 1 data set has the following features: Data collection is ended at point 1 for the groups in lines 6, 7 and 8, which have at that point achieved the 65% rate. At point 2, termination occurs for the group in line 4. Data collection continues until the very end for the remaining four groups. Still, for the line 1 group, the response rate at the end is only 44.6%, far from 65%.

Table 6.5
Response rate *P* (in per cent) at three points in the entire LCS 2009 data collection for the eight groups formed by the experimental x-vector

Group characteristic			Response rate *P* (per cent)			
Education	Property ownership	Origin	Attempt 12 ordinary	Attempt 2 follow-up	Final	Individuals in sample
Not high	Non-owner	Abroad	37.5	41.8	44.6	847
Not high	Non-owner	Sweden	54.6	59.8	64.6	3210
Not high	Owner	Abroad	58.5	62.3	66.8	171
Not high	Owner	Sweden	63.0	67.6	73.2	2036
High	Non-owner	Abroad	39.4	44.9	48.7	236
High	Non-owner	Sweden	66.8	71.6	77.6	816
High	Owner	Abroad	68.1	73.6	81.9	72
High	Owner	Sweden	72.2	77.4	81.5	832
Total			57.7	62.6	67.4	8220

For the Strategy 1 data collection, Table 6.6 shows the progression of the terms C_j and of their total $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100). Data collection has ended at point 1 for the high-responding groups in lines 6, 7 and 8. The desired decrease in the value of C_j for lines 6 and 8 is a result of a increasing denominator (but an unchanging numerator) in $W_{jr} = \sum_{r_j} d_k / \sum_r d_k$. The low response line 1 group accounts for the largest term $100 \times C_j$, going from 1.26 at point 1 to end at a still fairly high 0.94. Nevertheless, over the course of the Strategy 1 data collection, $100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is greatly reduced from 2.78 to 1.39, with an improved balance and reduced distance as a result. Table 6.7 shows the relative importance of each group. At each point, the dominance of the line 1 group becomes more and more accentuated, accounting at the end (column Final) for 67.4 % of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$.

Even with the simple interventions in Strategy 1, both the balance and the distance evolve in desired directions. Table 6.8 shows a balance BI_1 increasing from 0.805 to 0.843 and a distance $dist_{r|nr}$ shrinking from 0.394 to 0.326. These encouraging results contrast with the more stationary pattern in Table 6.2 for the whole LCS 2009 data collection.

Table 6.6
Experimental strategy 1; the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100) at three points in the data collection

Group characteristic			Value of $100 \times C_j$ at point		
Education	**Property ownership**	**Origin**	**Attempt 12 ordinary**	**Attempt 2 follow-up**	**Final**
Not high	**Non-owner**	**Abroad**	1.26	1.06	0.94
Not high	**Non-owner**	**Sweden**	0.11	0.03	0.00
Not high	**Owner**	**Abroad**	0.00	0.00	0.00
Not high	**Owner**	**Sweden**	0.21	0.24	0.08
High	**Non-owner**	**Abroad**	0.29	0.21	0.16
High	**Non-owner**	**Sweden**	0.25	0.07	0.02
High	**Owner**	**Abroad**	0.03	0.01	0.00
High	**Owner**	**Sweden**	0.64	0.31	0.17
$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$			2.78	1.93	1.39

Table 6.7
Experimental strategy 1; column-wise percentages of values in Table 6.6

Group characteristic			Point in data collection		
Education	**Property ownership**	**Origin**	**Attempt 12 ordinary**	**Attempt 2 follow-up**	**Final**
Not high	**Non-owner**	**Abroad**	45.2	55.1	67.4
Not high	**Non-owner**	**Sweden**	3.9	1.5	0.3
Not high	**Owner**	**Abroad**	0.0	0.0	0.3
Not high	**Owner**	**Sweden**	7.6	12.4	6.0
High	**Non-owner**	**Abroad**	10.4	10.9	11.6
High	**Non-owner**	**Sweden**	8.8	3.7	1.5
High	**Owner**	**Abroad**	1.0	0.5	0.3
High	**Owner**	**Sweden**	23.1	15.8	12.5
Total			100.0	100.0	100.0

Table 6.8
Experimental strategy 1; the response rate *P* (in per cent), BI_1, cv_{ts} and $dist_{r|nr}$ computed at three points in the data collection

Point in data collection	100×*P*	BI_1	cv_{ts}	$dist_{r\|nr}$
Attempt 12 ordinary	57.7	0.805	0.167	0.394
Attempt 2 follow-up	61.5	0.824	0.139	0.361
Final	63.9	0.843	0.118	0.326

6.4 Experimental strategy 2 and its results

Experimental strategy 2 uses the same **x**-vector, of dimension eight, as earlier in this section. The stopping rule is to consider data collection terminated (in the ordinary data collection or in the follow-up) for a group as soon as its response has reached 60%. As a result, Strategy 2 has five intervention points as shown in Table 6.9: Five groups terminate during the ordinary data collection at the indicated number of attempts, one group terminates at follow-up attempt 3. The problematic line 1 and line 5 groups continue to the end but still do not come close to 60%.

Table 6.9
Experimental strategy 2; termination point for data collection and achieved response rate (in per cent) for the eight groups

Group characteristic			Data collection feature	
Education	**Property ownership**	**Origin**	**Termination point**	**Achieved response rate**
Not high	**Non-owner**	**Abroad**	Final	44.6
Not high	**Non-owner**	**Sweden**	Attempt 3 follow-up	61.0
Not high	**Owner**	**Abroad**	Attempt 15 ordinary	60.2
Not high	**Owner**	**Sweden**	Attempt 9 ordinary	60.0
High	**Non-owner**	**Abroad**	Final	48.7
High	**Non-owner**	**Sweden**	Attempt 7 ordinary	60.2
High	**Owner**	**Abroad**	Attempt 8 ordinary	62.5
High	**Owner**	**Sweden**	Attempt 7 ordinary	63.5
Total data collection				58.9

Table 6.10

Experimental strategy 2; the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100) at six points in the data collection

Group characteristic			Value of $100 \times C_j$ at data collection point					
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**	1.39	1.40	1.29	0.99	0.78	0.60
Not high	**Non-owner**	**Sweden**	0.12	0.09	0.05	0.00	0.07	0.05
Not high	**Owner**	**Abroad**	0.00	0.00	0.00	0.01	0.00	0.00
Not high	**Owner**	**Sweden**	0.25	0.33	0.33	0.13	0.02	0.01
High	**Non-owner**	**Abroad**	0.35	0.31	0.31	0.22	0.15	0.09
High	**Non-owner**	**Sweden**	0.33	0.21	0.14	0.05	0.01	0.00
High	**Owner**	**Abroad**	0.02	0.03	0.02	0.01	0.00	0.00
High	**Owner**	**Sweden**	0.61	0.44	0.33	0.18	0.07	0.06
$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$			3.07	2.81	2.49	1.59	1.09	0.82

Table 6.10 shows the terms C_j (multiplied by 100) of the quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{8} C_j$. A comparison with Table 6.6 for Strategy 1 shows that Strategy 2 is an improvement. For all but the problematic line 1 group, C_j is reduced to low levels at the end (the column Final), and $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ ends at a considerably lower value, 0.82, as compared with 1.39 in Strategy 1.

Table 6.11
Experimental strategy 2; column-wise percentages of values in Table 6.10

Group characteristic			Data collection point					
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	45.3	49.8	52.1	62.4	71.2	73.7
Not high	Non-owner	Sweden	3.8	3.1	2.1	0.0	6.2	6.2
Not high	Owner	Abroad	0.0	0.0	0.0	0.7	0.2	0.1
Not high	Owner	Sweden	8.1	11.6	13.4	8.0	1.4	1.1
High	Non-owner	Abroad	11.5	11.2	12.6	13.5	13.5	10.4
High	Non-owner	Sweden	10.6	7.5	5.7	3.4	0.7	0.6
High	Owner	Abroad	0.6	1.1	0.9	0.7	0.4	0.4
High	Owner	Sweden	20.0	15.7	13.2	11.2	6.5	7.5
Total			100.0	100.0	100.0	100.0	100.0	100.0

Table 6.12
Experimental strategy 2; response rate *P* (in per cent), BI_1, cv_{ts} and $dist_{r|nr}$ computed at six points in the data collection

Data collection point	100×*P*	BI_1	cv_{ts}	$dist_{r\|nr}$
Attempt 7 ordinary	50.9	0.822	0.175	0.357
Attempt 8 ordinary	52.5	0.824	0.168	0.353
Attempt 9 ordinary	53.8	0.830	0.158	0.341
Attempt 15 ordinary	56.0	0.858	0.126	0.287
Attempt 3 follow-up	58.6	0.876	0.104	0.252
Final	58.9	0.892	0.091	0.220

6.5 Experimental strategy 3 and its results

Experimental strategy 3 uses the same **x**-vector, of dimension eight, as earlier in this section. The stopping rule is defined by calling data collection terminated in a group when its response rate (in the ordinary data collection or in the follow-up) has reached 50%. Table 6.13 shows strategy 3 data collection as terminated before the very end for all but the problematic line 1 and line 5 groups. The effect of lowering the stopping rule to 50% gives more pronounced improvement: Better balance, and decreased distance, compared with Strategy 2. Tables 6.13 to 6.16 are the counterparts for Strategy 3 of Tables 6.9 to 6.12 for Strategy 2.

Table 6.13
Experimental strategy 3; termination point for data collection and achieved response rate (in per cent) for the eight groups

Group characteristic			Data collection feature	
Education	**Property ownership**	**Origin**	**Termination point**	**Achieved response rate**
Not high	**Non-owner**	**Abroad**	Final	44.6
Not high	**Non-owner**	**Sweden**	Attempt 8 ordinary	50.0
Not high	**Owner**	**Abroad**	Attempt 7 ordinary	51.5
Not high	**Owner**	**Sweden**	Attempt 6 ordinary	52.6
High	**Non-owner**	**Abroad**	Final	48.7
High	**Non-owner**	**Sweden**	Attempt 5 ordinary	50.5
High	**Owner**	**Abroad**	Attempt 6 ordinary	52.8
High	**Owner**	**Sweden**	Attempt 4 ordinary	50.1
Total				50.3

Table 6.14
Experimental strategy 3; the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100) at six points in the data collection

Group characteristic			Value of $100 \times C_j$ at data collection point					
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**	1.51	1.39	1.23	1.10	1.05	0.13
Not high	**Non-owner**	**Sweden**	0.05	0.03	0.02	0.00	0.03	0.00
Not high	**Owner**	**Abroad**	0.01	0.00	0.00	0.01	0.01	0.00
Not high	**Owner**	**Sweden**	0.26	0.30	0.46	0.25	0.16	0.05
High	**Non-owner**	**Abroad**	0.59	0.38	0.35	0.27	0.22	0.00
High	**Non-owner**	**Sweden**	0.27	0.30	0.12	0.06	0.03	0.01
High	**Owner**	**Abroad**	0.00	0.01	0.02	0.01	0.01	0.00
High	**Owner**	**Sweden**	0.72	0.21	0.07	0.02	0.01	0.00
$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$			3.42	2.62	2.26	1.73	1.51	0.20

Table 6.15
Experimental strategy 3; column-wise percentages of values in Table 6.14

Group characteristic			Data collection point					
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	44.1	52.8	54.3	63.5	69.3	66.0
Not high	Non-owner	Sweden	1.6	1.2	0.9	0.1	1.9	0.5
Not high	Owner	Abroad	0.4	0.1	0.0	0.7	0.4	0.6
Not high	Owner	Sweden	7.6	11.6	20.1	14.6	10.6	27.6
High	Non-owner	Abroad	17.3	14.3	15.6	15.9	14.7	1.4
High	Non-owner	Sweden	7.9	11.5	5.3	3.3	2.1	2.8
High	Owner	Abroad	0.0	0.4	0.7	0.5	0.4	1.1
High	Owner	Sweden	21.1	8.0	2.9	1.3	0.6	0.1
Total			100.0	100.0	100.0	100.0	100.0	100.0

Table 6.16
Experimental strategy 3; the response rate P (in per cent), BI_1, cv_{ts} and $dist_{r|nr}$ computed at six points in the data collection

Point in data collection	$100\times P$	BI_1	cv_{ts}	$dist_{r\|nr}$
Attempt 4 ordinary	39.6	0.850	0.185	0.306
Attempt 5 ordinary	43.8	0.857	0.162	0.288
Attempt 6 ordinary	46.4	0.860	0.150	0.281
Attempt 7 ordinary	47.8	0.874	0.131	0.252
Attempt 8 ordinary	48.7	0.880	0.123	0.240
Final	50.3	0.955	0.044	0.089

6.6 The experimental strategies compared with the actual LCS 2009 data collection

In Table 6.17, we compare the actual LCS 2009 data collection with the three successive experimental strategies (which are based on the LCS 2009 data in the censored forms described in Sections 6.3, 6.4 and 6.5). For comparability, all results in Table 6.17 are computed on the standard auxiliary vector defined in Section 4, since it closely resembles the one used in practice to produce the LCS estimates. The entries for the actual LCS 2009 data collection (the first line) are taken from the bottom line, "Final", in Tables 4.1 and 5.2. Table 6.17 shows that each experimental strategy improves on the one in the preceding line.

For all three study variables, RDF_{CAL} is reduced from one strategy to the next (if we disregard a slightly higher value for the variable Employment in Strategy 2). For Income and Employment, the major reduction in RDF_{CAL} occurs in the step from Strategy 2 to Strategy 3.

Both the distance $dist_{r|nr}$ between respondents and nonrespondents and the balance BI_1 improve in each step. The distance $dist_{r|nr}$ drops from 0.603 to 0.383. The balance BI_1 increases from 0.717 to 0.808, the greatest change occurring from the actual LCS 2009 data (0.717) to Strategy 1 (0.765). (The balance shown in Table 6.17 is not surprisingly lower than the figure for the corresponding experimental strategy in Table 6.8, 6.12 or 6.16. This is because of different **x**-vectors; it is harder to achieve high balance with respect to a more extensive vector.)

But the most striking benefit from the experimental strategies lies in an implicit reduction of survey cost through significantly fewer call attempts. To reach the 67.4% response in the complete 2009 LCS data collection, 53,258 attempts were used, but to reach the 63.9% response in experimental Strategy 1, only 48,883 attempts are used. The number of call attempts is reduced by 8.2%. The reduction in call attempts is even more striking for the other two experimental strategies: 20.2% for Strategy 2 and 36.4% for Strategy 3. Not only do we get closer to the unbiased estimate (a lower RDF_{CAL}) and improved balance, the cost of data collection is also considerably lower, particularly for Strategies 2 and 3.

Table 6.17
The three experimental strategies compared with the actual LCS 2009 data collection; response rate *P* (in per cent), RDF_{CAL}, BI_1, $dist_{r|nr}$ and reduction (in per cent) of the number of call attempts. Computations based on the standard x-vector explained in Section 4

At end of data collection		RDF_{CAL}					
	100×*P*	Sickness benefits	Income	Employ-ment	BI_1	$dist_{r\|nr}$	*Reduction in %*
Actual LCS 2009	67.4	-3.6	2.9	3.1	0.717	0.603	0.0
Strategy 1	63.9	-1.6	2.7	3.0	0.765	0.489	8.2
Strategy 2	58.9	-1.2	2.6	3.2	0.787	0.433	20.2
Strategy 3	50.3	1.0	1.0	2.0	0.808	0.383	36.4

7 Discussion and implications for the future

In this article we have used process data from the Swedish CATI-system to examine the data collection in the 2009 Swedish Living Conditions Survey (LCS). Earlier studies at Statistics Sweden had already cast doubt on the merits of conducting a follow-up in the LCS; our results in Sections 4 and 5 confirm those earlier findings. In Section 4 we found that the estimation error for important variables may be smaller before than after the follow-up.

In Section 5 we applied alternative indicators of balance (or representativity) of the respondents. In Table 5.1 these indicators show a decreasing trend over the course of the LCS 2009 data collection. Contrary to reasonable expectations, the set of respondents is less balanced after the follow-up than at the end of the ordinary fieldwork. Another interesting indicator to monitor during the data collection is the distance $dist_{r|nr}$ between respondents and nonrespondents. A sign of a good data collection is a progressively decreasing distance. But Table 5.1 shows instead that the current data collection leads to increased rather than decreased distance. These findings add to the doubts about the efficiency of the current form of the LCS data collection.

The lack of balance, defined mathematically by $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ in Section 5, is an important tool. As Section 5 explains, it determines important notions such as the *balance of the set of respondents* and the *distance between respondents and nonrespondents*. The lack of balance is a function of the auxiliary vector denoted $\mathbf{x}_k$. When $\mathbf{x}_k$ codifies membership in one of J mutually exclusive and exhaustive sample subgroups, the lack of balance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is a sum of non-negative terms, $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} C_j$, where C_j is the contribution to lack of balance of the *j*:th group. This representation allows us to focus in the data collection on each specified group. Problematic are those groups for which C_j remains high at the end of the data collection. Examples of this were seen in Section 6. It is desirable to direct the data collection in a way so as to make all terms C_j small in the end.

Section 6 described three experiments carried out by interventions in the LCS 2009 data file. A set of important sample subgroups was defined, and data collection was deemed terminated at specified points in the data inflow. These experiments showed that appropriate interventions in the data collection can bring considerable improvement - better balance, smaller distance – compared with the traditional LCS data collection.

To use the conclusions from these experiments in practice, we must anticipate a "reasonable expectations" response rate, say 60% or even 50%, to be used as a stopping rule for data collection in a group. In a regularly repeated survey such prior information is usually available, but it may not be readily obtainable in a survey carried out for the first time. But an assessment would be necessary.

An objective for the near future is to improve the data collection for future editions of the LCS. To evaluate the potential advantages of a responsive design, plans are underway to conduct an embedded experiment in the data collection for the upcoming LCS 2011. One option is to let say one half of the LCS 2011 sample follow the traditional routines for data collection, while for the other half, data collection would considered terminated, for designated sample groups, at suitably chosen points in the data inflow, along the lines of our experiments.

An issue of relevance to the LCS survey is the frame over-coverage; certain sample sub-groups contain highly mobile people, some of whom may no longer reside in the country. It is clear that groups with chronically low response rate require particular attention in the data collection. In particular, improvements are needed to reach immigrants and younger persons whose style of living and interest in the survey may differ substantially from a majority of the population. Future savings could be realized by transferring interviewer time from "easy-to-catch" respondents to the more problematic groups.

A central question is the choice of auxiliary variables to enter into the vector $\mathbf{x}_k$ that determines the lack of balance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. This question needs to be addressed in the future. Auxiliary variables are used first during the data collection and then with a somewhat different perspective at the estimation stage. In the data collection, the selected auxiliary variables serve to monitor the balance of the response set and the distance $dist_{r|nr}$ between respondents and

nonrespondents, as the data collection proceeds. Thus at the data collection stage, the auxiliary vector should be one that lends itself well to contrasting respondents with nonrespondents. At the estimation stage, on the other hand, the auxiliary vector serves the purpose yield the most accurate estimates, particularly for the most important survey variables. This latter auxiliary vector is likely to contain more variables than the one used in monitoring the data collection.

References

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.

Japec, L. and Hörngren, J. (2005). *Effects of Field Efforts on Nonresponse Bias and Costs in the Swedish Labour Force Survey.* In: Japec, L. (2005). Quality Issues in Interview Surveys, Some contributions. Ph.D. thesis, Stockholm university.

Laflamme, F. (2009) *Experiences 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., Hörngren, J., Löfgren, T. and Lööw, H. (2007). *Delrapport I: Utveckling av system för kontaktstrategier i intervjuundersökningar med individer och hushåll*. (In Swedish) Internal Document, Statistics Sweden.

Lundquist, P. (2008). *Slutrapport: D2:3, Bortfallsanalys i ULF 2007*. (In Swedish) Internal Document, Statistics Sweden.

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.

Petric, M. (2009). *On the effects of the follow-up in the Statistics Sweden Survey on Household Finances.* Student Thesis, Örebro University.

Schouten, B. and Bethlehem, J. (2009). *Representativeness indicator for measuring and enhancing the composition of survey response. RISQ deliverable*, www.R-indicator.eu .

Schouten, B., Cobben, F. and Bethlehem, J. (2009). *Indicators for the representativeness of survey response*. Survey Methodology, 35, 101-113.

Särndal, C.E. and Lundström, S. (2005). *Estimations in Surveys with Nonresponse*. New York: Wiley.

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.

Särndal, C.E. (2011a). *The 2010 Morris Hansen Lecture: Dealing with Survey Nonresponse in Data Collection, in Estimation*. Journal of Official Statistics, 27, 1-21.

Särndal, C.E. (2011b). *Three factors to signal nonresponse bias – With applications to categorical auxiliary variables.* To appear, International Statistical Review.

Westling, S. (2008). *Delrapport II: Utveckling av system för kontaktstrategier i intervjuundersökningar med individer och hushåll. Analys av HEK 2006*. (In Swedish) Internal Document, Statistics Sweden.

ISSN 1653-7149 (online)

All officiell statistik finns på: **www.scb.se**
Statistikservice: tfn 08-506 948 01

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