Journal of Official Statistics, Vol. 27, No. 1, 2011, pp. 1–21

The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation

Carl-Erik Särndal[1]

In dealing with survey nonresponse, statisticians need to consider (a) measures to be taken at the data collection stage, and (b) measures to be taken at the estimation stage. One may employ some form of responsive design. In the later stages of the data collection in particular, one tries to achieve an ultimate set of responding units that is "better balanced" or "more representative" than if no special effort is made. The concept of "balanced response set" introduced in this article extends the well-known idea of "balanced sample." A measure of "lack of balance" is proposed; it is a quadratic form relating to a multivariate auxiliary vector; its statistical properties are explored. But whether or not good balance has been achieved in the data collection, a compelling question remains at the estimation stage: How to achieve the most effective reduction of nonresponse bias in the survey estimates. Balancing alone may not help. The nonresponse adjustment effort is aided by a bias indicator, a product of three factors involving selected powerful auxiliary variables.

Key words: Balanced response set; representative response set; responsive design; lack of balance; auxiliary information; calibrated adjustment weighting.

1. Representative Set of Units, Balanced Set of Units

Although probability sampling and design-based inference were dominating traditions in survey sampling during the 20th century – particularly from the 1930's and on following path-breaking advances by Jerzy Neyman, by Morris Hansen and his associates – purposive sampling still attracted attention and interest, among theoreticians as among practitioners. Purposive sample selection was practiced already in the late 19th century, as in "the representative method" used at the Norwegian Central Bureau of Statistics and attributed to its head at the time, Anders Kiaer. It consisted roughly speaking in the selection of units such that the sample would have the same or almost the same characteristics as the whole population.

"Same characteristics" later received a more specific meaning. The term "balanced sample" came to be used for a sample whose characteristic feature is the equality, or near-equality, of the sample mean with the corresponding population mean, for observable control variables. A balanced sample can thus be described in such appealing terms as "representative of the finite population" or "a miniature of the whole population."

Around 1970, the dominating probability sampling tradition faced a serious challenge from the model-based, or model-dependent, approach. It was shown for example that if a

[1] Professor emeritus, Statistics Sweden, S-701 89 Örebro, Sweden. Email: carl.sarndal@scb.se

certain model is true, an extreme sample, such as selecting purposively the n units with the largest values on a continuous auxiliary variable x, would be optimal, in a purely model-based assessment of precision. Not surprisingly, such selections (and the inferences they generate) are highly sensitive to the truth of one particular model, and such practices met with vigorous criticism from the design-based school.

Proponents of the model-based approach realized that there was a need to protect the inferences against model breakdown. One avenue was to post a hypothesized model where y has a polynomial dependence on the continuous auxiliary variable x, rather than just a simple linear one. A balanced sample in that connection is one for which the first few moments of the x-variable are equal for the sample and for the whole population.

Some early references to the model-based developments are Royall (1970) and Royall and Herson (1973a, b). Among the critics who defended the probability sampling (design-based) outlook were Hansen, Madow, and Tepping (1983). For objective review and discussion, see Brewer (1994;1999). Balancing of samples for different models is discussed extensively in the book by Valliant, Dorfman, and Royall (2000).

Proponents of the model-based approach did not deny that randomized selection could provide added protection for their model-based estimates and inferences. But balanced sampling belonged for a long time in the realm of purposive or non-probability sampling methods and was viewed with skepticism by the design-based school.

A decisive step was taken by Deville and Tillé (2004) through their cube method. They argued that auxiliary information can be used to advantage in the drawing of the sample itself, more precisely so as to select a balanced sample that is also a probability sample, with known and positive inclusion probabilities for all units. Two desirable goals, balancing *and* probability sampling, are thus achieved simultaneously. If $\mathbf{x}_k$ denotes the auxiliary vector value for unit k, balancing the sample to the population was defined in the cube method as the equality (or near-equality) between the unbiased Horvitz-Thompson (HT) estimator of the population total of $\mathbf{x}_k$ and that known total itself. The mechanics of the procedure can be described as a series of randomized moves, under restrictions, inside an N-dimensional cube, N being the population size.

Every sample that can result from the cube method is balanced – for a given auxiliary vector – without being purposive, and the known inclusion probabilities leave the field open for those traditional uses of design-based estimation where the inverses of those probabilities act as weights for unbiased or nearly unbiased estimation. Furthermore, the auxiliary information realizes already at the data collection stage an objective – a reduced design-based variance in the estimates – that would otherwise be realized later, at the estimation stage, by a regression adjustment to the HT estimator, as in a generalized regression (GREG) estimator, or by calibrated weighting. Since any sample resulting from the cube method is balanced, the regression adjustment in the usual GREG estimator is always zero. The HT estimator is, in a manner of speaking, all that is needed.

In its traditional usage, the term balancing involves the relationship between a sample and the entire population, for example so that means for the sample agree with means for the whole population, for the specified available variables. An extended idea of balancing is needed in this article dealing with nonresponse. This leads to the following formulation.

General principle of balancing: A subset of a given larger set of units is balanced when realized subject to an advantageous conformity or resemblance with the larger set. An example is to realize the subset so as to make means for appropriate auxiliary variables equal to corresponding means for the larger set.

In the context of surveys with nonresponse, this principle requires the ultimate set of respondents to be balanced to the probability sample of which the respondents are a subset. It is clearly advantageous to realize a response set with "the same characteristics" as the whole probability sample, because the latter is a selection not skewed by nonresponse. In this extended use of the term, balancing refers to a second phase of a selection, the first phase being the drawing of a sample from the population.

In the cube method, the auxiliary information serves at the sample selection stage to draw a balanced probability sample. Regression adjustment is then superfluous at the estimation stage (unless more information becomes available at that stage). Alternatively we can – with essentially the same benefit in terms of variance reduction – draw any, not necessarily balanced, probability sample and postpone the role of that same information until the estimation stage, to form a GREG estimator or a calibration estimator.

This article makes the point that a similar "before-or-after" option exists in regard to nonresponse: Either the auxiliary information can serve at the data collection stage, where it can be instrumental in realizing a well-balanced set of respondents, or the use of it can be postponed until the estimation stage and then take the form of a nonresponse weighting adjustment. But despite this similarity we shall find, in the nonresponse situation, some differences between the two aspects: A well-balanced response set does not eliminate the need to seek effective adjustment at the estimation stage.

2. Background

We assume having arrived at that stage of the survey process where a probability sample has been drawn from the finite population, with known positive inclusion probabilities for all units. The resulting set of identified units is targeted for data collection. But nonresponse occurs. Only a subset will respond and supply the value(s) of the study variable(s). Depending on how far the process has advanced, we consider two scenarios:

i) Action taken during the data collection, "prior action," particularly in its later phases, to achieve an ultimate response set that is "better balanced" or "more representative" than if no special effort is made;
ii) Action taken after the termination of the data collection, "posterior action," consisting in nonresponse adjustment at the estimation stage.

Both kinds involve the use of an auxiliary vector, denoted generically by $\mathbf{x}_k$, on which information exists over and beyond the set of responding units. That information may serve either at the data collection stage or at the estimation stage.

In many surveys, data collection extends over quite a large length of time, perhaps several weeks or months. It proceeds in an initial phase according to "an original plan." But at a certain point, one may wish to intervene, in the interest of efficiency, and take initiatives to achieve in the end a better balanced ultimate set of respondents.

Prior action includes different forms of responsive design. Its general objectives are formulated in Groves and Heeringa (2006). They use the term “phase capacity” for “the stable condition of an estimate in a specific design phase.” When phase capacity has been reached in a given phase, it is no longer effective to continue data collection in the same mode or phase; there is an incentive to modify the design, if data collection is to be continued at all. Options for responsive design in a Canadian setting are discussed in Mohl and Laflamme (2007) and Laflamme (2009). With a view to improving the response rate, different kinds of prior action have long been practiced. They include the use of more skilled interviewers, different forms of call-back, subsampling of nonrespondents, stronger incentives to participate, and so on.

In other forms of prior action, the interviewers are instructed to direct attention during follow-up to specific sample categories, identified by age, sex and other important characteristics, that turn out underrepresented in the earlier phases of the data collection. Such efforts may not in the end maximize the particular interviewer’s own response rate, nor the global response rate, because the targeted groups are hard to reach or unwilling to respond. But more importantly, a changed emphasis in the data collection may bring the benefit of a better balanced ultimate response set, and possibly a reduced bias in the survey estimates. To make this work, auxiliary information must exist on every unit k in the sample; the auxiliary vector value $\mathbf{x}_k$ must identify the categories to be pursued.

As the data collection unfolds, there is a need to evaluate whether the responsive design is effective, that is, if there is a momentum towards a better balanced set of respondents. The rate of response is by itself insufficient as a measure of the state of the data collection. There is a need for a more informative measure that can be computed continuously and monitored during the data collection including the follow-up. Schouten, Cobben, and Bethlehem (2009) propose an indicator of “representativity” of a response set. Their “*R*-indicator” is defined as a function of the variance of estimated response probabilities. They illustrate the indicator in the setting of the 2005 Dutch Labour Force Survey, where they compute the *R*-indicator for the basic LFS response and then also for the larger data set that combines the LFS response with the response from a call-back directed to a subsample of nonrespondents. They note a somewhat improved representativity – a higher value of the *R*-indicator – for the “LFS plus call-back” alternative.

A point of departure for this article is the concept of *balanced response set*: If means for measurable auxiliary variables are the same for respondents as for all those sampled, we call the response set *perfectly balanced*. Then, on average for the measured variables, the respondents agree with the whole set of sampled units. When those means differ, the response set is more or less imbalanced.

Once the data collection is terminated, the estimation phase begins. The study variable values, observed only for the ultimate set of respondents, form together with auxiliary data the material for estimating population totals and other parameters. The principal objective in this posterior action is to produce estimates with low bias, through weighting adjustment by calibration on selected auxiliary variables. Simple adjustment cell weighting is too limited in scope; we require more extensive auxiliary information.

Nonresponse causes both bias and increased variance. We focus here on the bias. Its square is typically the dominant portion of the Mean Squared Error (MSE). We address primarily surveys on individuals and households with quite large sample sizes, as is typical for government surveys; consequently, the variance contribution to MSE is low by comparison. Increased variance due to nonresponse is nevertheless an issue; striking a balance between variance increase and bias reduction is considered, for example, in Little and Vartivarian (2005).

Efficient adjustment requires powerful auxiliary information. The Scandinavian countries and The Netherlands – and increasingly other countries especially in Europe – are privileged. The available administrative registers guarantee a vast supply of potential auxiliary variables especially for surveys on individuals and households.

Many published articles discuss adjustment weighting and the question of how to select the best auxiliary variables. Kalton and Flores-Cervantes (2003) give many references. Rizzo, Kalton, and Brick (1996) and Slud and Bailey (2010) discuss weighting adjustment for panel nonresponse. Several recent contributions discuss adjustment though *calibration*, for example Beaumont (2005), Crouse and Kott (2004), Deville (2002), Kott (2006; 2008). Calibration, propensity to respond and the value of auxiliary data are topics in Brick and Jones (2008).

The bias cannot be estimated or otherwise made known; this is the dilemma of nonresponse. The realistic position in this article is that the bias is never entirely eliminated, no matter how powerful the auxiliary information. Some bias always remains after an adjustment. In this predicament we need an indicator to signal when a sizeable reduction of the bias may have taken place. The indicator serves to select the most effective or most promising auxiliary variables, in a perhaps long list of available ones. Bias indicators were proposed in Särndal and Lundström (2005; 2008; 2010), while Schouten (2007) uses a different perspective to motivate an indicator.

The role of the auxiliary vector(s) is different under the two scenarios:

i) Prior action focuses on a fixed auxiliary vector, that is, one with a fixed composition of auxiliary variables, and we try to achieve a better balanced ultimate set of respondents for that particular **x**-vector;

ii) Posterior action consists in building an **x**-vector from a supply of available *x*-variables; in this process we need to compare alternative **x**-vectors and select one that is likely to bring effective bias reduction in all or most estimates.

The contents of this article are arranged as follows. Section 3 presents the notation and the survey background, in the form of probability sampling followed by nonresponse. Section 4 reviews the features of the multidimensional auxiliary vector and the associated auxiliary information. Section 5 introduces the concepts of balanced response set and lack of balance for that set. Section 6 shows how lack of balance is related to the idea of estimating the unknown response probabilities θ_k with the aid of the auxiliary data. Three alternative indicators of balance, all contained in the unit interval, are proposed in Section 7; the idea is to monitor the phases of the data collection with the aid these indicators. Section 8 addresses "the dual estimation problem," that of estimating the inverse response probabilities $\phi_k = 1/\theta_k$. This leads to the nonresponse adjusted estimators described in Section 9; they are identified alternatively as calibration estimators. A sizeable nonresponse adjustment indicates that there is substantial bias needing to be compensated for. As Section 9 also shows, the adjustment is a product of three factors, each with simple interpretation. One of those is, as one might expect, the degree to which the auxiliary vector explains the study variable. But in itself this factor is insufficient to produce a sizeable adjustment. The levels of other two factors are also important. Section 10 focuses on the stepwise construction of the auxiliary vector, given the objective of effective nonresponse adjustment for *all* of the often numerous study variables in the survey. Empirical data from Statistics Sweden are used to illustrate the auxiliary variable selection.

3. Probability Sample and Responding Subset

A probability sample s is drawn from $U = \{1, \ldots, k, \ldots, N\}$. Out of a realized sample s, only a subset r of units responds; the value y_k of the study variable y, which may be continuous or categorical, is observed only for $k \in r$. Conceptually, r is the result of an unknown response mechanism operating on s. The data $\{y_k\colon k \in r\}$ provide, together with auxiliary data, the material for estimating 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 $\sum_A$). Under the sampling design used to draw s, the unit k has the known *inclusion probability* $\pi_k = \Pr(k \in s) > 0$ and the known *design weight* $d_k = 1/\pi_k$. The (design-weighted) survey response rate and its inverse value are, respectively,

$$P = \sum\nolimits_r d_k \Big/ \sum\nolimits_s d_k; \quad Q = 1\Big/P = \sum\nolimits_s d_k \Big/ \sum\nolimits_r d_k \tag{3.1}$$

We assume $0 < P < 1$. We call $(1 - P)/P = Q - 1$ the (empirical) *nonresponse odds*. For example, when the response rate is $P = 80\%$, the nonresponse odds are 1:4, or 0.25; for $P = 60\%$, the nonresponse odds are 1:1.5, or 0.67.

Denote further the unknown *response probability* of k by $\theta_k = \Pr(k \in r|s)$. Its inverse, $\phi_k = 1/\theta_k$, also unknown, is called *the influence* of k. We prefer this term to "weight," because "weight" is something known, or at least computable. For all k, both θ_k and $\phi_k = 1/\theta_k$ are conceptually defined, nonrandom, nonobservable entities.

The response indicator I is a binary random variable, observed for $k \in s$, with value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$. Then $E(I_k|s) = \theta_k$. Let a_k be a nonrandom (scalar or vector) value tied to unit k; consider two linear forms in I_k:

$$\sum\nolimits_s d_k(I_k\phi_k - 1)a_k; \quad \sum\nolimits_s d_k(I_k - \theta_k)a_k \tag{3.2}$$

Both are theoretical entities, not computable, since θ_k and $\phi_k = 1/\theta_k$ are unknown. Both have expected value equal to zero, given s, because $E(I_k|s)\phi_k - 1 = E(I_k|s) - \theta_k = 0$ for all k. Both have frequency interpretations with respect to repeated realizations of response sets r, given s: On average over such repetitions, $\sum_r d_k\phi_k a_k$ is equal to $\sum_s d_k a_k$, and $\sum_r d_k a_k$ is equal to $\sum_s d_k\theta_k a_k$. We can let $a_k = y_k$ in (3.2), because the study variable values y_k, observed only for $k \in r$, are viewed as nonrandom, as in traditional probability sampling theory.

4. Auxiliary Vector and Auxiliary Information

Auxiliary information plays a central role in dealing with survey nonresponse, both at the data collection stage and at the estimation stage. The auxiliary vector value

$\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ is assumed available for $k \in s$, where x_{jk} is the value for unit k of the jth auxiliary variable, x_j. The values x_{jk} may be continuous measurements or category indicators, equal to 1 or 0 to code presence or absence of a given trait or property.

An important special case involves a single categorical auxiliary variable defined by $J \geq 2$ mutually exclusive and exhaustive traits, as when the variable Age is defined by $J = 3$ traits, Young, Middle-aged, and Elderly. The trait of unit k is coded by the J-vector $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})' = (0, \ldots, 1, \ldots, 0)'$ (with a single entry "1"), or equivalently by the J-vector $\mathbf{x}_k = (1, \gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{J-1,k})'$, where $\gamma_{jk} = 1$ if k has the trait j and $\gamma_{jk} = 0$ otherwise.

In practice the auxiliary vector $\mathbf{x}_k$ often serves to code several categorical auxiliary variables. If the ith variable has J_i traits, $i = 1, \ldots, I$, and the I variables are arranged "side by side" in $\mathbf{x}_k$ (rather than as cross-classified), then the dimension of $\mathbf{x}_k$ is $J = 1 + \sum_{i=1}^{I} (J_i - 1)$. One trait, regardless of which one, is omitted in each categorical variable, to avoid a singular matrix. The number of variables I is often 10 or more, the dimension J often 50 or more. The vector $\mathbf{x}_k$ is typically built (in Scandinavia) from a vast supply of x-variables, including income class, level of education, level of indebtedness, marital status, country of birth, unemployment pattern; the "usual variables" age and sex may not be among the most effective ones. The vector $\mathbf{x}_k$ may be coded to also include relevant interactions.

We consider auxiliary vectors of the following type: For some constant vector $\boldsymbol{\mu} \neq 0$,

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k \in U \tag{4.1}$$

This is not a severe restriction on $\mathbf{x}_k$. Most vectors of interest in practice are covered. For example, if $\mathbf{x}_k = (1, x_k)'$, where x_k is a continuous variable value, then take $\boldsymbol{\mu} = (1, 0)'$; if $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})' = (0, \ldots, 1, \ldots, 0)'$, where the one and only "1" codes class membership of k, then take $\boldsymbol{\mu} = (1, \ldots, 1, \ldots, 1)'$.

We define two computable J-dimensional mean vectors and two computable $J \times J$ cross product matrices (or weighting matrices), assumed nonsingular:

$$\bar{\mathbf{x}}_{r;d} = \sum\nolimits_r d_k \mathbf{x}_k \Big/ \sum\nolimits_r d_k; \quad \boldsymbol{\Sigma}_r = \sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum\nolimits_r d_k \tag{4.2}$$

$$\bar{\mathbf{x}}_{s;d} = \sum\nolimits_s d_k \mathbf{x}_k \Big/ \sum\nolimits_s d_k; \quad \boldsymbol{\Sigma}_s = \sum\nolimits_s d_k \mathbf{x}_k \mathbf{x}_k' \Big/ \sum\nolimits_s d_k \tag{4.3}$$

Notation for weighted means and other aggregates obeys to the following rule: Out of two indices separated by semi-colon, the first specifies the set of units in the sum, the second specifies the weighting. The simpler notation for the matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$ does not fully respect that rule but is sufficiently clear. We need several quadratic forms and bilinear forms in the mean vectors $\bar{\mathbf{x}}_{r;d}$ and $\bar{\mathbf{x}}_{s;d}$. By (4.1) we have for all outcomes (s,r)

$$\bar{\mathbf{x}}'_{r;d} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}'_{r;d} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}'_{r;d} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = \bar{\mathbf{x}}'_{s;d} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{s;d} = 1 \tag{4.4}$$

Properties shown and used later are:

$$\bar{\mathbf{x}}'_{r;d} \boldsymbol{\Sigma}_s^{-1} \bar{\mathbf{x}}_{r;d} \geq 1; \quad \bar{\mathbf{x}}'_{s;d} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{s;d} \geq 1 \tag{4.5}$$

5. Balanced Response and Measuring Lack of Balance

The important J-vector $\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} = (D_1, \ldots, D_j, \ldots, D_J)'$ reflects similarity, or balance, between the response set r and the sample set s. The computable difference $D_j = \bar{x}_{j|r;d} - \bar{x}_{j|s;d}$ contrasts the respondent mean of the auxiliary variable x_j, $\bar{x}_{j|r;d} = \sum_r d_k x_{jk} / \sum_r d_k$, with the full sample mean of the same variable, $\bar{x}_{j|s;d} = \sum_s d_k x_{jk} / \sum_s d_k$. If $\mathbf{D} = \mathbf{0}$, the null vector, so that $\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}_{s;d}$, we say that the response set r is *perfectly balanced* with the probability sample s, for the specific vector $\mathbf{x}_k$. Then the respondent set r mirrors the sample s, in so far as the means of the variables in $\mathbf{x}_k$ goes. We can expect perfect balance to be an asset in dealing with nonresponse. Normally, $\mathbf{D} \neq \mathbf{0}$, suggesting a more or less pronounced departure from balance. We need to transform the multivariate $\mathbf{D}$ into a univariate statistic to quantify the lack of balance or imbalance, for given outcome (s, r) and given composition of the vector $\mathbf{x}_k$. This need is filled by the quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. Perfect balance, that is $\mathbf{D} = \mathbf{0}$, gives $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = 0$. Increased differences D_j between respondent mean and full sample mean tend to increase the value of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. We assume that $\boldsymbol{\Sigma}_s$ is full rank, so that $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \geq 0$.

To illustrate the quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ in an important special case, consider a single categorical auxiliary variable defined by J mutually exclusive and exhaustive traits coded by the J-vector $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})' = (0, \ldots, 1, \ldots, 0)'$, where $\gamma_{jk} = 1$ or 0 indicates presence or absence for unit k of the trait j. Let s_j and r_j be the subsets (out of the whole sample s and out of the whole response set r, respectively) of units with trait j; $r_j \subseteq s_j$. For that trait, $W_{js} = \sum_{s_j} d_k / \sum_s d_k$ and $W_{jr} = \sum_{r_j} d_k / \sum_r d_k$ represent the respective shares. Then $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is akin to a chi-squared statistic,

$$\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} (W_{jr} - W_{js})^2 / W_{js}$$

Zero lack of balance implies, for this particular $\mathbf{x}$-vector, that each trait's share of the response set r equals its share of the whole sample s. Equivalently, it implies that the group response rate $\sum_{r_j} d_k / \sum_{s_j} d_k$ is the same for all groups $j = 1, \ldots, J$. If Gender were the only available auxiliary variable, then $\mathbf{x}_k$ has dimension $J = 2$ to indicate male or female, and the response set is perfectly balanced with respect to Gender if men and women have the same response rate. In practice, the auxiliary vector is usually much more extensive, so as to code a number of categorical auxiliary variables.

We note that $\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} = (1 - P)(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})$, where $\bar{\mathbf{x}}_{s-r;d} = \sum_{s-r} d_k \mathbf{x}_k / \sum_{s-r} d_k$. Using also (4.4) we can then write

$$0 \leq \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \bar{\mathbf{x}}'_{r;d}\boldsymbol{\Sigma}_s^{-1}\bar{\mathbf{x}}_{r;d} - 1 = (1 - P)^2(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d}) \quad (5.1)$$

The first part of (4.5) now follows from (5.1). For $\mathbf{D} \neq \mathbf{0}$ and a fixed dimension $J \geq 2$, (5.1) shows that the lack of balance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ increases (i) with increased rate of nonresponse $1 - P$, and (ii) with increased separation between respondents and nonrespondents, considering that $(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})$ is a form of Mahalanobis distance squared between the two groups. The lack of balance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ has important properties shown in Section 6:

- Property 1: $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is bounded from above by the nonresponse odds: $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \leq Q-1$ for any outcome (s,r) and any composition of the vector $\mathbf{x}_k$;
- Property 2: $(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}$ is the coefficient of variation (standard deviation divided by mean) of response probabilities estimated by least squares.

Consequently, $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}/(Q-1)$ measures *lack of balance on a unit interval scale*, for the given vector specification $\mathbf{x}_k$; the complement $1-\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}/(Q-1)$ measures *balance*.

A very similar lack of balance statistic, $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$, is also important starting in Section 8. It differs from $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ only in regard to the weighting matrix, $\boldsymbol{\Sigma}_r$ instead of $\boldsymbol{\Sigma}_s$. Their numeric difference is often small. Pronounced mean differences $D_j = \bar{x}_{j|r;d} - \bar{x}_{j|s;d}$, $j = 1, \ldots, J$, tend to increase the value of both $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ and $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$. The upper bound $Q-1 = (1-P)/P$ on $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is a function only of the response rate P. For example, regardless of the choice of $\mathbf{x}$-vector, the upper bound is 0.25 for $P = 80\%$ and 0.67 for $P = 60\%$. For data encountered in practice, $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ usually lies considerably below its upper bound. We cannot specify an upper bound on $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$, as discussed in Section 8. In the example with a single categorical auxiliary variable with J traits, $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})'$, we have $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D} = \sum_{j=1}^{J} (W_{jr} - W_{js})^2/W_{jr}$. It is zero if all J trait groups have the same response rate. On the other hand, $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$ can be very large if some category shares W_{jr} are very small; one should avoid groupings that give rise to small W_{jr}.

6. Estimated Response Probabilities

We derive estimated response probabilities $\hat{\theta}_k$ from two requirements: (i) they are to be linear in $\mathbf{x}_k$, so that $\hat{\theta}_k = \boldsymbol{\lambda}'\mathbf{x}_k$ for some constant vector $\boldsymbol{\lambda}$, and (ii) they verify the constraint inspired by the second statistic in (3.2) with $a_k = \mathbf{x}_k$ and $\theta_k = \hat{\theta}_k$, that is, $\sum_s d_k(I_k - \hat{\theta}_k)\mathbf{x}_k = 0$, or $\sum_r d_k\mathbf{x}_k = \sum_s d_k\hat{\theta}_k\mathbf{x}_k$. These conditions determine $\boldsymbol{\lambda}$, and we obtain $\hat{\theta}_k = t_k$, where

$$t_k = \left(\sum\nolimits_r d_k\mathbf{x}_k\right)'\left(\sum\nolimits_s d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\mathbf{x}_k \tag{6.1}$$

There is no guarantee that $0 \leq t_k \leq 1$ for all units k, outcomes (s,r) and vector specifications $\mathbf{x}_k$, but it is not a drawback for this article if a small number of t_k fall outside the unit interval, because the t_k appear in the form of always nonnegative aggregates.

We obtain the same result, $\hat{\theta}_k = t_k$, by least squares fit, treating I_k as the dependent variable: Determine $\boldsymbol{\lambda}$ to minimize the weighted sum of squares $\sum_s d_k(I_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$. The estimating equation is $\sum_s d_kI_k\mathbf{x}_k' - \boldsymbol{\lambda}'\sum_s d_k\mathbf{x}_k\mathbf{x}_k' = \mathbf{0}'$; solving for $\boldsymbol{\lambda}$ leads to $\hat{\theta}_k = \boldsymbol{\lambda}'\mathbf{x}_k = t_k$ as given in (6.1).

The t_k are computable for $k \in s$. We need the mean over r, and the mean and variance over s:

$$\bar{t}_{r;d} = \left(\sum\nolimits_r d_kt_k\right)\Big/\left(\sum\nolimits_r d_k\right); \quad \bar{t}_{s;d} = \left(\sum\nolimits_s d_kt_k\right)\Big/\left(\sum\nolimits_s d_k\right);$$

$$S^2_{t|s;d} = \sum\nolimits_s d_k(t_k - \bar{t}_{s;d})^2\Big/\sum\nolimits_s d_k$$

A development making use of (4.1) and (4.4) and the fact that $\sum_s d_k t_k^2 = \sum_r d_k t_k$ gives

$$\bar{t}_{r;d} = P \times \bar{\mathbf{x}}'_{r;d} \mathbf{\Sigma}_s^{-1} \bar{\mathbf{x}}_{r;d}; \quad \bar{t}_{s;d} = P; \quad S^2_{t|s;d} = \bar{t}_{s;d}(\bar{t}_{r;d} - \bar{t}_{s;d}) = P^2 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \tag{6.2}$$

where $\bar{\mathbf{x}}_{r;d}$ and $\mathbf{\Sigma}_s$ are defined in (4.2) and (4.3). From (6.2) follows Property 2 in Section 5 concerning the coefficient of variation of $\hat{\theta}_k = t_k$:

$$cv_{t|s;d} = S_{t|s;d}/\bar{t}_{s;d} = (\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2} \tag{6.3}$$

To prove Property 1 in Section 5, consider the sum of squares $\sum_s d_k(I_k - P)^2$, and let $I_k - P = (t_k - P) + (I_k - t_k)$, where I_k is the response indicator with mean $P = \bar{I}_{s;d} = \sum_s d_k I_k / \sum_s d_k$. A development gives the orthogonal components decomposition

$$\sum_s d_k(I_k - P)^2 = \sum_s d_k(t_k - P)^2 + \sum_s d_k(I_k - t_k)^2 \tag{6.4}$$

where $\sum_s d_k(t_k - P)^2$ represents "variance explained". The cross product term is zero because $\sum_s d_k(I_k - t_k)\mathbf{x}'_k = \mathbf{0}'$. Dividing through in (6.4) by $\sum_s d_k$ we get

$$P(1 - P) = S^2_{t|s;d} + \left(P(1 - P) - S^2_{t|s;d}\right) \tag{6.5}$$

For another useful representation, divide through by P^2 and use (6.2) to obtain

$$Q - 1 = \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} + \left(Q - 1 - \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}\right) \tag{6.6}$$

or, in words, *Nonresponse odds* = *Lack of balance* + *Residual*. This proves Property 1:

$$0 \leq \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq Q - 1 \tag{6.7}$$

It also follows that

$$0 \leq S^2_{t|s;d} = P^2 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq P(1 - P) \leq 1/4 \tag{6.8}$$

7. Implications for a Responsive Design

Data collection extends in many surveys over some period of time. One may wish to modify the original design during the course of the data collection, with a view, in the later stages, to getting a better balanced ultimate response set. We may decide to monitor the follow-up effort and direct the data inflow in the direction of better balance. We need a computable indicator to direct such responsive action. Since it is unlikely that a full 100% response will be attained at a reasonable cost, one may have to stop data collection entirely at a suitable moment. The response set r and the response rate $P = \sum_r d_k / \sum_s d_k$ change as the data collection proceeds, as a result of interventions in the original design. Even though improved balance may be achieved, the ultimate nonresponse rate can still be considerable, causing unknown bias in the survey estimates, as further discussed in Sections 9 and 10.

A balance indicator (abbreviated *BI*), measured on the unit interval scale, is a useful tool for responsive design. For any given composition of $\mathbf{x}_k$, the indicator should attain the maximal value of unity for the perfect balance $\mathbf{D} = \mathbf{0}$. We consider three such balance indicators, all of them functions of $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$. By (6.2), they are corresponding functions of the variance $S^2_{t|s;d}$ of the response probability estimates $\hat{\theta}_k = t_k$. Define first

$$BI_1 = 1 - \frac{\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}}{Q-1} = 1 - \frac{S^2_{t|s;d}}{P(1-P)} \tag{7.1}$$

By (6.7), $0 \le BI_1 \le 1$ for all outcomes (s,r) and any specification of the auxiliary vector $\mathbf{x}_k$. It is akin to $1 - R^2$ in ordinary regression analysis, where R^2 is the coefficient of determination, measuring the proportion of variation explained.

Balance indicators also contained in the unit interval can be built from various other functions of $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$. It follows from (6.8) that such alternatives include

$$BI_2 = 1 - 4P^2\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 1 - 4S^2_{t|s;d} \tag{7.2}$$

$$BI_3 = 1 - 2P\left(\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}\right)^{1/2} = 1 - 2S_{t|s;d} \tag{7.3}$$

We have $0 \le BI_1 \le BI_2 \le 1$ and $0 \le BI_3 \le BI_2 \le 1$ for any survey outcome (s,r) and any specification of $\mathbf{x}_k$. Under the perfect balance $\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d} = \mathbf{0}$, all three balance indicators take the maximal value of unity. All three are near unity if the lack of balance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ is near zero, or, equivalently, if the variance $S^2_{t|s;d}$ of the estimated response probabilities is small. The choice in practice between the three is essentially a matter of taste; one may prefer BI_1 because it arises from of the orthogonal components decomposition. A link to BI_3 is mentioned at the end of the section. Table 1 illustrates the three indicators for three selected situations.

Table 1 illustrates the following: (1) For one and the same situation (table column), differences between the three balance indicators can be quite pronounced; (2) The two columns with $1 - P = 40\%$ illustrate that, by construction, the balance decreases when the lack of balance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ increases, and that when the nonresponse rate P is not far from 0.5, BI_1 and BI_2 are close; (3) It is not contradictory that the balance increases (by all three indicators) despite a higher nonresponse, as in the comparison of the two columns with the same lack of balance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 0.2$. The explanation lies in Equation (5.1) for $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$: although nonresponse has gone up from 20% to 40%, it is a considerably reduced squared distance, $(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})'\mathbf{\Sigma}_s^{-1}(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s-r;d})$, between respondents and nonrespondents that has kept $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ unchanged at 0.2. This emphasizes that the nonresponse rate is not an adequate tool for assessing the consequences of nonresponse.

Table 1. Values of three balance indicators, as a function of the nonresponse rate $1 - P$ and of the lack of balance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$

Balance indicator	$1 - P = 20\%$ $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 0.2$	$1 - P = 40\%$ $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 0.2$	$1 - P = 40\%$ $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 0.4$
BI_1	0.20	0.70	0.40
BI_2	0.49	0.71	0.42
BI_3	0.28	0.46	0.24

Interpreting the value of the balance indicator – whether the balance achieved so far in the data collection is good or not so good, on the unit interval scale – is directly linked to the choice of the **x**-vector used in $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. We have decided at the outset – perhaps somewhat arbitrarily – on a certain fixed vector composition $\mathbf{x}_k$ on which to assess balance. It may contain "traditional" variables such as age and sex, and/or x-variables considered as good predictors of an important study variable y. The objective is to influence the data collection in the direction of improved balance, *for that particular vector*. One can continuously monitor the balance indicator, compute it several times, before and after modifications of an original data collection design. To keep the **x**-vector composition fixed is essential in comparing the balance achieved in different surveys conducted in the same country, or in comparing different countries with respect to one and the same survey, such as the Labour Force Survey. But achieving high balance for the chosen **x**-vector is no guarantee of low or negligible bias in the estimates made in the survey. The bias adjustment involves lack of balance as one out of three factors, as seen when we come to the estimation stage in Sections 8 and 9.

We need to examine how the balance indicators behave when several alternative **x**-vectors are at hand for defining $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, and thereby the three indicators. Suppose the choice is between two auxiliary vectors $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, where the more extensive one, $\mathbf{x}_{2k}$, contains all the variables in the weaker one, $\mathbf{x}_{1k}$, plus one or more additional ones, linearly independent of those in $\mathbf{x}_{1k}$. Then $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is greater for $\mathbf{x}_{2k}$ than for $\mathbf{x}_{1k}$, because the proportion "variance explained" in (6.4) is greater. Thus, at a given response rate P, the balance, by any one of the three indicators, is greater for the weaker vector $\mathbf{x}_{1k}$ than for the stronger vector $\mathbf{x}_{2k}$. Although in a sense contradictory, this is logical, because $\mathbf{x}_{2k}$ demands more for achieving balance: it involves more means for which equality is required. It is easier to obtain balance for a less extensive **x**-vector than for a more extensive one. The most striking illustration of this is that the weakest possible auxiliary vector, $\mathbf{x}_k = 1$, gives an "automatic perfect balance," $BI_j = 1$ for $j = 1, 2, 3$, and for any outcome (s,r), yet that vector is essentially useless for nonresponse adjustment.

An indicator similar to (and sometimes identical to) BI_3 in (7.3) was proposed by Schouten, Cobben, and Bethlehem (2009) and called by them "representativity indicator." Their derivation is based on the variability of estimated response probabilities, and on the notion that a small variability of such estimates might indicate a "representative set of respondents," for a given **x**-vector. (As Section 1 points out, there are historical aspects on the relation between "representativity" and "balance.") The estimates $\hat{\theta}_k$ may be derived in different ways; the authors use a logistic regression fit to obtain first $\hat{\boldsymbol{\beta}}$, then $\hat{\theta}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})/[1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})]$ for $k \in s$, and their variance, $S^2_{\hat{\theta}}$, is computed. The authors propose the "R-indicator" $R = 1 - 2S_{\hat{\theta}}$, in close resemblance with (7.3). For one and the same vector $\mathbf{x}_k$, the numerical difference between R and BI_3 is often inconsequential; sometimes the two indicators are equal, as in the case of simple random sampling when the logistic model is estimated by maximum likelihood, and $\mathbf{x}_k$ is the classification vector, $\mathbf{x}_k = \gamma_k = (0, \ldots, 1, \ldots, 0)'$. The authors illustrate the use of the R-indicator for a follow-up study for the Dutch Labour Force Survey (LFS) in 2005. Two follow-up samples of the LFS nonrespondents were "approached once more using either a call-back approach . . . or a basic question approach." They compute the R-indicator for three data sets: LFS alone, LFS plus call-back, and LFS plus basic question. It is concluded that the basic question approach is likely to "give more of the same," consequently the R-indicator does not improve (does not get larger). But an increase, although modest, in the R-indicator was noted when the call-back respondents are added to the LFS respondents. In that study, 14 categorical auxiliary variables were used in estimating the logistic regression model for the response probabilities.

8. Estimating Influences

A possibly improved balance at the data collection stage does not alter the fact that adjustment weighting is necessary at the estimation stage. A duality exists between "estimating the response probabilities θ_k" as in Section 6, and "estimating the influences $\phi_k = 1/\theta_k$" as in this section. We expect a correspondence between the two. The rationale for estimating the ϕ_k is that if they were known, $\sum_r d_k \phi_k y_k$ would be a preferred estimator, unbiased for the total $Y = \sum_U y_k$. But since they are unknown, $\sum_r d_k \phi_k y_k$ is an unattainable ideal. Hence, we seek estimates $\hat{\phi}_k$ of the ϕ_k, then construct $\hat{Y} = \sum_r d_k \hat{\phi}_k y_k$. Unavoidably, this compromise is biased, but possibly only to a modest extent, if a powerful auxiliary vector can be specified to derive $\hat{\phi}_k$.

Predictions $\hat{\phi}_k$ of the unknown ϕ_k are produced from two requirements, (i) a linear form, $\hat{\phi}_k = \boldsymbol{\lambda}'\mathbf{x}_k$, and (ii) a calibration to the sample level, inspired by the first part of (3.2) with $a_k = \mathbf{x}_k$ and $\phi_k = \hat{\phi}_k$, that is, $\sum_r d_k \hat{\phi}_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. We get $\hat{\phi}_k = m_k$, where

$$m_k = \left(\sum\nolimits_s d_k \mathbf{x}_k\right)' \left(\sum\nolimits_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \tag{8.1}$$

The same predictions $\hat{\phi}_k = m_k$ are also obtainable by a least squares fit. Consider $\phi_k = 1/\theta_k$ as a conceptually defined number tied to unit k. Determine $\boldsymbol{\lambda}$ to minimize the weighted sum of squares $\sum_r d_k(\phi_k - \boldsymbol{\lambda}'\mathbf{x}_k)^2$. The estimating equation (by differentiating with respect to $\boldsymbol{\lambda}$) is $\sum_r d_k \phi_k \mathbf{x}_k' - \boldsymbol{\lambda}' \sum_r d_k \mathbf{x}_k \mathbf{x}_k' = \mathbf{0}'$. It cannot be solved for $\boldsymbol{\lambda}$ as is, because $\sum_r d_k \phi_k \mathbf{x}_k$ contains the unknown ϕ_k. We replace that sum by its computable expected value $\sum_s d_k \mathbf{x}_k$ to obtain the equation $\sum_s d_k \mathbf{x}_k' - \lambda' \sum_r d_k \mathbf{x}_k \mathbf{x}_k' = \mathbf{0}'$; solving now for $\boldsymbol{\lambda}$ we get again the estimated influences $\hat{\phi}_k = m_k$. The resulting calibration estimator of the total $Y = \sum_U y_k$ is therefore

$$\hat{Y}_{CAL} = \sum\nolimits_r d_k m_k y_k \tag{8.2}$$

where the m_k given by (8.1) have the interpretation of estimates of the unknown influences $\phi_k = 1/\theta_k$; furthermore, they satisfy the calibration $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$.

Remark For some x-variables, information may exist up to the level of the population. That is, for the vector $\mathbf{x}_k^*$ composed of these variables, the population total $\sum_U \mathbf{x}_k^*$ is known. The auxiliary vector is then

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$$

where $\mathbf{x}_k^\circ$ contains the x-variables whose unknown population total is estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$. The calibration equation is then

$$\sum\nolimits_r d_k m_k^* \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$$

The bias of the resulting estimator $\sum_r d_k m_k^* y_k$ is, in the leading term, the same as that of (8.2) (see Särndal and Lundström 2005), but the variance may be significantly smaller. Since we focus on bias, it is sufficient here to consider the factors (8.1). □

Different $\mathbf{x}$-vectors are more or less effective for reducing the bias remaining in $\hat{Y}_{CAL}$. The primitive $\mathbf{x}$-vector, $\mathbf{x}_k = 1$ for all k, serves as a benchmark; it gives $m_k = Q = 1/P$ for all k, and $\hat{Y}_{CAL}$ becomes $\hat{Y}_{EXP} = \hat{N}\bar{y}_{r;d}$, where $\bar{y}_{r;d} = \left(\sum_r d_k y_k\right)/\sum_r d_k$, $\hat{N} = \sum_s d_k$, and EXP stands for "expansion" (of the respondent mean).

The weight factor m_k given by (8.1) is computable for $k \in s$ (but used in $\hat{Y}_{CAL}$ only for $k \in r$). Moments can be computed. We need the mean over s, and mean and variance over r:

$$\bar{m}_{s;d} = \sum\nolimits_s d_k m_k \Big/ \sum\nolimits_s d_k; \quad \bar{m}_{r;d} = \sum\nolimits_r d_k m_k \Big/ \sum\nolimits_r d_k;$$

$$S^2_{m|r;d} = \sum\nolimits_r d_k (m_k - \bar{m}_{r;d})^2 \Big/ \sum\nolimits_r d_k$$

We have $\sum_r d_k m_k = \sum_s d_k$ by (4.1), and $\sum_r d_k m_k^2 = \sum_s d_k m_k$. We develop to obtain

$$\begin{aligned} &\bar{m}_{s;d} = Q \times \bar{\mathbf{x}}'_{s;d} \boldsymbol{\Sigma}_r^{-1} \bar{\mathbf{x}}_{s;d}; \quad \bar{m}_{r;d} = Q = 1/P; \\ &S^2_{m|r;d} = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}) = Q^2 \times \mathbf{D}' \boldsymbol{\Sigma}_r^{-1} \mathbf{D} \end{aligned} \tag{8.3}$$

where $\bar{\mathbf{x}}_{s;d}$ and $\boldsymbol{\Sigma}_r$ are defined in (4.3) and (4.2), and $\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d}$ as before. It follows that $0 \le \mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D} = \bar{\mathbf{x}}'_{s;d}\boldsymbol{\Sigma}_r^{-1}\bar{\mathbf{x}}_{s;d} - 1$, which proves the second part of (4.5). We have

$$cv_{m|r;d} = S_{m|r;d}/\bar{m}_{r;d} = \left(\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}\right)^{1/2} \tag{8.4}$$

The two coefficients of variation, $cv_{m|r;d} = \left(\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}\right)^{1/2}$ in (8.4) and $cv_{t|s;d} = \left(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}\right)^{1/2}$ in (6.3), are strikingly similar. Only the weighting matrices differ in $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$ and $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. Both measure lack of balance. But whereas $cv_{t|s;d} \le (Q-1)^{1/2}$, we cannot state an upper bound on $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$ or on $cv_{m|r;d}$.

The decomposition (6.4) has a counterpart here, namely when we expand the sum of squares $\sum_s d_k(I_k m_k - 1)^2$, which measures the variability of the quantities $I_k m_k$ for $k \in s$ around their mean $\sum_s d_k I_k m_k / \sum_s d_k = 1$. Let $I_k m_k - 1 = I_k(m_k - Q) + (I_k Q - 1)$, and develop the square to obtain the orthogonal components decomposition

$$\sum\nolimits_s d_k(I_k m_k - 1)^2 = \sum\nolimits_s d_k I_k (m_k - Q)^2 + \sum\nolimits_s d_k (I_k Q - 1)^2 \tag{8.5}$$

The cross product term is zero because $\sum_s d_k I_k (m_k - Q)(I_k Q - 1) = 0$. Simplifying each term and dividing through by $Q \times \left(\sum_s d_k\right)$, we can write the decomposition as

$$1 - P + \mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D} = \mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D} + (1 - P) \tag{8.6}$$

There is no upper bound apparent on $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$, but (8.5) and (8.6) do lead to an upper bound, explained at the end of Section 9, on the deviation of $\hat{Y}_{CAL}$ from the unbiased but unattainable estimate $\hat{Y}_{FUL}$.

9. Consequences for the Estimation Stage

When the estimation stage begins, the objective has shifted from "obtaining a better balanced response set" to "realizing an effective nonresponse bias adjustment," given the ultimate response set r. We can view this as a process starting with the primitive, or benchmark, estimator $\hat{Y}_{EXP} = \hat{N}\bar{y}_{r;d}$, generated by $\mathbf{x}_k = 1$, and proceeding in the direction of improved (less biased) estimators, $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$ with weight factors m_k given by (8.1), calibrated on increasingly powerful $\mathbf{x}$-vectors. The ideal but unattainable unbiased estimator for full response is $\hat{Y}_{FUL} = \sum_s d_k y_k$. Conceptually, two deviations are of interest:

$$\hat{Y}_{CAL} - \hat{Y}_{FUL} = \sum\nolimits_s d_k(I_k m_k - 1)y_k; \quad \hat{Y}_{EXP} - \hat{Y}_{FUL} = \sum\nolimits_s d_k(I_k Q - 1)y_k$$

If computable, $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ would be an estimate of the bias remaining in $\hat{Y}_{CAL}$, and $\hat{Y}_{EXP} - \hat{Y}_{FUL}$ would be an estimate of the usually larger bias of the benchmark $\hat{Y}_{EXP}$. We define the (empirical) *bias ratio*, for given outcome (s,r), given y-variable and given $\mathbf{x}$-vector, as

$$\text{bias ratio} = \frac{\hat{Y}_{CAL} - \hat{Y}_{FUL}}{\hat{Y}_{EXP} - \hat{Y}_{FUL}}$$

A third important deviation measures the change in the estimate,

$$\hat{Y}_{EXP} - \hat{Y}_{CAL} = \sum\nolimits_s d_k I_k (Q - m_k) y_k = \hat{N}\bar{y}_{r;d} - \sum\nolimits_r d_k m_k y_k$$

Then

$$\text{bias ratio} = 1 - \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{Y}_{EXP} - \hat{Y}_{FUL}}$$

with a computable numerator $\hat{Y}_{EXP} - \hat{Y}_{CAL}$, called *the adjustment*. It serves to compare alternative $\mathbf{x}$-vectors, one of which will be finally used for $\hat{Y}_{CAL}$. When the $\mathbf{x}$-vector becomes more powerful, the adjustment $\hat{Y}_{EXP} - \hat{Y}_{CAL}$ tends to increase (in absolute value), indicating a shrinking bias; the unknown denominator $\hat{Y}_{EXP} - \hat{Y}_{FUL}$ stays constant. The adjustment divided by $\hat{N} = \sum_s d_k$ can be written with the aid of (4.1) and (4.4) as the bilinear form

$$\frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N}} = \mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{C} \tag{9.1}$$

where the component C_j of $\mathbf{C} = (C_1, \ldots, C_j, \ldots, C_J)'$ is the covariance between auxiliary variable x_j and the study variable y,

$C_j = Cov(x_j, y) = \sum_r d_k(x_{jk} - \bar{x}_{j|r;d})\,(y_k - \bar{y}_{r;d})/\sum_r d_k$.

Perfect balance, $\mathbf{D} = \mathbf{0}$, implies $\hat{Y}_{EXP} = \hat{Y}_{CAL}$, and no adjustment occurs: the calibration estimator will not distance itself from the primitive one. But this does not guarantee a small bias; the unknown deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ may still be far from zero. The adjustment is also null if $\mathbf{C} = \mathbf{0}$, that is, in a complete absence of y-to-$\mathbf{x}$ relationship.

It is convenient to measure the adjustment in standardized units. Let $S_y^2 = S_{y|r;d}^2 = \sum_r d_k(y_k - \bar{y}_{r;d})^2 / \sum_r d_k$, where $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$. The *standardized adjustment* is defined as

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} \tag{9.2}$$

It measures the effectiveness for bias reduction of the auxiliary vector $\mathbf{x}_k$ used in $\hat{Y}_{CAL}$. More specifically, *StAdj* is the number of standard deviations that the improved mean estimate $\hat{Y}_{CAL}/\hat{N}$ has moved away from the primitive mean estimate $\hat{Y}_{EXP}/\hat{N}$. For example, if $StAdj = 0.16$, then $\hat{Y}_{CAL}/\hat{N} = \hat{Y}_{EXP}/\hat{N} - 0.16S_y$, stating that $\hat{Y}_{CAL}/\hat{N}$ has become 0.16 standard deviations removed (to the negative side) from the primitive mean estimate, and there is reason to believe that part of the distance to the ideal estimate $\hat{Y}_{FUL}/\hat{N}$ has been bridged. Although $0.16S_y$ may seem modest, it should be assessed in relation to an indication of standard error of the estimated mean; a conservative one is $S_y/\sqrt{n}$, which for $n = 10,000$ takes the much smaller value $0.01S_y$. In such a situation, the squared bias is likely to be the completely dominating component of the MSE.

We write the standardized adjustment as a product of three easily interpreted and computable factors,

$$StAdj = \frac{\hat{Y}_{EXP} - \hat{Y}_{CAL}}{\hat{N} \times S_y} = cv_{m|r;d} \times R_{y\mathbf{x}} \times R_{DC} \tag{9.3}$$

where

$$cv_{m|r;d} = \left(\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{D}\right)^{1/2}; \quad R_{y\mathbf{x}} = \frac{\left(\mathbf{C}'\mathbf{\Sigma}_r^{-1}\mathbf{C}\right)^{1/2}}{S_y};$$

$$R_{DC} = \frac{\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{C}}{\left(\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{D}\right)^{1/2}\left(\mathbf{C}'\mathbf{\Sigma}_r^{-1}\mathbf{C}\right)^{1/2}}$$

Their product equals $\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{C}/S_y$, confirming (9.1). In practice, a typical range for the first factor, $cv_{m|r;d} = (\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{D})^{1/2}$, is $0 \le cv_{m|r;d} \le 0.8$. As noted earlier, $\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{D} \ge 0$ is a measure of lack of balance with value zero under the perfect balance $\mathbf{D} = \mathbf{0}$. As for the second factor, it can be shown that $R_{y\mathbf{x}}^2 = \mathbf{C}'\mathbf{\Sigma}_r^{-1}\mathbf{C}/S_y^2$ is the coefficient of determination (the proportion of variance explained) for the (d_k-weighted) multiple regression fit of y_k on $\mathbf{x}_k$, $k \in r$. Hence $0 \le R_{y\mathbf{x}} \le 1$. Finally, $\left(\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{C}\right)^2 / \left(\mathbf{D}'\mathbf{\Sigma}_r^{-1}\mathbf{D}\right)\left(\mathbf{C}'\mathbf{\Sigma}_r^{-1}\mathbf{C}\right) = R_{DC}^2$ can be interpreted as the coefficient of determination in a weighted least squares fit of a regression through the origin of D_j on C_j, $j = 1, 2, \ldots, J$. We have $-1 \le R_{DC} \le 1$, and $|R_{DC}| = 1$ if the perfect proportionality $D_j = K \times C_j$, $j = 1, 2, \ldots, J$, holds for some constant K. Hence $|R_{DC}|$ measures the degree to which large differences $D_j = \bar{x}_{j|r;d} - \bar{x}_{j|s;d}$ between respondents and full sample are matched with large correlations between y and x_j. These properties were shown in Särndal and Lundström (2010).

For data encountered in practice, $|StAdj|$ is seldom greater than 0.30. To illustrate with fairly typical numbers, if $cv_{m|r;d} = 0.4$, $R_{y\mathbf{x}} = 0.8$, $R_{DC} = 0.5$, then $StAdj = 0.16$.

In order for *StAdj* to be numerically important, each of its three factors needs to be distinctly nonzero. Even if y and $\mathbf{x}$ are strongly related, with a coefficient of determination $R^2_{y\mathbf{x}}$ of say 0.90, $|StAdj|$ is small if the other two factors are near zero.

We use the sum of squares on the left-hand side of (8.5) to establish an upper bound on the deviation of $\hat{Y}_{CAL}$ from the unbiased estimate $\hat{Y}_{FUL}$ which we would prefer but cannot obtain. Although the bound is not sharp, it shows the maximum deviation that can occur given the survey outcome (s,r), and it can be of a certain guidance. To establish the bound, define the full sample moments $\bar{y}_{s;d} = \sum_s d_k y_k / \sum_s d_k$ and $S^2_{y|s;d} = \sum_s d_k (y_k - \bar{y}_{s;d})^2 / \sum_s d_k$, and the regression residuals for the (hypothetical) full sample least squares fit, $e_k = y_k - \mathbf{x}'_k \mathbf{B}_s$ with $\mathbf{B}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$. Then $\bar{e}_{s;d} = \sum_s d_k e_k / \sum_s d_k = 0$ by (4.1), and $\sum_s d_k (I_k m_k - 1) \mathbf{x}'_k \mathbf{B}_s = 0$. Hence

$$\hat{Y}_{CAL} - Y_{FUL} = \sum\nolimits_s d_k (I_k m_k - 1) y_k = \sum\nolimits_s d_k (I_k m_k - 1) e_k$$

By the Cauchy-Schwartz inequality

$$\left[\sum\nolimits_s d_k (I_k m_k - 1)^2\right] \left[\sum\nolimits_s d_k \left(e_k^2\right)\right] \geq \left[\sum\nolimits_s d_k (I_k m_k - 1) e_k\right]^2 = (\hat{Y}_{CAL} - \hat{Y}_{FUL})^2$$

Let $S^2_{e|s;d} = \sum_s d_k e_k^2 / \sum_s d_k$ be the residual variance. Then with the use of (8.5) and (8.6)

$$\frac{\left|\hat{Y}_{CAL} - \hat{Y}_{FUL}\right|}{\hat{N} \times S_{y|s;d}} \leq \left(Q - 1 + Q \times \mathbf{D}' \boldsymbol{\Sigma}_r^{-1} \mathbf{D}\right)^{1/2} \times \frac{S_{e|s;d}}{S_{y|s;d}} \tag{9.4}$$

Suppose the nonresponse rate is $1 - P = 20\%$, the lack of balance $\mathbf{D}' \boldsymbol{\Sigma}_r^{-1} \mathbf{D} = 0.08$, and the explained portion of the y-variance

$$1 - \frac{S^2_{e|s;d}}{S^2_{y|s;d}} = 90\%$$

Then

$$\frac{\left|\hat{Y}_{CAL} - \hat{Y}_{FUL}\right|}{\hat{N} \times S_{y|s;d}} \leq (0.25 + 1.25 \times 0.08)^{1/2} \times (0.1)^{1/2} = 0.19$$

This states that for any y-variable, the computable mean estimate $\hat{Y}_{CAL}/\hat{N}$ and the ideal but unattainable mean estimate $\hat{Y}_{FUL}/\hat{N}$ differ by at most 0.19 standard deviations $S_{y|s;d}$. The size of the nonresponse is decisive. If the nonresponse rate increases to $1 - P = 40\%$ and the lack of balance to $\mathbf{D}' \boldsymbol{\Sigma}_r^{-1} \mathbf{D} = 0.32$, while the explained portion of the y-variance remains at 90%, then the upper bound in (9.4) increases considerably to 0.35.

For any given survey outcome (s,r), the upper bound in (9.4) applies to any y-variable of interest in the survey, including for example those with markedly skewed distribution. The bound conveys a certain message. To make it a computable indicator of the survey situation, it is suggested that one replaces all sums over s in $S^2_{y|s;d}$ and $S^2_{e|s;d}$ by sums over r and changes the weighting from d_k into $d_k m_k$.

10. Empirical Illustration

A realistic goal for the data collection is a high degree of balance in the ultimate response set, as measured in Section 7 by $BI_1 = 1 - \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}/(Q-1)$, or two alternatives. But whether or not a satisfactory balance has been achieved at that stage, the estimation stage requires adjusting all y-variable estimates in the survey. To that end, we need to build an effective $\mathbf{x}$-vector on which to compute the weights m_k in the estimator (8.2). Formula (9.3) shows the standardized adjustment *StAdj* as a product of three factors, the first of which, $cv_{m|r;d} = (\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D})^{1/2}$, has the advantage of being a function of the chosen $\mathbf{x}$-vector but not of the study variable y. We seek a vector $\mathbf{x}_k$ that brings an important adjustment *StAdj*. In practice, a value $|StAdj|$ greater than 0.30 is rare, so 0.30 must be considered large. Factors that contribute to a sizeable $|StAdj|$ are:

i) a high lack of balance to be compensated for (the factor $cv_{m|r;d} = \left(D'\Sigma_r^{-1}D\right)^{1/2}$);
ii) a high degree of relationship y-to-$\mathbf{x}$ (the factor $R_{y\mathbf{x}}$);
iii) a high correlation between D_j and C_j (the factor $|R_{DC}|$).

Suppose that the $\mathbf{x}$-vector is built from scratch, starting with the trivial $\mathbf{x}_k = 1$, and adding successively more x-variables from a perhaps extensive list of available ones. The effect of adding one or more variables to any given $\mathbf{x}$-vector is to increase both $cv_{m|r;d}$ and $R_{y,\mathbf{x}}$, but the product *StAdj* does not necessarily increase in absolute value. It increases up to a point. More specifically, the pattern when new x-variables are added to $\mathbf{x}_k$ is that $|R_{DC}|$ shows a decreasing tendency; in the beginning, $|StAdj|$ increases, but at a certain point the admission of new x-variables will cause $|StAdj|$ to decrease, namely when the increases in $cv_{m|r;d}$ and $R_{y\mathbf{x}}$ are more than offset by the decrease in $|R_{DC}|$.

The stepwise forward selection of auxiliary variables to build the vector $\mathbf{x}_k$ requires a *selection criterion* whose computed value serves to trigger the selection of a new x-variable in each step. Three such criteria have been used in experiments at Statistics Sweden:

$$H_1 = cv_{m|r;d} \times R_{y\mathbf{x}} \times |R_{DC}|; \quad H_2 = cv_{m|r;d} \times R_{y\mathbf{x}};$$
$$H_3 = cv_{m|r;d} = \left(\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}\right)^{1/2} \tag{10.1}$$

An advantage with the criterion H_3 is its independence of the y-variable(s). It produces a selection of x-variables that can be viewed as a compromise for the perhaps numerous y-variables addressed in the survey. In Step 1, H_3 selects the x-variable that is responsible for the greatest lack of balance $\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D}$; added in Step 2 is the variable accounting for the greatest lack of balance, given the first entered variable, and so forth. The criteria H_2 and H_3 increase in each step. Typically, the increments are at first large; when they become trivially small, it is a signal to stop. The criterion $H_1 = |StAdj|$ is tailor-made for a specific y-variable and can be expected to be more efficient than H_3 for that particular y-variable, but not for all the rest. In the stepwise entering of one variable at a time, the typical pattern is that H_1 increases for a number of steps but eventually turns around and starts to decline, for the reason given above, namely that the decrease in $|R_{DC}|$ more than outweighs the increases in $cv_{m|r;d}$ and $R_{y,\mathbf{x}}$. For the criterion H_1, one might use the turning point as a stopping rule.

Table 2. Stepwise forward selection, criterion H_3, study variable Employed. Successive values of $H_3 \times 10^3$, of $\hat{Y}_{CAL} \times 10^{-4}$, and of $RDEV = (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{Y}_{FUL} \times 10^2$. For reference, $\hat{Y}_{FUL} \times 10^{-4} = 427$. In parenthesis the number of categories of the auxiliary variable

Step	Auxiliary variable entered	$H_3 \times 10^3$	$\hat{Y}_{CAL} \times 10^{-4}$	*RDEV*
0	(Trivial)	0	472	10.7
1	Education level (3)	186	452	6.0
2	Cluster of postcode areas (6)	250	451	5.6
3	Country of birth (2)	281	450	5.5
4	Income class (3)	298	437	2.4
5	Age class (4)	354	440	3.1
6	Sex (2)	364	438	2.8
7	Urban centre dwelling (2)	374	438	2.6
8	Indebtedness (3)	381	436	2.3

To illustrate these issues we consider features of an experiment with Swedish data described in further detail in Särndal and Lundström (2010). The data consist of one single simple random sample s of $n = 2,000$ persons drawn from the Swedish Register of the Total Population. The overall response rate was 50.8%. Auxiliary variable selection was carried out in the experiment from a list of 12 categorical x-variables with values known for $k \in s$. Another variable, also with a value known for $k \in s$, served as the study variable in this experiment, namely the dichotomous variable *Employed* with value 1 or 0. Hence we can compute the full sample (unbiased) estimate, which is $\hat{Y}_{FUL} = 427 \times 10^4$, and use it as a benchmark. At Step 0, the auxiliary vector is the trivial $\mathbf{x}_k = 1$, and the benchmark estimate is $\hat{Y}_{EXP} = \hat{N}\bar{y}_{r;d} = 472 \times 10^4$, a severe overestimation of 10.7% compared with $\hat{Y}_{FUL}$. At each subsequent step, the criterion is computed for every one of the categorical x-variables remaining in the list; the variable corresponding to the highest value of the criterion is selected and entered into the $\mathbf{x}$-vector; the weights $d_k m_k$ and the estimate $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$ for that new $\mathbf{x}$-vector are computed. The important changes in $\hat{Y}_{CAL}$ typically occur in the first few steps; the changes quickly taper off.

Table 2 shows the stepwise selection of the first eight variables with the criterion H_3, which reaches a level of around 0.38, a fairly typical value. Table 3 shows the first eight to be selected with H_1, which reaches a level of around 0.12, also fairly typical.

Table 3. Stepwise forward selection, criterion H_1, study variable Employed. Successive values of $H_1 \times 10^3$, of $\hat{Y}_{CAL} \times 10^{-4}$, and of $RDEV = (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{Y}_{FUL} \times 10^2$. For reference, $\hat{Y}_{FUL} \times 10^{-4} = 427$. In parenthesis the number of categories of the auxiliary variable

Step	Auxiliary variable entered	$H_1 \times 10^3$	$\hat{Y}_{CAL} \times 10^{-4}$	*RDEV*
0	(Trivial)	0	472	10.7
1	Income class (3)	76	446	4.5
2	Education level (3)	107	435	2.0
3	Presence of children (2)	114	433	1.4
4	Urban centre dwelling (2)	118	431	1.1
5	Sex (2)	123	430	0.7
6	Marital status (2)	125	429	0.5
7	Days unemployed (3)	121	430	0.9
8	Months with sickness benefits (3)	120	431	1.0

Table 3 illustrates the turn-around of H_1, occurring at Step 6. In Table 3 the estimates $\hat{Y}_{CAL}$ stabilize after four steps at around 430×10^4, which is closer to the unbiased estimate $\hat{Y}_{FUL} = 427 \times 10^4$ than what is obtained with H_3 in Table 2, where the estimate stabilizes at around 438×10^4. This can be expected since H_1 is tailor-made for the y-variable.

11. Concluding Remarks

This article has extended the idea of *balancing* to the context of survey nonresponse: The set of respondents should be balanced vis-à-vis the whole probability sample. The notion of *lack of balance* is central; it is defined as a quadratic form in the differences in auxiliary variable means between the response set and the whole sample. Our measure of *balance* is "lack of balance with the opposite sign," and confined to the unit interval. At the data collection stage one may use aspects of responsive design to achieve good balance in an ultimate set of respondents. A pressing objective remains nevertheless for the estimation stage: To adjust for the bias that still affects the estimates. The size of the adjustment has remaining lack of balance as one of its factors; two others are also important. One is the strength of the relationship between the study variable y and the auxiliary vector $\mathbf{x}$, the other is the degree to which large auxiliary variable mean difference between respondents and full sample is matched with large correlation between that auxiliary variable and the study variable. Further work would include more in-depth study of the three factors and of how they interact.

12. References

Beaumont, J.F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment. Survey Methodology, 31, 227–231.

Brewer, K.R.W. (1994). Survey Sampling Inference: Some Past Perspectives and Present Prospects. Pakistan Journal of Statistics, 10A, 213–233.

Brewer, K.R.W. (1999). Design-based or Prediction Inference? Stratified Random vs Stratified Balanced Sampling. International Statistical Review, 67, 35–47.

Brick, J.M. and Jones, M.E. (2008). Propensity to Respond and Nonresponse Bias. Metron, 66, 51–73.

Crouse, C. and Kott, P.S. (2004). Evaluation Alternative Calibration Schemes for An Economic Survey with Large Nonresponse. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Deville, J.C. (2002). La correction de la nonréponse par calage généralisé. Actes des Journées de Méthodologie, I.N.S.E.E., Paris. [In French]

Deville, J.C. and Tillé, Y. (2004). Efficient Balanced Sampling: The Cube Method. Biometrika, 91, 893–912.

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.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model Dependent and Probability Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, 776–793.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. Journal of Official Statistics, 19, 81–98.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology, 32, 133–142.

Kott, P.S. (2008). Some Research on Calibration Estimators at the National Agricultural Statistics Service. Pakistan Journal of Statistics, 24, 145–161.

Laflamme, F. (2009). Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada. Proceedings, 57th Session of the ISI, South Africa.

Little, R.J. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? Survey Methodology, 31, 161–168.

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.

Rizzo, L., Kalton, G., and Brick, J.M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. Survey Methodology, 22, 43–53.

Royall, R.M. (1970). On Finite Population Sampling Theory under Certain Linear Regression Models. Biometrika, 57, 377–387.

Royall, R.M. and Herson, J. (1973a). Robust Estimation in Finite Populations, I. Journal of the American Statistical Association, 68, 880–889.

Royall, R.M. and Herson, J. (1973b). Robust Estimation in Finite Populations, II: Stratification on a Size Variable. Journal of the American Statistical Association, 68, 890–893.

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, 167–191.

Särndal, C.-E. and Lundström, S. (2010). Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias. Survey Methodology, 36, 131–144.

Schouten, B. (2007). A Selection Strategy for Weighting Variables under a Not-missing-at-Random Assumption. Journal of Official Statistics, 23, 51–68.

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

Slud, E.V. and Bailey, L. (2010). Evaluation and Selection of Models for Attrition Nonresponse Adjustment. Journal of Official Statistics, 26, 127–143.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference – A Prediction Theory. New York: Wiley.

Received October 2010