RISQ

Representativity Indicators for Survey Quality

How to use R-indicators?

Work package 4
Deliverable 3

Barry Schouten, Mattijn Morren & Jelke Bethlehem
Centraal Bureau voor de Statistiek, Netherlands

Natalie Shlomo & Chris Skinner
University of Southampton, United Kingdom

17 March 2009

The RISQ Project is financed by the 7th Framework Programme (FP7) of the European Union.
Cooperation Programma, Socio-economic Sciences and the Humanities, Provision for Underlying Statistics

How to use R-indicators?

1. Introduction

The RISQ (Representativity Indicators for Survey Quality) project is funded from the 7th EU Framework Programme (FP7) and runs from March 2008 to June 2010. RISQ was set up in order to fill the gap of indicators that measure the representativeness of the response to survey and register requests. We call these indicators Representativity indicators or R-indicators. The main objectives of the project are to elaborate and develop R-indicators, to explore the statistical characteristics of these indicators, and to show how to implement them in a practical data collection environment. With these indicators the project attempts to support the comparison of the quality of different surveys and registers, and to facilitate the efficient allocation of data collection resources.

R-indicators may be used in the following settings:

- To compare the response to different surveys that share the same target population, e.g. households or businesses
- To compare the response to a survey longitudinally, e.g. monthly, quarterly or annually
- To compare the response to a survey during data collection, e.g. after various days, weeks or months of fieldwork

The first part of the RISQ project (March 2008 – June 2009) is devoted to the development of indicators. The second part concentrates on the application of these indicators in data collection (January 2009 – June 2010). The first RISQ paper (deliverable 2.1, Shlomo et al 2009) describes the statistical properties of two potential R-indicators; the indicator R proposed by Schouten, Cobben, Bethlehem (2009) and the variable selection measure q^2 proposed by Särndal and Lundström (2008). In this first RISQ paper we assumed that R-indicators are computed relative to a fixed set of auxiliary variables that are identified beforehand. The paper did not discuss questions like: what auxiliary variables should we include, should we fit a model to describe the non-response missing-data-mechanism, if so what models should we use, and what is the role of the sample size. These important questions were raised in the RISQ description of work (Bethlehem and Schouten 2008) and were assigned to a separate work package, WP 4. The same issues were also addressed in the review of deliverable 2.1 (deliverable 15, Biemer 2008). The current paper tries to give answers to these questions and gives a set of recommendations for the use of R-indicators in practical survey settings.

In Schouten and Cobben (2007), a response is called strongly representative when all individual response probabilities are equal. They call a response weakly representative with respect to some categorical variable X when the average response probabilities over the classes of X are equal. A response probability is not a straightforward concept, as becomes clear from the discussion in Shlomo et al (2009), but even without a rigorous definition it is intuitively clear that strong representativity is not a measurable characteristic of a response. Weak representativity, however, can be made mathematically more rigorous, and R-indicators may measure deviations from weak representativity with respect to a vector of available X's. Two characteristics of any indicator that attempts to measure representativeness of response then become clear: No statement about the representativeness is possible without any information that is auxiliary to the survey itself and there is a strong dependence on the size of the survey sample. Non-response leads to missing data. Without any knowledge of the missing units, there is no way to draw conclusions about their similarity to units that are observed. However, even with knowledge about the missing units, the strength of conclusions may be severely weakened when sample sizes are small. A sample size of, say, ten units simply does not allow for any statistically significant inferences about population characteristics. Both aspects are important to bear in mind and they are the focal points of this paper.

In this paper we investigate the dependence on the selected set of auxiliary variables, we compare models with a fixed set of variables to models where we employ variable selection, and we look into the relation to the type of model or more specifically the link function between covariates and response. In these analyses the dependence on the sample size follows from standard errors and confidence intervals, as R-indicators themselves are random variables. We restrict ourselves to the bias-adjusted indicator R of Shlomo et al (2009) and leave Särndal and Lundström's q^2 out of the scope of this paper. In Shlomo et al (2009) we found that both indicators lead to similar conclusions about the representativeness of response, although they stem from different objectives. Furthermore, we believe that the issues raised in this paper can be translated quite easily to any R-indicator.

The analyses are supported by data sets from the University of Southampton and Statistics Netherlands, which were selected and documented in RISQ work package 2. The documentations of the data sets are available on www.r-indicator.eu .

In section 2, we start with the background from previous papers (Schouten, Cobben, Bethlehem 2008) and Shlomo et al (2009). We give a brief description of our data sets in section 3. Next, in section 4, we relate R-indicators to the choice of auxiliary variables. In section 5, we investigate the role of different models for non-response. Finally, in section 6, we give recommendations for the use of R-indicators.

2. R-indicators revisited

In this section we start with notation and definitions from Shlomo et al (2009). Next, we discuss the relation between R-indicators and nonresponse bias. Finally, we repeat the statistical properties that were derived in Shlomo et al (2009).

2.1 Notation and definitions

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 or inclusion weight.

We suppose that the survey is subject to unit non-response. 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 vector of survey items, 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 item vector, 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 assume that values of x_i are observed for all respondents. In 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.

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. However, we shall postpone considering this possibility in detail until the forthcoming Deliverable 3.2 of RISQ work package 3. However, in this setting the same issues as are discussed in this paper play an important role, but the dependence on sample size is even more pronounced. Any statement about the representativeness of response needs to be adjusted first for the survey sampling design $p(.)$, which will generally lead to a loss of precision.

We assume a superpopulation model with independent and identically distributed random vectors $(R_i, X_i, Y_i)_{1 \le i \le N}$. In this paper we make again the assumption that the response behaviour of different units is independent. This may not be a valid assumption in some survey settings, but it can be assumed in the examples that we investigate. Furthermore, we do not explicitly include the sampling design in the joint probability but we assume that non-response is stable, i.e. R_i exists regardless of the unit being sampled or not, and that the sampling design is not confounded with the response mechanism.
We only observe (R_i, X_i, Y_i) for respondents in the sample and (R_i, X_i) for non-respondents in the sample.

Let $\rho_X(x) = E(R = 1 \mid X = x) = P(R = 1 \mid X = x)$ be the propensity to respond for a population unit carrying value x on the auxiliary vector. Thus, the response propensity is the conditional distribution of R on X. Furthermore, let $\rho_X = (\rho_X(x_1), \rho_X(x_2), \ldots, \rho_X(x_N))^T$ denote the vector of response propensities in the population. Since in this paper we will make distinctions between different sets of auxiliary variables, we will not omit the subscript like in Shlomo et al (2009). For each set of auxiliary variables, we, thus have different response propensities.

Literature often postulates the existence of individual response probabilities without a reference to a specific auxiliary vector, see the discussion in Shlomo et al (2009). Given that we only have one replication of response for each sample unit, such a probability cannot be measured and is a hypothetical construct. One may conjecture the existence of an auxiliary vector $\aleph$ that fully explains differences in response behaviour between individuals in the population for each possible survey conducted within that population. The corresponding response propensities $\rho_{\aleph}$ may then be interpreted as response probabilities; individuals with the same scores on this 'super' auxiliary vector show the same response behaviour. Schouten and Cobben (2007) give two definitions for representative response. We will translate them to the current setting.

Definition (strong): A response to a survey is strongly representative when $\rho_{\aleph}$ is a constant function.

Definition (weak): A response to a survey is weakly representative with respect to X when $\rho_X(.)$ is a constant function.

Strong representativity is a hypothetical property that cannot be determined in any practical survey setting. However, weak representativity can be evaluated. We will return to these concepts in section 4.2, when we discuss the role and types of auxiliary information.

R-indicators measure deviations from representativity. We give the following definition

$$R(\rho_X) = 1 - 2S(\rho_X)\,, \tag{2.1}$$

where $S(\rho_X)$ is the population standard deviation of ρ_X defined by

$$S(\rho_X) = \sqrt{\frac{1}{N-1}\sum\nolimits_U (\rho_X(x_i) - \bar{\rho}_X)^2}\,, \tag{2.2}$$

where

$$\bar{\rho}_X = \frac{1}{N}\sum\nolimits_U \rho_X(x_i)\,. \tag{2.3}$$

It is not difficult to show that $S(\rho_X) \leq \frac{1}{2}$ for any X, and, as a consequence, $R(\rho_X) \in [0,1]$.

Since also the marginal distribution $\rho_X(.)$ is unknown, we have to resort to estimators $\hat{\rho}_X(.)$. We then get

$$\hat{R}(\hat{\rho}_X) = 1 - 2\hat{S}(\hat{\rho}_X)\,, \tag{2.4}$$

with

$$\hat{S}(\hat{\rho}_X) = \sqrt{\frac{1}{N-1}\sum\nolimits_s d_i(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_X)^2}\,, \tag{2.5}$$

and

$$\hat{\bar{\rho}}_X = \frac{1}{N}\sum\nolimits_s d_i \hat{\rho}_X(x_i)\,. \tag{2.6}$$

It is most straightforward to apply a general linear model for the estimators $\hat{\rho}_X(.)$. Logistic regression, probit regression or linear regression are special cases of such models.

Note that $R(\rho_X) \geq R(\rho_Z)$ when variable Z is nested in X. The more refined the "resolution", the more variation is observed. As a consequence the R-indicator for the ideal auxiliary vector $\aleph$ will also be smaller $R(\rho_X) \geq R(\rho_\aleph)$.

2.2 Maximal absolute bias

The value of R-indicators only have a meaning when they can be related to nonresponse bias, as ultimately it is the bias that we are interested in. Schouten, Cobben and Bethlehem (2008) relate the R-indicator (2.2) to the bias of the Horvitz-Thompson estimator

$$\hat{\bar{y}}_{HT} = \frac{1}{N}\sum\nolimits_s d_i Y_i \,. \tag{2.7}$$

The bias B of the Horvitz-Thompson estimator can be bounded from above by

$$|B(\hat{\bar{y}}_{HT})| = \frac{|Cov(Y,\rho_Y)|}{E(\rho_Y)} \leq \frac{S(Y)S(\rho_Y)}{E(\rho_Y)} = \frac{S(Y)(1-R(\rho_Y))}{2E(\rho_Y)} \leq \frac{S(Y)(1-R(\rho_\aleph))}{2E(\rho_\aleph)}\,, \tag{2.8}$$

where $Cov(Y,\rho_Y)$ is the covariance (vector) between the survey items and the response propensities for those items. The second bound in (2.8) is again hypothetical when the ideal auxiliary vector $\aleph$ would be available.

In surveys with many survey items, it may be valuable to look at the standardized bias

$$\frac{|B(\hat{\bar{y}}_{HT})|}{S(Y)} = \frac{|Cov(Y,\rho_Y)|}{E(\rho_Y)S(Y)} \leq \frac{S(\rho_Y)}{E(\rho_Y)} \leq \frac{1-R(\rho_\aleph)}{2E(\rho_\aleph)}\,, \tag{2.9}$$

leading to an upper bound that holds for any survey item.

We now return to the original objectives of the RISQ, namely that we want to have a measure that enables comparison of the representativeness of response in different surveys or the same survey in time. In such a setting we are interested in the general representativeness of a survey, i.e. not the representativeness with respect to a single set of survey items. Also, we can investigate representativeness with respect to known auxiliary information only. We propose to use as a surrogate for (2.9)

$$B_m(\rho_X) = \frac{1-R(\rho_X)}{\bar{\rho}_X} = \sqrt{\frac{1}{N-1}\sum\nolimits_U \left(\frac{\rho_X(x_i)}{\bar{\rho}_X}-1\right)^2} \leq 1-\bar{\rho}_X\,. \tag{2.10}$$

Again the upper bound in (2.10) cannot be computed, but it may be replaced by

$$\hat{B}_m(\hat{\rho}_X) = \frac{1-\hat{R}(\hat{\rho}_X)}{\hat{\bar{\rho}}_X} = \sqrt{\frac{1}{N-1}\sum\nolimits_s d_i \left(\frac{\hat{\rho}_X(x_i}{\hat{\bar{\rho}}_X}-1\right)^2} \leq 1-\hat{\bar{\rho}}_X\,, \tag{2.11}$$

with $\hat{\bar{\rho}}_X$ as in (2.6). For any choice of X, it holds that the standardized maximal absolute bias is smaller than the non-response rate of the survey.

It is important to remark that $B_m(\rho_{\aleph}) \geq B_m(\rho_X)$, i.e. the (standardized) maximal absolute bias of the HT-estimator is underestimated by (2.11). Also, we do not know whether $B_m(\rho_Y)$ is smaller, equal or bigger than $B_m(\rho_X)$. This means we should always handle (2.11) with care, but it can be used in trade-offs between precision and costs.

2.3 Statistical properties

In Shlomo et al (2009) it is shown that $\hat{R}(\hat{\rho}_X)$ is a biased but consistent estimator of $R(\rho_X)$.

Suppose a link function h is used in a general linear model for the estimation of the response propensities $\rho_X(.)$. Hence, $h(x_i^T\beta)$ is used as a predictor for $\rho_X(x_i)$ with β a vector that needs to be estimated. Let $\hat{\beta}$ be the estimator and ∇h be the gradient, i.e. the vector with first order derivatives with respect to β.

Shlomo et al (2009) propose the following bias-adjusted estimator for $R(\rho_X)$ when the sampling design is a simple random sample without replacement

$$\hat{R}_B(\hat{\rho}_X) = 1 - 2\sqrt{(1 + \frac{1}{n} - \frac{1}{N})\hat{S}^2(\hat{\rho}_X) - \frac{1}{n}\sum_{i \in s} z_i^T \left[\sum_{j \in s} z_j x_j^T\right]^{-1} z_i}\ , \qquad (2.12)$$

with $z_i = \nabla h(x_i^T\hat{\beta})x_i$.

For stratified simple random samples without replacement the following estimator may be used

$$\hat{R}_B(\hat{\rho}_X) = 1 - 2\sqrt{\hat{S}^2(\hat{\rho}_X) + \sum_{h=1}^{H} \frac{N_h^2}{N^2}(\frac{1}{n_h} - \frac{1}{N_h})\hat{S}_h^2(\hat{\rho}_X) - \frac{1}{N}\sum_{i \in s} \frac{N_{h(i)}}{n_{h(i)}} z_i^T \left[\sum_{j \in s} z_j x_j^T\right]^{-1} z_i} \qquad (2.13)$$

where $h = 1,2,\ldots,H$ denote the strata, n_h is the (fixed) stratum sample size, N_h is the population stratum size, $h(i)$ is the stratum to which unit i belongs, and

$$\hat{S}_h^2(\hat{\rho}_X) = \frac{1}{n_h - 1}\sum_{s_h} (\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,h})^2\ ,$$

$$\hat{\bar{\rho}}_{X,h} = \frac{1}{n_h}\sum_{s_h} \hat{\rho}_X(x_i)\ ,$$

with s_h the sampled units in stratum h.
Note that as the sample size n or the stratum samples sizes n_h tend to infinity, that (2.12) and (2.13) reduce to (2.5). For other sampling designs, the bias-adjusted estimators have different forms. We will not discuss them here, but refer to Shlomo et al (2009) for general formulas for the bias-adjusted estimators.

Using either (2.12) or (2.13), we define the bias-adjusted estimator for the maximal absolute bias by

$$\hat{B}_{m,B}(\hat{\rho}_X) = \frac{1-\hat{R}_B(\hat{\rho}_X)}{\hat{\bar{\rho}}_X}, \tag{2.14}$$

Shlomo et al (2009) propose a linearization variance estimator for the variance $V(\hat{R}(\hat{\rho}_X))$ of the unadjusted estimator $\hat{R}(\hat{\rho}_X)$. Under simple random sampling without replacement it has the form

$$\hat{V}(\hat{R}(\hat{\rho}_X)) = \frac{1}{\hat{S}^2(\hat{\rho}_X)}\left(4\frac{1}{n^2}A^T\Sigma A + 2\frac{1}{n^2}tr(B\Sigma B\Sigma) + \frac{1}{n^2}\left(1-\frac{n}{N}\right)C\right), \tag{2.15}$$

and

$$A = \sum_{i\in s}(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_X)(z_i - \frac{1}{N}\sum_{j\in s} z_j) \tag{2.16}$$

$$B = \sum_{i\in s}(z_i - \frac{1}{N}\sum_{j\in s} z_j)(z_i - \frac{1}{N}\sum_{j\in s} z_j)^T \tag{2.17}$$

$$C = \sum_{i\in s}\left((\hat{\rho}_{X,i} - \hat{\bar{\rho}}_X)^2 - \frac{1}{n}\sum_{j\in s}(\hat{\rho}_{X,j} - \hat{\bar{\rho}}_X)^2\right)^2 \tag{2.18}$$

$$\Sigma = \left[\sum_s z_i^T x_i\right]^{-1} \tag{2.19}$$

For stratified simple random sampling without replacement, (2.15) can be rewritten in a fashion similar to (2.13). In the following we will always use the bias-adjusted estimators and omit the subscript “B”.

It is important to remark that (2.13) provides an approximation to the variance of the unadjusted R-indicator. The variance of the bias-adjusted estimator may be different. In this paper we, therefore, also resort to resampling methods for the estimation of confidence intervals (e.g. Wolter, 2007). In those cases we recompute $\hat{R}_B(\hat{\rho}_X)$ for M bootstrap samples $m = 1,2,\ldots,M$ and form a $100(1-\alpha)\%$ confidence interval estimate by ordering the estimates for the different bootstrap replicates and define the confidence interval in terms of the $\alpha/2$ and $1-\alpha/2$ quantiles. In this paper we use $M = 1000$ and $\alpha = 0.05$, i.e. we omit the smallest 25 estimates and the 25 biggest estimates.

3. Data sets

We base our findings on both real and simulated data sets. The real data sets are four surveys collected by Statistics Netherlands and one survey collected by the Office of National Statistics; in total three social surveys and two business surveys.
The selected social surveys are the 2005 Dutch Consumer Satisfaction Survey (CSS), the 2005 Dutch Health Survey (HS), and the 2001 UK Labour Force Survey (LFS). The business surveys are the 2007 Dutch Short Term Statistics for retail and industry (STS). Table 3.1 gives some characteristics for each of the household surveys surveys. The business surveys are described in table 3.2.

In table 3.2, for the business surveys response rates are given for 15, 30, 45 and 60 days of fieldwork. After 30 days STS needs to provide data for monthly statistics. The CSS has a three stage sampling design in which municipalities are primary sampling units, addresses are

secondary sampling units and persons belonging to the household core are tertiary sampling units. The selection of one person in the household core is not randomized but is left to the household. We ignore the impact on the design weights and treat them as equal.
Detailed documentation can be found in RISQ (2008) available through the RISQ website, www.r-indicator.eu .

Table 3.1: Description of household surveys.

CSS 2005	*HS 2005*	*LFS 2001*
n=17,908	n=15,411	n=3,253
Response=66,9%	Response=67,3%	Response=82,8%
Persons belonging to household core	Persons > 4 years	Persons >15 years
Three stage design (municipality, address, person)	Two stage design (municipality, person)	Systematic sampling of addresses within geographical regions
Equal design weights address	Equal design weights person	Equal design weights address
Fieldwork 10 days	Fieldwork 30 days	Fieldwork 7+7+2 days spread over 13 weeks
CATI[1]	CAPI[2]	CAPI

Table 3.2: Description of business surveys.

STS retail 2007	*STS industry 2007*
n=93,799 Response=49,5% (15d) Response=78,0% (30d) Response=85,8% (45d) Response=88,2% (60d)	n=64,413 Response=48,8% (15d) Response=78,7% (30d) Response=85,7% (45d) Response=88,3% (60d)
All businesses retail	All businesses industry
Stratified design on size class and business type	Stratified design on size class and business type
Unequal design weights	Unequal design weights
Fieldwork 90 days	Fieldwork 90 days
Paper + web	Paper + web

The two Dutch social surveys were linked to the same set of auxiliary variables from frame data, registers and administrative data. The same was done for the business surveys. The available set of auxiliary variables for the UK social survey is slightly different, but similar. Table 3.3 gives an overview of the available auxiliary variables. All auxiliary variables are categorical. For CSS the auxiliary variables are aggregated to the household core.

In the analysis we employ one simulated data set, denoted by SIM1, based on the 1995 Israel Census Sample of individuals aged 15 and over (N=753,711).
In the simulation population propensities were calculated using a 2 step process:

[1] CATI = Computer Assisted Telephone Interviewing
[2] CAPI = Computer Assisted Personal Interviewing

1. Probabilities of response were defined with the variables number of persons, child indicator, income, age, education, locality type, education, ethnicity, gender and marital status. Based on these probabilities, a response indicator was generated.
2. Using the response indicator as the response variable, we fit a logistic regression model on the population using the above explanatory variables. The predictions from this model serve as the 'true' response propensities for our simulation.

The 'true' R indicator based on the population propensities is 87,67%.

Table 3.3: Auxiliary variables. For CSS auxiliary variables all individual variables are aggregated to the household core.

CSS and HS 2005	Gender (for CSS mix is additional category) Age (for CSS age is averaged) Marital status Urbanization Average value of houses on postal code Yes or no a paid job (for CSS at least one job in household core) Type of household Ethnicity (for CSS mix is additional category)
LFS 2001	Region x Urbanization Age Gender Long-term illness indicator Tenure Marital status Economic activity Accommodation Ethnicity
STS 2007	Business size class (based on number of employees) Business subtype VAT 2006 as collected by Tax Board

4. The choice of auxiliary variables

In this section we investigate the role of the auxiliary variables. Intuitively it is clear that they play an essential role in measuring the representativeness of a response. We, first, start with a discussion on response probabilities and response propensities in section 4.1. Next in section 4.2, we elaborate on types of auxiliary information. Then, in section 4.3 we investigate the impact of the size of the set of auxiliary variables. Here, we do, however, fix models and do not select variables. In section 4.4, we briefly look at the role of interaction effects. Finally, in section 4.5 we move to variable selection.

4.1 Response probabilities and response propensities

In section 2.1 we briefly discussed the existence of a 'super' auxiliary vector $\aleph$ with corresponding response propensities $\rho_{\aleph}$. This vector $\aleph$ would contain all auxiliary information necessary to predict the response behaviour of individuals in any survey setting, i.e. $\rho_i = \rho_{\aleph}(\aleph_i)$ may be interpreted as the response probability of unit i. In practice we will only have a subset X and we will have to resort to response propensities ρ_X. It is, therefore, imperative that we first make clear why we need auxiliary variables, before we start investigating the role of auxiliary variables.

How could we interpret the underlying response propensity $\rho_\aleph$? We may adopt a frequentist approach and regard the response probability as the long run proportion of successes in obtaining response. But what would be the long run in the setting of a survey. Or better what is the numerator and the denominator, what variation is left. The numerator is obvious, that would be the number of successes. The denominator is less obvious, as it consists of everything that may change after the start of the fieldwork and until the end of the fieldwork. Changes may occur for the sample unit and its environment, e.g. the mood, mental or physical state of the person or household in combination with the weather, it could be vacations of a person, his job status during the fieldwork, or the state of the dwelling or neighbourhood. It may also be changes in the survey organisation, e.g. the choice of an interviewer, the timing of contact by interviewers within the rules of the contact strategy, the mood of interviewers.
What does not change during the fieldwork is the data collection strategy. By data collection strategy we mean the interview mode, the contact and refusal conversion strategy, the use of pre-notification letters or incentives, the group of interviewers that are employed by the survey organisation, the allocation of interviewers, the name of the survey, and so on. Hence, the data collection strategy is a set of rules that is set beforehand, and any different strategy leads to a different response probability.
What also does not change during fieldwork is any characteristic the sample unit had at a time point prior to the start of the fieldwork. In practice such characteristics usually have a demographic or socio-economic nature, but they may also consist of observations in a preceding survey or recruitment interview.
For the interpretation of a response probability it is not necessary that we are able to replicate responses for the same population unit in practice. Clearly, there is no practical experiment, like an interview and re-interview, where we can assume that the replications follow a binomial distribution with the response probability as a parameter. However, there are many settings for which the same is true, like medicine treatment studies or dose response studies.

There are three causes that limit inference about the response probabilities $\rho_\aleph$ and necessitate a shift to response propensities ρ_X. First and most important, we obviously have only one replication of the response of a sample unit. That means we cannot estimate response probabilities directly and individually. Even if we would be able to replicate response of an individual, i.e. be able to obtain independent, identically distributed responses, then we would still experience some limitation as there would a (small) loss of precision. Second, we have a finite sample. And third, we only have a limited amount of auxiliary information about the sample and population units, i.e. we do not know $\aleph$.
If we could ask each sample unit in a finite sample for its response probability, then we would be able to unbiasedly estimate R-indicators $R(\rho_\aleph)$. We simply observe the individual response probabilities. There is some loss of precision as we do not observe the full population.
We may expect that there exist auxiliary variables for which the response probabilities are smooth functions and that reduce the prediction error, like is done in regression estimation. If we would let the amount of auxiliary information grow to $\aleph$, then this error would disappear. If we would let the sample grow infinitely large, then we would be able to estimate any response propensity function ρ_X. In such settings we do not need to select variables, since all interactions terms would significantly contribute.

Clearly, since we do not know $\aleph$ in practical survey settings but only have a subset of auxiliary variables, we must accept that estimated R-indicators are biased; even if we would either let the sample size or the number of auxiliary variables grow large. This has important consequences. Since different surveys obviously have different $R(\rho_{\aleph})$'s, the estimators $\hat{R}(\hat{\rho}_X)$ have different biases with respect to $R(\rho_{\aleph})$.

Figure 4.1: Population R-indicators and their estimators for two different surveys that share auxiliary vector X. Z_1 and Z_2 are subsets of X obtained through a variable selection method.

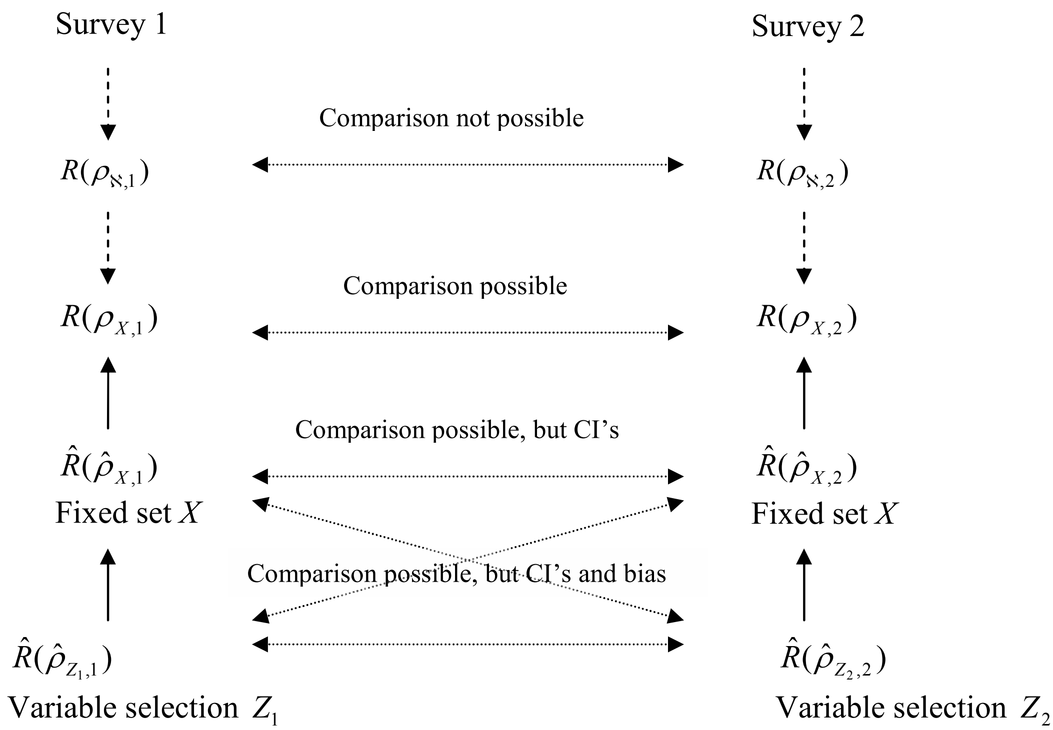

Figure 4.1 gives a schematic overview of R-indicators and estimators for two different surveys with underlying R-indicators $R(\rho_{\aleph,1})$ and $R(\rho_{\aleph,2})$. Survey data does not allow for a comparison of these two quantities. For this reason we move to the estimable quantities $R(\rho_{X,1})$ and $R(\rho_{X,2})$, based on response propensities for an auxiliary vector X that is linked to both survey samples, and their unbiased estimators $\hat{R}(\hat{\rho}_{X,1})$ and $\hat{R}(\hat{\rho}_{X,2})$. The comparison between the two resulting estimates is only to some extent hampered by a loss of precision which we can represent by confidence intervals. When a variable selection method is applied with X as input, then the two surveys may lead to two different selections Z_1 and Z_2, respectively. The corresponding estimators $\hat{R}(\hat{\rho}_{Z_1,1})$ and $\hat{R}(\hat{\rho}_{Z_2,2})$ are biased for $R(\rho_{X,1})$ and $R(\rho_{X,2})$. As a consequence the comparison suffers both from bias and loss of precision.

Given the previous discussion, we ask ourselves the following questions in the subsequent subsections:

- Do R-indicators depend strongly on the number of auxiliary variables, and if so how to choose variables?

- Is it sufficient to use main effect models or should we also include interaction terms?
- Is variable selection useful as a method to minimize the mean square error with respect to R-indicator based on the full set of auxiliary variables?

4.2 Types of auxiliary variables

By auxiliary variables we mean variables that are observed for all sample units outside the survey questionnaire. We distinguish four types of auxiliary variables:

1. Auxiliary variables that become available from a source other than the survey or the survey data collection, and that are constant during the fieldwork.
2. Auxiliary variables that are collected by the interviewer or survey organisation during the fieldwork but that are constant during the fieldwork.
3. Auxiliary variables collected by the interviewer or survey organisation that have changed since the start of the fieldwork but are independent on the interaction between the survey organisation and the respondent.
4. Auxiliary variables collected by the interviewer or survey organisation that depend on the interaction between the survey organisation and the respondent.

Examples of the different types of variables are:

1. Demographic or socio-economic characteristics available through frame data, censuses, registrations or administrative data.
2. Interviewer observations about the dwelling or the neighbourhood, the answers given by the sample unit in a preceding survey or interview.
3. Interviewer observations about the mood of the respondent or the propensity to respond, the result of a contact attempt.
4. The type of non-response, the number of contact attempts or refusals, the mode in which the interview is conducted in a mixed-mode survey.

Type 2 to 4 information are often called paradata or process data.

Type 1 and 2 auxiliary information are the same except for the scope of the information. Type 1 information will usually be available for the whole of the target population, while type 2 information is typically only available for the sample. Särndal and Lundström (2002), for instance, also distinguish these two types in calibration estimators.

Type 3 auxiliary information describes characteristics of sample units that change the underlying response probability, and may even lead to a zero or one response probability. If an interviewer observes that a sample unit was very busy or not in the mood, or was not at home during six subsequent days, then the response probability is different. Such information cannot be used to model that response probability. Another example would be the age or experience of the interviewer that was allocated to the household, person or business in case allocation of interviewers is subject to randomness. The average age or experience of interviewers of the survey organisation in the region of the sample unit would, however, be type 1 information.

The same is true for type 4 auxiliary information which is not constant either but is also the consequence of the fieldwork and the interaction with the sample unit. It does not make sense to model the response probability with variables that causally follow the attempt to get a response. The most extreme is the response indicator itself. This variable cannot be used as an explanatory variable for response probabilities. However, the same holds for example for the number of contact attempts.

Type 3 and 4 variables cannot be used as independent or explanatory variables for response, they can, however, be used as dependent variables in models. For instance, one may distinguish different model equations for non-contact and refusal. RISQ work package 6 is devoted to the use of R-indicators for different types of non-response.

4.3 Fixed variable sets

In this section we use logistic regression to model the response propensities. In all cases we include an intercept and model parameters are fitted based on a fixed model, i.e. we do not add or remove variables using a selection criterion. In section 5 we look at the type of model.

For each of the surveys we compute R-indicators for a reduced set of auxiliary variables and for the full set of available auxiliary variables given by table 3.3. For CSS and HS the reduced set is gender, age, marital status and urbanization. For both STS surveys the reduced set is business size class and business type. Hence, for the business surveys we omit the VAT information.

The logistic regression models for the social surveys contain mostly main effects, except for age and marital status in CSS and HS and region and type of locality in LFS. These variables are crossed and some categories are combined as the variables are strongly multi-collinear. For the two business surveys VAT is crossed with business size class (number of employees) for the same reason. We exclude all other interactions between the auxiliary variables, although some of them may be multi-collinear too. In section 4.4 we briefly address models were interaction terms are added. In the following we will refer to the two variable sets as the small and full set. Note that the small set is a subset of the full set, and, as a consequence the R-indicator values are always higher.

Table 4.1: Bias-adjusted R-indicators and maximal bias for CSS and HS using small and full sets of auxiliary variables. 95% confidence intervals are estimated for the R-indicators.

Survey		*Small set*	*Full set*
HS 2005 full sample (n=15411)	R	83,2%	80,8%
	95% CI	81,9% - 84,7%	79,4% - 82,3%
	B	12,5%	14,3%
CSS 2005 full sample (n=17908)	R	83,3%	82,1%
	95% CI	81,8% - 84,8%	80,7% - 83,4%
	B	12,5%	13,4%
HS 2005 1:4 sample (n=3852)	R	82,5%	78,2%
	95% CI	79,7% - 85,4%	75,4% - 81,1%
	B	13,0%	16,2%
CSS 2005 1:4 sample (n=4477)	R	81,9%	82,7%
	95% CI	79,1% - 84,6%	80,2% - 85,3%
	B	13,6%	12,9%
HS 2005 1:9 sample (n=1712)	R	82,6%	77,3%
	95% CI	78,4% - 86,6%	73,5% - 80,9%
	B	12,9%	16,8%
CSS 2005 1:9 sample (n=1989)	R	79,3%	75,8%
	95% CI	75,4% - 83,3%	72,1% - 79,8%
	B	15,5%	18,1%

Tables 4.1 and 4.2 contain the bias-adjusted R-indicators and maximal absolute biases for, respectively, the social and business surveys. For the business survey we compute R-indicators after 15, 30, 45 and 60 days of fieldwork. Especially, the result after 30 days is interesting as the STS surveys provide the input to monthly statistics that need to be published after one month of fieldwork.

Tables 4.1 and 4.2 show that, as expected, R-indicators are smaller and biases bigger as the set of auxiliary variables is bigger. However, patterns can be quite different.
For HS and CSS the small and full sets have R-indicators that are reasonably close, given the sizes of the confidence intervals. For the CSS the two R-indicators attain values that lay in each others confidence intervals. The R-indicators for CSS and HS are similar with respect to the small set. When information on jobs, average house value, household type and ethnic background is added, the picture changes and the CSS scores better. In all cases the maximal absolute biases show that non-response bias may run up to 10 to 15% of the variation in the survey item. Recall that for CSS and HS, the non-response rates were 33,1% and 32,7%, respectively, which are bounds for $\hat{B}_m$.

Table 4.2: Bias-adjusted R-indicators and maximal bias for business surveys using small and full sets of auxiliary variables. The R-indicators are computed after 15, 30, 45 and 60 days fieldwork. 95% confidence intervals are estimated for the R-indicators and non-response rates are given.

		Small				*Full*			
Survey		*15d*	*30d*	*45d*	*60d*	*15d*	*30d*	*45d*	*60d*
STS industry	R	92,1%	93,3%	94,0%	94,2%	90,5%	91,8%	93,1%	93,3%
	CI	91,3-92,8	92,7-94,0	93,5-94,4	93,8-94,6	89,7-91,3	91,3-92,2	92,6-93,5	92,8-93,8
	B	8,1%	4,2%	3,5%	3,3%	9,7%	5,2%	4,1%	3,8%
	$1-\hat{\bar{\rho}}_X$	51,5%	21,3%	14,3%	11,7%	51,5%	21,3%	14,3%	11,7%
STS retail	R	96,1%	94,6%	94,0%	94,1%	88,1%	87,9%	88,3%	89,0%
	CI	95,4-96,7	94,0-95,2	93,5-94,5	93,6-94,6	87,3-88,8	87,3-88,6	87,6-88,9	88,3-89,6
	B	3,9%	3,5%	3,5%	3,3%	12,0%	7,7%	6,8%	6,2%
	$1-\hat{\bar{\rho}}_X$	50,5%	22,0%	14,2%	11,8%	50,5%	22,0%	14,2%	11,8%

The differences for the business surveys are more pronounced. As these surveys are much bigger, the confidence intervals are small with widths between 1 and 1,5%. The R-indicator for STS retail after 30 days fieldwork drops almost 7% when VAT is added to the auxiliary information. For STS industry the decrease is much smaller. Apparently, the size of VAT in the previous year does not relate to response very strongly. Without the VAT information the retail respondents have a higher R-indicator than the industry respondents. When VAT is added this picture changes drastically; now the retail respondents score worse.
The main survey item of the STS surveys is monthly turnover (subdivided over different activities). As VAT in a previous year can be expected to correlate strongly to turnover in the running year, it is important that representativeness is good with respect to VAT. Hence, application to the STS reveals that it is important to select auxiliary variables carefully.
The standardized maximal biases in table 4.2 are small relative to the non-response rate. Allready after 15 days of fieldwork, with a non-response rate of around 50%, the biases are in most cases smaller than 10%. The only exception is full auxiliary set for STS retail which

leads to a 12% bias. These results are promising given that VAT is correlating strongly to STS key statistics.

Table 4.1 also contains the R-indicators and maximal biases for sub-samples of the HS and CSS samples. For both surveys 1:4 and 1:9 simple random samples are drawn. As a consequence the sample sizes drop from around 15000 units to around 4000 and 2000 units. The inverse of the width of the 95% confidence intervals turns out to be proportional to the square root of the sample size, as expected from (2.15). For the full samples the width is close to 3%, while for the 1:9 sub-samples the width is around 8%. Hence, the confidence intervals may be quite large for sample sizes that are used in practical settings.
Remarkably, the size of the auxiliary variable set does not have a noticeable influence on the confidence intervals in the cases we investigated. They have approximately the same size for the small and full sets. We conclude that it is especially the sample size that restricts statements about the representativeness and only to a lesser extent the number of auxiliary variables selected.

Next, we have investigated the role of the number of auxiliary variables in simulated data set SIM1. We drew 200 samples using simple random sampling with three different sampling fractions, 1:50, 1:100, 1:200. For each sample, we generated a sample response variable using the 'true' population response propensity based on an outcome of a random uniform number.
We ran four logistic regression models on the samples:

1. Full Model - The explanatory variables that were used to generate the population propensities (see section 3). In this model we know $\aleph$, the set of auxiliary variables that predict non-response behaviour.
2. Simple Model A - Number of persons, child indicator, Income group, age.
3. Simple Model B - Type of locality, age, number of persons and child indicator.
4. Simple Model C - Type of locality, gender, child indicator, age.

For each of the 200 samples, we calculated R-indicators with and without bias corrections and estimated standard deviation following (2.15). The results are presented in tables 4.3 through 4.5, and contain the means of the values across each of the 200 samples. We included the proportion of samples that covered the true R-indicator within the 95% confidence interval. We also calculated the empirical standard deviation of the bias-adjusted R-indicators across the samples to compare it to the analytical expression of the variance.

In tables 4.3 to 4.5 we see that the bias adjustment depends on the sample size, and varies between 0,3% and 2,3%. In general the bias adjustment increases when the sample size decreases. This is as expected. The bias adjustments vary considerably over the different models, e.g. models B and C have smaller adjustments than the full model and model A.
The standard errors in (2.15) and the empirical standard errors are in all cases similar in size. The empirical confidence intervals tend to be slightly bigger than those that follow from the approximation in (2.15).
The simulations confirm that the selected set of auxiliary variables plays an important role. The smaller models A to C lead to R-indicator values that are outside the 956% confidence intervals of the full model.

We conclude that the choice of auxiliary variables is important when computing R-indicators. If an R-indicator value is disseminated, then it should be accompanied with the auxiliary variable set that it originates from. This conclusion is not at all surprising, but has some interpretational consequences. We should be aware that different target populations or universes have, by nature, different auxiliary characteristics. Although, the R-indicator has a meaning through the maximal bias, it does not make much sense to compare the R-indicator of a social survey to that of a business survey. The same is true for surveys that have (a subset of) households as a target population and surveys that have (a subset of) persons as a target population. It does make sense to compare surveys of which the target populations are subsets of one larger population. However, one should be aware that it is easier to get a representative response in a survey among young, single households in large cities, then in a survey among all households. In the first case the target population by itself is much more homogeneous.

Table 4.3: The mean values of unadjusted and adjusted R-indicators for 1:50 samples applied to SIM1. The true R-indicator equals 87,67%. Also given are estimated standard deviations, percentage of 95% confidence intervals covering the true R-indicator and empirical standard deviations.

	R-indicator		*Standard deviation*	*% samples with true R-indicator in 95% CI*	*Empirical standard deviation*
	unadjusted	*adjusted*			
Full Model	86,98%	87,59%	0,693%	93%	0,754%
Model A	89,57%	90,20%	0,683%	3%	0,704%
Model B	89,81%	90,19%	0,668%	5%	0,711%
Model C	90,91%	91,24%	0,667%	0%	0,696%

Table 4.4: The mean values of unadjusted and adjusted R-indicators for 1:100 samples applied to SIM1. The true R-indicator equals 87,67%. Also given are estimated standard deviations, percentage of 95% confidence intervals covering the true R-indicator and empirical standard deviations.

	R-indicator		*Standard deviation*	*% samples with true R-indicator in 95% CI*	*Empirical standard deviation*
	unadjusted	*adjusted*			
Full Model	86,40%	87,60%			
Model A	88,95%	90,17%	0,991%	29%	0,977%
Model B	89,44%	90,18%	0,960%	26%	0,963%
Model C	90,62%	91,26%	0,956%	3%	0,959%

Table 4.5: The mean values of unadjusted and adjusted R-indicators for 1:200 samples applied to SIM1. The true R-indicator equals 87,67%. Also given are estimated standard deviations, percentage of 95% confidence intervals covering the true R-indicator and empirical standard deviations.

	R-indicator		*Standard deviation*	*% samples with true R-indicator in 95% CI*	*Empirical standard deviation*
	unadjusted	*adjusted*			
Full Model	85,63%	87,99%			
Model A	87,76%	90,08%	1,455%	64%	1,486%
Model B	89,00%	90,48%	1,393%	47%	1,439%
Model C	90,04%	91,27%	1,383%	25%	1,411%

4.4 Main effects and interaction effects

One may argue whether the inclusion of interaction effects leads to a different model or to different auxiliary information (or both). Here, we consider interactions as an extension of the available auxiliary information. In other words, when we include interactions between auxiliary variables, then the set of auxiliary information is increased. The important question then is whether it makes a difference when some or all interactions are omitted. In the light of the previous section, we should say that it may make a difference. As the R-indicators are dependent on the set of auxiliary variables, they are also dependent on the classifications of those variables and of the interactions between the variables. Hence, when disseminating R-indicators it should be stated how the variables are combined. One may, however, hope that it suffices to restrict oneself to main effects.

Table 4.6 contains R-indicator values for HS and CSS based on different crossings of the small variable set gender (2 classes), age x marital status (26 classes) and urbanization degree (5 classes). The simplest model is a main effect model, while the most complex is a saturated model that includes all interactions. The saturated model has 260 classes.

Table 4.6: Bias-adjusted R-indicators and maximal biases for CSS and HS based on different main effect and interaction effect terms from the small set. 95% confidence intervals are given. The auxiliary variables are abbreviated to A = gender, B = age x marital status, and C = urbanization.

	HS 2005			CSS 2005		
	R	95% CI	B	R	95% CI	B
A + B + C	83,2%	81,9 - 84,7	12,5%	83,3%	81,8 - 84,8	12,5%
AxB + C	82,2%	80,8 - 83,6	13,2%	82,2%	80,9 - 83,6	12,3%
AxC + B	83,2%	81,7 - 84,6	12,5%	83,0%	81,7 - 84,5	12,7%
BxC + A	81,2%	79,8 - 82,6	13,9%	81,9%	80,5 - 83,2	13,5%
AxB + AxC + BxC	80,3%	79,0 - 81,7	14,6%	80,6%	79,2 - 81,9	14,5%
AxBxC	79,2%	77,9 - 80,6	15,5%	78,9%	77,6 - 80,2	15,8%

From table 4.6 a number of conclusions can be drawn. As expected the R-indicator value decreases when model complexity is increased and interactions are included. The decrease is for both cases considerable, a drop of 4%, and the confidence intervals of the main effect and saturated models do not overlap. The confidence intervals are, however, quite stable in size. When interactions are added, the confidence interval is shifted but does not get wider.
The R-indicators of the saturated models in table 4.6 are smaller than those of the main effect models in table 4.1 that include the extended auxiliary variable set. Again the confidence intervals are similar in size.

We conclude that the inclusion of interaction terms may strongly affect the value of the R-indicators. Therefore, it is important to consider which interaction terms to include in models for non-response and which not. In general the values of R-indicators depend strongly on the complexity of the model. Remarkably, confidence interval sizes only mildly depend on the complexity of the model.

4.5 Variable selection

Given the discussions in the previous sections, it is clear that variable selection may not be suitable when comparing different surveys. Even when significance levels are fixed, variable selection leads to different sets of auxiliary variables depending on the sample size and the non-response missing-data-mechanism. In fact, when applying variable selection one may seek to estimate either a true underlying R-indicator or the R-indicator that corresponds to a saturated model including all available auxiliary variables. In both cases the resulting estimated R-indicator may be biased. For different surveys these biases may be different, so that any comparison between them is hampered.

One may, however, investigate whether variable selection is suitable from a mean square error (MSE) point of view. In other words, does it make sense to restrain to a smaller model in order to balance bias and variance. However, if we regard MSE as a criterion then we first have to decide what R-indicator we want to estimate, i.e. what is the reference for the bias. There are two possible choices of reference for the computation of MSE

$$MSE_1(\hat{R}(\hat{\rho}_Z)) = V(\hat{R}(\hat{\rho}_Z)) + \left(E(\hat{R}(\hat{\rho}_Z) - R(\rho_X)\right)^2 \tag{4.1}$$

$$MSE_2(\hat{R}(\hat{\rho}_Z)) = V(\hat{R}(\hat{\rho}_Z)) + \left(E(\hat{R}(\hat{\rho}_Z) - R(\rho_{\aleph})\right)^2, \tag{4.2}$$

where Z represents an auxiliary vector for some submodel of the saturated model with all available auxiliary variables X. MSE_1 takes the R-indicator based on the response propensities ρ_X as the reference for bias. MSE_2 considers the response probabilities originating from the 'super' set, $\rho_{\aleph}$, as a benchmark. In (4.1) and (4.2) we consider the bias-adjusted R-indicators, but have omitted the index 'B'.

We know that $R(\rho_Z) \geq R(\rho_X)$ and $\hat{R}(\hat{\rho}_Z) \geq \hat{R}(\hat{\rho}_X)$, and that $MSE_1(\hat{R}(\hat{\rho}_X)) = V(\hat{R}(\hat{\rho}_X))$. Clearly, we cannot measure MSE_2, but it can easily be seen that MSE_2 is smallest when $R(\rho_X) = R(\rho_{\aleph})$. In other words, if X represents all auxiliary variables that explain response behaviour, then it would be the least favourable to shift from the saturated model to a smaller model. Hence, we can view MSE_1 as a worst-case measure for MSE_2. In the following we will estimate the MSE_1 for different models by

$$\hat{MSE}_1(\hat{R}(\hat{\rho}_Z)) = \hat{V}(\hat{R}(\hat{\rho}_Z)) + \left(\hat{R}(\hat{\rho}_Z) - \hat{R}(\hat{\rho}_X)\right)^2, \tag{4.3}$$

$$\hat{MSE}_1(\hat{R}(\hat{\rho}_X)) = \hat{V}(\hat{R}(\hat{\rho}_X)). \tag{4.4}$$

We performed several analyses in which we select variables. We used three strategies: CHAID classification trees, and forward and backward search in logistic regression. In the analyses we adjusted the bias of R-indicators as if the resulting model was a fixed model. This is not correct. In general variable selection will lead to a slight overfitting and to R-indicators that are slightly smaller than the R-indicators that result from fixed models. Bias adjustment for R-indicators based on variable selection is a topic of further research.

Table 4.9 gives R-indicators and biases that follow from a CHAID analysis (see Kass 1980) of HS and CSS. In both cases the small set of auxiliary variables, gender, age x marital status and urbanization, were input to CHAID classification. CHAID repeatedly splits a population into sub-groups based on the auxiliary variables offered and and Chi-square statistics. The

result is a tree in which the leaves represent strata. The strata together form a saturated model, i.e. the strata are disjoint and the union of the strata is the population. In order to investigate the dependence on sample size, again 1:4 and 1:9 sub-samples were taken from the survey samples. The maximum number of levels in the CHAID tree was three, as there are only three classification variables.
The maximum number of nodes equals the product of the numbers of categories of the three variables, which is 2 x 26 x 5 = 260. The number of nodes produced by CHAID are much smaller than this maximal number. In the full samples around 15 nodes were selected, while for the 1:4 and 1:9 samples the number of nodes drops below ten. The R-indicator values for do not show a strong increase when the sampling fraction is decreased. The only exception is the 1:9 sample for HS, for which the R-indicator is much higher and outside the confidence intervals of the full sample and 1:4 sample.

The R-indicators for the saturated model (gender x age x marital status x urbanization) are given in table 4.7. For HS and CSS the R-indicator is 79,2% and 78,9%, respectively. The corresponding MSE_1 of (4.4) equals 4,74 x10^{-5} for HS and 4,40 x10^{-5} for CSS. The full sample MSE_1 of (4.5) for the CHAID classification is equal to 1,42 x10^{-3} for HS and 1,90 x10^{-3} for CSS. Hence, MSE_1 is for both surveys much larger for the saturated models that follow from CHAID. From a MSE point of view we would favour the full interaction model to the CHAID classifications.

Table 4.9: Bias-adjusted R-indicators and maximal biases for the social surveys following a CHAID fit using the small set of auxiliary variables. 95% confidence intervals are given. Bias adjustment and confidence intervals are made based on the CHAID classification as saturated model.

Survey		*Full sample*	*1:4 sample*	*1:9 sample*
HS 2005	R	82,9%	82,8%	90,2%
	95% CI	81,5% - 84,3%	80,1% - 85,6%	85,8% - 94,4%
	B	12,7%	12,8%	6,7%
	Strata	14	5	3
CSS 2005	R	83,2%	84,6%	84,5%
	95% CI	81,8% - 84,7%	81,7% - 87,4%	80,3% - 88,7%
	B	12,6%	11,5%	11,6%
	Strata	15	6	2

Table 4.10: The mean unadjusted and adjusted R-indicators following from CHAID classifications based 500 on samples with fractions 1:50, 1:100 and 1:200. The means are given for a small and an extended set of auxiliary variables. The true R-Indicator value is 87,67%. For each series of 500 simulations the minimum, mean and maximum number of nodes is given.

	Extended set R-indicator			*Small set R-indicator*		
	unadjusted	*adjusted*	*Nodes*	*unadjusted*	*adjusted*	*Nodes*
1:200	89,71%	90,15%	range = (2,8) mean = 4,8	91,06%	91,36%	range = (2,5) mean = 3,0
1:100	89,01%	89,40%	range = (3,15) mean = 8,4	90,56%	90,85%	range = (2,8) mean = 4,9
1:50	88,51%	88,85%	range = (4,29) mean = 15,6	90,50%	90,73%	range = (3,13) mean = 7,7

Using simulated data set SIM1, we carried out another CHAID classification. We drew 500 samples each for different sample rates 1:50, 1:100 and 1:200. For each sample, we calculated response propensities within each terminal node of the tree by the number of respondents divided by the sample size, i.e. for a saturated model. We used a small set of variables based on those that were used to generate the propensities in the population: number of persons, gender and age and the extended set of variables: number of persons, gender, age, income, locality type and child indicator.

The results are given in table 4.10. The algorithm produced varying sizes of trees depending on the sample size and the size of auxiliary variable set. No tree produced more than 30 terminal nodes. The minimum, maximum and mean number of nodes are given in table 4.10. Especially for the 1:50 samples the difference between the minimum and maximum number of nodes is great, ranging from 4 to 29 for the full set and from 3 to 13 for the small set. This indicates the variation in the classifications.

Both models on all sample sizes overestimate the R-indicator, even for the extended variable set. When the sampling fraction is increased, then in all cases the R-indicator values decrease and approach the true R-indicator value 87,67%. The decrease in R-indicator values was expected as a larger sample size allows for more significant interactions and, hence, for larger classification trees.

Table 4.11: Bias-adjusted R-indicators and maximal biases for HS following a forward Wald selection using the full set of auxiliary variables. 95% confidence intervals are given. Bias adjustment and confidence intervals are made based on the selected variables. The auxiliary variables are abbreviated to A = gender, B = age x marital status, C = urbanization, D = average house value, E = paid job, F =household type, and G = ethnic background.

Survey	*Full sample*	*1:4 sample*	*1:9 sample*
HS 2005	A+B+C+D+E+F+G 80,8% (79,4 – 82,3) MSE = 5,47 x 10^{-5}	A+B+C+D+E+F+G 78,8% (75,9 – 81,6) MSE =1,44 x 10^{-4}	A+B+C+D+E+F+G 74,1% (70,0 – 78,0) MSE = 4,16 x 10^{-4}
	B 85,5% (84,0 – 87,0) MSE = 2,27 x 10^{-3}	B 83,7 % (81,1 – 86,5) MSE = 2,59 x 10^{-3}	B 81,7% (77,6 – 86,1) MSE =6,25 x 10^{-3}
	B+G 82,9% (81,4 – 84,2) MSE = 4,92 x 10^{-4}	B+G 81,0% (78,0 – 84,1) MSE = 7,26 x 10^{-4}	B+A 79,9% (76,0 – 83,8) MSE = 3,76 x 10^{-3}
	B+G+C 81,7% (80,3 – 83,2) MSE = 1,36 x 10^{-4}	B+G+F 80,0% (77,1 – 82,9) MSE =3,63 x 10^{-4}	B+A+C 78,0% (74,2 – 81,9) MSE =1,91 x 10^{-3}
	B+G+C+F 81,2% (79,7 – 82,8) MSE = 7,85 x 10^{-5}	B+G+F+E 79,9% (76,7 – 82,7) MSE =3,55 x 10^{-4}	B+A+C+E 76,9% (72,5 – 80,9) MSE =1,24 x 10^{-3}
	B+G+C+F+E 81,0% (79,6 – 82,4) MSE = 5,50 x 10^{-5}	B+G+F+E+C 79,2% (76,6 – 82,0) MSE =2,06 x 10^{-4}	B+A+C+E+F 75,7% (71,3 – 79,8) MSE =7,26 x 10^{-4}
		B+G+F+E+C+A 79,0% (76,3 – 81,9) MSE =2,08 x 10^{-4}	

We also investigated variable selection in a logistic regression setting. We applied a forward selection using Wald statistics (see Agresti 2001) for the full sets of auxiliary variables for HS and CSS. Tables 4.11 and 4.12 have the resulting bias-adjusted R-indicators, maximal absolute biases and mean square errors. Again, 1:4 and 1:9 sub-samples were drawn from the survey samples. The first row of the table contains the outcomes for the full model. The subsequent rows give the R-indicator values corresponding to the forward searches. The last model shown is the model for which no additional variable was significant on a 5% level.

Table 4.12: Bias-adjusted R-indicators and maximal biases for CSS following a forward Wald selection using the full set of auxiliary variables. 95% confidence intervals are given. Bias adjustment and confidence intervals are made based on the selected variables. The auxiliary variables are abbreviated to A = gender, B = age x marital status, C = urbanization, D = average house value, E = paid job, F =household type, and G = ethnic background.

Survey	*Full sample*	*1:4 sample*	*1:9 sample*
CSS 2005	A+B+C+D+E+F+G 82,1% (80,7 – 83,4) MSE = 4,74 x 10^{-5}	A+B+C+D+E+F+G 78,7% (75,8 – 81,3) MSE = 1,97 x 10^{-4}	A+B+C+D+E+F+G 75,5% (71,7 – 79,5) MSE = 3,96 x 10^{-4}
	B 84,6% (83,2 – 86,0) MSE = 6,76 x 10^{-4}	F 84,4% (82,8 – 88,0) MSE = 3,42 x 10^{-3}	A 84,2% (79,6 – 88,4) MSE = 8,07 x 10^{-3}
	B+F 83,2% (81,8 – 84,6) MSE = 1,72 x 10^{-4}	F+C 84,0% (81,3 – 86,8) MSE = 3,00 x 10^{-3}	A+C 82,8% (78,7 – 86,9) MSE = 5,77 x 10^{-3}
	B+F+G 82,8% (81,4 – 84,2) MSE = 1,00 x 10^{-4}	F+C+E 82,8% (80,2 – 85,9) MSE = 1,89 x 10^{-3}	A+C+E 82,2% (78,1 – 86,6) MSE = 4,96 x 10^{-3}
	B+F+G+E 82,5% (81,2 – 84,0) MSE = 6,70 x 10^{-5}	F+C+E+B 80,1% (77,4 – 82,7) MSE = 3,79 x 10^{-4}	
	B+F+G+E+A 82,4% (81,0 – 83,8) MSE = 6,00 x 10^{-5}	F+C+E+B+G 79,4% (76,7 – 82,1) MSE = 2,39 x 10^{-4}	

Table 4.13: The unadjusted and adjusted R-indicators for LFS 2001 that follow from a backward selection employing all auxiliary variables and interactions long- term illness x age, marital status x gender, economic activity x marital status, long-term illness x marital status. Also given are standard errors and the mean square error computed with respect to the full model.

	R-indicator		*SE*	*MSE*	*B*
	unadjusted	*adjusted*		(x 10^{-4})	
Full auxiliary set	87,5%	91,5%	1,04%	1,07	5,1%
- tenure	87,6%	91,2%	1,03%	1,14	5,3%
- economic activity x marital status	88,3%	90,8%	1,01%	1,42	5,6%
- long-term illness x marital status	88,3%	90,7%	1,00%	1,63	5,6%
- long-term illness x age	88,5%	90,6%	1,00%	1,78	5,7%
- long-term illness	88,5%	90,5%	0,99%	1,83	5,7%
- age	88,7%	90,5%	0,99%	1,95	5,7%
- gender x marital status	88,9%	90,5%	0,98%	1,96	5,7%
- gender	88,9%	90,4%	0,98%	2,06	5,8%
- ethnicity (final model)	89,3%	90,7%	0,96%	1,51	5,6%

The patterns in tables 4.11 and 4.12 are the same in all cases; the R-indicator and MSE's decrease when variables are selected, and the maximal absolute bias goes up. The final selected variables produce R-indicators and MSE's that are in most cases close to those of the full model. In many cases the R-indicator confidence intervals overlap and MSE's have the same order of magnitude. For the full sample HS and CSS one may prefer the parsimonious models based on variable selection.

Finally, we applied two backward variable selections for LFS 2001 and for SIM1. The results for the LFS are shown in table 4.13 and for SIM1 in tables to 4.14 to 4.16. The LFS non-response rate was 17,2%. Standardized biases are given in table 4.13.

In the backward selection of the LFS, we included all available auxiliary variables, see table 3.3, plus the four interaction terms long- term illness x age, marital status x gender, economic activity x marital status, long-term illness x marital status. The results are shown in table 4.13. In the backward selection nine variables or interactions were dropped. The final model consists of five variables region x urbanization, marital status, economic activity and accommodation. We also calculated the mean square errors in (4.3) and (4.4), i.e. relative to the model that contains all main effects plus the four interaction effects.
From table 4.13 we can draw several conclusions. First of all, the bias corrections get smaller when the models get smaller; they range from 4,0% for the full model to 1,4% for the final model. Standard errors also drop, but the decrease is very gradual and confirms that the number of auxiliary variables does not have a strong effect on the size of confidence intervals. As expected, the adjusted R-indicators decrease when variables are deleted from the model. This decrease is, however, small relative to the standard error. The difference in R-indicator values between the largest and smallest model is only 0,8%, which is smaller than the standard error.
The variable selection for LFS leads to the same conclusion as for HS and CSS. Smaller models that follow from variable selection may have comparable values for R-indicators, standard errors and, consequently, MSE. As the smaller models are parsimonious, one may favour them to the full models. However, from a MSE point of view, there is no strong incentive to search for smaller models.

In the backward variable selection for SIM1 we started with a model including nine variables, number of persons, locality type, child indicator, income, education, gender, age, ethnicity, marital status and four interactions, number of persons x type of locality, number of persons x marital status, gender x marital status and child indicator x marital status. The backward selection was performed for the full data set and resulted in the stepwise deletion of four main effects and four interaction effects. The final model consisted of number of persons, type of locality, child indicator, age, education, and number of persons x type of locality. Next, we computed R-indicators for several the selection steps on 200 simple random samples with fractions 1:50, 1:10 and 1:200. Tables 4.14 to 4.16 give the results for each of the fractions.

Tables 4.14 to 4.16 give a picture that resembles table 4.13, but in this case we know the true R-indicator. In all cases we again see that the size of the bias adjustment decreases when variables and interactions are deleted from the model. After adjustment all values are close to the true R-indicator, even for the 1:200 sample, and the proportion of confidence intervals that does not contain the true R-indicator is negligible. The MSE is computed with respect to the full model, not with respect to the true R-indicator which would be unknown in practical settings. For the 1:50 and 1:100 samples there is a turning point for the third model; from the full model to this model the MSE drops but increases when more terms are removed from the model. For 1:200 samples there is no turning point.

Table 4.14: Mean adjusted and adjusted R-indicators for 200 1:50 samples for SIM1. The true R-indicator equals 87,67%. Also given are the standard error, the fraction of confidence intervals that contains the true R-indicator and the MSE.

	R-indicator		*Standard error*	*% CI's with true R-indicator*	*MSE* (x 10^{-4})
	unadjusted	*adjusted*			
Full set	86,33%	87,66%	0,706%	100%	0,498
- ethnicity	86,53%	87,67%	0,703%	100%	0,494
- child x marital status					
- # persons x marital status	86,82%	87,68%	0,697%	100%	0,486
- gender x marital status					
- gender	87,04%	87,84%	0,696%	100%	0,517
- marital status					
Final set	87,49%	88,05%	0,685%	100%	0,621

Table 4.15: Mean adjusted and adjusted R-indicators for 200 1:100 samples for SIM1. The true R-indicator equals 87,67%. Also given are the standard error, the fraction of confidence intervals that contains the true R-indicator and the MSE.

	R-indicator		*Standard error*	*% CI's with true R-indicator*	*MSE* (x 10^{-4})
	unadjusted	*adjusted*			
Full set	85,21%	87,76%	1,029%	100%	1,06
- ethnicity	85,57%	87,75%	1,021%	100%	1,04
- child x marital status					
- # persons x marital status	86,11%	87,78%	1,012%	100%	1,02
- gender x marital status					
- gender	86,39%	87,94%	1,009%	100%	1,05
- marital status					
Final set	87,06%	88,18%	0,988%	100%	1,15

Table 4.16: Mean adjusted and adjusted R-indicators for 200 1:200 samples for SIM1. The true R-indicator equals 87,67%. Also given are the standard error, the fraction of confidence intervals that contains the true R-indicator and the MSE.

	R-indicator		*Standard error*	*% CI's with true R-indicator*	*MSE* (x 10^{-4})
	unadjusted	*adjusted*			
Full set	83,02%	87,09%	1,450%	96%	2,10
- ethnicity	83,66%	87,24%	1,449%	98%	2,12
- child x marital status					
- # persons x marital status	84,63%	87,75%	1,485%	100%	2,64
- gender x marital status					
- gender	84,98%	87,89%	1,475%	100%	2,82
- marital status					
Final set	86,02%	88,17%	1,437%	100%	3,23

Hence, we can conclude that variable selection based on MSE may lead to smaller models and that parsimonious models may have an MSE that is close to that of the full model. The use of parsimonious models and the focus on MSE may be especially welcome in surveys were the number of available auxiliary variables is large relative to the sample size. In such settings there may be a turning point in the MSE. When one would like to compare different surveys, then the set of variables that is shared may be small so that a trade off between bias and variance will not be necessary. However, when comparing a survey in time or when monitoring response during data collection, there may be many auxiliary variables and such trade-offs may be very realistic.
One should realise that even when monitoring the data collection of a single survey, variable selection leads to R-indicators with different biases for different moments during the fieldwork. Obviously, the sample size is the same, and variable selection may be very helpful to approximate the R-indicator that corresponds to a complete crossing of the available auxiliary variables.

5. Non-response models with different link functions

Apart from the selection of explanatory, auxiliary variables one needs to choose a type of model to link these variables to response. In this section we investigate the possible impact on R-indicators for two types of models; linear regression and logistic regression. In other words, we employ the following link functions

$$\text{linear regression:} \quad h(x^T\beta) = x^T\beta$$

$$\text{logistic regression:} \quad h(x^T\beta) = \frac{\exp(x^T\beta)}{1+\exp(x^T\beta)}.$$

In all models we include intercepts. For the bias-adjustment we need that

$$\text{linear regression:} \quad \nabla h(x^T\beta) = 1$$

$$\text{logistic regression:} \quad \nabla h(x^T\beta) = \frac{\exp(x^T\beta)}{(1+\exp(x^T\beta))^2}.$$

We did not investigate probit models, which may be an extension to our research. Also, classification tree methods like CHAID or CART are often used to model response propensities. However, these methods correspond to saturated linear regression models, i.e. to a regression on a single categorical variable following from the classification. This implies that it is sufficient to consider linear regression.

We must again note that in this paper we only employ demographic and socio-economic auxiliary variables. We do this because we investigate the use of R-indicators for the comparison of different surveys. When one is interested in monitoring the representativeness of the survey response in time or during data collection, then the objective for the use of R-indicators is different. It is then interesting to include survey specific auxiliary variables, and one would most likely fix the model when making such a comparison within a single survey.
In a different RISQ work package (WP6) we look at non-response types and include fieldwork paradata. When non-response types are distinguished or paradata are included, then

models are usually more complex and may consist of multiple model equations. Examples are multilevel models, sample selection models and pattern mixture models.

Tables 5.1 and 5.2 contain estimated R-indicators with their corresponding 95% confidence intervals and maximal absolute bias for the four survey data sets. In each survey we estimated R-indicators for the small set and the extended set. Also, for the business surveys we look at 30 days and 60 days of fieldwork. The STS response rate after 60 days of fieldwork is approximately 92% and the differences between linear and logistic regression may be more pronounced.

Table 5.1: Bias-adjusted R-indicators and maximal bias for social surveys using small and full sets of auxiliary variables and a linear and logistic regression. 95% confidence intervals are estimated for the R-indicators.

	Small set		*Full set*	
Data set	*Linear*	*Logistic*	*Linear*	*Logistic*
CSS 2005	83,3%	83,3%	82,1%	82,1%
	81,9% - 84,7%	81,8% - 84,8%	80,6% - 83,5%	80,7% - 83,4%
	25,0%	25,0%	26,8%	26,8%
HS 2005	83,3%	83,2%	80,9%	80,8%
	81,8% - 84,7%	81,9% - 84,7%	79,5% - 82,4%	79,4% - 82,3%
	24,8%	25,0%	28,4%	28,5%

Table 5.2: Bias-adjusted R-indicators and maximal bias for business surveys using small and full sets of auxiliary variables and a linear and logistic regression. The R-indicators are computed after 30 and 60 days fieldwork. 95% confidence intervals are estimated for the R-indicators.

	Small set		*Full set*	
Data set	*Linear*	*Logistic*	*Linear*	*Logistic*
STS industry 2007				
After 30 days	93,3%	93,3%	91,8%	91,8%
	92,7% - 93,8%	92,7% - 94,0%	91,3% - 92,4%	91,3% - 92,2%
	8,5%	8,5%	10,4%	10,4%
After 60 days	94,2%	94,2%	93,3%	93,3%
	93,8% - 94,6%	93,8% - 94,6%	92,9% - 93,8%	92,8% - 93,8%
	6,6%	6,6%	7,6%	7,6%
STS retail 2007				
After 30 days	94,6%	94,6%	88,0%	87,9%
	94,0% - 95,2%	94,0% - 95,2%	87,3% - 88,7%	87,3% - 88,6%
	6,9%	6,9%	15,4%	15,5%
After 60 days	94,3%	94,1%	89,0%	89,0%
	93,8% - 94,7%	93,6% - 94,6%	88,4% - 89,6%	88,3% - 89,6%
	6,5%	6,7%	12,5%	12,5%

All calculations show that there is little or no difference between R-indicator and maximal bias values for linear and logistic regression. This is true even for the STS surveys after 60 days. The differences all lay well within the 95% confidence intervals. The confidence intervals themselves are also very similar and differ at most 0,2%. This is a very promising result. We do not need to worry too much about the link function when estimating R-indicators.

6. The use of R-indicators in practice; recommendations

Based on the various analyses and discussions we give the following recommendations for the presentation, implementation and use of R-indicators and standardized maximal absolute bias:

How to present R-indicators?

- Maximal absolute biases and R-indicators cannot be evaluated or presented separately from the auxiliary variables that were used for the prediction of response propensities.
- When comparing different surveys, one should use the same set of auxiliary variables, with the same classifications and with the same interactions between those variables.
- R-indicators and maximal absolute biases should always be given together with a confidence interval.

How to model non-response and estimate response propensities when computing R-indicators?

- The number of selected auxiliary variables has only a mild effect on the size of confidence intervals for R-indicators. In the survey data we have investigated, variable selection led to models that were inferior to models including all variables from a mean square error point of view. We, therefore, recommend to use fixed sets of auxiliary variables.
- Interaction terms may decrease R-indicators considerably. Again we found that from a mean square error point of view, models that include all interactions are often superior to less complex models that omit some of the interactions. We recommend to fix beforehand what interaction effects to include.
- The choice of link function in simple non-response models only has a minor influence, at least in the cases we have investigated. Further research is, however, needed to assess whether a change of link function might affect the impact of interaction terms in the model.
- The inclusion of response-unrelated auxiliary variables leads to an increase of the standard error of R-indicators, but not to a decrease of the bias of the R-indicators with respect to any reference. We, therefore, recommend to restrict analysis to auxiliary variables for which it is known from the literature that they relate to response behaviour.

How to deal with the sample size?

- R-indicators and maximal absolute biases are random variables and confidence intervals can be quite wide, even for surveys with sample sizes of 5000. Small samples do not allow for strong conclusions about the representativeness of the response.
- The size of confidence intervals is only mildly affected by the number of auxiliary variables.
- Only when the number of auxiliary variables is large relative to the sample size, we recommend to use variable selection with the mean square error as criterion. For the simulated data we constructed settings where parsimonious models are preferable to the full models with respect to MSE.

How to use R-indicators?

- R-indicators measure the distance to a fully representative response; they do not reflect the impact of non-response on the bias of (weighted) means of survey variables, and nor does the response rate. The standardized bias combines the response rate and the R-indicator and is designed to make comparisons of non-response bias under worst case scenarios.
- When comparing different surveys, we recommend to fix a number of sets of auxiliary variables beforehand (including interactions) and to add all variables to the models. One should restrict to demographic and socio-economic characteristics that are generally available in many surveys.
- When comparing a survey in time, we again recommend to fix a number of sets of auxiliary variables in the computation of the R-indicators. However, now the sets need not be restricted to general population characteristics but may include variables that correlate to the main survey items, and variables that relate to the data collection (paradata). When many variables are available, parsimonious models may be favoured to full models with respect to MSE.
- The use of R-indicators during data collection will be discussed in forthcoming RISQ papers.

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, Øyvin Kleven and Li-Chun Zhang 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., Schouten, B. (2008), Representativity Indicators for Survey Quality (RISQ), collaborative project, 7th Framework Programme FP7SSH20071, Socioeconomic sciences and the humanities Part 8, CBS, Voorburg, The Netherlands.

Biemer, P. (2008), A review of “The statistical properties of R-indicators”, Technical review, RISQ Deliverable 15, available on request at jg.schouten@cbs.nl .

Cobben, F. and Schouten, B. (2007), An empirical validation of R-indicators, Discussion paper 08006, CBS, Voorburg, available at

www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2008 .

Cochan, W.G. (1977) Sampling Techniques, 3rd Ed, New York: Wiley.

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap. New York: Chapman and Hall.

RISQ (2008), RISQ Data set documentations, Deliverable 1, available at www.r-indicator.eu .

Särndal, C-E. and Lundström, S. (2005), Estimation in Surveys with Nonresponse, John Wiley & Sons, Chichester, England.

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. (2007), R-indexes for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, CBS, Voorburg, available at

www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007 .

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.C. (2009), The statistical properties of R-indicators, RISQ deliverable 2.1, available at www.r-indicator.eu .

Wolter, K.M. (2007), Introduction to Variance Estimation. 2nd Ed. New York: Springer.