Aspects of Responsive Design for the Swedish Living Conditions Survey

Lundquist, Peter
Statistics Sweden
Box 24 300
SE 104 51 Stockholm, Sweden
E-mail: Peter.Lundquist@scb.se

Särndal, Carl-Erik
Statistics Sweden
Klostergatan 23
SE 701 89 Örebro, Sweden
E-mail: Carl.Sarndal@scb.se

1. Introduction

In this article we analyze the data collection as currently carried out in the periodic Swedish Living Conditions Survey (LCS), and we seek forms for this data collection which may be more appropriate in the future. The LCS nonresponse rate is high, about 33%. Types of responsive design are being considered. Groves and Heeringa (2006) 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 incentive to modify the design, if data collection is to be continued at all.

Responsive design can take different forms. Our objective in this article is to change the orientation of the data collection, especially in its later stages, so as to achieve an ultimate response that is "better balanced" or "more representative" than if no special effort had been made. Indicators are then needed to monitor the data collection as it unfolds. Suitable indicators computed on process data and register data are proposed in Särndal (2011) and in Schouten, Cobben, and Bethlehem (2009). Options for responsive design in a Canadian setting are discussed in Mohl and Laflamme (2007) and Laflamme (2009).

A number of earlier studies at Statistics Sweden have illustrated that it is inefficient to continue the data collection under a predetermined scenario driven primarily by the simple motivation to obtain the best possible ultimate response rate, or to reach a predefined response rate. These studies strongly suggest that Statistics Sweden is spending valuable resources on efforts that seem to have little or no effect on the estimates or on the representativity of the ultimate set of respondents.

Section 2 gives a brief description of the 2009 version of the Swedish LCS. We analyze in Section 3 the data collection in the LCS 2009, with the use of register data and process data from Statistics Sweden's telephone interviewing system, WinDATI. The final Section 4 presents results from an "experiment in retrospect" carried out on the LCS 2009 data. We suggest an embedded experiment to be used in the LCS 2011.

2. The Swedish Living Conditions Survey (LCS)

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 article 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.

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 contains a series of "call attempts", which play an important role in our analysis. The LCS 2009 ordinary field work lasted five weeks, at the end of which the response rate was 60.4%; for some sampled persons as many as 20 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 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.

3. Indicators computed on the LCS 2009 data collection

We computed several indicators, using process data from LCS 2009 as well as data from Swedish population registers. We use balance indicators defined in Särndal (2011) and the *R*-indicator proposed by Schouten, Cobben and Bethlehem (2009). We also follow the progression over the data collection of population total estimates for three register variables considered to be correlated with the main study variables in the survey. Being register variables, they have known values for all sampled units, not only for respondents. Consequently, we can follow the development these three estimates as a function of the number of call attempts.

The theoretical background is as follows: The population $U = \{1,\ldots,k,\ldots,N\}$ consists of N units (individuals) indexed $k = 1,2,\ldots,N$. A probability sample s is drawn from U; the unit k has the known inclusion probability $\pi_k = \Pr(k \in s) > 0$ and the known design weight $d_k = 1/\pi_k$. Nonresponse occurs. Let r be the set of units (individuals) having responded at a given point in the data collection. The value y_k of the study variable y is observed for $k \in r$. As the data collection progresses, the size of r increases, but we do not expect r to reach the full sample s. The (design-weighted) survey response rate and its inverse value are, respectively,

$$P = \sum_r d_k / \sum_s d_k \quad ; \quad Q = 1/P = \sum_s d_k / \sum_r d_k$$

(If A is a set of units, $A \subseteq U$, we write $\Sigma_{k\in A}$ as Σ_A.)

The unknown response probability of unit k is denoted $\theta_k = \Pr(k \in r|s)$. The response rate P is an estimate of the (unknown) mean response probability in the population, $\bar{\theta}_U = \sum_U \theta_k / N$.

The use of auxiliary information is crucially important. Denote by $\mathbf{x}_k$ the auxiliary vector value for unit k, assumed available at least for every unit $k \in s$, possibly for every $k \in U$. If $J \geq 1$ auxiliary variables are used, then $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$, where x_{jk} is the value for unit k of the j^{th} auxiliary variable, x_j. The calibration estimator of the total $Y = \sum_U y_k$ used here is

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

where $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ is the nonresponse adjustment factor for unit k. The weights $d_k m_k$ for $k \in r$ are consistent with the design unbiased right hand side of the calibration equation

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

This is one type of nonresponse adjustment by calibration, as explained for example in Särndal and Lundström (2005). It will generally reduce the nonresponse bias, perhaps considerably, depending on the strength of $\mathbf{x}_k$, but some nonresponse bias always remains. Expressions for the remaining bias are given in Särndal and Lundström (2005). When many auxiliary variables are available, as is typically the case in surveys of individuals and household in Scandinavia, the selection of suitable auxiliary variables becomes an important topic, discussed for example in Särndal and Lundström (2008, 2010).

The three register variables that play the role of study variables in our analysis are: *Sickness allowance* (a categorical variable equaling 1 for a recipient of allowance; 0 if not), *Income* (a continuous variable

including employment as well as retirement income), and *Employed* (a categorical variable equaling 1 for an employed person; 0 if not). Since y_k is available for $k \in s$, we compute the unbiased full sample estimate for each of the three variables, $\hat{Y}_{FUL} = \sum_s d_k y_k$.

At each step in the series of call attempts, we compute the percentage relative difference between $\hat{Y}_{FUL}$ and $\hat{Y}_{CAL}$,

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

The adjustment factor m_k in $\hat{Y}_{CAL}$ is based on a fixed auxiliary vector $\mathbf{x}_k$ 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); *Residence ownership* (equaling 1 for owner of residence; 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 an *experimental* **x**-*vector* needed in the next Section 4). We considered it to be suitable for studying the evolution of the estimates over the data collection. The variables are a subset of those used to produce the calibration estimates in the LCS 2009. In Tables 1 and 2, the entries for Attempt *a* (where *a* = 1, 2, 3 or 8) are computed on the union of the sets of persons having responded at attempts 1, 2, …, *a*. 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 at the end of the follow-up period.

Table 1 prompts the following conclusions: (1) At the very end of the data collection (the row "Final"), the *RDF* remains disappointingly large, -3.6%, 2.9%, and 3.1%, respectively. Here, pursuing the data collection according to an unchanging original plan does not make the estimation error small in the end; (2) For all three study variables, *RDF* is much smaller (in fact near zero) at earlier stages of the data collection (full detail not shown here); (3) The numerically important changes in the RDF occur early in the series of attempts.

Table 1. The LCS 2009 data collection: Progression of the response rate P (in per cent) and of RDF (in per cent) for three selected register variables. The auxiliary vector for the computations is the standard x-vector explained in this section.

Step in the data collection	*P*	*RDF*		
		Sickness allowance	Income	Employed
Attempt 1	12.8	10.5	-0.05	-1.3
Attempt 2	24.6	3.3	-1.1	-2.0
Attempt 3	32.8	1.6	-0.4	0.2
Attempt 8	53.0	1.0	2.4	2.4
End ordinary field work	60.4	-0.9	3.3	2.9
Final	67.4	-3.6	2.9	3.1

We also studied how the degree of balance, or representativity, changes as the LCS 2009 data collection progresses. To this end we used indicators whose theoretical background we briefly explain. A more detailed description is given in Särndal (2011). The computable difference $D_j = \bar{x}_{jr} - \bar{x}_{js}$ contrasts the respondent mean $\bar{x}_{jr}$ for the j^{th} auxiliary variable with the full sample mean $\bar{x}_{js}$, where

$$\bar{x}_{jr} = \sum_r d_k x_{jk} / \sum_r d_k , \qquad \bar{x}_{js} = \sum_s d_k x_{jk} / \sum_s d_k .$$

(Means here and in the following are *d*-weighted over the indicated set of units.) If $D_j = 0$ for all *J* auxiliary variables, then we call *r* a *perfectly balanced response set*, for the specified auxiliary vector $\mathbf{x}_k$. The respondents are then on average equal to all those sampled, for every variable in the auxiliary vector $\mathbf{x}_k$. Because the auxiliary vector $\mathbf{x}_k$ is usually multivariate, matrix language becomes necessary. Let

$\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (D_1,...,D_j,...,D_J)'$. Under perfect balance, $\mathbf{D} = \mathbf{0}$. But normally $\mathbf{D} \neq \mathbf{0}$, suggesting departure from balance. We transform the multivariate $\mathbf{D}$ into a univariate statistic, $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, where $\boldsymbol{\Sigma}_s = \Sigma_s d_k \mathbf{x}_k \mathbf{x}_k' / \Sigma_s d_k$. The quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ measures lack of balance, for given survey outcome (*s*, *r*) and given composition of $\mathbf{x}_k$. Increased differences D_j tend to increase the lack of balance. As shown in Särndal (2011), $0 \leq (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})/(Q-1) \leq 1$. Two balance indicators taking values in the unit interval for any outcome (*s*, *r*) and any given composition of $\mathbf{x}_k$ are given by

$$BI_1 = 1 - \sqrt{(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})/(Q-1)} \quad , \quad BI_2 = 1 - 2P\sqrt{\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}} \tag{3.1}$$

For perfect balance, both equal unity. They can also be interpreted with reference to the concept of variance of estimated response probabilities $\hat{\theta}_k$ for $k \in s$,

$$S_{\hat{\theta}}^2 = \sum_s d_k (\hat{\theta}_k - \bar{\hat{\theta}}_s)^2 / \sum_s d_k \tag{3.2}$$

In particular, if ordinary linear least squares is used, the estimates are $\hat{\theta}_k = t_k$ for $k \in s$, where $t_k = \mathbf{x}_k'\mathbf{b}$ with $\mathbf{b} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1}(\sum_r d_k \mathbf{x}_k)$. The variance (3.2) computed with $\hat{\theta}_k = t_k$ is denoted S_t^2. We have $S_t^2 = P^2 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, and therefore

$$BI_1 = 1 - S_t \Big/ \sqrt{P(1-P)} \quad , \qquad BI_2 = 1 - 2S_t$$

The *R*-indicator (where *R* stands for "representativity") of Schouten, Cobben and Bethlehem (2009) was built on the concept of estimated response probabilities. These authors use 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 $k \in s$, and finally the variance $S_{\hat{\theta},\log}^2$, computed by (3.2) with $\hat{\theta}_k = \hat{\theta}_{k,\log}$. Their (unadjusted) *R*-indicator is

$$R = 1 - 2\,S_{\hat{\theta},\log} \tag{3.3}$$

These authors also suggest an "adjusted R-indicator". Its objective is to reduce a bias that (3.3) may have when viewed as an estimate of a corresponding population quantity. Empirical work has shown that the indicators BI_1, BI_2 and R (unadjusted and adjusted) behave very similarly, as illustrated in Table 2 later in this section.

Another important indicator is based on the concept of distance between the response set *r* and the nonresponse set *s*-*r*. For a specified vector $\mathbf{x}_k$ it is given by

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

This distance changes during the course of the data collection, in simple relationship with the balance indicators BI_1 and BI_2: $BI_1 = 1 - \sqrt{P(1-P)} \times dist$; $BI_2 = 1 - 2P(1-P) \times dist$. Table 2 shows the progression of BI_1, BI_2, unadjusted R, adjusted R and $dist$ at the same steps of the LCS 2009 data collection as in Table 1. Ideally, the balance should increase, and the distance $dist$ between respondents and nonrespondents should decrease, as the data collection unfolds. But Table 2 shows the opposite for the LCS 2009 data; the indicators "go the wrong way." The balance decreases (by all four measures), and the distance increases. (Only a few selected steps are shown in Table 2, but the patterns are consistent.) Thus Table 2 reinforces the impression in Table 1, namely, that there is no strong motivation for the follow-up procedure as used in the LCS 2009. Moreover, it is questionable whether the ordinary filed work should continue for as long as it does, rather than stop after say 12 or 15 attempts; these steps are not shown in Table 2.

Table 2. 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, and the distance measure dist. As in Table 1, computations are based on the standard x-vector explained in this section.

Step in data collection	*P*	*BI₁*	*BI₂*	*R unadj.*	*R adjusted*	*dist*
Attempt 1	12.8	0.855	0.904	0.902	0.905	0.433
Attempt 2	24.6	0.802	0.829	0.829	0.831	0.460
Attempt 3	32.8	0.779	0.793	0.794	0.796	0.470
Attempt 8	53.0	0.751	0.752	0.758	0.760	0.499
End ordinary field work	60.4	0.738	0.744	0.752	0.754	0.536
Final	67.4	0.717	0.735	0.742	0.743	0.603

4. An experiment with LCS 2009 data

We plan to introduce an embedded experiment in the data collection for the upcoming LCS 2011. One form that such an experiment may take is to end data collection for designated sample groups after a suitable number of call attempts, while for other groups the data collection would continue for yet some time, whereas for remaining groups it would continue until the very end of the data collection period.

We can to some degree anticipate the results of such an embedded experiment by working with the existing LCS 2009 data file. Although we cannot add more data to that file, we can use it for different "experiments in retrospect," involving interventions in the data collection, thus permitting us to see the effects of such interventions. In particular, we want to follow the balance indicators and the distance *dist* between respondents and nonrespondents as the data collection proceeds. Hence, we delete data in the existing LCS data file in an organized fashion, pretending that data collection has been terminated for specified groups. Thus we sacrifice some of the available LCS data by pretending that, at one or more intermediate "points of intervention," data collection has ended for specified sample subgroups.

Results from one of several "experiments in retrospect" are shown here. For this *experimental strategy*, we worked with an *experimental* **x**-*vector* defined by the complete crossing of three dichotomous auxiliary variables: *Education level* (high, not high), *Residence ownership* (owner, not owner), *Country of birth* (Sweden, other). The auxiliary vector, of dimension $2^3 = 8$, is $\mathbf{x}_k = (\gamma_{1k},...,\gamma_{8k})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ otherwise, $j = 1,...,8$. We used two intervention points, placed before the ultimate end of data collection. At each intervention point we used an attained group response rate of 65 % as a deemed stopping rule for data collection. Attempt 12 of the ordinary data collection was used as Intervention point 1; this implied that data collection was deemed terminated for three of the eight groups, which at that point had achieved response rates greater than 65%. Follow-up attempt 2 was used as Intervention point 2; this led to declaring data collection terminated for one other group. For the remaining four groups, data collection continued until the very end. For the least responsive group, the response rate at the end was only 44.6%, still far from 65%.

Table 3. The experimental strategy: Response rate (in per cent), balance indicator BI₁ and distance measure dist, computed with the experimental x-vector.

Step in the data collection	*P*	*BI₁*	*dist*
Attempt 12	57.7	0.805	0.394
Follow-up attempt 2	61.5	0.824	0.361
Final	63.9	0.843	0.326

In Table 3 we note that the negative impression from Table 2 (based on the actual LCS 2009 data collection) has now been broken. The interventions give a desired favorable result. Both BI_1 and *dist* improve: The balance BI_1 increases, and the distance *dist* decreases, as the data collection unfolds. In

Table 4, we compare results computed on the experimental data collection strategy with results computed on the actual LCS 2009 data collection, in both cases with the standard auxiliary vector, because it is close to the one likely to be used in practice. The experimental strategy is seen to be much better: *RDF* improves (decreases in absolute value) for all three register variables examined, although the improvement is modest for Income and Employed. The balance BI_1 improves markedly from 0.717 to 0.765, and *dist* is significantly reduced from 0.603 to 0.489. But the most striking improvement realized by the experimental strategy (although not explicit in Table 4) is the large reduction in the number of call attempts: 48,883 attempts are used to reach the 63.9% final response rate in the experimental strategy, as compared with 53,258 attempts used to reach the 67.4% final response rate in the actual LCS 2009 data collection. Hence the experimental strategy reduces the number of call attempts by 8.2%. Thus we can obtain improvement (in accuracy of estimates and in balance) at a substantially lower data collection cost.

Table 4. The experimental data collection strategy compared with the actual LCS 2009 data collection: RDF (in per cent) for three register variables, and values of indicators BI_1 and dist. The computations are based on the standard auxiliary vector.

Data collection	*P*	RDF			*BI*1	*dist*
		Sickness allowance	Income	Employed		
Experimental strategy	63.9	-1.6	2.7	3.0	0.765	0.489
Actual LCS 2009	67.4	-3.6	2.9	3.1	0.717	0.603

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**, 439-457.

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.

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.

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

Schouten, B. and Bethlehem, J. (2009). Representativeness Indicator for Measuring and Enhancing the Composition of Survey Response. RISQ deliverable, www.R-indicator.eu .

Särndal, C.-E. and Lundström, S. (2005). Estimation 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. (2011). Dealing with Survey Nonresponse in Data Collection, in Estimation. Journal of Official Statistics, **27**, 1-21.