Does balancing of survey response reduce nonresponse bias?

Barry Schouten and Fannie Cobben

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201226)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*****	provisional figure
******	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/’12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/’10– 2011/’12	crop year, financial year, etc. 2009/’10 to 2011/’12 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201200 X-10

DOES BALANCING OF SURVEY RESPONSE REDUCE NONRESPONSE BIAS?

Summary: Recently, representativeness indicators, or R-indicators, have been proposed as indirect measures of nonresponse error in surveys. The indicators employ available auxiliary variables in order to detect nonrepresentative response. They may be used as quality objective functions in the design of survey data collection. Such designs are called adaptive survey designs as different subgroups receive different treatments. The obvious question is whether the decrease in nonresponse bias caused by adaptive survey designs could also be achieved by nonresponse adjustment methods that employ the same auxiliary variables.

In this paper, we discuss this important question. We provide theoretical and empirical considerations on the role of both the survey design and nonresponse adjustment methods to make response representative. The empirical considerations are supported by a wide range of household and business surveys from Statistics Netherlands.

Keywords: Nonresponse; R-indicators; Nonresponse adjustment.

1. Introduction

Response rates alone are not a good indication of survey quality (Stoop 2005, Groves and Peytcheva 2008). There are examples of survey statistics that showed larger nonresponse bias after additional response was collected (Cobben and Schouten 2008). Therefore, research in the survey methodological area has focussed on proposing alternative indicators for survey quality. One of those indicators is the so-called R-indicator (Schouten, Cobben and Bethlehem 2009) in which the 'R' stands for representativity. This indicator measures the deviation from representative response. The key ingredient to the R-indicator is the estimated response probability. Response probabilities are calculated with a pre-determined set of registry variables and paradata observations and, hence, the R-indicator cannot be viewed separately from these variables. This holds true, however, for any indicator that attempts to measure traces of non-representative response beyond the response rate. Alternative indicators that have recently been proposed by Särndal and Lundström (2008 and 2010), Särndal (2011a and 2011b), Wagner (2010) and Andridge and Little (2011), share the same dependence on the set of covariates.

R-indicators may be used as quality objective functions in nonresponse reduction procedures. Schouten, Shlomo and Skinner (2011) propose partial R-indicators that detail the analysis of representativity of response to individual variables and to categories of those individual variables. These indicators may be used to identify subpopulations based on registry data and paradata that may be targeted in the nonresponse reduction. When different subgroups receive different treatments, then the design is called an adaptive survey design (Wagner 2008, Schouten, Calinescu and Luiten 2011). One important and legitimate criticism to such balancing and optimization of data collection is that the same could be achieved by nonresponse adjustment afterwards on the same set of variables. Beaumont and Haziza (2011), for instance, propose to focus on the minimization of standard errors in adaptive survey designs for this reason. We believe, however, that stronger nonrepresentative response on a set of relevant registry data and paradata is indicative of stronger nonrepresentative response on other variables as well. Hence, balancing response relative to a number of registry and paradata variables will in general also improve representativeness with respect to other variables. It will not remove the need to adjust afterwards, but we conjecture that remaining bias is smaller also after adjustment. So R-indicators are viewed as process quality indicators; lower values are indications of a low process quality that transfers to the survey target variables. In this paper, we seek empirical evidence to support this conjecture.

The main research question to this paper is: Does balancing of data collection treatments on auxiliary variables lead to less nonresponse bias on survey target variables, even after adjustment using the same auxiliary variables? In other words, if we choose design features differently, like the survey mode or the timing and number of interviewer calls, for say age and income, would the reduction of nonresponse bias be larger than for nonresponse adjustment of a single treatment design using age and income as weighting variables.

First, we make one important side remark. By design, the R-indicator has no direct relation to any specific population parameter, e.g. the mean or total, or estimator of that parameter. As a result it cannot be expected to be the best indicator for all parameters. For this reason, we also focus attention to the coefficient of variation of response propensities, termed the maximal bias by Schouten, Cobben and Bethlehem (2009), as this measure links to the nonresponse bias of population means and totals. In this paper, we restrict ourselves to the estimation of population means.

There is no easy way to answer the main research question, as in most cases nonresponse biases on survey target variables are unknown. We circumvent this complexity by dividing available auxiliary variables into two sets: a set to be used in the assessment and improvement of indicators of nonrepresentative response and a set to be used in the evaluation of remaining nonresponse bias. We can do so in two ways. First, we can apply indicators to growing sets of auxiliary variables and investigate whether patterns are consistent, i.e. whether worse indicator values on small models go together with worse values on large models. The non-selected auxiliary variables in the small models function as surrogates of survey target variables. Since the indicators employ multivariate models, consistency of the patterns corresponds to smaller remaining bias after nonresponse adjustment. Second, and alternatively, we can pick one auxiliary variable, treat it artificially as a survey target variable, perform nonresponse adjustment using all auxiliary variables but the selected auxiliary variables and compare remaining bias to indicator values. Worse indicator values should then relate to smaller remaining bias.

In this paper, we apply both approaches and break down the main research question into two sub questions:

1. Do worse indicator values on a set of variables X coincide with worse indicator values on other variables?
2. Do worse indicator values on a set of variables X coincide with larger remaining nonresponse bias on other variables after calibration on X?

It is important to stress that both questions are mostly empirical questions. One can easily construct examples where the answer to both questions is no. An affirmative answer to the questions, therefore, does not imply that the indicators have the feature that they detect nonresponse bias on other variables. It merely means that lower quality survey data collection, in the majority of cases, tends to affect the full range of potential variables and that the indicators signal this tendency. This issue has also been debated by Särndal (2011).

We have selected a wide range of survey data sets, both household and business surveys, to find empirical support. We compare the representativeness of response for growing sets of auxiliary variables and compare the patterns for different data sets. In all cases, the set X is exactly the same. Comparisons are made over different surveys, over different waves of a survey, during data collection and after different survey process steps like establishing contact and obtaining cooperation.

We discuss one data set in more detail, an adaptive survey design pilot study linked to the 2009 Dutch Survey of Consumer Sentiments. In this pilot study, the design was adapted to six socio-demographical variables in an attempt to improve R-indicators while having similar costs and response rates. To this data set other auxiliary variables were linked that were not available during the design of the pilot. The pilot was performed within the 7th EU Research Framework Programme project RISQ, see www.risq-project.eu.

In this paper, we focus on nonresponse bias. We assume that sample sizes are large, so that precision is not an issue. The restriction to the nonresponse error may, however, be too naïve, especially when multiple survey modes or intensive refusal conversion procedures are considered. Calinescu, Schouten and Bhulai (2012) generalize adaptive survey designs to nonresponse and measurement errors. Since adjustment methods for measurement error are quite different from adjustment methods for nonresponse error, we leave a debate about balancing or adjusting measurement error to a future paper.

The paper is organized as follows. In section 2, we briefly describe the various indicators for representative response. Next, in section 3, we provide theoretical considerations on representativeness of response and nonresponse adjustment. We show that the indicators are important components of nonresponse bias, but that, theoretically, none of the answers to the two research questions is affirmative for all settings. In section 4, we move to empirical evaluations of the role of the survey design and nonresponse adjustment in obtaining representative response. In section 5, we end with a summary and conclusion.

2. Indicators for nonresponse error

In this section, we briefly revisit the R-indicators and the coefficient of variation of response propensities. We link them to balance indicators proposed by Särndal and Lundström (2008 and 2010) and Särndal (2011), and to adaptive survey designs. We refer to Schouten, Cobben and Bethlehem (2009), Schouten, Shlomo and Skinner (2011) and Shlomo, Skinner and Schouten (2012) for detailed accounts of the R-indicators.

2.1 R-indicators and partial R-indicators

Let *i=1,2,…,N* be the labels of the units in the population. By s_i we denote the 0-1-sample indicator, i.e. in case unit *i* is sampled it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit *i*. If unit *i* is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is *n*. Next, π_i denotes the first-order inclusion probability of unit *i*, and ρ_i is the probability that unit *i* responds in case it is sampled, i.e. $\rho_i = P[r_i = 1 | s_i = 1]$. Let $\tilde{\rho} = (\rho_1, \rho_2, \dots, \rho_N)'$ be the vector of response probabilities.

As we do not observe ρ_i, we have to estimate its value. We do so by using a vector of auxiliary information x_i that is available for all units *i* in the sample. We let $\hat{\rho}_i$ denote an estimator for ρ_i that uses all or a subset of the available auxiliary variables contained in x_i. By $\hat{\bar{\rho}}$ we denote the weighted sample average of the estimated response probabilities, i.e.

$$\hat{\bar{\rho}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\rho}_i \frac{s_i}{\pi_i}, \tag{1}$$

where we use the inclusion weights π_i.

Now, the R-indicator is defined as

$$\hat{R} = 1 - 2\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \frac{s_i}{\pi_i} (\hat{\rho}_i - \hat{\bar{\rho}})^2} = 1 - 2\hat{S}(\hat{\rho}) \tag{2}$$

In (2) the variation in the estimated response probabilities is weighted using the inclusion probabilities and is computed with respect to the average weighted response probability given by (1).

The R-indicator $\hat{R}$ is an estimator of the population R-indicator that is based on the population standard deviation $S(\rho)$ of the 'true' response probabilities

$$R=1-2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i-\bar{\rho})^2}=1-2S(\rho). \qquad (3)$$

The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of x_i. Hence, the R-indicator cannot be seen separately from the variables *X* that are used to estimate response probabilities. Therefore, we denote the R-indicator by R(*X*).

In words, the response to a survey is called representative for a variable *X* when the response propensity function $\rho_X(x)$ is constant in x. Likewise, we call the response conditional representativity for Z given X, when the (joint) response propensity function $\rho_{ZX}(z,x)$ depends only on x.

The ratio of the standard deviation $S(\rho)$ of the response propensities over the response rate is called the maximal bias by Schouten, Cobben and Bethlehem (2009). A better name, perhaps, is the coefficient of variation of the response propensities. It is defined as

$$cv(\rho)=\frac{S(\rho)}{\bar{\rho}}, \qquad (4)$$

and provides an upper bound to the bias of the response mean as we will show in section 3.1.

The definitions of representative and conditionally representative response do not focus on survey variables that are missing, but they can be rephrased in terms of the well-known missing data mechanisms Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR); see Little and Rubin (2002). When response is representative for X then the missing data is Missing Completely At Random within classes defined by X. We denote this by MCAR(X). When response is conditionally representative for Z given X, then the missing data is MAR(Z,X). When the latter is not true, then the missing data is NMAR(Z,X). Hence, MCAR is an unconditional feature of the missing data on a variable and must hold regardless of any additional information available. MAR is a conditional feature of missing data on a variable and requires sufficient additional information. NMAR applies when missing data affects the distribution of a variable, but additional information is not powerful enough.

R-indicators and partial R-indicators are based on the Euclidean distances to representative response and conditionally representative response. Partial R-indicators are defined for categorical variables only. They decompose the variance in response propensities into between and within components. The unconditional partial R-indicator $P_u(X)$ equals the square root of the between variance given a stratification based on X. The conditional partial R-indicator $P_c(Z,X)$ is the square root of the within variance for the same stratification attributable to Z.

Partial R-indicators are Euclidean distances to MCAR and MAR. $P_u(X)$ measures the distance to MCAR(X), and $P_c(Z,X)$ the distance to MAR(Z,X). Hence, the conditional partial R-indicator reflects the extent to which missing data is NMAR for Z given X.

Details about the definition and properties of R-indicators and partial R-indicators can be found in Schouten, Cobben and Bethlehem (2009), Schouten, Shlomo and Skinner (2011) and Shlomo, Skinner and Schouten (2012).

It is straightforward to show that

$$P_c(Z,X) = S^2(\rho_{XZ}) - S^2(\rho_X) \leq S^2(\rho) - S^2(\rho_X), \tag{5}$$

where ρ represents the individual response propensity. If one would not like to assume the concept of individual propensities, then the propensities can be replaced by the response propensity function $\rho_{\aleph}$ of some supervector $\aleph$. For a recent discussion see also Olson and Groves (2012). Regardless of this conceptual choice, there can always be found a variable Z that exactly fulfils (5), i.e. a variable that takes up the remaining variance in the propensities.

Relation (4) also presents the well-known nonresponse paradox. A variable X that relates weakly to nonresponse, leaves more room for NMAR nonresponse behaviour, but at the same time does not contradict that nonresponse follows arbitrary, random patterns. Although lower R-indicators correspond to more predictive variables X and, thus, to a smaller, remaining, maximal impact of NMAR missing-data, it is indicative of an imperfect data collection process. This holds especially true because the explanatory variables are usually not selected because of their relation to nonresponse, but because of their general relation to many survey topics.

2.2 Alternative indicators

R-indicators are not the only measures put forward by the recent literature to measure nonresponse error and to adapt survey data collection.

Andridge and Little (2011) and Wagner (2010) propose to use the Fraction of Missing Information (FMI) as indicator for the quality of survey response. The FMI originates from the multiple imputation literature where missing data is imputed using available auxiliary information. It is, however, applicable to any missing data setting and to any estimation method. The FMI quantifies the amount of information that is still missing given the available, auxiliary information for some estimator of the sample mean. Since the FMI is essentially a measure of precision rather than of bias, it can be used as a measure that is complementary to R-indicators.

A variety of indicators was introduced by Särndal and Lundström (2008 and 2010) and Särndal (2011). The indicators in Särndal (2011), termed balance indicators, are very similar to R-indicators. In fact, if response propensities are estimated through linear regression, than one of the balance indicators is the R-indicator and other indicators are subtle variations using different scaling constants to transform the variance of response propensities to the [0,1] interval. Another indicator proposed by Särndal (2011), the distance measure, estimates the contrast between respondents and nonrespondents. It is equal to the coefficient of variation of the response propensities divided by the nonresponse rate, when using linear regression. Särndal and Lundström (2008 and 2010) introduce several indicators that rank covariates X in their ability to adjust for nonresponse. Särndal and Lundström (2010) factor the nonresponse adjustment in calibration into three terms. One of the terms isolates the component of the nonresponse adjustment that is independent of the target variable, i.e. the part that is present in any calibration adjustment. This component equals the coefficient of variation of the adjustment weights (termed response influences in their 2008 paper) in calibration estimators. The coefficient of variation is, however, restricted to respondents. It can be shown, using Taylor linearization, that this measure is very similar to the coefficient of variation of the estimated response propensities. In general, clearly, the coefficient of variation for respondents will be different from that for the full population. Bethlehem (2012) provides some discussion of the relation between the two coefficients. Given the strong resemblance between the indicators proposed by Särndal and Lundström (2008 and 2010), Särndal (2011) and the R-indicators, we conjecture that they provide very similar and consistent pictures when performing the evaluations of section 4.

2.3 R-indicators, partial R-indicators and adaptive survey designs

In this paper, we focus on adaptive survey designs (e.g. Schouten, Calinescu and Luiten 2011), i.e. we assume that response probabilities can be identified from previous waves of the same survey or from similar surveys that have been conducted. This contrasts responsive survey designs (Groves and Heeringa 2006), where evaluation time points are explicitly defined. These time points define so-called design phases. Each design phase could in fact be an adaptive survey design.

Adaptive survey designs may be viewed as extensions of sampling designs. Whereas standard sampling designs assume a uniform data collection strategy, adaptive survey designs employ multiple strategies. Instead of inclusion probabilities, there are strategy allocation probabilities. Another extension lies in the quality objective function. Sampling designs typically focus on precision, but adaptive survey designs also attempt to account for non-sampling errors through indirect quality indicators. Adaptive survey designs make an explicit trade off between quality and costs and other constraints. As such they need not restrict attention to nonresponse error, but up to now most studies limited themselves to nonresponse.

Essentially, adaptive survey designs differentiate efforts for different population subgroups. These subgroups may be defined from registry data but also during data collection based on interviewer observations or other forms of paradata. It is by no means trivial how the subgroups and effective data collection strategies are to be identified. Schouten, Shlomo and Skinner (2011) propose partial R-indicators to distinguish subgroups that contribute most to nonrepresentative response. Still, the response probabilities and costs associated with application of strategies to these subgroups need to be estimated in a robust way. Generally, adaptive survey designs should be modest, therefore, in their number of strategies and subgroups. However, the designs of surveys may benefit from a structured look and an explicit use of the strong quality-cost differential between strategies (most markedly the survey mode).

The important question arises whether any reduction of nonrepresentative response through adaptive survey designs could be achieved as well through nonresponse adjustment. In the previous sections, we explained that there is no mathematical reason to support the assertion that nonresponse adjustment has this property. There is, however, also not a statistical argument that shows that stronger nonrepresentative response on some variables transfers to other variables. In the next section, we provide some intuition behind the use of the R-indicator and coefficient of variation as indicators of process quality.

3. Some theoretical considerations on non-representative response

In this section, we link the various indicators to nonresponse adjustment and give a rationale why higher R-indicators on auxiliary variables should correspond to higher R-indicators on all variables.

3.1 R-indicators and nonresponse adjustment

We start by deriving expressions for the bias of simple response means and weighted response means. Since the R-indicators are not specific to any survey target variable Y, any population parameter of that survey variable or any estimator, it can be stated beforehand that the indicators cannot be ideal indicators in any specific setting. The indicators do allow for flexibility in choosing the covariates X, but for specific variables, parameters and estimators, other indicators may be preferred. Nonetheless, we show that the variance of response propensities and the response rate are important components of the bias of unweighted and weighted response means.

It is known that the bias of the response mean $\bar{y}_R$ is approximately equal to

$$B(\bar{y}_R) = \frac{\operatorname{cov}(Y,\rho)}{\bar{\rho}} = \frac{cor(Y,\rho)}{\bar{\rho}} S(\rho)S(Y), \tag{6}$$

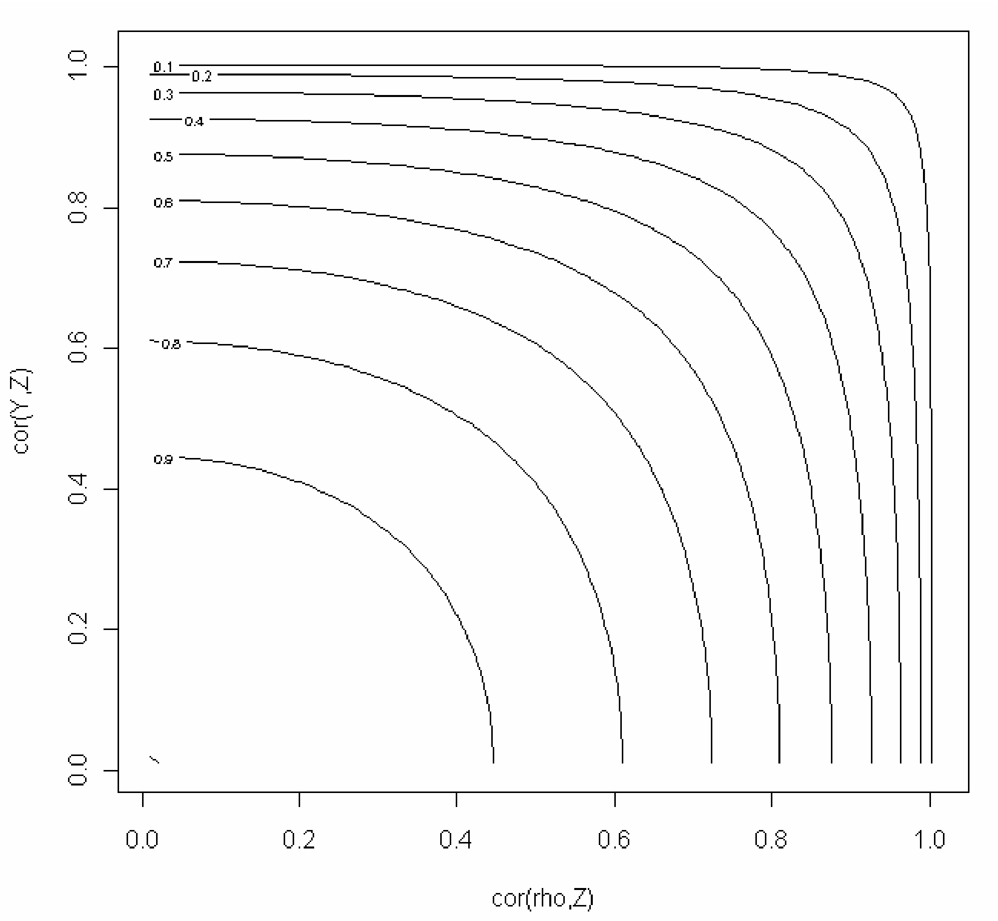
with $\bar{\rho}$ being the population average propensity. Using simple arguments, as proposed in Schouten (2007), it can be shown that for any variable Z it holds that

$$B(\bar{y}_R) \geq \frac{S(\rho)S(Y)}{\bar{\rho}} \left(cor(Y,Z)cor(Z,\rho) - \sqrt{1-cor^2(Y,Z)}\sqrt{1-cor^2(\rho,Z)} \right) \tag{7a}$$

$$B(\bar{y}_R) \leq \frac{S(\rho)S(Y)}{\bar{\rho}} \left(cor(Y,Z)cor(Z,\rho) + \sqrt{1-cor^2(Y,Z)}\sqrt{1-cor^2(\rho,Z)} \right). \tag{7b}$$

The intervals in (7a) and (7b) are sharp, meaning that a Y can be constructed that satisfies the upper or the lower bound. Figure 3.1.1 shows that correlations need to be large in order to get small bias intervals in (7a-b).

Figure 3.1.1: Contourplot of $\sqrt{1-cor^2(Y,Z)}\sqrt{1-cor^2(\rho,Z)}$.



Hence, any choice for a variable Z sets an interval for the bias that is centered around

$$\frac{S(\rho)S(Y)}{\bar{\rho}}cor(Y,Z)cor(Z,\rho). \tag{8}$$

The maximal bias in absolute sense can be derived as

$$\frac{S(\rho)S(Y)}{\bar{\rho}}\left(|cor(Y,Z)||cor(Z,\rho)|+\sqrt{1-cor^2(Y,Z)}\sqrt{1-cor^2(\rho,Z)}\right). \tag{9}$$

Schouten, Cobben and Bethlehem (2009) call $S(\rho)/\bar{\rho}$ the maximal bias. It is unknown, but used as a rationale by the authors to use $S(\rho_X)/\bar{\rho}$ as a quantity to derive acceptable levels for the R-indicator. It is easy to show that

$$\frac{S(\rho)}{\bar{\rho}}\leq\sqrt{\frac{1-\bar{\rho}}{\bar{\rho}}}.$$

A candidate choice for Z in (7) is the propensity score ρ_X, as introduced by the Rosenbaum and Rubin (1983 and 1984) papers, which is the projection of X to the plane of the propensity ρ. The propensity score maximizes $cor(\rho, Z)$ using linear combinations of the X, possibly through a link function like logistic or probit regressions. It can easily be seen that

$$cor(\rho_X, \rho) = \frac{S(\rho_X)}{S(\rho)}, \tag{10}$$

and, hence, the correlation between the propensity score and true propensity is always positive.

When $Z = \rho_X$, the absolute value of the displacement (8) can, then, be rewritten and bounded by the R-indicator when fixing the response rate

$$S(Y)\,|\,cor(Y, \rho_X)\,|\,\frac{S(\rho_X)}{\bar{\rho}} \leq \frac{S(\rho_X)S(Y)}{\bar{\rho}} = \frac{(1 - R(X))S(Y)}{2\bar{\rho}}, \tag{11}$$

i.e. the displacement is bounded by coefficient of variation, and, when we fix the response rate, by the R-indicator.

The maximal absolute bias in (9) translates to

$$S(Y)\left(|\,cor(Y, \rho_X)\,|\,\frac{S(\rho_X)}{\bar{\rho}} + \sqrt{1 - cor^2(Y, \rho_X)}\sqrt{\frac{S^2(\rho)}{\bar{\rho}^2} - \frac{S^2(\rho_X)}{\bar{\rho}^2}}\right). \tag{12}$$

It is important to note that $S(\rho)/\bar{\rho}$ changes from one survey to the other and from one time point during data collection to another time point. Therefore, for two surveys with the same R-indicator and response rate, the maximal absolute bias will, in general, be different. Let us for the moment view Y as unspecified and $S(\rho)/\bar{\rho}$ as unknown. We can treat (12) as a function of $S(\rho_X)/\bar{\rho}$, while keeping the other terms fixed. It is easy to show that (12) is convex and attains its maximum at

$$|\,cor(Y, \rho_X)\,|\,\frac{S(\rho)}{\bar{\rho}}, \tag{13}$$

and the maximum value is $S(Y)\dfrac{S(\rho)}{\bar{\rho}}$.

Hence, the coefficient of variation and the R-indicator show up in the upper bound to maximal absolute bias.

We, now, move to bias expressions for weighted response means. We focus on the general regression estimator, see e.g. Bethlehem, Cobben and Schouten (2011).

The general regression estimator employs $Z = \beta_{XY}^{/} X$, originating from the objective to predict Y from the covariate vector X. Conveniently, the estimator can be rewritten to a weighted sum of Y for respondents where weights are independent of the Y. It can be shown that the regression estimator using the true β_{XY} would center the interval in (7) around zero, but does not affect the width of the interval

$$B(\bar{y}_{GREG}) \geq -\frac{S(\rho)S(Y)}{\bar{\rho}}\sqrt{1-cor^2(Y,\beta'_{XY}X)}\sqrt{1-cor^2(\rho,\beta'_{XY}X)} \qquad (14a)$$

$$B(\bar{y}_{GREG}) \leq \frac{S(\rho)S(Y)}{\bar{\rho}}\sqrt{1-cor^2(Y,\beta'_{XY}X)}\sqrt{1-cor^2(\rho,\beta'_{XY}X)}\ . \qquad (14b)$$

This is intuitively clear as the regression estimator assumes MAR(Y,X) to hold while the interval is valid for any missing-data-mechanism.

In practice, the true slope parameter is unknown and needs to be estimated as well. If we ignore the sampling variation in the estimated slope parameter, then instead of $Z = \beta'_{XY}X$ the predictor $\tilde{Z} = \tilde{\beta}'_{XY}X$ is used, where $\tilde{\beta}_{X,Y}$ is the slope parameter for respondents.

The general regression estimator can be rewritten to

$$\bar{y}_{GREG} = \bar{y}_R - \tilde{\beta}'_{X,Y}\bar{x}_R + \tilde{\beta}'_{X,Y}\bar{X} = \overline{\left(y-\tilde{\beta}'_{X,Y}x\right)}_R + \tilde{\beta}'_{X,Y}\bar{X}\ , \qquad (15)$$

with $\bar{X}$ the population mean for the auxiliary vector, and ignoring variation in the estimated slope parameter.

Using similar arguments as for (7a) and (7b), it can be shown that

$$B(\bar{y}_{GREG}) \geq \frac{S(\rho)S(Y-\tilde{\beta}'_{X,Y}X)}{\bar{\rho}}\left(cor(Y-\tilde{\beta}'_{X,Y}X,Z)cor(Z,\rho)-\sqrt{1-cor^2(Y-\tilde{\beta}'_{X,Y}X,Z)}\sqrt{1-cor^2(\rho,Z)}\right)$$

$$B(\bar{y}_{GREG}) \leq \frac{S(\rho)S(Y-\tilde{\beta}'_{X,Y}X)}{\bar{\rho}}\left(cor(Y-\tilde{\beta}'_{X,Y}X,Z)cor(Z,\rho)+\sqrt{1-cor^2(Y-\tilde{\beta}'_{X,Y}X,Z)}\sqrt{1-cor^2(\rho,Z)}\right)$$

(16)

Using again $Z = \rho_X$, we get for the maximal absolute bias

$$S(Y-\tilde{\beta}'_{X,Y}X)\left(|\,cor(Y-\tilde{\beta}'_{X,Y}X,\rho_X)\,|\frac{S(\rho_X)}{\bar{\rho}}+\sqrt{1-cor^2(Y-\tilde{\beta}'_{X,Y}X,\rho_X)}\sqrt{\frac{S^2(\rho)}{\bar{\rho}^2}-\frac{S^2(\rho_X)}{\bar{\rho}^2}}\right)$$

. (17)

Expression (17) is interesting as it shows how variation in propensities ρ_X relates to the bias. Two counteracting forces can be observed. The larger $S(\rho_X)/\bar{\rho}$, the smaller the second term in (17) and the smaller the width of the interval. This is intuitively clear; the better we understand the nonresponse the less room for NMAR nonresponse. It should, however, again be noted that for different surveys or different data collection time points, the term $S(\rho)/\bar{\rho}$ changes as well. This is again the well-known nonresponse paradox.

Contrary, the larger $S(\rho_X)/\bar{\rho}$, the bigger the displacement and the first term of (17). Consequently, there is an increasing impact of correlation between regression residuals and response propensities. When the nonresponse is MAR, then $\tilde{\beta}_{X,Y} = \beta_{X,Y}$, and the residuals are uncorrelated with the response propensities ρ_X, i.e. the first term in (17) vanishes. If the nonresponse is NMAR, then $\tilde{\beta}_{X,Y} \neq \beta_{X,Y}$, and the residuals and propensities may be correlated.

From (17), it is not true, in general, that larger $S(\rho_X)/\bar{\rho}$ in one survey or at one time point imply a smaller maximal absolute bias as $S(\rho)/\bar{\rho}$ may change. However, if, analogous to (12), we view Y as unspecified and $S(\rho)/\bar{\rho}$ as unknown, then (17) can be considered as a function of $S(\rho_X)/\bar{\rho}$. It is again a convex function of $S(\rho_X)/\bar{\rho}$. The point where it attains a maximum is

$$|\,cor(Y-\tilde{\beta}_{X,Y}^{/}X,\rho_X)\,|\,\frac{S(\rho)}{\bar{\rho}}, \tag{18}$$

and the maximal value is $S(Y-\tilde{\beta}_{X,Y}^{/}X)\dfrac{S(\rho)}{\bar{\rho}}$.

From (17), it cannot be concluded that smaller $S(\rho_X)/\bar{\rho}$ will in general lead to smaller bias of the general regression estimator. There is a risk that the increase in displacement is larger than the decrease in interval width. The contrary cannot be concluded either, however. Nonresponse adjustment does not fully remove traces of nonresponse bias on the auxiliary variables under NMAR nonresponse; $S(\rho_X)/\bar{\rho}$ still impacts the size of the bias even after adjustment.

We left Y unspecified. For this reason, we can lift the correlation in (18) to 1 for a specific choice of Y. This would, however, be an academic choice of target variable. In the next sections, we turn to empirical evaluations where we treat some of the auxiliary variables as target variables.

Summarizing, we can say that $S(\rho_X)/\bar{\rho}$ turns out to be an important component of nonresponse bias of response means with or without nonresponse adjustment. Hence, we expect that when the population mean is the parameter of interest, then the coefficient of variation provides a more focussed, and, hence, more consistent view than the R-indicator which is not linked to a specific population parameter or estimator.

3.2 The R-indicator and coefficient of variation as process quality indicators

We would like to formalize the utility of the R-indicator and coefficient of variation as process quality indicators. More specifically, we would like to formalize the intuition that larger variation of response propensities for X corresponds to larger variation of the true individual response probabilities. Doing so, we capitalize on the existence of an individual response probability.

In section 3.1, we view the vector X as fixed and given. All derivations and conclusions that we have made so far, do apply, however, to any arbitrary vector. Here, we view auxiliary variables themselves as being sampled from the population of all possible random variables.

Suppose a large population consists of G fully homogeneous and equally sized groups, labelled by $g = 1,2,\dots,G$. All units in group g behave exactly the same in every way, and they have the same response probability for any given survey design. The stratification into the groups itself is not observed, but we do observe categorical variables X_k, $k = 1,2,\dots,K$, that cluster groups into smaller numbers of groups.

Let us for simplicity look at an 0-1 indicator variable X. Assume that X was constructed by a simple random sample without replacement of size G_X from the set of G groups. Let s_g be the 0-1 indicator that group g was selected. We then have the following definition of X

$$X = \begin{cases} 1 & \forall g, s_g = 1 \\ 0 & \forall g, s_g = 0 \end{cases}, \tag{19}$$

i.e. X is one for all selected groups g and zero otherwise. Since the groups have equal size, the probability that $X = 1$ is equal to G_X / G.

Now, let ρ_g be the response probability of group g, so that the response propensity function $\rho_X(x)$ for X is defined as

$$\rho_X(x) = \begin{cases} \frac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g & if \quad x = 1 \\ \frac{1}{G - G_X}\sum_{g=1}^{G} (1 - s_g)\rho_g & if \quad x = 0 \end{cases}. \tag{20}$$

In order to investigate the relation between the indicators based on X and those based on the full stratification with the G groups, we consider the expected mean and the expected variance of the response propensity function ρ_X.

The mean response propensity can be derived as

$$\bar{\rho}_X = \frac{G_X}{G}\frac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g + (1 - \frac{G_X}{G})\frac{1}{G - G_X}\sum_{g=1}^{G} (1 - s_g)\rho_g = \frac{1}{G}\sum_{g=1}^{G} \rho_g = \bar{\rho}\,. \tag{21}$$

From (21) we can conclude that the mean response propensity $\bar{\rho}_X$ is always equal to the mean individual response probability $\bar{\rho}$. Clearly, the expected mean response propensity is then also equal to $\bar{\rho}$. Hence, regardless of the choice of X, the mean response propensity is the mean of the individual probabilities.

The variance of ρ_X, $S^2(\rho_X)$, is equal to

$$S^2(\rho_X) = \frac{G_X}{G}(\rho_X(1) - \bar{\rho})^2 + (1 - \frac{G_X}{G})(\rho_X(0) - \bar{\rho})^2\,. \tag{22}$$

The expectation of $\rho_X(x)$ is always equal to $\bar{\rho}$, and, hence, (22) can be rewritten to

$$S^2(\rho_X) = \frac{G_X}{G}Var(\rho_X(1)) + (1 - \frac{G_X}{G})Var(\rho_X(0))\,, \tag{23}$$

where $Var(\rho_X(x))$ is the variance of ρ_X with respect to the sampling design. Since X is constructed using a simple random sample without replacement, $Var(\rho_X(x))$ is equal to

$$Var(\rho_X(x)) = \begin{cases} \frac{1}{G_X}(1-\frac{G_X}{G})S^2(\rho) & if \quad x=1 \\ \frac{1}{G-G_X}\frac{G_X}{G}S^2(\rho) & if \quad x=0 \end{cases} . \tag{24}$$

Combining (23) and (24) gives

$$S^2(\rho_X) = \frac{1}{G}S^2(\rho), \tag{25}$$

so that the expected variance is equal to the variance of the individual response probabilities times the population diversity constant $1/G$.

With similar arguments, it can be reasoned that if X is a categorical variable with C categories, then

$$S^2(\rho_X) = \frac{C-1}{G}S^2(\rho). \tag{29}$$

So for all X, the variance of the response propensity function ρ_X is proportional to the variance of the underlying variance of individual response probabilities. This is a useful finding as it implies that, if for some survey design the R-indicator is smaller or the coefficient of variation is larger than for another survey design, then also the variance of the individual response probabilities is larger. As a consequence, the expected variance of the propensity function resulting from any random draw of subgroups g would be larger for that design too. Although it would not be true that the variance of all propensity functions is larger, there may in fact be various variables that lead to a smaller variance, it must hold that for an arbitrary variable the variance is larger. This conclusion supports the intuition that for surveys with many target variables, one would prefer larger R-indicators or smaller coefficients of variation. It also shows that for single topic surveys, it may actually be the survey target variable itself that is one of the exceptional variables.

If we could assume that the set of auxiliary variables X_k, $k=1,2,\ldots,K$, consists of independent random draws of subgroups, then we could estimate G and $S^2(\rho)$. The parameter G may be estimated from the maximal covariance found among the X_k or using the first eigenvalue in a factor model. The variance $S^2(\rho)$ can be estimated by the average of the propensity function variances.

Of course, the discussion in this section is counterfactual, auxiliary variables cannot be considered as independent, random draws of population subgroups. However, models for nonresponse are often being critised for the lack of relevant, explanatory variables; standard variables like age or gender may have proved to be indicative of homogeneity in the population, they were certainly not picked to model response.

4. Empirical considerations on non-representative response

In the previous section, we have discussed the impact of NMAR nonresponse on nonresponse bias and enumerated a number of indicators that have been proposed in the recent literature. The only measurable term in the nonresponse bias derivations, is the coefficient of variation $S(\rho_X)/\overline{\rho}$ for propensities based on X. We have evaluated how the nonresponse bias relates to this indicator. However, since the true $S(\rho)/\overline{\rho}$ is unknown and may change from one survey to the other and during data collection, changes in $S(\rho_X)/\overline{\rho}$ can only be interpreted when making assumptions on $S(\rho)/\overline{\rho}$. This holds true even under general regression estimation, where there is a trade off between displacement and interval width under NMAR nonresponse.

Even when some theoretical considerations, as laid out in the previous section, would advise to aim at decreasing $S(\rho_X)/\overline{\rho}$, it would make a much stronger case if empirical results support such an endeavour. For this reason, we have evaluated a wide range of survey data.

In section 4.1. we present the various data sets in the evaluation. In sections 4.2 and 4.3, we answer the two research questions.

4.1 Data sets in the evaluation

Our empirical illustration is based on a variety of survey data sets. The following data sets were evaluated:

- Health survey (HS) 2010 for three mode designs: web only, CAPI only , and web → CATI + CAPI.
- Crime Victimization survey (CVS) 2006 for three mode designs: web only, CATI + CAPI, and web → CATI + CAPI
- Survey of Consumer Sentiments (SCS) 2005 and 2009 after different numbers of calls and after contact and full response
- Labour Force Survey (LFS) 2009 and 2010 for two mode designs: CAPI only (2009 and 2010), and CATI+CAPI (2010),
- Survey of Consumer Sentiments (SCS-RISQ) under adaptive survey design and under regular CATI design: this is a pilot study conducted under project RISQ, www.risq-project.eu .
- Short Term Business statistics (STS) 2007 for Manufacturing and Retail after 25, 30 and 60 days of data collection
- LISS-panel for different processing steps: obtaing contact for recruitment interview, participation in recruitment interview, willing to be panel member, registered as a panel member, active in panel after one year, active in panel after two years and active in panel after three years.

We distinguish four types of comparisons: 1) a comparison of different surveys or different survey designs with the same target population, 2) a comparison of the same survey over different time periods, 3) an evaluation of a survey during data collection, and 4) an evaluation of a survey after different processing steps. An evaluation during data collection can be based on time but also on numbers of visits or calls. An evaluation of processing steps refers to obtaining contact, obtaining participation, recruitment for a panel, etc. Table 4.1.1 summarizes the comparisons we have made using the different data sets.

Table 4.1.1: Overview of comparisons.

	Type of comparison			
	Different surveys	*Survey in time*	*Data collection*	*Processing steps*
HS 2010	x			
CVS 2006	x			
HS – CVS	x			
LFS		x		
SCS		x	x	x
SCS-RISQ	x			
STS			x	
LISS				x

Table 4.1.2: Overview of available auxiliary variables.

Data sets	*Auxiliary variables used*
HS 2010	Employment status, etnicity, age, urbanization, type of household and zip code house value
CVS 2006	Etnicity, age, urbanization and type of household
HS – CVS	Etnicity, age, urbanization and type of household
LFS	Employment status, etnicity, age, urbanization, type of household and zip code house value
SCS	Gender, etnicity, income, age, zip code percentage nonnative, urbanization and type of household
SCS-RISQ	Owns car from company, type of business, number and size of jobs
STS	Business size, NACE, VAT of reference month in previous year, VAT of reference month
LISS	Employment status, etnicity, age, urbanization, type of household and zip code house value

Table 4.1.2 contains the selected auxiliary variables in each comparison. The choice of auxiliary variables is somewhat arbitrary; the survey data sets were taken as they are regularly produced by the statistical departments. In the social survey data sets, the potential number of auxiliary variables is much larger, as various additional registry data exist. However, for illustrational purposes the range and number of variables suffices. The LFS, STS and SCS-RISQ data sets contain auxiliary variables that are closely related to the survey topics. Employment status, age and house value are all strong predictors for unemployment. VAT of the previous year is

a strong predictor for the STS turnover in a year. The selected auxiliary variables for the RISQ pilot relate to consumers expectations of economy.

4.2 Do larger R-indicators transfer to other variables?

In this section, we answer the first research question; do worse indicator values on a set of variables X coincide with worse indicator values on other variables?

The approach taken to answer the first research question is the following: For each data set, we sort the available auxiliary variables in a random order and compute the coefficient of variation $S(\rho_X)/\rho$ and the R-indicator R(*X*) for growing models of auxiliary variables. We start by computing the indicators for the first auxiliary variable. We then add the second variable to the model, then the third, and so on, until the full set of auxiliary variables is included. The random order that we have chosen is alphabetical by the labels of the variables.

Clearly, the available set of auxiliary variables may represent only weak proxy variables to the survey target variables. Nonetheless, by employing cumulating models, we simulate NMAR nonresponse on the smaller models. Also, some of the data sets (LFS, STS, SCS) we have selected do have auxiliary variables that are strong predictors of the survey target variables.

We present the patterns for the R-indicator and coefficient of variation in section 4.2.1, except for the RISQ pilot study that is discussed separately in section 4.2.2.

4.2.1 Empirical patterns of the R-indicator and coefficient of variation

We computed the R-indicator and coefficient of variation for all data sets in table 4.1 for growing sets of auxiliary variables, as specified by table 4.2. For the sake of brevity, we show only part of the results. In order to quantify consistency in the patterns we, first, derive a useful measure.

In general, a data set comprises of D different designs, different waves, different time points or different data collection steps, and of M auxiliary variables. In total, we, thus, get DM different values of the indicators. As we compare changes in the preferred designs of the indicator, we effectively have $M-1$ comparisons of the ranking of D indicator values. The changes in rankings can essentially be viewed as pairwise inversions, and the total number of pairwise inversions needed to make all rankings the same is a measure of the consistency of the pattern. For every data set we compute the total number of pairwise inversions and compare it to the expected number of pairwise inversions under the assumption that different auxiliary variables provide completely independent pictures of representative response. It can be shown that the expected number of inversions for a data set is $(M-1)D(D-1)/4$.

It is important to remark that the expected number of inversions holds only when auxiliary variables are uncorrelated and standard errors of the indicators can be ignored. Neither is true of course, the auxiliary variables do show multicollinearity, and some designs have indicator values that are not significantly different and lead to ties. In most cases, however, the correlation between auxiliary variables is very modest. Furthermore, the sampling variation in the indicators tends to masks signals, so that the number of inversions is a conservative measure. For this reason, we view it as a useful measure.

We give four explicit illustrations: one where we compare different surveys (HS – CVS), one where we compare different waves of the same survey (LFS), one where we compare different data collection steps (LISS) and one where we compare different time points (STS).

Figure 4.2.1 shows the patterns for the comparison of the CVS and HS survey data sets under different mode designs. The CVS has three different designs: web only, interviewer modes only (CATI and CAPI) and full mixed-mode (web, CATI and CAPI), with respective response rates equal to 30.2%, 68.9% and 64.7%. The HS has also three designs: web only, CAPI only and full mixed-mode (web, CATI and CAPI). Here, the response rates are: 35.6%, 61.0% and 63.2%. Especially, the differences between the response rates of web only and the other designs are very large. We did not label the different lines in the plot, as the plots function merely to give a visual presentation of the number of pairwise inversions. For the R-indicator, it is hard to observe by sight how large this number is. It turns out to be equal to 14, where 22.5 inversions would be expected. For the coefficient of variation it is much smaller; there are two inversions only. The web only designs have the largest coefficient of variation, to a large extent caused by the low response rates. The web only CVS design has the highest R-indicator in the smallest model, but then deteriorates and remains the weakest design.

Figure 4.2.2 shows the indicators for the LISS panel after 1) contact for the recruitment interview, 2) participation in the contact interview, 3) for those recruited respondents that are willing to be panel member, 4) for those willing respondents that actually register into the panel, 5) after one year, 6) after two years, and 7) three years. The response rates after the seven steps are: 90.9%, 75.0%, 53.5%, 48.5%, 41.2%, 35.7% and 32.9%. So the response rates show a gradual decrease over the various steps. In the LISS data set, the number of inversions is small for both the R-indicator, one inversion, and the coefficient of variation, five reversions, where 52.5 were expected. For this data set the indicators are extremely consistent.

Table 4.2.1 contains the patterns for the LFS, including standard errors. We considered the LFS of 2009 and 2010 in CAPI, and the CATI plus CAPI design in 2010. The response rates are: 61.9%, 59.0% and 54.0%. The mixed-mode design with CATI and CAPI performs always worst for the coefficient of variation. The number of inversions is small for both indicators, one for the R-indicator and two for the coefficient of variation. A number of 7.5 was expected. Two other conclusions can also be drawn from this comparison: R-indicators are often close to each other and within each others confidence intervals, while for the coefficient of variation differences tend to be stronger. Still, both indicators show consistent pictures.

Figure 4.2.1: Coefficient of variation and R-indicator for three CVS and three HS designs for four nested models: ethnicity, ethnicity + age, ethnicity + age + urbanization and ethnicity + age + urbanization + type of household.

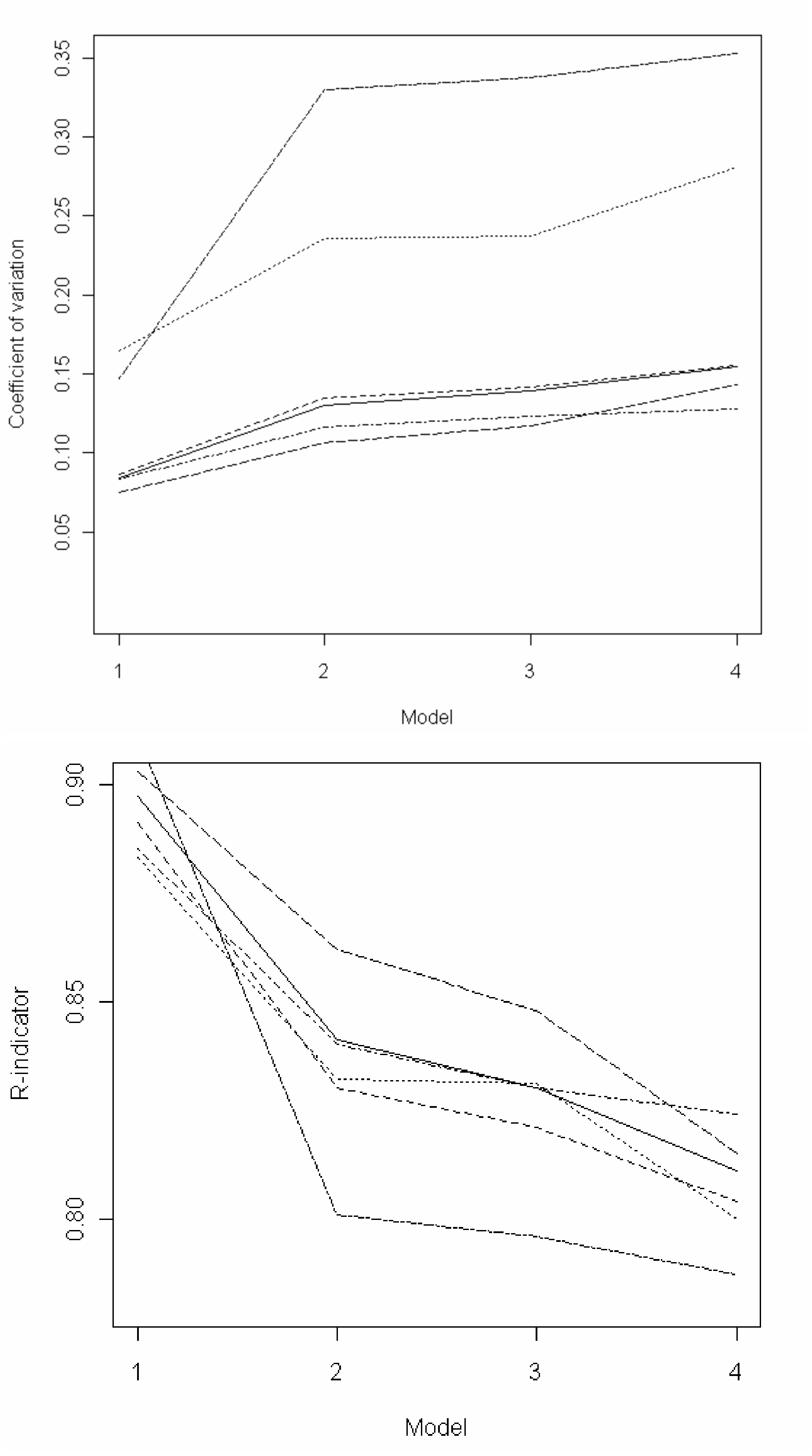
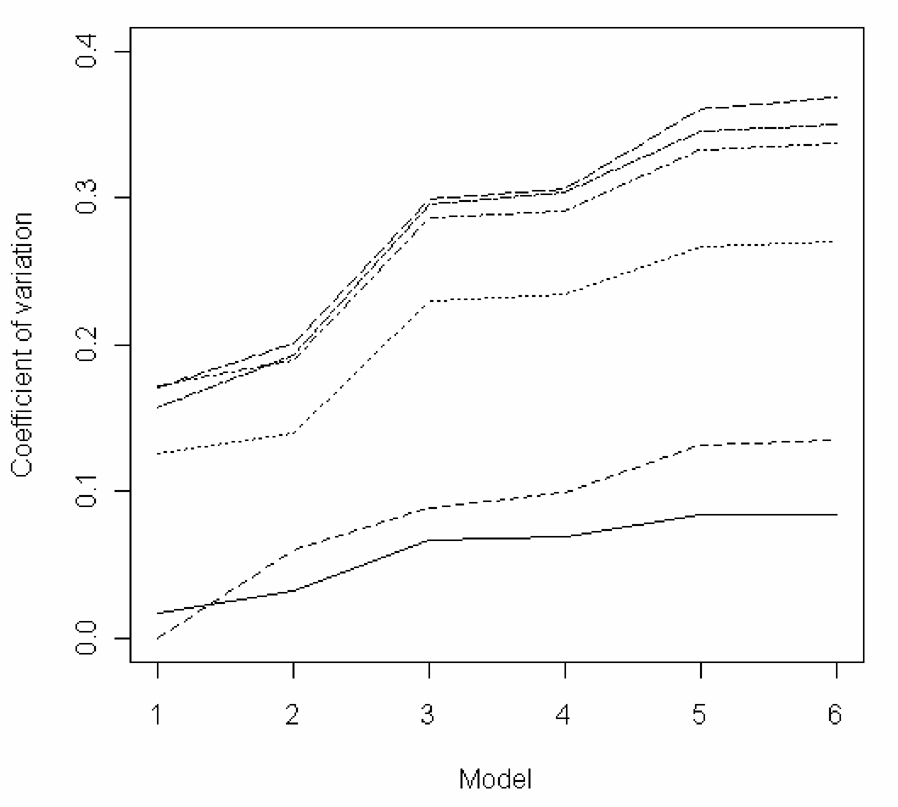


Figure 4.2.2: Coefficient of variation and R-indicator for different process steps in the LISS panel for six nested models: employment status, employment status + ethnicity, employment status + ethnicity + age, employment status + ethnicity + age + urbanization, employment status + ethnicity + age + urbanization + type of household, employment status + ethnicity + age + urbanization + type of household + average house value.



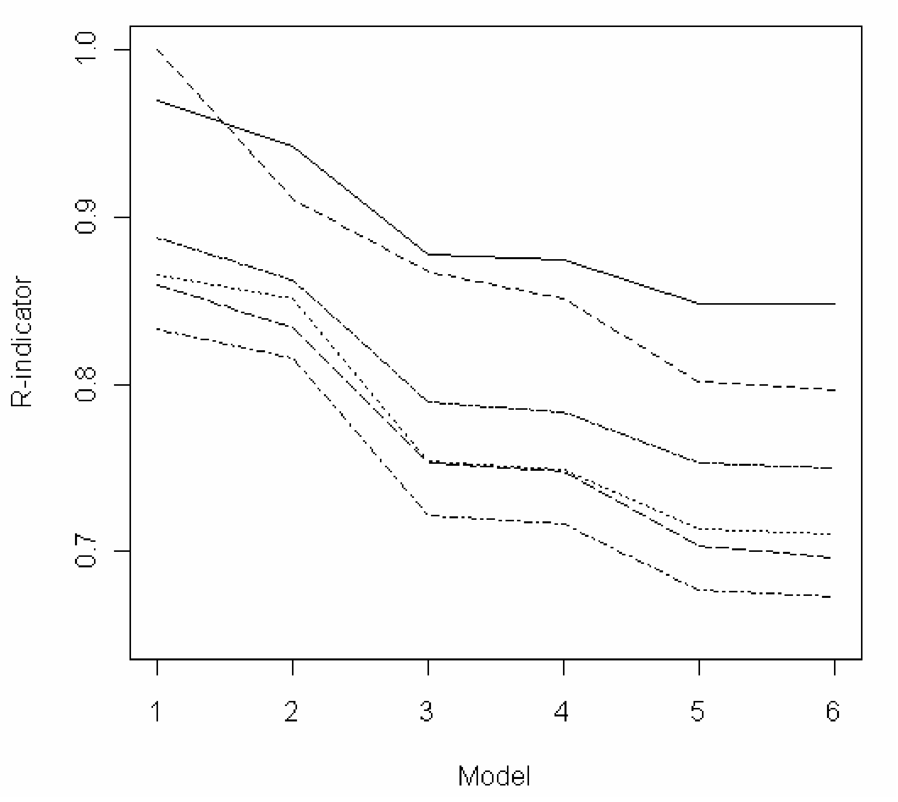


Table 4.2.1: Indicator values for LFS 2009 CAPI, LFS 2010 CAPI and LFS 2010 CATI+CAPI for six nested models. Standard errors are given within brackets.

	Coefficient of variation			*R-indicator*		
Model	*CAPI09*	*CAPI10*	*MM10*	*CAPI09*	*CAPI10*	*MM10*
Employment status	0,043 (0,002)	0,033 (0,003)	0,049 (0,004)	0,947 (0,003)	0,961 (0,004)	0,947 (0,004)
+ ethnicity	0,099 (0,002)	0,103 (0,003)	0,119 (0,004)	0,877 (0,003)	0,878 (0,004)	0,872 (0,004)
+ age	0,100 (0,002)	0,105 (0,003)	0,123 (0,004)	0,876 (0,003)	0,877 (0,004)	0,868 (0,004)
+ %nonnative	0,119 (0,002)	0,114 (0,003)	0,129 (0,004)	0,853 (0,003)	0,866 (0,004)	0,861 (0,004)
+ urbanization	0,120 (0,002)	0,118 (0,003)	0,135 (0,004)	0,851 (0,003)	0,861 (0,004)	0,855 (0,004)
+ household type	0,125 (0,002)	0,124 (0,003)	0,137 (0,004)	0,845 (0,003)	0,854 (0,004)	0,852 (0,004)

Table 4.2.2: Indicator values for STS 2007 Manufacturing after 25, 30 and 60 days for four nested models. SE's for R-indicators and coefficients of variation are stable: 0,003 and 0,002 for all time points, respectively.

	Coefficient of variation			*R-indicator*		
Model	*25 days*	*30 days*	*60 days*	*25 days*	*30 days*	*60 days*
Business size	0,043	0,036	0,026	0,941	0,946	0,957
+ NACE	0,074	0,063	0,053	0,898	0,907	0,911
+ VAT (t-12)	0,077	0,065	0,053	0,894	0,904	0,911
+ VAT (t)	0,078	0,067	0,056	0,892	0,901	0,906

Table 4.2.3: Indicator values for STS 2007 Retail after 25, 30 and 60 days for four nested models. SE's for R-indicators and coefficients of variation are stable: 0,003 and 0,002 for all time points, respectively.

	Coefficient of variation			*R-indicator*		
Model	*25 days*	*30 days*	*60 days*	*25 days*	*30 days*	*60 days*
Business size	0,038	0,045	0,051	0,947	0,931	0,912
+ NACE	0,057	0,063	0,068	0,921	0,904	0,882
+ VAT (t-12)	0,080	0,083	0,080	0,889	0,874	0,862
+ VAT (t)	0,089	0,092	0,090	0,876	0,860	0,845

Tables 4.2.2 and 4.2.3 contain the indicator values for the STS 2007 Manufacturing and Retail, respectively, after different numbers of data collection days. The fourth STS auxiliary variable in table 4.2, VAT of reference month, is a surrogate for the main STS survey variable, total monthly turnover. VAT of the reference month is available only after 18 months. Hence, this data set has auxiliary variables that strongly relate to the main target variable. The auxiliary variables show, however, also more multicollinearity. Especially, reported VAT in the previous year and reported VAT in the current year are related, despite the fact that part of the businesses did not exist in the previous year. The response rates after 25 days, 30

days and 60 days are 69.1%, 74.4% and 83.4% for Manufacturing, and 69.8%, 76.3% and 86.3% for Retail. The number of inversions is zero in all cases, except for the coefficient of variation for the STS Retail that shows one inversion. The expected number of inversions is 4.5.

We did not show results for the SCS and the separate analysis for the HS using six nested models. The SCS is a CATI survey and response was evaluated after one call, three calls, six calls, nine calls, twelve calls and after all calls. The response rate grows from 31.7% after one call to 67.4% after all calls.

Table 4.2.4 summarizes the observed and expected numbers of inversions for all data sets. In all cases, the observed numbers are smaller than the expected numbers and in general the coefficient of variation shows smaller numbers than the R-indicator. In the three studies where indicators are evaluated during data collection (SCS, LISS and STS), the patterns for both indicators are more stable than for the comparisons between surveys, between survey designs and over survey waves.

Table 4.2: Summary of total observed and expected numbers of pairwise inversions for the various data sets.

Dataset	*Number of designs*	*Number of variables*	*Number of pairwise inversions*		
			Expected	*R-indicator*	*Coefficient of variation*
HS	3	6	7.5	3	1
CVS 2006	3	4	4.5	3	1
HS – CVS	6	4	22.5	14	2
LFS	3	6	7.5	1	2
SCS	8	7	84	19	6
LISS	7	6	52.5	1	5
STS-IND	3	4	4.5	0	0
STS-RET	3	4	4.5	0	1

Overall, we conclude that the coefficient of variation shows relatively stable patterns when models are expanded with more auxiliary variables. This means that indeed this indicator functions as a process quality indicator; relatively small models provide signals that are confirmed under bigger models. This holds especially true in comparisons during data collection and after different data collection steps. The R-indicator shows a more volatile picture, but still the empirical evidence suggests a consistent picture.

4.2.2 The SCS adaptive survey design study

In this section, we address the SCS adaptive survey design study. The pilot study was linked to the SCS 2009. For details we refer to Luiten and Schouten (2013). A sample of approximately 6000 addresses was randomly divided over a control and experimental group. The control group received the regular CATI design, while the experimental design was subject to an adaptive survey design. The adaptive survey design distinguished sample units based on age, ethnicity, income, type of household and urbanization and assigned different strategies in order to equalize response rates for the corresponding subgroups. The strategies employed different survey modes (web, paper and CATI), different prioritizations of calls and different interviewer allocations. As a constraint it was demanded that both costs and response rate should be maintained.

It is obvious that an adaptive survey design may be successful in improving the indicator values for the auxiliary variables that form the subgroups in the design. For this reason, we selected four, new auxiliary variables: ownership of a company car, business type of person in household with biggest job, and number and sizes of jobs in household. These variables were not involved in the adaptive survey design, but clearly do have some association to the selected adaptive design variables age, ethnicity, income, type of household and urbanization. But so will any survey target variable. If adaptive survey designs are to be promising extensions of sampling designs, then the indicator values should also be better for variables that were not involved in the adaptation.

Table 4.2.5 provides the coefficient of variation and R-indicator for the control group and experimental group. Again, we added the variables ownership of a company car, business type of person in household with biggest job, and number and sizes of jobs in household one at a time. Except for the smallest model with ownership of a company car, all models lead to more representative response for the experimental group. Testing the one-sided null-hypotheses of a smaller R-indicator and larger coefficient of variation leads to p-values of 6,3% and 5,3%, respectively, i.e. close to the standard level of 5%. We conclude that for the pilot study the adaptive survey design also improved representativity of other variables. These variables are associated with the topics of the SCS. In the next section, we investigate whether this gain is preserved under weighting.

Table 4.2.5: Indicators for RISQ control group and experimental group. Standard errors are given within brackets.

Model	*Coefficient of variation*		*R-indicator*	
	Control	*Experimental*	*Control*	*Experimental*
Car	0,000	0,018	1,000	0,977
	(0,019)	(0,014)	(0,024)	(0,018)
+ business type	0,066	0,034	0,916	0,956
	(0,014)	(0,014)	(0,018)	(0,018)
+ # of jobs	0,087	0,051	0,889	0,934
	(0,014)	(0,015)	(0,018)	(0,019)
+ sizes of jobs	0,096	0,063	0,878	0,918
	(0,014)	(0,015)	(0,018)	(0,019)

4.3 Is nonresponse bias of other variables after adjustment larger when R-indicators are larger?

We now turn to the second research question; Do worse indicator values on a set of variables X coincide with larger remaining nonresponse bias on other variables after calibration on X?

We, again, restrict the analysis to the general regression estimator. We consider three examples: 1) the three mode designs of the HS 2010, 2) the CVS and HS survey, and 3) the RISQ pilot study. For the first two examples we weighted the response for the last auxiliary variables that was added to the models using the other auxiliary variables as weighting variables. In other words, the last variable is treated as the target variable. For the HS survey this is the zip code house value in categories and for the CVS-HS comparison this is the type of household. For the RISQ pilot we performed weighting using exactly the same variables as were used in the set up of the design. The target variables for this study were the four extra auxiliary variables that were linked later on and that were evaluated in section 4.2.2.

For the three HS designs and the CVS-HS comparison we computed sample means, unweighted response means and weighted response means for each category of the artificial target variables. Subsequently, the impact of weighting was measured by taking the Euclidean distance between the category sample means and the (un)weighted category response means. Again, we weighted the variables by cumulating models of auxiliary variables. The order in which variables were added was the same as for the computation of the coefficient of variation and the R-indicators. The only difference being that the last variable was not added but is treated as a target variable. The resulting net distances were compared for the various data sets. They are given in tables 4.3.1 and 4.3.2.

Table 4.3.1: Euclidean distances for the three mode designs in the HS 2010 between sample mean and unweighted response mean, and between sample mean and weighted response mean using cumulating models.

	Web	*CAPI*	*Full mixed-mode*
unweighted	0,0540	0,0235	0,0255
employment	0,0543	0,0238	0,0253
+ ethnicity	0,0436	0,0179	0,0203
+ age	0,0387	0,0161	0,0172
+ urbanization	0,0377	0,0139	0,0165
+ household type	0,0281	0,0104	0,0142

Table 4.3.2: Euclidean distances for the various CVS and HS designs between sample mean and unweighted response mean, and between sample mean and weighted response mean using cumulating models.

	HS			*CVS*		
	CAPI	*Web*	*MM*	*Regular*	*Web*	*MM*
unweighted	0,0394	0,0675	0,0282	0,0440	0,0913	0,0436
ethnicity	0,0410	0,0660	0,0297	0,0447	0,0866	0,0442
+ age	0,0320	0,0614	0,0205	0,0331	0,0638	0,0199
+ urbanization	0,0285	0,0598	0,0191	0,0295	0,0643	0,0177

The results for the HS 2010 in table 4.3.1 tell us that the web only design leads to the largest remaining bias. The other two designs are much closer to each other in terms of remaining bias, but the CAPI design always has a slightly smaller bias. This picture is completely in line with the coefficient of variation that showed the same ranking of the three designs.

The results for the CVS-HS comparison are different. The CVS mixed-mode design starts as the third best design, but becomes the best design when more variables are added. Some rankings of data sets, as they were identified by the coefficient of variation, are also found for the remaining bias, e.g. the HS web design is always better than the CVS web design and the web only designs are always outperformed by the other designs. However, the ranking of the remaining bias for the other designs is not consistent with the coefficient of variation. Figure 4.3.1 presents the Euclidean distances of table 4.3.2.

Figure 4.3.1: Euclidean distances for the various CVS and HS designs between sample mean and unweighted response mean (model 1), and between sample mean and weighted response mean using cumulating models (model 2 to 4).

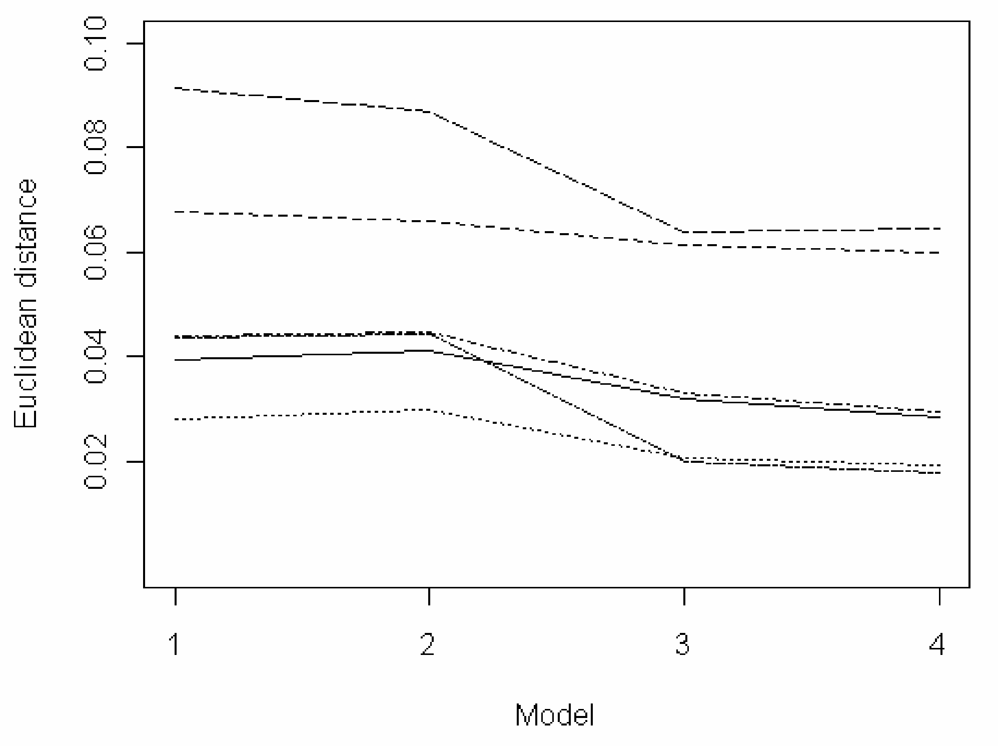


Table 4.3.3: Euclidean distances for the experimental group and control group between sample mean and adjusted response mean.

	Control	*Experimental*
Car	0,0042	0,0042
Business type	0,0175	0,0123
Number of jobs	0,0239	0,0062
Number × size of jobs	0,0187	0,0062

The final comparison is for the RISQ pilot. In section 4.2.2, we showed that the experimental group performed better on auxiliary variables that were not included in the adaptive design than the control group. It is important to validate whether the experimental group remains to perform better, when we weight the response for these variables using the adaptive survey design subgroups. We treated the four auxiliary variables ownership of company car, business type of largest job, and number and sizes of jobs as target variables. Again we computed the Euclidean distances with respect to the sample mean for both the experimental and control group. Table 4.3.3 contains the distances. The results indicate that the remaining nonresponse bias for the control group is always larger than or equal to that of the experimental group.

Taking the results of this section together, we get a mixed picture. For two of the three examples we found that stronger signals of nonrepresentative response lead to more remaining nonresponse bias after nonresponse adjustment. One of the two examples is the adaptive survey design pilot, which is a promising result. However, the third example showed that patterns identified in the indicators do not need to be reproduced necessarily under weighting. More empirical studies are key as the number of target variables we considered in this section was relatively small.

5. Conclusions

We draw the following conclusions:

- There is no strong, theoretical ground to state that stronger signals of nonrepresentative response, lead to more nonrepresentative response and more nonresponse bias on other variables. However, the coefficient of variation and the R-indicator may serve as process quality indicators; employing the intuitively appealing idea that where there is smoke there is fire.
- Overall, we find empirical evidence that patterns of nonrepresentative response, as measured by the coefficient of variation, are persistent in larger models. That means that relatively small models rank different surveys, different designs and different data collection time points similar to larger models. The picture for the R-indicator is more volatile. Here, ranking of different data sets more often changes when models get larger. But still the indicator leads to smaller numbers of changes in ranking of designs, waves, etc than expected.
- The adaptive survey design we investigated was successful in reducing nonrepresentative response on other variables than were used in the differentiation over subgroups.
- In two out of the three cases we investigated, the remaining nonresponse bias after calibration followed exactly the same pattern as the coefficient of variation. When there are stronger traces of nonrepresentative response on a set of auxiliary variables, then the bias after adjustment on these variables remains larger.

There are clearly some limitations to the study we have performed. The surveys were all conducted by Statistics Netherlands and with Dutch businesses and households as the target populations, and we restricted ourselves to auxiliary variables that are available at Statistics Netherlands. Although in most cases we selected auxiliary variables that show little correlation, the variables do show multicollinearity. Clearly, if we apply indicators to copies of the same auxiliary variables, the values will be consistent. Alternatively, we could have performed a factor analysis and could have used the strongest factors as input to the evaluation of the R-indicator and coefficient of variation. Effectively, this would have meant that the number of auxiliary variables in the comparisons would have been reduced. Since this is a cumbersome operation and since sampling variation obscures the signals shown by the indicators, we decided to stick to the original auxiliary variables. Nevertheless, given the range of surveys and the explanatory power of some of the auxiliary variables, we feel that the study does provide generalizable results.

To return to the main question in the title: Does balancing of data collection treatments on auxiliary variables lead to less nonresponse bias on survey target variables, even after adjustment using the same auxiliary variables? The empirical results show that it pays off to employ designs that lead to more representative response on auxiliary variables. This result suggests that, from a nonresponse bias point of view, it is better to pick a design that has a lower value on the coefficient of variation, and less strongly, that has a higher value on the R-indicator. This provides an incentive to use adaptive survey designs, but one may also use a uniform design that has proven to be more representative. Of course, a total survey error view would make this choice much more complex as standard errors and measurement errors need to be accounted for as well.

References

Andridge, R.R., Little, R.J.A. (2011), Proxy pattern-mixture analysis for survey nonresponse, Journal of Official Statistics, 27 (2), 153 – 180.

Beaumont, J.F. , Haziza, D. (2011), A theoretical framework for adaptive collection designs, Paper presented at 5th International Total Survey Error Workshop, June 21 – 23, Quebec, Canada.

Bethlehem, J. (2012), Using response probabilities for assessing representativity, Discussion paper 201212, Statistics Netherlands, available at www.cbs.nl .

Bethlehem, J.G., Cobben, F., Schouten, J.G. (2011), Handbook of Nonresponse in Household Surveys, Handbook, Wiley Series in Survey Methodology, USA.

Calinescu, M., Schouten, B., Bhulai, S. (2012), Adaptive survey designs for the Labour Force Survey that minimize nonresponse and measurement risk, Discussion paper 201224, Statistics Netherlands, available at www.cbs.nl .

Cobben, F., Schouten, B. (2008), An empirical validation of R-indicators, Discussion paper 08006, Statistics Netherlands, available at www.cbs.nl.

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, Vol. 169, Issue 3, pp. 439 – 457.

Groves, R.M., Peytcheva, E. (2008), The impact of nonresponse rates on nonresponse bias, Public Opinion Quarterly, 72, 167 – 189.

Luiten, A., Schouten, B. (2013), Adaptive fieldwork design to increase representative household survey respons. A pilot study in the Survey of Consumer Satisfaction, Journal of Royal Statistical Society, Series A, 176 (1).

Olsen, K., Groves, R.M., (2012), An examination of within-person variation in response propensity over the data collection period, Journal of Official Statistics, 28 (1), 29 – 51.

Särndal. C.E. (2011), The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation, Journal of Official Statistics, 27 (1), 1 – 21.

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 (2), 131 – 144.

Schouten, B. (2007), A selection strategy for weighting variables under a not-missing-at-random assumption, Journal of Official Statistics, 23 (1), 1 – 19.

Schouten, J.G., Bethlehem, J., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N., Skinner, C. (2012), Indicators for evaluating, comparing, monitoring and improving survey response, To appear in International Statistical Review.

Schouten, J.G., Calinescu, M., Luiten, A. (2011), Optimizing quality of response through adaptive survey designs, Discussion paper 201118, CBS, Den Haag.

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

Schouten, J.G., Shlomo, N., Skinner, C. (2011), Indicators for monitoring and improving representativeness of response, Journal of Official Statistics, 27(2), 231 – 253.

Shlomo, N., Skinner, C., Schouten, J.G. (2012), Estimation of an indicator of the representativeness of survey response, Journal of Statistical Planning and Inference, 142, 201 – 211.

Stoop, I. (2005), Surveying nonrespondents, Field Methods 16, 23 – 54.

Wagner, J. (2008), Adaptive survey design to reduce nonresponse bias, PhD thesis, University of Michigan, USA.

Wagner, J. (2010), The fraction of missing information as a tool for monitoring the quality of survey data, Public Opinion Quarterly, 74 (2), 223 – 243.