Measuring Survey Quality Through Representativeness Indicators Using Sample and Population Based Information

Chris Skinner[1], Natalie Shlomo[1], Barry Schouten[2], Li-Chun Zhang[3], Jelke Bethlehem[2]

[1]Southampton Statistical Sciences Research Institute, University of Southampton,
e-mail: C.J,Skinner@soton.ac.uk, N.Shlomo@soton.ac.uk,
[2] Central Bureau of Statistics, Netherlands,
e-mail: jg.schouten@cbs.nl; j.bethlehem@cbs.nl
[3] Statistics Norway, e-mail: mailto:xxx@xxx.xx li.chun.zhang@ssb.no

Abstract

The RISQ (Representativity Indicators for Survey Quality) project, funded by the European 7th Framework Programme, is a joint effort of the NSI's of Norway, The Netherlands and Slovenia, and the Universities of Leuven and Southampton to develop quality indicators for survey response. The response rate alone is insufficient to measure the potential impact of non-response. These Representativity Indicators (*R-indicators*) are developed to be used as tools at different stages of the data collection process: monitoring field strategies, targeting field resources and comparing strategies for increasing response rates. The auxiliary information for these indicators depend on prior information about respondents and non-respondents and paradata that becomes available during fieldwork. In many countries, prior auxiliary information for non-respondents may be limited. However, marginal distributions at the population level are often available through registers and population estimates. In this talk, we present and compare possible *R-indicators* and evaluate their sampling properties. We also propose *R- indicators* that are based entirely on population totals and the respondent set, and compare their properties to sample-based *R-indicators* through a simulation study.

Keywords: response rate, non-response bias, quality indicators

1. Introduction

One of the largest source of non-random error in surveys is due to non-response. Research has shown, see e.g. Bethlehem (1988) that the non-response bias of estimates is determined by two factors:

- The amount to which respondents and non-respondents differ, on average, with respect to the target variable. The more they differ, the larger the bias will be. This is sometimes called the contrast between response and non-response;

- The amount of response in the survey. The response rate sets a bound to the maximal impact of non-response; the lower the response rate of the survey, the larger the potential impact of bias.

To assess the effects of non-response on the quality of estimators, one needs to measure both the response rate and the contrast between response and non-response.

The response rate alone is an insufficient quality indicator to measure the potential impact of non-response. Statistical indicators of the contrast should measure the degree to which respondents and non-respondents differ from each other.

The project RISQ (Representativity Indicators for Survey Quality), funded by the European 7th Framework Programme, is a joint effort of the NSI's of Norway, the Netherlands and Slovenia, and the Universities of Leuven and Southampton to develop quality indicators for survey response. These indicators measure the degree to which the group of respondents of a survey resembles the complete sample. When this is the case, the response is called representative. In survey practice, response rates are almost always computed. However, an indication of the contrast is seldom given explicitly since information is needed on characteristics of households or enterprises that did not respond to the survey. Nonetheless, when information is available that is auxiliary to the survey one can indirectly measure part of the contrast. It is the objective of the RISQ project to translate auxiliary information to Representativity Indicators, to develop these quality indicators, to explore their characteristics and to show how to implement and use them in a practical survey.

In this paper, we describe briefly two Representativity Indicators (henceforth called *R-indicators*). More discussion and theory of these *R-indicators* is provided in Shlomo, Skinner, Schouten, Bethlehem and Zhang (2008).

We focus on two *R-indicators*:

1. a measure based upon the variance of estimated response probabilities, as discussed in the papers by Cobben and Schouten (2005, 2007) and Schouten et al. (2008).
2. a related measure proposed by Särndal and Lundström (2008), in the context of selecting auxiliary variables for weighting adjustment.

The estimation of the *R-indicators* is very much dependent on the nature of available auxiliary information. We initially assume that auxiliary information is available at the sample level. In many cases, however, auxiliary information may only be available in aggregated form at the population level. We introduce theory for estimating *R-indicators* when aggregated auxiliary information is known and compare them to the sample level indicators in a simulation study.

Section 2 contains a formulation of the theoretical framework for the *R-indicators*. The two *R-indicators* are defined formally at the population level in Section 3 and their estimation is discussed in Section 4. Section 5 contains a simulation study on the properties of sample based and population based estimated *R-indicators* and we conclude in Section 6 with future work.

2. Theoretical Framework for R-Indicators

2.1 General notation and nature of available information

We suppose that a sample survey is undertaken, where a sample s is selected from a finite population U. The sizes of s and U are denoted n and N, respectively. The units in U are labelled $i = 1, 2, \ldots, N$. The sample is assumed to be drawn by a probability sampling design $p(.)$, where the sample s is selected with probability $p(s)$. The first order inclusion probability of unit i is denoted π_i and $d_i = \pi_i^{-1}$ is the design weight. In some cases, we shall assume simple random sampling without replacement.

We suppose that the survey is subject to unit nonresponse. The set of responding units is denoted . Thus, we have $r \subset s \subset U$. We denote summation over the respondents, sample and population by Σ_r, Σ_s and Σ_U, respectively. We let R_i be the response indicator variable so that $R_i = 1$ if unit i responds and $R_i = 0$, otherwise. Hence, $r = \{i \in s; R_i = 1\}$. We shall suppose that the typical target of inference is a population mean $\bar{Y} = N^{-1}\sum_U y_i$ of a survey variable, taking value y_i for unit i.

We suppose that the data available for estimation purposes consists first of the values $\{y_i; i \in r\}$ of the survey variable, observed only for respondents. Secondly, we suppose that information is available on the values of $x_i = (x_{1,i}, x_{2,i}, \ldots, x_{K,i})^T$, a vector of auxiliary variables. We shall usually suppose each $x_{k,i}$ is a binary indicator variable, where x_i represents one or more categorical variables, since this will be the case in the applications we consider, but our presentation allows for general $x_{k,i}$ values. We assume that values of x_i are observed for all respondents. For the majority of this document we shall also assume that x_i is known for all sample units, i.e. for both respondents and non-respondents. We refer to this as *sample-based auxiliary information*. This is a natural assumption if, for example, the variables making up x_i are available on a register. However, in many countries and survey settings the availability of auxiliary information on non-respondents may be very limited, e.g. because of the absence of a register. In such circumstances, *aggregate population-based auxiliary information* may be available. This might take the form of a (finite) population total and/or mean and/or covariance matrix of x_i.

2.2 Response propensities

We define the *response propensity* as a conditional expectation of the response indicator variable R_i given the values of specified variables and survey conditions (Little, 1986, 1988): $\rho_X(x_i) = E(R_i \mid x_i)$, where the vector of auxiliary variables is defined as in section 2.1. For simplicity, we shall usually write $\rho_i = \rho_X(x_i)$ and hence denote the response propensity just by ρ_i. A detailed discussion of the notion of response propensities and their properties is presented in Shlomo, et al. (2008). In this discussion it was argued that it is desirable to select the auxiliary variables constituting x_i in such a way that the *missing at random*, denoted MAR (Little and Rubin, 2002) holds as closely as possible and that our definition of $\rho_i = \rho_X(x_i)$ relates to a specific choice of auxiliary variables x_i. A different choice would generally result in a different ρ_i. We note also that we define the response propensity conditional on the survey conditions that apply when the data are collected in order to be able to compare the representativeness of different surveys.

2.3 Non-response models

In order to estimate *R-indicators*, we shall first estimate the response propensities, where these are defined as $\rho_i = E(R_i \mid x_i)$ as discussed in the previous section. In this paper, we shall use parametric modelling assumptions about how ρ_i depends on x_i. In Shlomo, et. al, 2008, we also discuss non-parametric models and some implications of the complexity of the model for the variability of the ρ_i.

A general class of models representing the dependence of ρ_i on x_i may be expressed in the form:

$$\mathrm{g}(\rho_i) = x_i'\beta, \tag{1}$$

where g(.) is a specified link function, β is a vector of unknown parameters to be estimated, and x_i may involve the transformation of the original auxiliary variables (e.g. by including interaction terms) for the purpose of model specification. A standard choice of link function is the logit function, leading to the logistic regression model:

$$\log[\rho_i/(1-\rho_i)] = x_i'\beta \tag{2}$$

Another link function with similar behaviour to the logit is the probit function. We shall also consider the use of the identity link function, which gives the 'linear probability model':

$$\rho_i = x_i'\beta, \tag{3}$$

since this will offer particular simplifications in the case of population-based auxiliary information.

Särndal and Lundström (2008) consider the reciprocal link function, which gives:

$$\rho_i^{-1} = x_i'\lambda, \tag{4}$$

and they refer to ρ_i^{-1} as the *influence* and denote it ϕ_i. They assume that the vector x_i is defined in such a way that there exists a constant vector c such that $c'x_i = 1$ for all $i \in U$. This restriction will in most practical situations be met and is effectively equivalent to assuming that a constant intercept term is included in the auxiliary information.

Särndal and Lundström (2008) view (4) as a hypothetical model which will not hold in practice and they instead focus on a finite population approximation to this model. This approximation is obtained by first defining a value λ_U of λ which achieves the best fit of model (4) in the finite population. For mathematical convenience, they define the fit as the weighted sum of squared differences $\sum_U \rho_i(\rho_i^{-1} - x_i'\lambda)^2$ and this is minimised when:

$$\lambda_U = (\textstyle\sum_U \rho_i x_i x_i')^{-1} \sum_U x_i, \tag{5}$$

provided x_i is defined so that the inverted matrix in (5) is non-singular. This implies that a finite population approximation to ϕ_i is given by:

$$\phi_{Ui} = x_i '(\sum_U \rho_i x_i x_i ')^{-1} \sum_U x_i \,. \qquad (6)$$

We refer to these quantities as the *approximate influences*.

3. Definition of R-Indicators at Population Level

Let $\boldsymbol{\rho} = (\rho_1, \rho_2, ..., \rho_N)'$ denote the vector of response propensities in the population. Following Schouten et al. (2008), the representativity of the response mechanism may be measured by the variation between the ρ_i and in particular by the standard deviation of the response propensities given by:

$$S(\boldsymbol{\rho}) = \sqrt{\frac{1}{N-1} \sum_U (\rho_i - \bar{\rho}_U)^2} \,, \qquad (7)$$

where $\bar{\rho}_U = \sum_U \rho_i / N$. It may be shown that: $S(\boldsymbol{\rho}) \leq \sqrt{\bar{\rho}_U (1-\bar{\rho}_U)} \leq \frac{1}{2}$. Hence, transforming $S(\boldsymbol{\rho})$ to:

$$R(\boldsymbol{\rho}) = 1 - 2S(\boldsymbol{\rho}) \qquad (8)$$

ensures that $0 \leq R(\boldsymbol{\rho}) \leq 1$ and, as discussed by Schouten et al. (2008), $R(\boldsymbol{\rho})$ defines an *R-indicator* which takes values on the interval [0,1] with the value 1 indicating the most representative response, where the ρ_i display no variation, and the value 0 indicating the least representative response, where the ρ_i display maximum variation.

Särndal and Lundström (2008) define the following *R-indicator*:

$$Q^2(\boldsymbol{\rho}) = [\sum_U \rho_i]^{-1} [\sum_U \rho_i (\phi_{Ui} - \bar{\phi}_{\rho U})^2] \qquad (9)$$

where $\bar{\phi}_{\rho U}$ is the ρ_i- weighted mean of the ϕ_{Ui} given by

$$\bar{\phi}_{\rho U} = (\sum_U \rho_i)^{-1} (\sum_U \rho_i \phi_{Ui}) \,. \qquad (10)$$

This quantity is a weighted variance of the approximate influences. We may expect its magnitude to be inversely related to the magnitude of $R(\boldsymbol{\rho})$. Thus, in very rough terms, we expect $R(\boldsymbol{\rho})$ to decrease and $Q^2(\boldsymbol{\rho})$ to increase as the variability of the ρ_i increases.

4. Estimation

4.1 Estimation of population totals from sample and respondent data

In the following sections, we shall use the fact that, for a given variable z_i, the design-weighted sample total $\sum_s d_i z_i$ is a design-unbiased estimator of the population total $\sum_U z_i$ and the design-weighted respondent total $\sum_r d_i z_i$ is an unbiased estimator of the ρ_i - weighted population total $\sum_U \rho_i z_i$.

4.2 Estimation of non-response models

The estimation of the models in section 2.3 depends on the nature of the auxiliary information. For sample-based auxiliary information, the model in (1) can be estimated from the data on respondents and non-respondents by maximum pseudo likelihood (Skinner, 1989) i.e. the parameter vector β in this model may be estimated by the value $\hat{\beta}$, which solves:

$$\sum_s d_i[R_i - g^{-1}(x_i'\beta)]x_i = 0 \tag{11}$$

where $g^{-1}(.)$ is the inverse of the link function. One reason for using the design weights here is because the objective is to estimate an *R-indicator* which provides a descriptive measure for the population.

The linear probability model in (3) can be estimated in closed form by ordinary least squares or by weighted least squares, where the weights are the design weights.

For the reciprocal link function model in (4), Särndal and Lundström (2008) approximate the model by $\rho_i^{-1} \approx x_i'\lambda_U$, where λ_U is defined in (5) and estimate this approximate model by estimating λ_U from the sample data by:

$$\hat{\lambda}_U = (\sum_r d_i x_i x_i')^{-1} \sum_s d_i x_i \ . \tag{12}$$

Note that this estimation follows the strategy in section 4.1 and that it also assumes sample-based auxiliary information.

4.3 Estimation of response propensities

For the generalized linear model in (1), the usual estimator of the response propensity ρ_i is:

$$\hat{\rho}_i = g^{-1}(x_i'\hat{\beta}), \tag{13}$$

where $\hat{\beta}$ is the estimator of β obtained as discussed in the previous section.

In the case of the linear probability model in (3), if β is estimated by (design-) weighted least squares, the implied estimator of ρ_i is given by:

$$\hat{\rho}_i^{OLS} = x_i'(\sum_s d_i x_i x_i')^{-1} \sum_s d_i x_i R_i , \tag{14}$$

which may also be expressed as:

$$\hat{\rho}_i^{OLS} = x_i'(\sum_s d_i x_i x_i')^{-1} \sum_r d_i x_i \ . \tag{15}$$

In the approach of Särndal and Lundström (2008) with the reciprocal link function, ϕ_i is estimated by:

$$\hat{\phi}_i = x_i'\hat{\lambda}_U , \tag{16}$$

where $\hat{\lambda}_U$ is defined in (12), so that:

$$\hat{\phi}_i = x_i'(\sum_r d_i x_i x_i')^{-1}\sum_s d_i x_i \tag{17}$$

and the resulting estimator of ρ_i is $\hat{\phi}_i^{-1}$.

For the logit or probit link function, the estimator $\hat{\rho}_i$ obtained from (13) must fall in the feasible interval [0,1]. This is not necessarily the case for either the estimator based on the linear probability model in (14) or the estimator $\hat{\phi}_i^{-1}$ of ρ_i based on (17).

In the case of population-based auxiliary information, we note that $\sum_s d_i x_i$ and $\sum_s d_i x_i x_i'$ are unbiased for $\sum_U x_i$ and $\sum_U x_i x_i'$, respectively and that in large samples we may expect that $\sum_s d_i x_i \approx \sum_U x_i$ and $\sum_s d_i x_i x_i' \approx \sum_U x_i x_i'$. It follows from (15) that we may approximate $\hat{\rho}_i^{OLS}$ by:

$$\tilde{\rho}_i^{OLS} = x_i'(\sum_U x_i x_i')^{-1}\sum_r d_i x_i , \tag{18}$$

and from (17) that we may approximate $\hat{\phi}_i$ by:

$$\tilde{\phi}_i = x_i'(\sum_r d_i x_i x_i')^{-1}\sum_U x_i . \tag{19}$$

Expressions (18) and (19) provide estimators of the response propensity for respondents when x_i is not available for individual non-respondents but aggregate population-level information is available. The estimator in (18) requires knowledge of the population sums of squares and cross-products $\sum_U x_i x_i'$ of the elements of x_i. If this is unknown, we can estimate $\sum_s d_i x_i x_i'$ in (15) by rewriting:

$$\sum_s d_i x_i x_i' = \sum_s d_i (x_i - \bar{x}_s)(x_i - \bar{x}_s)' + N\bar{x}_s\bar{x}_s' , \tag{20}$$

where $\bar{x}_s = \sum_s x_i / n$. $\bar{x}_s$ can be replaced by $\bar{X}_U$. The covariance matrix: $S_{xx} = N^{-1}\sum_s d_i (x_i - \bar{x}_s)(x_i - \bar{x}_s)'$ may be replaced by the observed covariance matrix:

$$S_{xxr} = (\sum_s d_i R_i)^{-1}\sum_s d_i R_i (x_i - \bar{x}_r)(x_i - \bar{x}_r)' , \tag{21}$$

where $\bar{x}_r = (\sum_s R_i x_i)/(\sum_s R_i)$. The estimator in (19) only requires knowledge of the population total of each of the elements of x_i.

4.4 Estimation of R-indicators

Let $\hat{\rho}_i$ be an estimator of the response probability ρ_i, as discussed in the previous section. Assuming that sample-based auxiliary information is available, $\hat{\rho}_i$ may be computed for each $i \in s$. An estimator of $\bar{\rho}_U$ is then given by

$$\hat{\bar{\rho}}_U = (\sum_s d_i \hat{\rho}_i)/N. \tag{22}$$

Alternatively, we could replace N in the denominator by $\sum_s d_i$. We estimate the *R-indicator* $R(\boldsymbol{\rho})$ by:

$$\hat{R}(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_s d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2} \tag{23}$$

Again, we could replace $N-1$ in this expression by $\sum_s d_i$.

If $\hat{\rho}_i$ is only available for respondents ($i \in$), as in the case of aggregated population-level auxiliary information described at the end of the previous section, a possible estimator of $R(\boldsymbol{\rho})$ is:

$$\hat{R}_r(\boldsymbol{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_r d_i \hat{\rho}_i^{-1}(\hat{\rho}_i - \hat{\bar{\rho}}_r)^2}$$

where $\hat{\bar{\rho}}_r = (\sum_r d_i)/N$. This corrects for non-response bias using $\hat{\rho}_i^{-1}$- weighting. The validity of this correction depends on the validity of the estimates $\hat{\rho}_i$.

We now turn to the estimation of $Q^2(\boldsymbol{\rho})$ in (9). Särndal and Lundström (2008) propose the following estimator:

$$q^2 = [\sum_r d_i]^{-1}[\sum_r d_i(\hat{\phi}_i - \bar{\phi}_r)^2], \tag{24}$$

where $\hat{\phi}_i$ is defined in (17) and $\bar{\phi}_r = (\sum_r d_i \hat{\phi}_i)/(\sum_r d_i)$. They note that in fact $\bar{\phi}$ can be re-expressed as $\bar{\phi}_r = (\sum_s d_i)/(\sum_r d_i)$.

The estimator in (24) is based only upon respondent data. However, $\hat{\phi}_i$ in (17) does depend on $\sum_s d_i x_i$ which may not be available in the case of aggregated population level information. In such cases, we may replace $\hat{\phi}_i$ in (24) by $\tilde{\phi}_i$ from (19). Shlomo, et. al. (2008) describe bias adjustments for $\hat{R}(\boldsymbol{\rho})$ and q^2 as well as confidence intervals for the *R-indicators*.

5. Simulation Study of the Properties of the Estimated R-indicators

5.1 Design of Simulation Study

In this section, we carry out simulation studies to assess the sampling properties of the two *R-indicators*: $\hat{R}(\boldsymbol{\rho})$ defined in (23) and q^2 defined in (24). The simulation study is based on samples drawn from Census data from the 1995 20% Israel Census Sample containing 753,711 individuals aged 15 and over in 322,411 households. The sample design is similar to a standard household survey carried out at NSIs. The sample units are households and all persons over the age of 15 in the sampled households are interviewed. Typically a proxy questionnaire is used and therefore there is no individual non-response within the household. In this study, we assume that every household has an equal probability to be included in the sample.

We carry out a two-step design to define response probabilities in the Census data. In the first step, we determine probabilities of response based on explanatory variables that typically lead to differential non- response based on our experiences of working with survey data collection. A response indicator was then generated for each unit in the Census from these probabilities. In the second step, we fit a logistic regression model, as in (2), to these Census data and thus determined a 'true' response propensity for each unit as predicted by this model fitted to the population. The dependent variable of the model is the response indicator and the independent variables of the model the explanatory variables used in the first step. This two-step design ensures that we have a known model generating the response propensities and therefore we can assess model misspecification besides the sampling properties of the indicators.

The explanatory variables used to generate the response probabilities are Type of locality (3 categories), number of persons in household (1,2,3,4,5,6+), children in the household indicator (yes, no). Samples of size n are drawn from the Census population of size N at different sampling fractions 1:50, 1:100, and 1:200. For each sample drawn, a sample response indicator is generated from the 'true' population response probability. The overall response rate is 82%. Response propensities and *R-indicators* are then estimated from the sample.

5.2 Results

Throughout this simulation we examine the sampling properties of the *R-indicators* as well as the impact of model misspecification on their properties. Because smaller sample sizes generally lead to the selection of a less complex model, we shall consider that misspecification is represented by a simpler model.

In Table 1, we examine samples drawn at different sampling rates, estimate response propensities for each sample and calculate the measure $\hat{R}(\boldsymbol{\rho})$ (defined in (23)) for both sample and population based auxiliary variables. We present results for both the true model and a less complex model.

Table 1 shows that the sample based estimator $\hat{R}(\boldsymbol{\rho})$ increases as the sample size increases. If the specified model is correct, there is some downward bias and this tends to increase as the sample size increases. This is as expected. Sampling error tends to lead to overestimation of the variability of the estimated response propensities and this leads to underestimation of the *R-indicator*. The degree of underestimation is, however, small. Under the less complex model, estimation of response propensities results in a 'smoothing' of the propensities and hence an

Table 1: Simulation Means of $\hat{R}(\boldsymbol{\rho})$ for Sample and Population Based Auxiliary Variables for 500 Samples ('True' R-Indicator = 0.8780)

Sampling Fraction	'True' Model (Number of Persons, Locality Type, Child Indicator) $\hat{R}(\boldsymbol{\rho})$			Less Complex Model (Number of Persons) $\hat{R}(\boldsymbol{\rho})$		
	Sample Based	Population Based		Sample Based	Population Based	
		Covariance Matrix Known	Covariance Matrix Unknown		Covariance Matrix Known	Covariance Matrix Unknown
1:200 (n=1,612)	0.8714	0.8251	0.8290	0.8791	0.8477	0.8456
1:100 (n=3,224)	0.8741	0.8531	0.8524	0.8801	0.8684	0.8652
1:50 (n=6,448)	0.8761	0.8683	0.8652	0.8810	0.8729	0.8680

overestimation in $\hat{R}(\boldsymbol{\rho})$. Results of $\hat{R}(\boldsymbol{\rho})$ with the bias correction (not shown here) have produced a more stabilized indicator across different sample sizes under the correct model, but for the less complex model with overestimation of $\hat{R}(\boldsymbol{\rho})$, the bias correction can exacerbate the overestimation. Comparing $\hat{R}(\boldsymbol{\rho})$ when using sample based auxiliary variables and population based auxiliary variables, Table 1 shows that the *R-indicator* is underestimated when using population based auxiliary variables. The variation of response propensities is larger than the variation under sample based auxiliary variables. Since we used simple random samples and assumptions of *MAR*, there seems to be little difference between the two population level indicators of $\hat{R}(\boldsymbol{\rho})$ based on a known population covariance matrix and an estimated covariance matrix from the response set.

In Table 2, we examine the properties of q^2 based on the variance of the estimated response influences. For this indicator, we expect low values to reflect good quality and small non-response bias. We compare the full set of explanatory variables in the model used in this simulation to a less complex model as before.

Table 2: Simulation Means of q^2 for Sample and Population Based Auxiliary Variables for 500 Samples

Sampling Fraction	'True' Model (Number of Persons, Locality Type, Child Indicator) $Q^2(\boldsymbol{\rho}) = 0.0087$		Less Complex Model (Number of Persons) $Q^2(\boldsymbol{\rho}) = 0.0082$	
	Sample Based	Population Based	Sample Based	Population Based
1:200 (n=1,612)	0.0102	0.0171	0.0092	0.0132
1:100 (n=3,224)	0.0096	0.0129	0.0088	0.0109
1:50 (n=6,448)	0.0089	0.0107	0.0083	0.0095

Results from Table 2 show the decrease in q^2 as the sample sizes increase. Shlomo, et al. (2008) discuss a bias correction for this indicator as well. The less complex model produces 'smoother' estimated influences and hence less variation. This is expected since the indicator was developed to assess the effectiveness of auxiliary variables on bias reduction. Using population based auxiliary variables have increased the variation of the estimated influences.

6. Future Work

From these initial simulations, we continue to develop theory for using population based auxiliary information. In particular, we will investigate an iterative OLS algorithm for the estimation of $\hat{\rho}_i$ and $\hat{\beta}$, and use propensity weighted totals in the estimation process. The final $\hat{\rho}_i$ can also be used to calculate propensity weighted totals for estimating q^2. We will also examine using a GLM model instead of the OLS model, although we have found no departures from the expected interval of [0,1] using the OLS model.

Other areas of work for the *R-indicators* are to apply bias corrections and consider the variances of the estimators, potential variance estimators and confidence intervals. In addition, we can calculate maximum bounds on the non-response bias through the use of these *R-indicators*. We will focus on more practical aspects of implementation, such as the choice of auxiliary variables and the specification of the model and compare the quality of surveys in various respects: subsequent surveys on the same topic within a country, different surveys within a country, and surveys on the same topic in different countries. The results of these comparisons will help to build a better European Statistical System.

Acknowledgements: We thank the members of the RISQ project: Katja Rutar from Statistični Urad Republike Slovenije, Geert Loosveldt and Koen Beullens from Katholieke Universiteit, Leuven, Oyvin Kleven from Statistisk Sentralbyrå, Norway, Fannie Cobben from Centraal Bureau voor de Statistiek, Netherlands, Ana Marujo and Gabi Durrant from the University of Southampton, UK for their valuable input.

References

Bethlehem, J.G. (1988) Reduction of nonresponse bias through regression estimation, *Journal of Official Statistics*, 4, 251 – 260

Cobben, F. and Schouten, B. (2005) Bias measures for evaluating and facilitating flexible fieldwork strategies, Paper presented at *16th International Workshop on Household Survey Nonresponse*, August 28-31, Tällberg, Sweden

Cobben, F. and Schouten, B. (2007), An empirical validation of R-indicators, *Discussion paper*, CBS, Voorburg

Little, R.J.A. (1986) Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157

Little, R.J.A. (1988) Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-301

Little, R.J.A. and Rubin,D.B. (2002) *Statistical Analysis with Missing Data*, Hoboken, New Jersey: 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

Schouten, B., Cobben, F. and Bethlehem, J. (2008) Indicators for the Representativeness of Survey Response. *Survey Methodology* (to appear)

Shlomo, N., Skinner, C.J., Schouten, B., Bethlehem, J., Zhang, L. (2008) Statistical Properties of R-indicators, Work Package 3, Deliverable 3.1, RISQ Project, 7th Framework Programme (FP7) of the European Union

Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In Skinner, C.J., Holt, D. and Smith, T.M.F. eds. *Analysis of Complex Surveys*, Chichester: Wiley