Journal de la Société Française de Statistique
Vol. 155 No. 4 (2014)

Balancing the response and adjusting estimates for nonresponse bias: complementary activities

Titre: Équilibrage de la réponse et ajustement des estimateurs pour biais de non-réponse : activités complémentaires

Carl-Erik Särndal[1] and Peter Lundquist[2]

Abstract: One objective of Responsive Design is to manage the data collection through appropriate planning and intervention, so as to promote in the end a well-balanced or well representative set of respondents. At that stage, auxiliary information, including paradata, plays a crucial role. But regardless of what can be accomplished during data collection, accurate estimation is the ultimate goal. The auxiliary variables play an important role at that stage as well, as when calibrated weights are used for adjustment in order to reduce the nonresponse bias that nevertheless affects the estimates.

The concept of imbalance of the survey response is central in this article. We define and measure its components, total, marginal and conditional imbalance. We propose methods based on response propensity, observed continuously throughout the data collection, for obtaining a well-balanced ultimate response. We apply the methods to data from a major Swedish survey, and we explore how a successful reduction of imbalance may contribute further to reducing the bias of estimates, over and beyond what calibration adjustment will accomplish in that regard.

Résumé : Un des objectifs d'une collecte adaptative de données est de profiter d'une planification et d'une intervention appropriées, afin d'obtenir au final un ensemble de répondants bien équilibré ou bien représentatif. A ce stade, l'information auxiliaire, qui inclue les paradonnées, joue un rôle central. Mais quoique l'on puisse accomplir durant la période de la collecte, le but ultime est d'obtenir des estimations précises. Au stade de l'estimation, les variables auxiliaires jouent également un rôle important, comme lorsque des poids calés sont utilisés pour réduire le biais de non-réponse qui affecte néanmoins les estimations.

Le concept de déséquilibre de la réponse est central dans cet article. Nous définissons et nous mesurons ses composantes, le déséquilibre total, marginal ou conditionnel. Nous proposons des méthodes basées sur la propension (ou l'intensité) de la réponse, observable de façon continue pendant la collecte de données, dans le but d'obtenir une réponse ultime bien équilibrée. Nous appliquons ces méthodes à des données d'une importante enquête suédoise, et nous examinons dans quelle mesure une réduction bien réussie du déséquilibre peut contribuer à réduire le biais, au-delà de ce qu'un ajustement par calage peut apporter.

Keywords: Auxiliary information, Household surveys, Imbalance, Nonresponse, Paradata, Responsive design
Mots-clés : Déséquilibre, Enquêtes ménage, Information auxiliaire, Non-réponse, Paradonnées, Sondage adaptatif
AMS 2000 subject classifications: 62D05, 62G05

1. Introduction and literature review

General objectives for Responsive Design were formulated in Groves (2006) and Groves and Heeringa (2006). A number of developments have followed. The terms *adaptive design* and *re-*

[1] Ph.D., Professor Emeritus, Statistics Sweden.
E-mail: carl.sarndal@telia.com

[2] Ph.D., Senior Methodologist, Statistics Sweden.
E-mail: peter.lundquist@scb.ce

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

Avenues for developing adaptive designs are reviewed in Wagner (2008). Responsive design in a Canadian setting is reviewed in Mohl and Laflamme (2007) and Laflamme (2009).

Responsive Design focuses, by definition, on the data collection phase of a survey. One prominent idea is that the data collection may be inspected at suitable decision points, in order to bring perhaps a change of direction or emphasis, in the hope that a better composition of the set of respondents will ultimately pave the way for less bias in the estimates.

At the end of the data collection, the final set of respondents should be representative, or well balanced. Different ways have been suggested to promote this goal. Case prioritization is considered in Peytchev et al. (2010). Stopping rules aimed at halting data collection attempts for designated sample units is considered in Rao et al. (2008) and in Wagner and Raghunathan (2010). Couper and Wagner (2012) discuss uses of paradata to manage the survey response.

Nonresponse methodology is examined in the recent *Handbook of Nonresponse in Household Surveys* by Bethlehem et al. (2011). This handbook proposes a measure of the representativity of the survey response, the R-indicator. It is derived from the idea of varying response probability among the population units. Since only a sample is available, estimated response probabilities are used to construct the basic R-indicator as $\hat{R} = 1 - 2\hat{S}$, where $\hat{S}$ is the standard deviation of response probability estimates. Related references are Schouten and Bethlehem (2009), Schouten et al. (2009) and Schouten et al. (2011).

At the estimation stage, estimates must nevertheless be produced with the response, more or less representative, that was finally realized and recorded. Understandably, there is some tendency in the literature to look separately at the two activities, realizing representativity at the data collection stage and reducing bias at the estimation stage. The two activities, both striving for accurate estimation, are of course interrelated.

The nonresponse bias in the estimates cannot be quantified or fully corrected, but indicators of the risk of bias can be useful, as reviewed in Wagner (2012) and in Kreuter et al. (2010). A basic idea in seeking evidence of bias is to observe how the response rate varies between demographic groups, as explored in Peytcheva and Groves (2009). Proxy pattern-mixture analysis is a method proposed by Andridge and Little (2011) for assessing non-response bias for the mean of a survey variable. The selection of the best auxiliary variables for reducing the bias as much as possible is discussed in Särndal (2011b), Särndal and Lundström (2008) and Särndal and Lundström (2010).

The concept of response propensity has been useful for nonresponse adjustment methods and is used in this article also. One well-known method is to compute a response propensity score by a logistic regression of the nonresponse indicator on the auxiliary variables, and then form adjustment cells based on this score. This is particularly efficient when there is ample auxiliary information. Vartivarian and Little (2002) consider adjustment cells based on joint classification by the response propensity and summary predictors of the outcomes, to exploit residual associations between the covariates and the outcome after adjusting for the propensity score. Response propensity is also a prominent theme in Brick and Jones (2008).

As often pointed out, two factors influence bias and variance of the estimates: the degree to which the auxiliary information explains the study variable and the degree to which it explains the response indicator. Such explanation, in both cases, is in practice only realized to a degree, and not perfectly, and there is an interaction, as recognized in Little and Vartivarian (2005).

How important is it really to seek better balance in the data collection? Is it worth the effort? Coming to the estimation stage, are the estimates really significantly improved by balancing having taken place during data collection? Might one not just carry out the data collection in a standardized, simpler and less expensive manner, and leave the adjustment-based on the auxiliary information - to the estimation stage? This article attempts to get some perspectives on these questions.

Thus we are led to ask: What can be done (or what should be done) at the data collection stage? What remains to be done at the estimation stage? The answers depend to some degree on the survey environment. Scandinavia and The Netherlands are privileged in that surveys, especially those on households and individuals, can rely on rich sources of auxiliary information for adjustment at the estimation stage. Many other countries are less well equipped and only simple forms of adjustment become possible.

The material in this article is presented as follows: Following an introduction (section 2), a measure of imbalance in the response is defined, relative to a chosen auxiliary vector (sections 3, 4, 5); that auxiliary variables serve in two ways - for directing the data collection and/or in calibration at the estimation stage - is made clear (section 6); methods for monitoring the data collection with the aid of response propensity are outlined (section 7); an analysis of imbalance (ANIMB) is formulated (section 8); response monitoring and ANIMB are illustrated empirically with data from the Swedish Living Conditions Survey (sections 9, 10); then the focus shifts to the estimation stage, and bias adjustment of estimates is discussed (sections 11, 12); the question whether balancing the response during data collection yields true advantages for the estimation is examined, theoretically and empirically (sections 13, 14).

2. The survey background

We consider a context of probability sampling, primarily for surveys on individuals and households (but not limited to those), supported by an ample supply of auxiliary variables, as is usually the case in Scandinavia. The target population $U = \{1,...,k,...,N\}$ consists of N units (individuals) indexed $k = 1,2,\dots,N$. A probability sample s is drawn from U; unit k has the known inclusion probability $\pi_k = \Pr(k \in s) > 0$, and the known design weight $d_k = 1/\pi_k$. These "d-weights" are used throughout this article in computing means, variances and other statistics.

The variables that enter into consideration are: The study variable (y-variable), the auxiliary variables (x-variables), and the response indicator variable denoted I.

Surveys of national importance usually involve many study variables. To present the reasoning, we focus on one of them. Denote y_k the value for unit k of the study variable y for which we wish to estimate the population total $Y = \sum_U y_k$. (A sum $\sum_{k \in A}$ over a set of units $A \subseteq U$ will be written as $\sum_A$.) If the response were complete, this estimation would be based on values y_k then available for all units $k \in s$. But the response is incomplete. At the end of the data collection period, the value y_k is available only for a subset r of the sample s.

The response indicator variable denoted I has value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$. Auxiliary variables (x-variables) play an important role. Some are used at the data collection stage, others enter the scene at the estimation stage. In Section 6 we elaborate on this division of the auxiliary variables into two categories.

An auxiliary vector is made up of a number of auxiliary variables. The generic notation for an auxiliary vector is $\mathbf{x}$; its value $\mathbf{x}_k$ is known at least for all units $k \in s$, possibly for all $k \in U$. The x-variables in the vector can be continuous or categorical; the latter is the case in many applications in statistical agencies. Section 9 contains an example of a list of categorical x-variables fairly typical for a Swedish survey on individuals. The dimension of $\mathbf{x}$, denoted $J \geq 1$, may be quite large, as when it incorporates a number of categorical variables, each with a number of classes.

For technical convenience, the $\mathbf{x}$-vectors we use satisfy the following requirement: There exists a constant vector $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all units k. It is not a major restriction; all that is required is one such vector $\boldsymbol{\mu}$. Most $\mathbf{x}$-vectors of importance are covered. When the $\mathbf{x}$-vector codes a set of mutually exclusive and exhaustive categories, it is of the type $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$, where the only "1" identifies the category of unit k. Then $\boldsymbol{\mu} = (1, \ldots, 1, \ldots, 1)'$ satisfies the requirement. When the $\mathbf{x}$-vector is used to code, say, four mutually exclusive and exhaustive age classes and, in addition, the univariate variable "sex" equal to 1 for male and 0 for female, the dimension is $J = 4 + 1 = 5$ (age and sex are not crossed), and $\boldsymbol{\mu} = (1, 1, 1, 1, 0)'$ satisfies the requirement. If x is a univariate continuous variable, and $\mathbf{x}_k = (1, x_k)'$, as for a regression with intercept, then $\boldsymbol{\mu} = (1, 0)'$ satisfies the requirement. (But $\mathbf{x}_k = x_k$, as in a regression without an intercept, would not be covered.) The reason why the requirement streamlines several derivations in this paper will be apparent at the first use made of it, the derivation of formula (5).

The time perspective is important. In many surveys of importance, the data collection extends over a period of days or weeks or even months. We follow the data collection as a function of some time related aspect, such as the data collection day or the call attempt number. These portray the data collection somewhat differently. For example, the tenth call attempt may occur on different days for two different sample units.

Here we use the call attempt number as the time dependent element. The process can be followed and monitored with the aid of data collection devices such as Statistics Sweden's WinDATI, which records all call attempts by a staff of interviewers, destined to establish a telephone contact, and then a completed interview with a selected sample unit.

There is a series of successively larger response sets $r^{(a)}$, where a refers to the time dimension, $a = 1, 2, \ldots$, and

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

Here $r^{(a)}$ is the set of units having delivered the value y_k at a certain point a (after a call attempts, or after a data collection days). For simpler notation, we let r refer to any one of the increasingly larger response sets. A recording device such as WinDATI allows us to intervene in the data collection and if necessary to redirect it, to realize in the end a better balanced final response set.

For the response r, the realized (design-weighted) response proportion of the sample s is

$$P = \sum_r d_k / \sum_s d_k \tag{2}$$

The proportion P increases as the data collection evolves. It is a basic descriptive characteristic of the response. In principle, the ultimate response set r satisfies $r \subseteq s \subseteq U$, but by practical necessity, data collection will almost always stop before r has reached the full probability sample s. The ending value P is the ultimate response rate for the survey. Then the values y_k for $k \in r$ are, together with auxiliary vector values $\mathbf{x}_k$ for $k \in s$, the material for estimating parameters such as the population total $Y = \sum_U y_k$.

3. Measuring the imbalance of a response set

The set of respondents r present at any given point in the data collection is a more or less well balanced representation of the probability sample s that contains r. The imbalance property is formulated in terms of $\mathbf{x}$-vector means. The computable (design weighted) $\mathbf{x}$-vector mean is $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$ for the response set r and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$ for the full sample s. If $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, we say that the response set is *perfectly balanced* with respect to the chosen $\mathbf{x}$-vector. Ordinarily we do not achieve this in a survey, at least not exactly, but during data collection we can strive to come close.

The mean difference vector $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ is composed of the differences $\bar{x}_{jr} - \bar{x}_{js}$, $j = 1, \ldots, J$, which is the difference for the jth x-variable between the respondent mean, $\bar{x}_{jr} = \sum_r d_k x_{jk} / \sum_r d_k$, and the full sample mean, $\bar{x}_{js} = \sum_s d_k x_{jk} / \sum_s d_k$.

A large difference $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ signifies that the response set is not well balanced. This difference is to some degree inflated by a high nonresponse rate $1 - P$, because $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (1 - P)(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r})$, where $\bar{\mathbf{x}}_{s-r} = \sum_{s-r} d_k \mathbf{x}_k / \sum_{s-r} d_k$ is the mean for the nonresponse set $s - r$. For a constant separation $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{s-r}$ between response and nonresponse means, the difference $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ is "unrealistically large" if the response rate P is low. We take this into account in defining the (scalar) imbalance statistic as

$$IMB(r,\mathbf{x}|s) = P^2 \times (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\Sigma_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s) \tag{3}$$

where the $J \times J$ weighting matrix, assumed non-singular, is

$$\Sigma_s = (\textstyle\sum_s d_k \mathbf{x}_k \mathbf{x}'_k)/(\sum_s d_k). \tag{4}$$

It is important to note that imbalance is measured with respect to a given **x**-vector, for the given sample s. For one and the same response set r, the imbalance can be numerically quite different depending on how many variables are included in the **x**-vector, and which ones. The notation $IMB(r,\mathbf{x}\,|s)$ reflects the fact that imbalance is a function of (i) the set of respondents r present at a certain point in the data collection and (ii) the specified **x**-vector, that is, the choice of x-variables for the vector. The value $IMB(r,\mathbf{x}|s)$ depends on the characteristics (the values $\mathbf{x}_k$) of the units k, respondents as well as non-respondents. The imbalance tells considerably more about the survey response than just its proportion of the sample, which is the response rate, $P = \sum_r d_k / \sum_s d_k$. The response rate alone is insufficient to describe the quality of the set of respondents.

Even in a case where the response r is just a small subset of s, $IMB(r,\mathbf{x}|s) = 0$ can happen (but is unlikely to do so), namely if the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ holds. A typical feature of the data collection is that the increasing response rate P causes $\bar{\mathbf{x}}_r$ to draw nearer (and in the limit become equal to) the fixed sample mean $\bar{\mathbf{x}}_s$. For high response, r is near s and $\bar{\mathbf{x}}_r \approx \bar{\mathbf{x}}_s$. The factor P^2 in the definition (3) regulates the tendency for the imbalance to be artificially high when response is low.

A reason for interposing the inverse of Σ_s in (3) is that a simple upper bound can then be stated on the imbalance: Given s, we have $0 \leq IMB \leq P(1-P)$, whatever r and the values $\mathbf{x}_k$ for $k \in s$. For example, for $1-P = 20\%$ nonresponse, $0 \leq IMB \leq 0.16$; for 50% nonresponse, $0 \leq IMB \leq 0.25$. For data encountered in practice, IMB is usually much below the upper bound.

Apart from the factor P^2, equation (3) defines the imbalance as a quadratic form in the difference vector $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$. This equation can also be written as

$$IMB(r,\mathbf{x}|s) = P^2 \times (\bar{\mathbf{x}}'_r\Sigma_s^{-1}\bar{\mathbf{x}}_r - 1) \tag{5}$$

This follows from (3) because $\mathbf{x}'_k\Sigma_s^{-1}\bar{\mathbf{x}}_s = 1$ for all k, which is a consequence of the **x**-vector form $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all k:

$$\begin{aligned}
\mathbf{x}'_k\Sigma_s^{-1}\bar{\mathbf{x}}_s &= \mathbf{x}'_k(\textstyle\sum_s d_k\mathbf{x}_k\mathbf{x}'_k)^{-1}(\sum_s d_k\mathbf{x}_k) \\
&= \mathbf{x}'_k(\textstyle\sum_s d_k\mathbf{x}_k\mathbf{x}'_k)^{-1}(\sum_s d_k\mathbf{x}_k\mathbf{x}'_k\boldsymbol{\mu}) \\
&= \mathbf{x}'_k\boldsymbol{\mu} = 1.
\end{aligned}$$

Unless emphasis is required, we use the compact notation IMB for $IMB(r,\mathbf{x}|s)$. The notion of imbalance was used in Särndal (2011a) and is related to the R-indicator of Schouten et al. (2009).

4. The concepts of distance and balance

For a given **x**-vector, the (weighted Euclidian) distance between respondents and nonrespondents is measured, at any given point in the data collection, by

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

where *nr* stands for nonresponse. It has a simple relation to the imbalance:

$$dist_{r|nr} = \sqrt{IMB}/P(1-P)$$

In an efficient data collection, the distance should decrease, or at least not get markedly greater, when the response set grows larger.

From $IMB \leq P(1-P)$ follows an upper bound on the distance, $dist_{r|nr} \leq 1/\sqrt{P(1-P)}$. For example, for 50% nonresponse, $dist_{r|nr} \leq 2$ for any response r and any vector specification **x**.

A measure of balance (the negation of imbalance) on a unit interval scale can be constructed in a variety of ways. One possibility, suggested by $IMB \leq P(1-P)$, is the balance indicator

$$BI = 1 - \sqrt{\frac{IMB}{P(1-P)}} = 1 - \sqrt{P(1-P)} \times dist_{r|nr}.$$

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

$$BI_{alt} = 1 - 2\sqrt{IMB} = 1 - 2P(1-P) \times dist_{r|nr}.$$

The indicator BI_{alt} is a special case of the R-indicator (with R for "representativity") of Schouten et al. (2009). They developed it from the idea of a variability in the (unknown) response probabilities of the population units. A computable R-indicator is derived by first getting response probability estimates $\hat{\theta}_k$ for $k \in s$, then obtain their standard deviation $S_{\hat{\theta}}$, and then let the R-indicator be defined as $R = 1 - 2S_{\hat{\theta}}$. When the estimates $\hat{\theta}_k$ are derived by linear regression of I_k on $\mathbf{x}_k$, this construction gives BI_{alt}. The literature on the R-indicator emphasizes the case where the $\hat{\theta}_k$ are derived by logistic (rather than linear) regression.

5. The case of mutually exclusive groups

The case of mutually exclusive and exhaustive groups is of special interest because in this case the imbalance (3) takes a particularly simple and transparent form. Then the vector $\mathbf{x}_k$ has $J-1$ entries "0" and one single entry "1" pointing out the group to which the unit k belongs. The imbalance statistic (3) is then a sum of J nonnegative terms:

$$IMB = \sum_{j=1}^{J} C_j$$

where $C_j = \ W_j \times (P_j - P)^2$ is the imbalance attributed to category j, $W_j = \sum_{s_j} d_k / \sum_s d_k$ is the sample proportion and $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ the response rate in that category, and P is the overall response rate (2). During data collection we can follow the evolution of the contributions C_j of the different groups to the total imbalance, as in Lundquist and Särndal (2013). In a data collection claimed to be efficient, we like to see a decreasing tendency in the terms C_j. In the terminology of (Bethlehem et al., 2011, p. 190), we can call C_j the unconditional partial imbalance for category j of the classification with J categories.

6. Auxiliary variables of two kinds

We make an important distinction in regard to the auxiliary variables. Potentially, there may be many, as is often the case in the Scandinavian countries. We assume here that a set of auxiliary variables has been identified for the process that includes a data collection stage and an estimation stage. The variables are of two kinds, with different functions. Some are designated for monitoring and steering the data collection, in order to obtain a final response set that is reasonably well balanced with respect to precisely those variables. They make up the *monitoring vector* denoted $\mathbf{x}_{MV}$, of dimension J. The monitoring vector is an instrument for the data collection; other available auxiliary variables stay neutral at that stage, but are important at the estimation stage where they enter, usually together with those in $\mathbf{x}_{MV}$, in the computation of calibrated adjustment weights in estimating parameters such as the population total $Y = \sum_U y_k$; the objective is then to control or reduce bias and variance.

For practical reasons, the monitoring vector is usually restricted to a selection of rather few x-variables. An important case is when this vector identifies a set of mutually exclusive and exhaustive sample subgroups. A practical advantage is that it is relatively easy for the survey manager to monitor a data collection directed to a modest number of groups. How do we select these groups? Arguably, they should be groups for which we expect large differences in response rate, because such differences can cause large imbalance.

In a regularly repeated survey, we may have a good idea at the outset what those groups might be. The choice is less obvious in a survey carried out for the first time. One possibility is then to identify suitable groups by an analysis. One such tool is classification tree analysis, CHAID. It may for example be carried out at the end of the ordinary data collection and serve to identify the groups that should receive particular emphasis in the follow-up. Tree analysis is described in (Bethlehem et al., 2011, p. 263-265).

7. Monitoring based on the response propensity

The data collection is monitored with the aid of response propensities computed on a chosen monitoring vector $\mathbf{x}_{MV}$ with value $\mathbf{x}_{MVk}$ for unit k, known for all $k \in s$. Hence $\mathbf{x}_{MV}$ is an auxiliary vector with a special function, namely, to direct the data collection. We assume for $\mathbf{x}_{MV}$ also the property $\boldsymbol{\mu}'\mathbf{x}_{MVk} = 1$ for all k and some constant vector $\boldsymbol{\mu}$. We measure response propensity at a given point in the data collection with the predicted values from the regression of the response indicator I on $\mathbf{x}_{MV}$. By least squares, we determine first the value $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ that minimizes $\sum_s d_k(I_k - \boldsymbol{\lambda}'\mathbf{x}_{MVk})^2$. This leads to the response propensity computable for all $k \in s$ by

$$\hat{P}_{MVk} = \hat{\boldsymbol{\lambda}}'\mathbf{x}_{MVk} = (\textstyle\sum_s d_k I_k \mathbf{x}_{MVk})'(\sum_s d_k \mathbf{x}_{MVk}\mathbf{x}'_{MVk})^{-1}\mathbf{x}_{MVk}. \tag{6}$$

Although $\hat{P}(r, \mathbf{x}_{MVk} | s)$ would be more informative, the simpler notation $\hat{P}_{MVk}$ will be used, unless special emphasis is needed. The mean response propensity is equal to the response rate for the response set r: $\sum_s d_k \hat{P}_{MVk} / \sum_s d_k = \sum_r d_k / \sum_s d_k = P$. This is shown with the aid of the requirement $\boldsymbol{\mu}' \mathbf{x}_{MVk} = 1$ for all k, similarly as in the proof of $\mathbf{x}'_k \Sigma_s^{-1} \bar{\mathbf{x}}_s = 1$ at the end of Section 3.

By definition, the variance of the response propensities is $S^2_{\hat{P}MVs} = \sum_s d_k (\hat{P}_{MVk} - P)^2 / \sum_s d_k$. Developing the square and some matrix manipulation reveals an important property of this variance: It is equal to the imbalance of the response r relative to the monitoring vector:

$$S^2_{\hat{P}MVs} = IMB(r, \mathbf{x}_{MV} | s)$$

Hence, $S^2_{\hat{P}MVs}$can be computed either as the variance of $\hat{P}_{MVk}$, or through the quadratic form (3) with $\mathbf{x} = \mathbf{x}_{MV}$.

The data collection is examined, and possible action is taken, at number of intervention points, preferably at least five. These are specified in advance and are defined here by the contact attempt number. In our experiments, contact attempts are considered discontinued for the units having attained, at each intervention point, a high (comparatively speaking) response propensity $\hat{P}_{MVk}$. These units are not further contacted; we shall say that they are "left cold". The rationale is that the subsequent contact attempts should focus on units that have so far shown less propensity to respond. Hence data collection attempts cease at different points for different units.

There are two variations of this procedure, the Threshold method and the Fixed proportion method. We now describe them more formally, assuming that a set of suitable intervention points and a suitable monitoring vector $\mathbf{x}_{DC}$have been identified in advance.

The Threshold method. At the first intervention point, the propensity $\hat{P}_{MVk}$ is computed for all units $k \in s$. The units (respondents and nonrespondents) with value $\hat{P}_{MVk}$ greater than a threshold fixed in advance, say 60%, are identified. Contact attempts are discontinued for those units; they are left cold. At the second intervention point, the $\hat{P}_{MVk}$ are recomputed for all units $k \in s$. Those left cold at the first point will have their values $\hat{P}_{MVk}$ somewhat changed (for certain $\mathbf{x}$-vectors not changed at all), but without consequence; they remain cold. Among the remaining units, those (respondents and nonrespondents) with new $\hat{P}_{MVk}$-values greater than the same fixed threshold value are identified, and they are now also left cold. In the same manner, at each of the following intervention points, the values $\hat{P}_{MVk}$ are again recomputed for all $k \in s$, and among the units still in contention, those (respondents and nonrespondents) with $\hat{P}_{MVk}$ greater than the threshold are left cold. Those remaining at the last intervention point continue to be contacted until the very end of the data collection period.

An example with a sample s of $n = 10$ units labelled $k = 1$ to 10, and two intervention points prior to the end of the data collection period, illustrates the procedure in more detail. At the first intervention point, suppose units $k = 1, 2, 3$ and 4 have values $\hat{P}_{MVk}$ that exceed the specified threshold. Of these, suppose $k = 1$ and 2 are nonrespondents, $k = 3$ and 4 respondents. At the second intervention point, $\hat{P}_{MVk}$ is recomputed for all 10 units. That is, recomputed for units $k = 1, 2, 3$ and 4 as well, but any action at point two is restricted to units $k = 5, 6, 7, 8, 9$ and 10.

Suppose that at that point $k = 5,6,7$ and 8 have new values $\hat{P}_{MVk}$ that exceed the same fixed threshold. Of these, suppose that $k = 6$ and 7 are nonrespondents, $k = 5$ and 8 respondents. For units $k = 9$ and $k = 10$, contact attempts continue until the end. There, suppose that $k = 9$ has responded, but not $k = 10$. The final response set in this example is therefore $r = \{3,4,5,8,9\}$ and the nonresponse set is $s - r = \{1,2,6,7,10\}$.

Setting the threshold value requires some insight and planning, based on knowledge about the same survey or about similar surveys. If the response rate for the survey is realistically assessed at around 65%, a threshold of 60% or of 55% may be used.

The Fixed proportion method. In this method a fixed proportion of the sample is identified and left cold at each of L intervention points, defined in advance. The values $\hat{P}_{MVk}$ are computed for all $k \in s$ at each of these points. Their mean $P = \sum_s d_k \hat{P}_{MVk} / \sum_s d_k = \sum_r d_k / \sum_s d_k$ is increasing at each point. The values are size ordered, and $100/(L+1)$ percent of the sample units, those with the largest $\hat{P}_{MVk}$ among those still in contention, are left cold at each point.

For example, if $L = 5$, then a fixed portion $1/6$ of the sample s is left cold at each of the five points, and $1/6$ continues to the very end. Thus at the first point, 16.67% of the sample turns cold; units in that part are no longer approached. At the second point, those 20% with the highest recomputed $\hat{P}_{MVk}$, out of the $5/6$ still in contention, are left cold, and so on. After the fifth point, $1/6$ of the sample remains in contention until the very end.

In both methods, at the end of data collection, the propensities $\hat{P}_{MVk}$ are computed a final time for all units $k \in s$. Their mean P at that point is the ultimate response rate for this monitored data collection. Their variance $S^2_{\hat{P}MVs} = IMB(r, \mathbf{x}_{MV}|s)$ may be substantially lower than in a traditional data collection, without interventions, because it is in the nature of both methods to reduce that variance.

Variations of the threshold method are obtained by setting different values for the threshold. The lower the threshold, the more stringent the data collection and the more uniform the final propensities. Whether the final response rate will be higher or lower than in an absence of interventions depends on how the survey resources are managed. Halting contact attempts for some units tends to lower the response rate; on the other hand, resources are freed by halting and should be used to intensify contact efforts for the less responsive sample members; the response rate may not in the end be any lower than in an absence of interventions.

But in our experiments, where more and more units in a fixed actual response are left cold, the ultimate response rate will be lower than in an absence of interventions, and so will ordinarily the final imbalance.

An advantage of the fixed proportion method is that there is no need to assess or guess a suitable threshold value for $\hat{P}_{MVk}$, something which may not be easy for a first-time survey.

In the empirical Section 10, we illustrate both methods on Swedish survey data. In the threshold method we compare three different monitoring vectors $\mathbf{x}_{MV}$, and for each of these, three different threshold values are used. Lower threshold means more stringent surveillance of the data collection, thereby a reduced variance in the propensities and a reduced final imbalance.

8. Total, marginal, and conditional imbalance

The imbalance $IMB(r,\mathbf{x}|s)$ defined by (3) can be computed for the response r present at any point in the data collection, and for any vector specification $\mathbf{x}$, given the sample s. (The only condition on $\mathbf{x}$ is that the matrix Σ_s in (3) be nonsingular.) It is of interest to see how the imbalance reacts to different choices of the $\mathbf{x}$-vector.

We assume here that a set of auxiliary variables has been identified to serve the two phases of the survey, the data collection and the estimation that follows later. These form a "total $\mathbf{x}$-vector" denoted $\mathbf{x}_{tot}$ with value $\mathbf{x}_{tot,k}$ known for $k \in s$. This "total supply" may be a selection from an even larger pool of potentially available x-variables.

At the termination of data collection, there is a final response set r whose degree of imbalance we wish to measure. In particular, r may be the result of interventions based on response propensity computed on a certain monitoring vector $\mathbf{x}_{MV}$, as described in Section 7. We assume that the variables in $\mathbf{x}_{MV}$ are all or some of those in $\mathbf{x}_{tot}$. If fewer than all are used, the remaining variables in $\mathbf{x}_{tot}$ come into play at the estimation stage, in the calibrated weight computation.

The imbalance in the final response set r relative to $\mathbf{x}_{tot}$ is the *total imbalance*, $IMB(r,\mathbf{x}_{tot}|s)$, which is $IMB(r,\mathbf{x}|s)$ given by (3) when computed on $\mathbf{x} = \mathbf{x}_{tot}$.

We may also wish to measure imbalance with respect to subsets of auxiliary variables. The imbalance relative to a vector $\mathbf{x}_b$ composed of some of the variables in $\mathbf{x}_{tot}$ is the *marginal imbalance* of $\mathbf{x}_b$, $IMB(r,\mathbf{x}_b|s)$, computed by (3) with $\mathbf{x} = \mathbf{x}_b$.

For example, we may wish to evaluate the imbalance relative to those variables in the total vector $\mathbf{x}_{tot}$ that are not active in the data collection, but reserved for use in calibrated weight computation at the estimation stage.

Let $\mathbf{x}_a$ be the complement vector, made up of those variables in $\mathbf{x}_{tot}$ that are not in $\mathbf{x}_b$. The *conditional imbalance* of $\mathbf{x}_a$ controlling for $\mathbf{x}_b$ is defined as the non-negative difference $IMB(r,\mathbf{x}_{tot}|s) - IMB(r,\mathbf{x}_b|s) = IMB(r,\mathbf{x}_a|s,\mathbf{x}_b)$. The analysis of imbalance (ANIMB) table, which can be computed for the response r present at any point in the data collection, takes the following form, illustrated empirically in Section 10:

Source of imbalance	Component of imbalance
Marginal of $\mathbf{x}_b$	$IMB(r,\mathbf{x}_b|s)$
Conditional of $\mathbf{x}_a$ given $\mathbf{x}_b$	$IMB(r,\mathbf{x}_a|s,\mathbf{x}_b) = IMB(r,\mathbf{x}_{tot}|s) - IMB(r,\mathbf{x}_b|s)$
Total	$IMB(r,\mathbf{x}_{tot}|s)$

9. Sweden's 2009 Living Conditions Survey: Generating alternative response sets

The 2009 Swedish Living Conditions Survey (called LCS 2009), a contributor to the European living conditions survey EUSILC, provides data suitable for illustrating the concepts in earlier sections. We start with the response set as actually recorded in that survey. From it we generate 12 new response sets using the two methods presented in Section 7. We then, in Section 10, illustrate total, marginal and conditional imbalance, by computing these quantities for each of the 13 response sets, Actual and 12 generated. It allows us to see how they react to different compositions of the response set.

LCS 2009 is described in Lundquist and Särndal (2013). This sample survey is designed to measure different aspects of social welfare in Sweden, in particular among different population subgroups. The LCS 2009 sample consists of a sample of individuals 16 years and older, drawn from the Swedish Register of Total Population. The data set used in the analysis in this report is a subsample of $n = 8,220$ individuals, taken from the entire LCS 2009 sample. This subsample can be regarded as a simple random sample.

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

"WinDATI events" include not only productive attempts but also call without reply, busy line, contact with household member other than the sampled person, and appointment booking for later contact. When contact and data delivery has occurred, the data collection effort is complete for the sample member in question. Every registered WinDATI event is a "(call) attempt" in the following.

The LCS 2009 ordinary field work lasted five weeks, at the end of which the response rate was 60.4%. For some sampled persons, 30 or more call attempts had then been recorded. This was followed by a three week break during which characteristics of non-interviewed individuals were examined, in order to prepare for the three week follow-up period, which concluded the data collection. All individuals considered by the survey managers to be potential respondents were included in the follow-up effort, which brought the response rate up to an ultimate 67.4%. However, there was no separate strategy or revised procedure for the follow-up. It followed the same routines as the ordinary field work. Hence, there were no attempts at responsive design such as for example a follow-up focusing on underrepresented groups.

For purposes of illustration, we assume that a set of auxiliary variables has been designated for use in the survey. Some or all of these x-variables may be used to monitor the data collection. One possibility is to use *all of them* for monitoring, then to use them again for the calibration at the estimation stage. Another possibility is to use only *some of them* for monitoring, and to use them also, together with the rest, in calibration at the estimation stage. We examine these alternatives. We wish to strike a balance between two objectives: (i) a well-balanced set of respondents when data collection ends, and (ii) getting accurate estimates at the estimation stage.

Many x-variables are potentially available and useful for LCS 2009. We consider here the following selection of x-variables, with notation and definition:
Educ (for Education level) equaling 1 for a person with high education; 0 otherwise;
Owner (for Property ownership) equaling 1 for a person who owns his residence; 0 otherwise;
Origin (for Country of origin) equaling 1 for a person born in Sweden; 0 otherwise;
Phone (for Phone access), equaling 1 for a person with phone number accessible at the start of the data collection; 0 otherwise;
Age (for Age group), coded by four zero/one variables according to the age brackets: -24, 25-64, 65-74, 75+;
Civil (for Civil status); equal to 1 for a married or widowed person; 0 otherwise;
Gender; equal to 1 for male, 0 otherwise.
All are dichotomous with the exception of Age which has four categories.

We illustrate the imbalance concept by comparing results for several different final response sets r. The first of these, used as a reference, is the data collection as actually carried out in the LCS 2009. There were no interventions. The data collection progressed with an essentially unchanging format, as described earlier in this section. The other 12 response sets are generated from the actual LCS 2009 response set by "interventions after the fact," using three different monitoring vectors $\mathbf{x}_{MV}$.

In this exercise, we chose five intervention points during the data collection: Attempts 3, 6 and 9 of the ordinary data collection, the end of the ordinary data collection, and attempt 3 of the follow-up. At each of these points, $\hat{P}_{MVk}$ is computed for all $k \in s$, where s is the LCS subsample of size $8,220$. In the threshold method, units with $\hat{P}_{MVk}$ greater than the specified threshold are identified and "left cold". That is, we pretend that data collection attempts have been stopped for these units. Their y-values are not used to compute estimates. In the equal proportions method, $1/6$ of the sample, those with the highest $\hat{P}_{MVk}$, is left cold at each point.

We consider three plans, depending on the monitoring vector for the data collection. In Plan 1, *Educ*, *Owner* and *Origin* are variables selected for monitoring; the vector is

$$\mathbf{x}_{MV1} = (Educ \times Owner \times Origin) \tag{7}$$

This crossing of three dichotomous variables gives eight mutually exclusive and exhaustive groups; $\mathbf{x}_{MV1}$ has dimension $J = 2^3 = 8$ and equally many possible values, $\mathbf{x}_{MV1k}= (\gamma_{1k}, \ldots, \gamma_{8k})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ otherwise. For sake of argument we pretend that Plan 1 reflects a professional judgment and desire to monitor the data collection through precisely those eight groups, so that the remaining variables, *Phone*, *Age*, *Civil* and *Gender*, are considered reserved for computing calibrated weights, together with those in the monitoring vector.

We consider two alternative plans. In Plan 2, the data collection is monitored by

$$\mathbf{x}_{MV2} = ((Educ \times Owner \times Origin) + Age) \tag{8}$$

Age has four classes and is coded as a three dimensional vector (with four possible values) to make possible the matrix inversion in (6) and other places. The dimension is $8+3=11$, and the number of possible values $\mathbf{x}_{MV2k}$ (the number of recognized properties among the sample units) is $8\times 4=32$.

In Plan 3, the monitoring vector for the data collection incorporates all the x-variables in the supply,

$$\mathbf{x}_{MV3} = ((Educ\times Owner\times Origin)+Phone+Age+Civil+Gender) \tag{9}$$

In this vector, *Educ*, *Owner* and *Origin* are crossed, while the variables *Phone*, *Age*, *Civil* and *Gender* enter in a "side-by-side" manner, giving $\mathbf{x}_{MV3}$ the dimension $2^3+1+3+1+1=14$, with $8\times 2\times 4\times 2\times 2=256$ possible values.

We consider 13 different response sets r: The actual LCS 2009 response and 12 generated ones. Nine of these are obtained by combining the monitoring vectors $\mathbf{x}_{MV1}$,$\mathbf{x}_{MV2}$ and $\mathbf{x}_{MV3}$, given in (7), (8) and (9), with three different thresholds, 65%, 55% and 50%. For each of the nine possibilities, we compute, at each intervention point and for all units $k\in s$, the propensities $\hat{P}_{MVk}$ given by (6) and those for which $\hat{P}_{MVk}$ exceeds the threshold are left cold, pretending that data collection attempts have been stopped for these units. This gives nine constructed response sets. For each vector, the threshold progression from 65% to 50% pushes in a direction of reduced variability in the final propensities $\hat{P}_{MVk}$; their variance $S^2_{\hat{P}MVs} = IMB(r,\mathbf{x}_{MV}|s)$ is progressively reduced.

We also used the equal proportions method to create three more response sets, with the same three monitoring vectors, and with the same intervention points.

With its 256 recognized categories of sample units, the vector $\mathbf{x}_{MV3}$ generates a smooth distribution, not far from normal in appearance, of the $8,220$ computed values $\hat{P}_{MVk}$. To illustrate theory, the plenitude of $\mathbf{x}_{MV3}$ makes it the preferred choice among the three suggested vectors, but a consideration for practice is that monitoring few categories is much easier, as in the case of only eight categories recognized by $\mathbf{x}_{MV1}$.

The interest lies now in comparing these 13 response sets r in regard to their balance properties, as done with an ANIMB analysis in the next section.

10. Empirical evidence: Analysis of imbalance for alternative response sets

We illustrate the concepts total, partial and conditional imbalance introduced in Section 8 by applying them to the 13 response sets r (Actual, and 12 generated) from the Swedish LCS 2009 described in the preceding section. The results are given in Table 1.

We choose to measure total imbalance with respect to the vector formed by all the x-variables:

$$\mathbf{x}_{tot} = ((Educ\times Owner\times Origin)+Phone+Age+Civil+Gender). \tag{10}$$

This vector with dimension 14 is identical to the monitoring vector $\mathbf{x}_{MV3}$ in (9). We pretend here that *Phone*, *Age*, *Civil* and *Gender* are variables intended primarily for calibrated weight computation in the estimation. Therefore we let $\mathbf{x}_b$ be the vector with dimension $1+4+1+1=7$ composed of these variables,

$$\mathbf{x}_b = (Phone + Age + Civil + Gender). \tag{11}$$

The remaining variables form the vector $\mathbf{x}_a$. For each of the 13 response sets, we compute and compare the ANIMB components $IMB(r, \mathbf{x}_{tot}|s)$ (the total imbalance), $IMB(r, \mathbf{x}_b|s)$ (the marginal imbalance of $\mathbf{x}_b$) and $IMB(\mathbf{x}_a, r|s, \mathbf{x}_b) = IMB(r, \mathbf{x}_{tot}|s) - IMB(r, \mathbf{x}_b|s)$ (the conditional imbalance, controlling for $\mathbf{x}_b$).

Certain patterns are predictable for the magnitude of the ANIMB components and we wish to see if these are confirmed empirically. For one and the same monitoring vector, we expect to see the imbalance $IMB(r, \mathbf{x}_{tot}|s)$ to decrease with the threshold, because the data collection becomes more stringent. Also, expanding the monitoring vector, as in the progression from (7) to (9), should be accompanied by a decreasing imbalance, for one and the same threshold.

TABLE 1. *Response rate, distance, and imbalance (total, marginal, conditional) for Actual LCS2009 response, and* $3 \times 4 = 12$ *derived response sets, four for each of three monitoring vectors.*

			Imbalance		
Response set	Response rate P	Distance $dist_{r\|nr}$	Total $100 \times IMB_{tot}$	Marginal $100 \times IMB_b$	Conditional $100 \times IMB_{a\|b}$
Actual	67.4	0.623	1.881	1.258	0.622
MV1; TH65	63.3	0.499	1.338	1.086	0.252
MV1; TH55	56.6	0.423	1.080	1.028	0.051
MV1; TH50	52.5	0.407	1.034	0.961	0.072
MV1; eql	53.4	0.462	1.320	0.989	0.331
MV2; TH65	63.3	0.478	1.230	0.982	0.248
MV2; TH55	57.7	0.401	0.956	0.839	0.117
MV2; TH50	53.5	0.348	0.747	0.727	0.020
MV2; eql	54.5	0.362	0.808	0.757	0.051
MV3; TH65	63.3	0.456	1.122	0.857	0.265
MV3; TH55	56.9	0.328	0.648	0.547	0.100
MV3; TH50	53.3	0.262	0.426	0.392	0.034
MV3; eql	54.6	0.284	0.495	0.441	0.054

The marginal imbalance $IMB(r, \mathbf{x}_b|s)$ is likely to change little as long as variables in $\mathbf{x}_b$ are left outside the monitoring vector (Plan 1), but when they are included (to some extent in Plan 2, more completely in Plan 3), the value of $IMB(r, \mathbf{x}_b|s)$ should react in a decreasing direction. The patterns for the conditional imbalance $IMB(\mathbf{x}_a, r|s, \mathbf{x}_b) = IMB(r, \mathbf{x}_{tot}|s) - IMB(r, \mathbf{x}_b|s)$ are less predictable ahead of time.

Table 1 is arranged to show the effect of changes in the $\mathbf{x}$-vector and changes in the threshold. The notation is as follows: $MV1$, $MV2$ and $MV3$ refer to the monitoring vectors, (7), (8) and (9).

TH refers to the threshold, 65%, 55% or 50%; eql refers to the equal proportions method, and $IMB_{tot} = IMB(r, \mathbf{x}_{tot}|s)$, $IMB_b = IMB(r, \mathbf{x}_b|s)$, $IMB_{a|b} = IMB(r, \mathbf{x}_{tot}|s) - IMB(r, \mathbf{x}_b|s)$. The table generates the following comments.

- All of the 12 monitored response sets show lower, in most cases considerably lower, figures than Actual response, on all accounts: distance and the three components of imbalance. The reduction of the imbalance from 1.881 (Actual) to 0.426 ($MV3$, $TH50$) is, in our experience large, although not reduced to near-zero levels; this would not be expected in practice.
- The total imbalance IMB_{tot} and the distance $dist_{r|nr}$ develop in perfect regularity: For one and the same MV, both drop in when the threshold decreases from 65% to 50%. For one and the same threshold, both drop in expanding the monitoring vector from $MV1$ to $MV3$.
- The equal proportions method also gives improved (lower) distance and imbalance in the progression from $MV1$ to $MV3$. The results for $MV2$ and $MV3$ are placed between those of $TH55$ and $TH50$.
- The vector $MV1$ is defined by rather few (eight) classes, each containing a fairly large number of sample units. The threshold method calls for setting aside an entire class when its response propensity meets the threshold. But in the equal proportions method, classes must be divided; a randomly selected subset of a class was identified to meet the exact $1/6$ called for at each intervention point. This explains the comparatively high total imbalance, 1.32, for $MV1$; eql. With few classes, the equal proportions method should be used with care; for an extensive vector such as $MV3$, it can be used with confidence.
- The marginal imbalance is expected to remain comparatively high as long as the variables in $\mathbf{x}_b$ do not participate in the monitoring. In line with this, the marginal imbalance IMB_b is at its highest for Actual. To obtain lower levels for IMB_b, theory leads us to expect that the variables in $\mathbf{x}_b$, *Phone*, *Age*, *Civil* and *Gender*, should be active in the monitoring. This starts to happen with $MV2$ and is accentuated in $MV3$. The expected decrease in marginal imbalance is confirmed.
- The conditional imbalance $IMB_{a|b}$ is, also as expected, at its highest level for Actual. In the other 12 cases, considerable reduction occurs. It is seen here that the feature that determines $IMB_{a|b}$ is not so much the monitoring vector as rather the threshold value.

11. Study variables and their estimates

In the preceding sections, study variables (y-variables) do not enter into consideration. The AN-IMB in Sections 9 and 10 is based entirely on auxiliary variables. But the ultimate goal for the survey is accurate estimation, for all of the often numerous y-variables, and in particular for the most important ones. We now look at the estimation stage, with an objective to estimate the total of a typical y-variable, $Y = \sum_U y_k$, with as little nonresponse bias as possible. To study this empirically, we need y-data for the full sample s. This is possible if we designate one or more register variables to play the role of a real y-variable. They are then "pseudo y-variables" with values y_k available for all $k \in s$.

Under full response, unbiasedness would be achieved by the Horvitz-Thompson estimator

$$\hat{Y}_{FUL} = \sum\nolimits_s d_k y_k. \qquad (12)$$

Under nonresponse, unbiasedness is not realized. The value y_k is recorded for $k \in r$, missing for $k \in s - r$. The simplest estimator, often considerably biased, is by straight expansion of the response mean,

$$\hat{Y}_{EXP} = \left(\sum\nolimits_s d_k\right) \frac{\sum_r d_k y_k}{\sum_r d_k} = \left(\sum\nolimits_s d_k\right) \bar{y}_r. \tag{13}$$

Reduced bias is usually realized by the calibration estimator

$$\hat{Y}_{CAL} = \sum\nolimits_r d_k m_k y_k \tag{14}$$

where the weight factors are $m_k = (\sum_s d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1} \mathbf{x}_k$. They have the calibration property $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. The right hand side is unbiased for the population total $\sum_U x_k$; consequently, we can count on $\hat{Y}_{CAL}$ to be less biased than $\hat{Y}_{EXP}$ when $\mathbf{x}$ is strongly related to y. The use of the calibrated weights $d_k m_k$ reduces the nonresponse bias, without eliminating it completely, despite any balancing that may have taken place at the data collection stage. The choice of calibration vector $\mathbf{x}$ is important. Normally it should include all x-variables deemed effective for bias reduction. In the empirical work in Section 13 we let $\mathbf{x}$ include the full supply of x-variables (see Section 9), so that $\mathbf{x} = \mathbf{x}_{tot}$ given by (10).

In the text that follows, the estimators are simply referred to as FUL, EXP and CAL.

The deviations of EXP and CAL from the unbiased estimate are measured by $\hat{Y}_{EXP} - \hat{Y}_{FUL}$ and $\hat{Y}_{CAL} - \hat{Y}_{FUL}$. We decompose the first of these as

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = (\hat{Y}_{EXP} - \hat{Y}_{CAL}) + (\hat{Y}_{CAL} - \hat{Y}_{FUL}) \tag{15}$$

or, in words,

Deviation of EXP $=$ Adjustment of EXP $+$ Deviation of CAL.

Here the term $\hat{Y}_{EXP} - \hat{Y}_{CAL}$ is the computable nonresponse adjustment that we apply to the simple estimator $\hat{Y}_{EXP}$ to arrive at the improved (less biased) estimator $\hat{Y}_{CAL}$. The other two terms in (15) would be unknown in a real survey, but we can compute them in experiments with pseudo y-variables.

12. A perspective through regression analysis

An alternative view of equation (15) is obtained by a regression perspective. As is well known, if the $\mathbf{x}$-vector well explains the y-variable, the nonresponse bias will be low. The regression relationship is a key element in understanding the bias. In the extreme case of perfect explanation, where $y_k = \boldsymbol{\beta}' \mathbf{x}_k$ holds exactly for all k and some vector $\boldsymbol{\beta}$, the bias is completely eliminated: $\hat{Y}_{CAL} = \hat{Y}_{FUL}$. In practice the explanation is partial, at best. Two different regressions need to be considered, the one based on the response and the one based on the full sample.

We have the y-means $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ for the response and $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ for the full sample, and the linear regression coefficient vectors $\mathbf{b}_r$ for the response and $\mathbf{b}_s$ for the full sample, where

$$\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_r d_k \mathbf{x}_k y_k\right) ; \qquad \mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum_s d_k \mathbf{x}_k y_k\right). \tag{16}$$

Of these, $\bar{y}_r$ and $\mathbf{b}_r$ are computable from the response; $\bar{y}_s$ and $\mathbf{b}_s$ are not, but are important for the theoretical context. The means $\bar{y}_r$ and $\bar{y}_s$ may differ considerably because of a non-random response. The same holds for the regression vectors $\mathbf{b}_r$ and $\mathbf{b}_s$. As is well-known in regression theory, nonrandom selection of cases can severely bias the regression. Heckman (1979) is an early reference; many articles followed in the setting of regression theory. Here, in the setting of nonresponse theory, the difference between $\mathbf{b}_r$ and $\mathbf{b}_s$ is also a key factor, because r is generally not a random selection from s.

Equation (15) when divided by $\hat{N} = \sum_s d_k$ reads

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s \tag{17}$$

The properties $\bar{\mathbf{x}}'_r\mathbf{b}_r = \bar{y}_r$ and $\bar{\mathbf{x}}'_s\mathbf{b}_s = \bar{y}_s$, needed in establishing (17), follow from the requirement $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all k. We show the first of them; the second is analogous:

$$\begin{aligned}\bar{\mathbf{x}}'_r\mathbf{b}_r &= \left(\sum_r d_k\right)^{-1}\left(\sum_r d_k(\boldsymbol{\mu}'\mathbf{x}_k)\mathbf{x}'_k\right)\left(\sum_r d_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\sum_r d_k\mathbf{x}_k y_k \\ &= \left(\sum_r d_k\right)^{-1}\boldsymbol{\mu}'\sum_r d_k\mathbf{x}_k y_k = \bar{y}_r\end{aligned}$$

Equation (17) is important because it emphasizes two critical differences, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ and $\mathbf{b}_r - \mathbf{b}_s$. The first term on the right hand side of (17), $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r$, is zero under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. The second term, $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$, is zero if the regression is consistent in the sense $\mathbf{b}_r = \mathbf{b}_s$. For data encountered in practice, both terms are usually non-zero.

Equation (17) shows the nearness of $\bar{y}_r$ to $\bar{y}_s$ (or of $\hat{Y}_{EXP}$ to $\hat{Y}_{FUL}$) as a function of two considerations: How close $\bar{\mathbf{x}}_r$ is to $\bar{\mathbf{x}}_s$ (the balance on the $\mathbf{x}$-vector), and how close $\mathbf{b}_r$ is to $\mathbf{b}_s$ (the regression bias aspect).

Another useful representation follows from (17) by substituting $\mathbf{b}_r = \mathbf{b}_s + (\mathbf{b}_r - \mathbf{b}_s)$ in the first term on the right hand side,

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_s + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s + (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s). \tag{18}$$

The last term, $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'(\mathbf{b}_r - \mathbf{b}_s)$, is a measure of interaction between $\bar{\mathbf{x}}_r$ and $\mathbf{b}_r$. Without being negligible, the interaction term is usually numerically less important compared with the other two terms in on the right hand side of (18). We shall further examine the terms of (17) and (18).

As Sections 10 and 11 have shown, we can reduce the imbalance with respect to $\mathbf{x}$ by a monitored data collection based on response propensity. This brings $\bar{\mathbf{x}}_r$ closer to $\bar{\mathbf{x}}_s$, and the adjustment term $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r$ in (17) is likely to be reduced as a result. If at the same time the other term, $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$, is unchanged, the difference $\bar{y}_r - \bar{y}_s$ and the deviation $\hat{Y}_{EXP} - \hat{Y}_{FUL}$ will decrease, as a result of reduced imbalance.

Although a closeness of $\hat{Y}_{EXP}$ to $\hat{Y}_{FUL}$ is desirable, more important in producing survey estimates is the closeness of $\hat{Y}_{CAL}$ to $\hat{Y}_{FUL}$, because in the presence of strong auxiliary information, the calibration adjusted $\hat{Y}_{CAL}$ is the one that would be used, not $\hat{Y}_{EXP}$.

The question arises: Does reducing the imbalance - relatively easy to accomplish - also bring about a reduction of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\ (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$? In a comparison of two different response sets, does the one with the lower imbalance give lower deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL}$?

We present a heuristic argument suggesting that this is the case. When the response r is generated with a monitoring vector $\mathbf{x}_{MV}$, as in the methods in section 7, then $\mathbf{x}_{MV}$ is in general different from the vector $\mathbf{x}$ used for the calibration estimator (14), and in equations (17) and (18). Typically, $\mathbf{x}$ contains at least as many x-variables as $\mathbf{x}_{MV}$.

Consider the term $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s = \mathbf{x}'_s\mathbf{b}_r - \bar{y}_s$, where r is a typical final response set, realized with or without interventions, and described by the indicator $I_{rk} = 1$ for $k \in r$ and $I_{rk} = 0$ otherwise, with mean equal to the response rate, $\sum_s d_k I_{rk} / \sum_s d_k = \sum_r d_k / \sum_s d_k = P$. In line with (16), we have

$$\mathbf{x}'_s\mathbf{b}_r = \bar{\mathbf{x}}'_s\left(\sum\nolimits_s d_k I_{rk}\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\left(\sum\nolimits_s d_k I_{rk}\mathbf{x}_k y_k\right).$$

This can be seen as a realization of the conceptual (not computable) scalar quantity

$$\bar{\mathbf{x}}'_s\mathbf{b}_{s\theta} = \bar{\mathbf{x}}'_s\left(\sum\nolimits_s d_k\theta_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\left(\sum\nolimits_s d_k\theta_k\mathbf{x}_k y_k\right). \tag{19}$$

Here the conceptual θ_k is the *intensity* with which unit k, described by $\mathbf{x}_k$, is present in s, and I_{rk} is a manifestation of that intensity.

If the intensities θ_k are all equal, with variance zero, then (19) equals $\bar{\mathbf{x}}'_s(\sum_s d_k\mathbf{x}_k\mathbf{x}'_k)^{-1}\ (\sum_s d_k\mathbf{x}_k y_k) = \bar{\mathbf{x}}'_s\mathbf{b}_s = \bar{y}_s$, and $(\mathbf{b}_{s\theta} - \mathbf{b}_s)'\bar{\mathbf{x}}_s = 0$. The variance of the intensities is the key to understanding whether or not$(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ is small.

The realized response r, coded by I_{rk}, embodies, for unit k described by the vector value $\mathbf{x}_k$, an empirical intensity measured by the least squares prediction for I_{rk}

$$\hat{\theta}_k = \left(\sum\nolimits_s d_k I_{rk}\mathbf{x}_k\right)'\left(\sum\nolimits_s d_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}\mathbf{x}_k$$

with mean equal to the response rate, $\sum_s d_k\hat{\theta}_k / \sum_s d_k = \sum_r d_k / \sum_s d_k = P$, and variance equal to the imbalance of r with respect to $\mathbf{x}$,

$$S^2_{\hat{\theta}s} = \sum\nolimits_s d_k(\hat{\theta}_k - P)^2 / \sum\nolimits_s d_k = IMB(r, \mathbf{x}|s).$$

(These properties are consequences of the requirement $\boldsymbol{\mu}'\mathbf{x}_{MVk} = 1$ for all k, used several times earlier in this article.) Replacing the unknown θ_k in (19) by their estimates $\hat{\theta}_k$, we get $\mathbf{b}_{s\hat{\theta}} = (\sum_s d_k\hat{\theta}_k\mathbf{x}_k\mathbf{x}'_k)^{-1}\sum_s d_k\hat{\theta}_k\mathbf{x}_k y_k$ as an estimate of $\mathbf{b}_{s\theta}$. The variance of the empirical intensities $\hat{\theta}_k$ over s (which is the imbalance of r) is a key element. If that imbalance is zero, then

$\bar{\mathbf{x}}_s'\mathbf{b}_{s\hat{\theta}} = \bar{\mathbf{x}}_s'\mathbf{b}_s$, and $(\mathbf{b}_{s\hat{\theta}} - \mathbf{b}_s)'\bar{\mathbf{x}}_s = 0$, and we expect that $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s \approx 0$.

We do not - by interventions or any other methods - reduce the imbalance to zero, neither in our experiments nor in actual practice. But a reduction of the imbalance is accomplished, by monitoring the data collection. We expect thereby a certain reduction of $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$, and a reduction of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL}$, not likely to zero, but to a degree. The effect may be modest, since several factors intervene, including the specific character of the y-variable. A few influential y_k-values may have considerable effect in one or the other direction.

13. Application to pseudo y-variables

We return to the Swedish LCS 2009 considered in sections 9 and 10. We illustrate the theoretical Section 12 by a use of "pseudo y-variables". In methodological study, it is useful to designate one or more register variables to play the role of y-variables. They are then pseudo y-variables, with values y_k known for $k \in s$. (They cannot serve as auxiliary variables (x-variables), thus restricting somewhat the supply of such variables.) This allows us to see how the more or less biased estimators under nonresponse, $\hat{Y}_{EXP}$ and $\hat{Y}_{CAL}$ given in (13) and (14), behave relative to the unbiased estimate under full response, $\hat{Y}_{FUL}$ given in (12). For a pseudo y-variable, all terms in (15), (17) and (18) are computable, and we can compare their relative sizes.

Three register variables available in LCS 2009 are used here as pseudo y-variables, with y_k recorded for $k \in s$: *Benefits* (a categorical variable equal to 1 for a person receiving sickness or other social 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 those because they are variables with quite different character, and they are similar to real study variables in the Swedish LCS.

We compute the unbiased full sample (Horvitz-Thompson) estimate $\hat{Y}_{FUL}$ for the three pseudo y-variables, as well as the nonresponse estimators $\hat{Y}_{EXP}$ and $\hat{Y}_{CAL}$. In Table 2, the terms in equation (15) are measured in more readily interpretable relative deviations (in per cent):

$$RDF_{EXP} = RADJ_{EXP} + RDF_{CAL} \tag{20}$$

where $RDF_{EXP} = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{FUL})/\hat{Y}_{FUL}$; $RDF_{CAL} = 100 \times (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{Y}_{FUL}$ and where $RADJ_{EXP} = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{CAL})/\hat{Y}_{FUL}$. The table shows the relative adjustment $RADJ_{EXP}$ and the relative deviation RDF_{CAL} for the CAL estimator; RDF_{EXP} is the sum of these two terms. The $\mathbf{x}$-vector for computing the weight factors m_k in (14) is $\mathbf{x} = \mathbf{x}_{tot}$ given by (10).

The 13 response sets in Table 2 are the same as those in Table 1. The first is the actual LCS 2009 response, resulting from the original data collection plan, without any interventions or attempts at balancing. The 12 generated response are those derived in Section 10, with the same six intervention points. In Table 2, the notation $MV1$, $MV2$ and $MV3$ refers to the monitoring vectors, (7), (8) and (9), and is followed by the threshold percentage, 65, 55 or 50, and, as the fourth alternative for each vector, *eql* for the equal proportions method.

TABLE 2. *$RADJ_{EXP}$ and RDF_{CAL} at the end of data collection, for three pseudo y-variables (columns), and* 13 *final response sets r; Actual LCS2009 data collection (row 1), and* 12 *generated (rows 2 to 13). The notation MV identifies the monitoring vector, 1, 2 or 3, and is followed by the threshold value, or by eql for equal proportions method.*

		Benefits		Income		Employment	
Response set r	$100 \times IMB$	$RADJ_{EXP}$	RDF_{CAL}	$RADJ_{EXP}$	RDF_{CAL}	$RADJ_{EXP}$	RDF_{CAL}
Actual	1.881	-4.85	-4.56	3.45	3.30	1.68	3.08
MV1; 65	1.338	-4.78	-2.26	1.23	3.17	0.30	3.05
MV1; 55	1.080	-4.89	-0.95	0.53	2.76	-0.90	2.96
MV1; 50	1.034	-6.78	0.73	0.58	2.10	-1.58	2.51
MV1; eql	1.320	-5.39	-0.15	-3.00	2.28	-2.42	2.95
MV2; 65	1.230	-3.63	-4.12	1.45	3.28	1.19	2.93
MV2; 55	0.956	-3.44	-2.67	1.79	3.19	1.25	2.79
MV2; 50	0.747	-2.17	-1.84	1.50	2.84	1.49	2.82
MV2; eql	0.808	-1.19	-2.44	0.78	2.63	2.30	2.40
MV3; 65	1.122	-3.64	-3.75	2.20	3.35	1.46	3.00
MV3; 55	0.648	-3.63	-1.41	1.61	2.75	0.79	2.94
MV3; 50	0.426	-3.07	-0.95	1.20	2.32	0.42	2.61
MV3; eql	0.495	-0.97	-1.91	1.41	2.70	2.58	2.83

For each of the three vectors, Table 1 showed that a clear trend towards lower imbalance occurs when the threshold goes from 65% via 55% to 50%. Still, the lowest imbalance, 0.426 for the case $MV3$; 50, is far from the ideal (but in practice unattainable) value of zero. The interest now lies in observing the effect of lower imbalance on $RADJ_{EXP}$ and RDF_{CAL}. (When commenting on the size of $RADJ_{EXP}$ and RDF_{CAL}, which can have either sign, we refer to their absolute values. For example, "decrease" is to be interpreted as "decrease in absolute value.")

In examining Table 2 one should keep in mind that the entries depend on the specific character of each y-variable. The magnitude of RDF_{CAL} and RDF_{EXP} can shift considerably from one y-variable to another. They are usually, but not necessarily, of the same sign. Nevertheless, there are clearly distinct patterns. Theory leads us to expect that both terms on the right hand side of (20), $RADJ_{EXP}$ and RDF_{CAL}, should decrease as a result of reduced imbalance. It should also be remembered that in a real survey, the adjustment is the only of the three terms in (15) that can be computed; in our exploration here, computation of the three terms is made possible by the use of pseudo y-variables.

An examination of Table 2 prompts the following comments:

- First, the importance of adjustment by calibration, leaving aside the question of monitoring, is clearly demonstrated. For Actual, a large portion of $RDF_{EXP} = RADJ_{EXP} + RDF_{CAL}$ is "adjusted away" by calibration; RDF_{CAL} is less than 50% of RDF_{EXP} for *Benefits* and *Income*. Adjustment by auxiliary information, here as in practice, can lead to no more than a partial fulfilment of the ideal "elimination of bias."
- Compared with Actual, all 12 generated response sets have lower imbalance, and, as an expected consequence, lower, even much lower, values in most cases for both $RADJ_{EXP}$ and RDF_{CAL}. That their magnitude differs considerably between the three y-variables is in order;

they reflect the specific character of each y-variable.

- The most consistent pattern for $RADJ_{EXP}$ and RDF_{CAL} occurs with the monitoring vector $MV3$, aided by its 256 different properties for the sample units and a smooth distribution of the propensities used to generate the response. The imbalance decreases when the threshold moves from 65% via 55% to 50% and is accompanied by a decrease in both $RADJ_{EXP}$ and RDF_{CAL}.
- For the vectors $MV1$ and $MV2$, there is also a definite decreasing trend for $RADJ_{EXP}$ and RDF_{CAL} as the threshold goes from 65% to 50%, although the pattern is not consistent. The results for $MV1$ are sensitive to the modest number (eight) of groups coded by that vector.
- The equal proportions method, labelled *eql*, works well with the vector $MV3$ for *Benefits* and *Income*, but is less satisfactory for *Employment*. But for $MV1$, where random subdivision of the rather small number of groups is a disturbing factor (see the fourth comment on Table 1), both imbalance and $RADJ_{EXP}$ are at comparatively high levels.

14. Discussion

In the first part of this article (Sections 3 to 10), we developed the concept of imbalance of the response, and measured empirically its components, total, partial and conditional imbalance, in the Swedish Living Conditions Survey. As part of that examination, we compared 13 response sets, one actually observed, and twelve that were generated from the actual response via the concept of response propensity. The fact that these represent different degrees of imbalance in the response allows us to compare and study how selected auxiliary variables can contribute to balance.

That the imbalance should be reduced to zero (or that the auxiliary vector should completely explain the response) is a faint hope, never realized in survey practice. Equally hypothetical is that the auxiliary vector will completely explain the survey variable.

Still, it is a fact that adjusting the simple estimate by calibrated weighting reduces the bias, up to a degree. This is well known and confirmed here. But in addition, we have presented some evidence (Sections 11 to 13) suggesting that the bias still remaining in the estimates after calibrated weighting adjustment can be further reduced, namely, if the underlying set of respondents was well balanced. That is, we have reason to believe that that efforts to balance the response during data collection - something which is not always simple to administer - will pay off at the estimation stage.

We have in this way, indirectly at least, asked: How do we in the most productive manner put the auxiliary variables to work in a survey; which ones should be active in the data collection, which ones at the estimation stage? There is no complete answer in this article to this broader question; further work is required.

References

Andridge, R. and Little, R. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27(2):153.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of nonresponse in household surveys*. John Wiley & Sons.

Brick, M. and Jones, J. (2008). Propensity to respond and nonresponse bias. *Metron*, 66(1):51–73.

Couper, M. and Wagner, J. (2012). Using paradata and responsive design to manage survey nonresponse. *Unpublished manuscript*.

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675.

Groves, R. and Heeringa, S. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R., and Raghunathan, T. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):389–407.

Laflamme, F. (2009). Experiences in assessing, monitoring and controlling survey productivity and costs at Statistics Canada. In *Proceedings from the 57th International Statistical Institute Conference*.

Little, R. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161–168.

Lundquist, P. and Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish living conditions survey. *Journal of Official Statistics*, 29(4):557–582.

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

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of nonresponse bias through case prioritization. *Survey Research Methods*, 4(1):21–29.

Peytcheva, E. and Groves, R. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, 25(2):193.

Rao, R., Glickman, M., and Glynn, R. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27(12):2196–2213.

Särndal, C.-E. (2011a). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27(1):1–21.

Särndal, C.-E. (2011b). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International statistical review*, 79(2):233–254.

Särndal, C.-E. and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 24(2):167.

Särndal, C.-E. and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36(2):131–144.

Schouten, B. and Bethlehem, J. (2009). Representativeness indicators for measuring and enhancing the composition of survey response. *RISQ deliverables, Work package 8, deliverable 9, www.risq-project.eu*.

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1):101–113.

Schouten, B., Shlomo, N., and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27(2):1–24.

Vartivarian, S. and Little, R. (2002). On the formation of weighting adjustment cells for unit nonresponse. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*.

Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias*. Ph.D. Thesis, University of Michigan, Ann Arbor.

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

Wagner, J. and Raghunathan, T. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29(9):1014–1024.