Representativity Indicators for Survey Quality

Representativeness indicators for measuring and enhancing the composition of survey response

RISQ-paper 9

Barry Schouten, Jelke Bethlehem
Centraal Bureau voor de Statistiek

December 17, 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

Representativeness indicators for measuring and enhancing the composition of survey response

Summary: Survey organisations have to make a steadily growing effort to achieve high response rates in household surveys. For this reason the question how to allocate fieldwork resources efficiently becomes more and more important. From the literature it is known that a focus on the response rate alone is not sufficient. The composition of the response to a survey is at least as important. Good indicators to measure, monitor and control the representativeness of response are, however, lacking.
The project RISQ (Representativity Indicators for Survey Quality) 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 may be used as tools to compare the response composition of surveys in time and of different data collection strategies. They may also serve as tools for the construction of data collection strategies that balance response rates given auxiliary information that is available beforehand and given paradata that becomes available during fieldwork.
We discuss different types of indicators for the quality of survey response and their properties. Furthermore, we illustrate and discuss their use in comparing and monitoring surveys by application of the indicators to several data sets.

1. Introduction

One of the most important factors affecting the quality of surveys of households or enterprises is nonresponse. The impact of nonresponse on survey quality is typically measured by the response rate. The response rate alone, however, is not sufficient as a quality indicator to capture the potential impact of nonresponse. The bias of estimates resulting from nonresponse also depends on the contrast between respondents and nonrespondents with respect to a target variable. The more they differ, the larger the bias will be. Good indicators that measure the degree to which the group of respondents of a survey still resembles the complete sample are currently lacking.

The RISQ (Representativity Indicators for Survey Quality) project is funded by the 7th EU Framework Programme (FP7). 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 supports the comparison of the quality of different surveys, both business and household surveys, and registers, and to facilitate the efficient allocation of data collection resources.

The indicators can be used in three different 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 monitor the response to a survey during data collection, e.g. after various days, weeks or months of fieldwork

Since we want the indicators to facilitate the evaluation of any survey, the indicators should not relate to the response quality of specific survey items. Different surveys have different survey items, which would make a comparison impossible a priori. Furthermore, indicators should not relate to a specific population parameter or a specific estimator or model. Again different surveys may aim at different statistics and may employ different estimators, estimation strategies and underlying models. These are important requisites to the definition of representative response and to indicators that measure a deviation from that definition.
R-indicators may help improving the quality of the response by targeting groups that are underrepresented, but they are not designed to be tools for the selection of weighting variables. The indicators may inform the construction of weights but that is not their primary purpose. They should, however, assess the extent to which weighting models are leaning on assumptions about the non-response mechanism. When a survey response is less representative, then survey researchers have to rely more strongly on nonresponse adjustment techniques. Indicators may thus be used to produce a more balanced response, but they will not make non-response adjustment methods redundant.

Representativeness is a property that is not defined in the survey literature. In order to avoid ambiguity, we, therefore, explicitly define representativeness and conditional representativeness. However, apart from the definition two remarks are important to make beforehand. First, our definition is dependent upon information on auxiliary variables. Hence, any indicator will have to be disseminated together with a statement about what auxiliary information was employed to evaluate representativeness. Second, our representativeness indicator is estimated from sample data and has a precision that depends on the sample size.

In section 2 we define representativeness and indicators. We illustrate the different types of indicators in section 3. Next, in section 4 we discuss the use of indicators in practical survey settings. In section 5, we address future research.

2. Representativeness and indicators for representativeness

For the sake of brevity we only give condensed descriptions of representativeness and corresponding indicators. We refer to Schouten et al (2009), and the RISQ deliverables Shlomo et al (2009a) and Shlomo et al (2009b) for details.

2.1 Representativeness

Ideally we would like to define representativeness based on individual response probabilities. Their interpretation is not straightforward, however, and has been open to extensive debates in the literature; see e.g. various chapters in Madow and Olkin (1983). Moreover, it is impossible to estimate such probabilities based on a single response for each sample unit without making strong assumptions. For these reasons we restrict ourselves to response propensities. Let ρ_X denote the response propensity function for variable X, say age or gender, i.e. $\rho_X(x)$ is the probability that a population unit carrying value $X = x$, say young people or females, will respond to the survey request. We suppose that X is a subset of a supervector $\aleph$ of auxiliary variables that explains response behaviour and for which the response propensities $\rho_\aleph$ can be viewed as individual response probabilities. This $\aleph$ may be viewed as the whole of characteristics of a person or business that determines their response behaviour given a survey design.

We propose two definitions for representativeness of survey response; representative response and conditional representative response.

Definition: A response to a survey is representative with respect to X when response propensities are constant for X, i.e. when $\rho_X(x)$ is a constant function.

Definition: A response to a survey is conditional representative with respect to X given Z when conditional response propensities given Z are constant for X, i.e. when $\rho_{X,Z}(x,z) = \rho_Z(z)$ for all x.

The two definitions can be measured for any auxiliary vectors X and Z, e.g. age and gender or business size and type of business. In order to do that we need a distance function or metric, say $d(\rho_1, \rho_2)$, that measures distance between two vectors of response propensities ρ_1 and ρ_2. For this purpose we use the Euclidean distance

$$d(\rho_1, \rho_2) = \sqrt{\frac{1}{N}\sum_U (\rho_{1,i} - \rho_{2,i})^2}\ , \qquad (2.1)$$

where N is the population size, U the population units and i a label for a population unit.

This definition of representative response is proposed in Schouten et al (2009). The motivation for the definition is that it conforms to random samples, or in other words response leads to equal selection probabilities and can be considered as an additional phase in the sampling design. It's interpretation is straightforward as a result of that. It does not relate to a specific survey item, a specific estimator or a specific model for response behaviour other than that we assume that response propensities exist. The definition of conditional representative response is new.

Both definitions relate to assumptions that are common in literature about missing data. Missing-data-mechanisms (like non-response to a survey) are often termed either Missing-Completely-at-Random (MCAR), Missing-at-Random (MAR) or Not-Missing-at-Random (NMAR), after the influential work of Little and Rubin (2002). The three mechanisms represent decreasingly strict assumptions about the missingness of data; with MCAR being the most favourable and NMAR being the least favourable setting. There is an essential conceptual difference between these mechanisms and the definition of representative response, that is important to stress. This difference arises from the objectives behind the definitions. The missing-data-mechanisms originate from the focus on estimation while the definition of representative response comes from the focus on data collection.

Somewhat confusingly, in the literature the missing-data-mechanisms are usually referred to without an explicit reference to what is missing. However, the mechanisms only have a meaning when they are connected to variables. A different mechanism may apply to different items Y in the same survey and different sets of auxiliary information X. MCAR(Y) means, in terms of response propensities, that $\rho_Y(y)$ is constant in y, i.e. response is representative with respect to Y. MAR(X,Y) means that $\rho_{X,Y}(x,y) = \rho_X(x)$ for all y, while NMAR(X,Y) implies that $\rho_{X,Y}(x,y) \neq \rho_X(x)$ for at least one possible outcome y. The distinction between Y and X is deliberate. Y is a variable of interest in a survey and X is an auxiliary variable. The three mechanisms cannot be tested formally for any survey item, but are underlying to

models that attempt to adjust for the impact of nonresponse. MCAR means that no adjustment is needed. MAR means that the distribution of Y is affected by nonresponse and parameters of that distribution like the mean may be biased as a result of that. Adjustment using X, if relations with Y and R are specified correctly, removes this bias. NMAR implies that X does not suffice to remove the bias of all parameters.

With the definition of representative response we do not have estimation in mind but data collection. We do not consider a specific Y nor a specific parameter of any distribution. We question whether data collection succeeded in obtaining a balanced response for a set of pre-selected variables X that is available before and during data collection. Of course the selected variables may be of a general, wide interest when multiple surveys are compared, or may consist of relevant variables for a particular survey when that survey is compared to itself. Hence, representative response with respect to X is the same as MCAR(X), but non-representative response does not conform to MAR or NMAR in any way. The most that can be said is that the more deviant from representative response, the stronger one has to rely on MAR assumptions in the estimation of parameters of interest.

2.2 Measuring deviations from representative response

Given (2.1) we define a representativeness indicator or R-indicator, as the transformed distance between ρ_X, the response propensity function for X, and the constant vector $\rho_0 = (\rho, \rho, \ldots, \rho)^T$, which equals the survey response rate ρ.

$$R(X) = 1 - 2d(\rho_X, \rho_0) = 1 - 2S(\rho_X) \tag{2.2}$$

It is easy to show that d is the standard deviation S of the response propensities for X. The transformation in (2.2) was made so that $R \in [0,1]$ and representative response is represented by a value of one (or 100%) for the indicator. A value of 0 indicates the largest possible deviation from representative response.

Note that $R(X_1) \geq R(X_2) \geq R(\aleph)$ when variable X_2 is nested in X_1. The more refined the "resolution", the more variation is observed. So one should not compare R-indicators based on different vectors of auxiliary variables.

In general X will be a vector of auxiliary variables like age, gender or urbanization for household surveys and business type and size for business surveys. If measuring representativeness is restricted to one auxiliary variable, say Z, then we call the indicator a partial representativeness indicator or partial R-indicator. At the variable level the partial R-indicator is defined as

$$P_u(Z) = d(\rho_Z, \rho) = S(\rho_Z), \tag{2.3}$$

the standard deviation of the response propensity function $\rho_Z(z)$ in the population.

The subscript u in (2.3) is given in order to distinguish partial R-indicators for unconditional representative response from those for conditional representative response that we will define in section 2.3. For any Z it holds that $P_u(Z) \in [0,1]$. Furthermore, $P_u(Z) \in [0, (1 - R(X))/2]$ when Z is an element of X.

Next, for categorical variables we define partial R-indicators for each category. Let Z be a categorical variable with categories $k = 1,2,\ldots,K$ and let Z_k be the 0-1 variable that indicates whether $Z = k$ or not. For example, Z represents age and Z_k is the indicator for being younger than 35 years of age. The partial R-indicator for a category k is defined as

$$P_u(Z,k) = \sqrt{\frac{N_k}{N}}(\rho_{Z_k} - \rho)\,, \tag{2.4}$$

with $N_k = \sum_U Z_k$ the number of population units in category k. $P_u(Z,k)$ originates from dividing $P_u(Z)$ over the strata of Z while maintaining the signs between the stratum response propensity ρ_{Z_k} and the overall response rate ρ. Negative values indicate underrepresentation while positive values indicate overrepresentation. We have that $P_u(Z,k) \in [-1,1]$ and

$$P_u(Z) = \sqrt{\sum_{k=1}^{K} P_u^2(Z,k)}\,.$$

Note that (2.3), the partial R-indicator at the variable level, is in fact the square root of the "between" variance for variable Z. As such it is a component of the total variance of response propensities in (2.2), and, hence, always smaller than or equal to that variance.

2.3 Measuring deviations from conditional representative response

In measuring conditional representativeness we want to adjust the impact for one variable for the impact of other variables. Based on (2.1) we propose

$$P_c(Z \mid X) = d(\rho_{X,Z}, \rho_X) = \sqrt{\frac{1}{N-1}\sum_U (\rho_{X,Z}(x_i,z_i) - \rho_X(x_i))^2}\,, \tag{2.5}$$

the distance between propensities based on X and Z, and based on X alone. For example, X could be a vector containing household composition, household income and province of residence while Z equals the age of the head of the household.

Again, we define partial R-indicators for classes of categorical variables by distributing (2.5) over the classes of Z.

$$P_c(Z,k \mid X) = \sqrt{\frac{1}{N-1}\sum_U Z_k(\rho_{X,Z}(x_i,z_i) - \rho_X(x_i))^2}\,. \tag{2.6}$$

Other than for the unconditional partial indicators, we cannot assign a positive or negative sign to the category level conditional partial indicators in (2.6). The reason is that the sign may be different for each subclass of X. In some subclasses a certain age of the head of the household may have a positive effect on response while in others it has a negative effect.

It can be shown that (2.5) is the square root of the "within" variance of the $\rho_{X,Z}$ propensities for a stratification of the population with X. In other words, it is the variation that is left

within the cells defined by X. In our example, it represents the variation in response behaviour due to the age of the head of the household given its household composition and income and the province in which the household lives. As the within variance is again a component of the total variance, the conditional partial indicators too cannot exceed the total variance that makes up the R-indicator in (2.2). Furthermore, the conditional partial R-indicator for Z is always smaller than the unconditional partial R-indicator for that variable. This makes sense; the impact on response behaviour is to some extent removed by accounting for other characteristics of the population unit. In many survey settings, for instance, the impact of gender on response behaviour is completely or considerably removed by accounting for the age of the person.

2.4 Maximal absolute contrast and maximal absolute bias

In order to enable R-indicators to be interpreted in terms of the impact of nonresponse on survey estimation, we consider the standardized bias of the design-weighted response mean $\hat{\bar{y}}_r$ of an arbitrary survey item y.

$$\frac{|B(\hat{\bar{y}}_r)|}{S(y)} = \frac{|Cov(y,\rho_Y)|}{\rho S(y)} = \frac{|Cov(y,\rho_{\aleph})|}{\rho S(y)} \leq \frac{S(\rho_{\aleph})}{\rho} = \frac{1-R(\aleph)}{2\rho}, \tag{2.7}$$

with ρ the average response propensity (or expected response rate). Clearly, we do not know $\rho_{\aleph}$. Moreover, we want to have a measure that enables comparison of the representativeness of response in different surveys or the same survey over time. In such a setting we are interested in the general representativeness of a survey, i.e. not the representativeness with respect to single survey items. We use as an approximation for (2.7)

$$B_m(X) = \frac{1-R(X)}{2\rho}. \tag{2.8}$$

B_m represents the maximal absolute standardized bias under the scenario that non-response correlates maximally to the selected auxiliary variables.

We let $\Delta B_m(X \mid Z) = B_m(X,Z) - B_m(Z)$ denote the difference in maximal absolute standardized bias when adding X to the vector Z of auxiliary variables. $\Delta B_m(X \mid Z)$ can be informative of the contribution of X.

Additionally, we consider the maximal contrast between respondents and non-respondents. The contrast for a variable Y is the expected difference between the response mean and nonresponse mean of that variable. The bias of the response mean can be rewritten as the product of the non-response rate $1-\rho$ and the contrast.

$$B(\hat{\bar{y}}_r) = (1-\rho)(E(\hat{\bar{y}}_r) - E(\hat{\bar{y}}_{nr})).$$

Hence, we may define the maximal absolute standardized contrast as the maximal absolute standardized bias divided by the non-response rate. We denote it by $C_m(X)$

$$C_m(X) = \frac{1 - R(X)}{2\rho(1-\rho)}. \qquad (2.9)$$

For convenience we will refer to B_m and C_m as the maximal bias and maximal contrast.

The R-indicator, the maximal bias and the maximal contrast provide means to evaluate the quality of response. Ideally, one would like to bound the R-indicator from below, i.e. to derive values of the R-indicator that are acceptable and values that are not. We construct three so-called response-representativity functions that can be used for deriving lower bounds for the R-indicator. They are a function of a threshold γ and the response rate ρ. The threshold γ represents a quality level. The response-representativity functions are defined as

$$RR_1(\gamma,\rho) = 1 - \frac{2}{\xi_{1-0.5\alpha}}\gamma \qquad \text{(maximal variation in response propensities)}$$
$$RR_2(\gamma,\rho) = 1 - 2\rho\gamma \qquad \text{(maximal bias)}$$
$$RR_3(\gamma,\rho) = 1 - 2\rho(1-\rho)\gamma, \qquad \text{(maximal contrast)}$$

with $\xi_{1-0.5\alpha}$ being the $1-0.5\alpha$ quantile of the standard normal distribution.

We will explain the background and interpretation of each of the functions. The functions originate from setting a threshold γ to $\xi_{1-0.5\alpha}S(\rho_X)$, $B_m(\rho_X)$ and $C_m(\rho_X)$, respectively.
The first function, RR_1, is the most general. It is based on the idea that R-indicators present the quality of response regardless of the estimators that the survey researcher is going to use and the population parameters that his or her survey is aiming at. In that setting the concepts nonresponse bias and contrast have little meaning and a lower bound can be based on the distribution of response propensities alone. A quality threshold γ may be derived by demanding that a specified proportion of the response propensities must have a maximal distance to the mean response propensity. More specifically, we may request that $100(1-\alpha)\%$ of the probability mass of the response propensities should be within a distance γ to the response rate ρ. For example if $\alpha = 0.05$ and $\gamma = 5\%$, we want 95% of the response propensities to be at most 5% away of the response rate. Clearly, we do not know the distribution that is underlying to the response propensities. We suggest, therefore, for the sake of simplicity, to assume that the propensities follow a normal distribution. Then the interval $[\rho - \xi_{1-0.5\alpha}S(\rho_X), \rho + \xi_{1-0.5\alpha}S(\rho_X)]$ contains $100(1-\alpha)\%$ of the probability mass. RR_1 follows easily from demanding that $\xi_{1-0.5\alpha}S(\rho_X) \le \gamma$.
RR_2 and RR_3 arise when it is demanded that the maximal bias and maximal contrast must not exceed a prescribed threshold γ. In the setting where the response quality of a single survey is evaluated, it becomes interesting to consider the estimators that are employed and the population parameters that are estimated. In many surveys the population parameters are population means or population totals. The maximal bias and maximal contrast then get a clear meaning; they reflect the quality of simple response means. RR_2 and RR_3 follow from $B_m(\rho_X) \le \gamma$ and $C_m(\rho_X) \le \gamma$, respectively. For instance, when $\gamma = 5\%$, we do not want the maximal absolute bias or the maximal absolute contrast to be bigger than 5%.

When other population parameters are targeted like population medians or population standard deviations, then other response-representativity functions may be more useful for quality assessment. We did not investigate such alternatives, however.

We will return to the response-representativity functions in section 4. The three functions can be used to plot traces of response rate and response representativeness over time or during data collection. As such they may be used to assess the number of days, weeks or months that is needed to get a response that satisfies a minimal quality level represented by the quality threshold.

2.5 Estimation of indicators and, contrast and bias

In the RISQ project we have proposed estimators for R, P_u, P_c, B_m and C_m. We refer to Shlomo et al (2009), Schouten et al (2009) and Skinner et al (2009) for the estimators and their details. The estimators replace population means by design-weighted sample and response means and response propensities by estimated propensities. Propensities are estimated by means of general linear models like linear regression, logistic regression or probit regression.

The estimators for the (partial) R-indicators, maximal bias and maximal contrast are random variables and depend on the sample. For this reason we have investigated the statistical properties of the estimators. They are described in detail in Shlomo et al (2009). Shlomo et al (2009) propose approximations to the standard errors that allow for computation of approximate confidence intervals. In section 4 we present R-indicators together with their confidence intervals.

Table 3.1: Description of household and business surveys.

Survey	Consumer Sentiments Survey (CSS) 2005	Health Survey (HS) 2005	Short-Term Statistics (STS) retail 2007	Short-Term Statistics (STS) industry 2007
Sample size	17,908	15,411	93,799	64,413
Response rate	66,9%	67,3%	49,5% (15days) 78,0% (30days) 85,8% (45days) 88,2% (60days)	48,8% (15days) 78,7% (30days) 85,7% (45days) 88,3% (60days)
Target population	Persons belonging to household core	Persons > 4 years	All businesses retail	All businesses industry
Design	Three stage design (municipality, address, person)	Two stage design (municipality, person)	Stratified design on size class and business type	Stratified design on size class and business type
Design weights	Equal	Equal	Unequal	Unequal
Fieldwork	10 days	30 days	90 days	90 days
Mode	CATI[1]	CAPI[2]	Web & paper	Web & paper

In the examples of this paper the auxiliary variable vector X is available at the sample level by means of direct linkage to frame data, registrations and administrative data. This is not feasible and realistic in many practical settings. Survey researchers may have access to population totals only. Within the RISQ project estimators based on population totals have been investigated. Skinner et al (2009) propose both sample-based and population-based estimators for response propensities and R-indicators. The population-based estimators employ population totals and no direct linkage is needed. Skinner et al (2009) distinguish two

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

situations: 1) all two-way interaction tables are available, 2) only marginal population tables are available. The first situation means that for instance the population tables age x gender, gender x type of household and age x type of household are available. The second situation refers to the setting where the frequency tables age, gender and type of household are known but no interactions are available.

3. Examples

We illustrate the possible uses of the indicators with two household surveys, two business surveys and one business register. The survey designs are summarized in table 3.1. For the household surveys we have the following auxiliary variables at the sample level: gender, age, marital status, urbanization, average value of houses in a postal code area, job status (yes or no a paid job), type of household and ethnicity. For the business surveys we could dispose over business type, business size and VAT reported to Tax Office in previous year.

Use 1: comparing the representativeness of two household surveys

Table 3.2 contains R-indicators and their corresponding 95% confidence intervals for the two selected household surveys. Response propensities are estimated using logistic regression with main effects only for are gender, age x marital status, urbanization, house value, paid job, household type, and ethnic background.

Table 3.2: R-indicators for the two household surveys HS 2005 and CSS 2005.

HS 2005	*CSS 2005*
R = 80,8% 95%CI = (79,4% – 82,3%)	R = 82,1% 95%CI = (80,7% – 83,4%)

The Consumer Sentiments survey performs slightly better than the Health survey, but the difference in R-indicator is not significant at the 5% level. One can, therefore, conclude that there is no evidence that the response to the two surveys differs strongly in terms of representativeness.

Use 2: evaluating the representativeness of a business register in time

We computed the R-indicator for the business register of VAT reports for the months January, June and December. Businesses have to report their VAT to the Tax Board on a monthly, quarterly or annual basis depending on their size. Small companies report only once a year while big companies have to submit VAT reports every month. Statistics Netherlands uses the VAT records as input to statistics about business turnover. For monthly statistics the VAT reports need to be available between 25 and 30 days after the end of the reference month. After 25 days processing data begins and after 30 days the statistics are made public. Since the reporting frequency depends on the size of the company, the months January, June and December are very different. For January only monthly reports are available, while for June and December also, respectively, the quarterly and annual reports can be used. We view the completion of the register as response and R-indicators as measures of the representativeness of available reports.

The completion rates and R-indicators are given in table 3.3. For the estimation of the completion probabilities we used VAT reported one year earlier in the same month and the

total wages of the reporting month. The total wages are also reported to the Tax Board and are available quickly after the end of the reporting month.

The completion rates are given after 25, 30 and 60 days. The completion rate for January is extremely low, only 20% of the businesses has submitted a tax report after 25 days. For June and December these rates are much higher. After 30 days more than 85% of the businesses has reported for December.

Table 3.3: The completion rate ρ, R-indicator and maximal bias for the VAT register of January, June and December after 25, 30 and 60 days of data collection.

	January			*June*			*December*		
	25 d	*30 d*	*60 d*	*25 d*	*30 d*	*60 d*	*25 d*	*30 d*	*60 d*
ρ	19,7%	26,1%	28,1%	64,1%	81,5%	83,2%	48,1%	84,1%	88,0%
$R(\rho)$	68,3%	60,4%	61,4%	73,9%	71,6%	73,1%	84,6%	76,9%	81,5%
B	80,4%	75,8%	68,7%	20,4%	17,4%	16,2%	16,0%	13,8%	10,5%

From table 3.3 we can conclude that the representativeness is lowest for January and highest for December. As the completion rate follows the same pattern, the maximal bias is highest for January and lowest for December. However, for each of the three months it does not pay off to wait longer than 25 days when it comes to representativeness.

Use 3: evaluating the representativeness of response during data collection

Table 3.4 contains R-indicators for the two business surveys for all available auxiliary variables and a restricted set where VAT is omitted. The R-indicators are given for response after 15, 30, 45 and 60 days of fieldwork. STS surveys need to provide statistics 30 days after the end of the reference month. The R-indicators show that for retail representativeness does not improve over time and is especially affected by VAT. The representativeness for industry improves over time and is only mildly related to VAT of the previous year.

Table 3.4: R-indicators, maximal bias and maximal contrast 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.

		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%
	C	15,8%	19,5%	24,6%	27,9%	19,0%	24,5%	28,2%	32,4%
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%
	C	7,8%	15,7%	24,6%	28,3%	23,8%	36,0%	47,7%	53,2%

Figure 3.1: RR_1 curves and response to STS industry and retail after 15, 30, 45 and 60 days for $\gamma = 0.05$ and $\gamma = 0.10$.

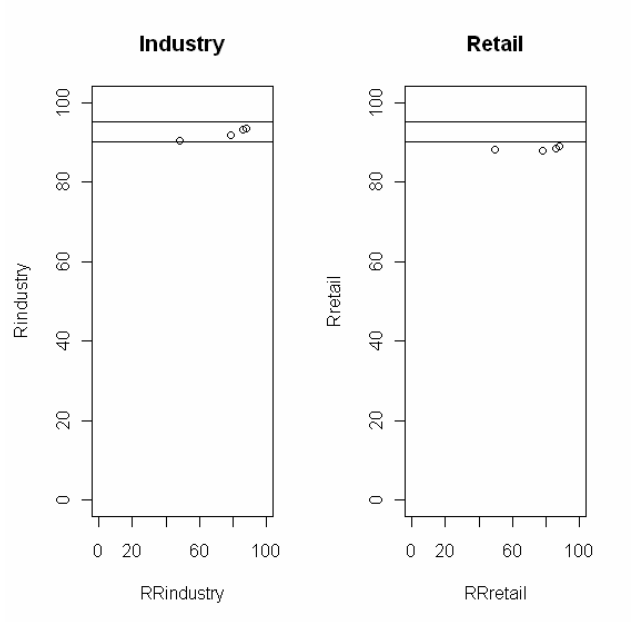
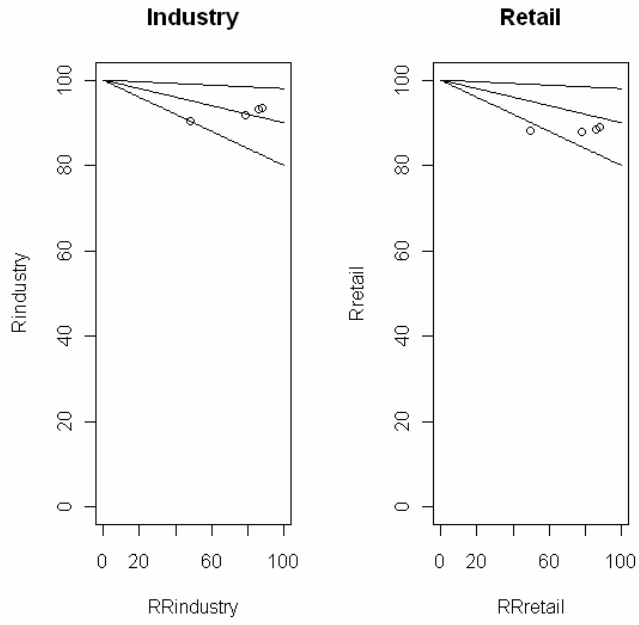


Figure 3.2: RR_2 curves and response to STS industry and retail after 15, 30, 45 and 60 days for $\gamma = 0.01$, $\gamma = 0.05$ and $\gamma = 0.10$.

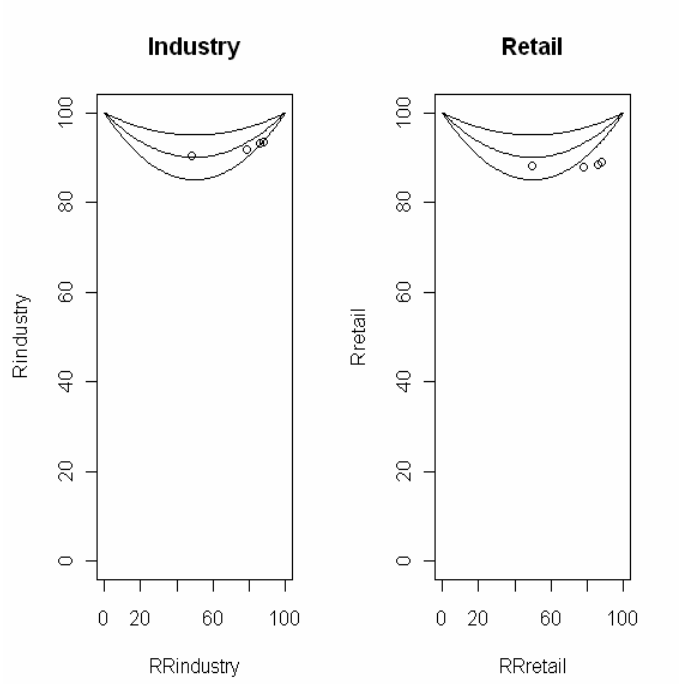


Figures 3.1 to 3.3 illustrate the response-representativity curves RR_1, RR_2 and RR_3 for response from the 2007 STS for Industry and Retail business using the extended model with busines type and business size x VAT. The response and R-indicator are taken from table 3.3 and are plotted for 15, 30, 45 and 60 days of fieldwork

For response in STS industry, the R-indicator is higher than the 10% RR_1 threshold after 15 days and is approaching the 5% RR_1 threshold after 60 days. For response in STS retail, the

R-indicator reaches the 10% RR_1 only after 60 days. RR_2 presents a similar picture for both surveys. The R-indicators for both STS industry and retail exceed the 10% RR_2 threshold after 30 days. However, both surveys never reach the 1% threshold and the STS retail does not reach the 5% after 60 days. The picture from RR_3 is different as quality is decreasing with the number of fieldwork days. The maximal contrast increases after 15 days. For STS industry it is approaches the 20% RR_3 level, while for STS retail it is considerably lower than the 20% threshold.

Figure 3.3: RR_3 curves and response tot STS industry and retail after 15, 30, 45 and 60 days for $\gamma = 0.05$, $\gamma = 0.10$ and $\gamma = 0.20$.



Figures 3.4 and 3.5 show unconditional and conditional partial R-indicators and the differences in maximal bias $\Delta B_m(X \mid Z)$ for the retail businesses after 15, 30, 45 and 60 days. The unconditional indicators are computed for business type and business size. The conditional indicator and differences in maximal bias are computed for business type only with respect to business size x VAT. Recall that unconditional partial R-indicators are always larger in size than the conditional partial R-indicators as for the conditional partial R-indicators the impact of the other variables is accounted for and removed. As such the conditional partial R-indicators reflect the impact of a single population characteristic adjusted for theo ther characteristics.

Figure 3.4 shows that the representativeness with respect to business type does not show a fixed pattern until 45 days. After 45 days the unconditional partial R-indicators are stable. Throughout the data collection business type 2 enterprises are overrepresented. However, business type 4 starts with a strong underrepresentattion but catches up after 30 days.

The unconditional partial R-indicators for business size are stable after 30 days and show that small businesses (GK equals 1) are strongly underrepresented.

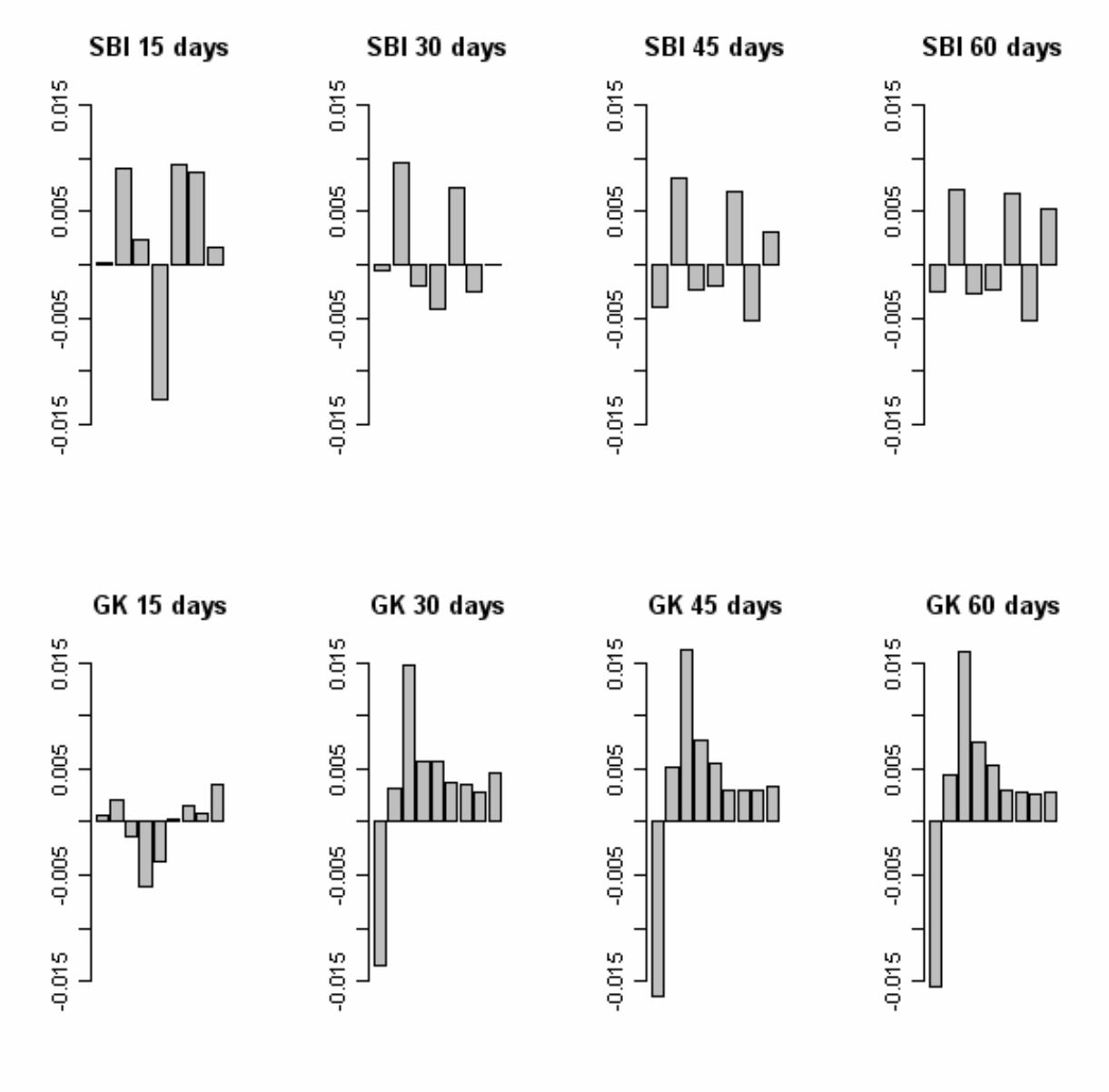
Figure 3.5 shows that conditioning on business size x VAT does not have a strong impact for businesses from type 5 but reduces the impact of businesses from types 2 and 6. Hence, one may conclude that type 5 businesses show a higher response even when conditioning for their size and turnover, while for type 2 and 6 businesses the overrepresentation and

underrepresentation is to some extent the result of the size and turnover composition of these subpopulations.

4. Discussion

From the examples in section 3 it becomes clear that R-indicators and partial R-indicators may be useful tools, but that they need to be evaluated carefully. First, the variables that are selected for the prediction of response play an important role. Second, the sample size reduces the strength of conclusions.

Figure 3.4: For t=15, 30, 45, 60 unconditional partial indicators for STS retail for Z = business type (SBI) and Z = business size (GK).

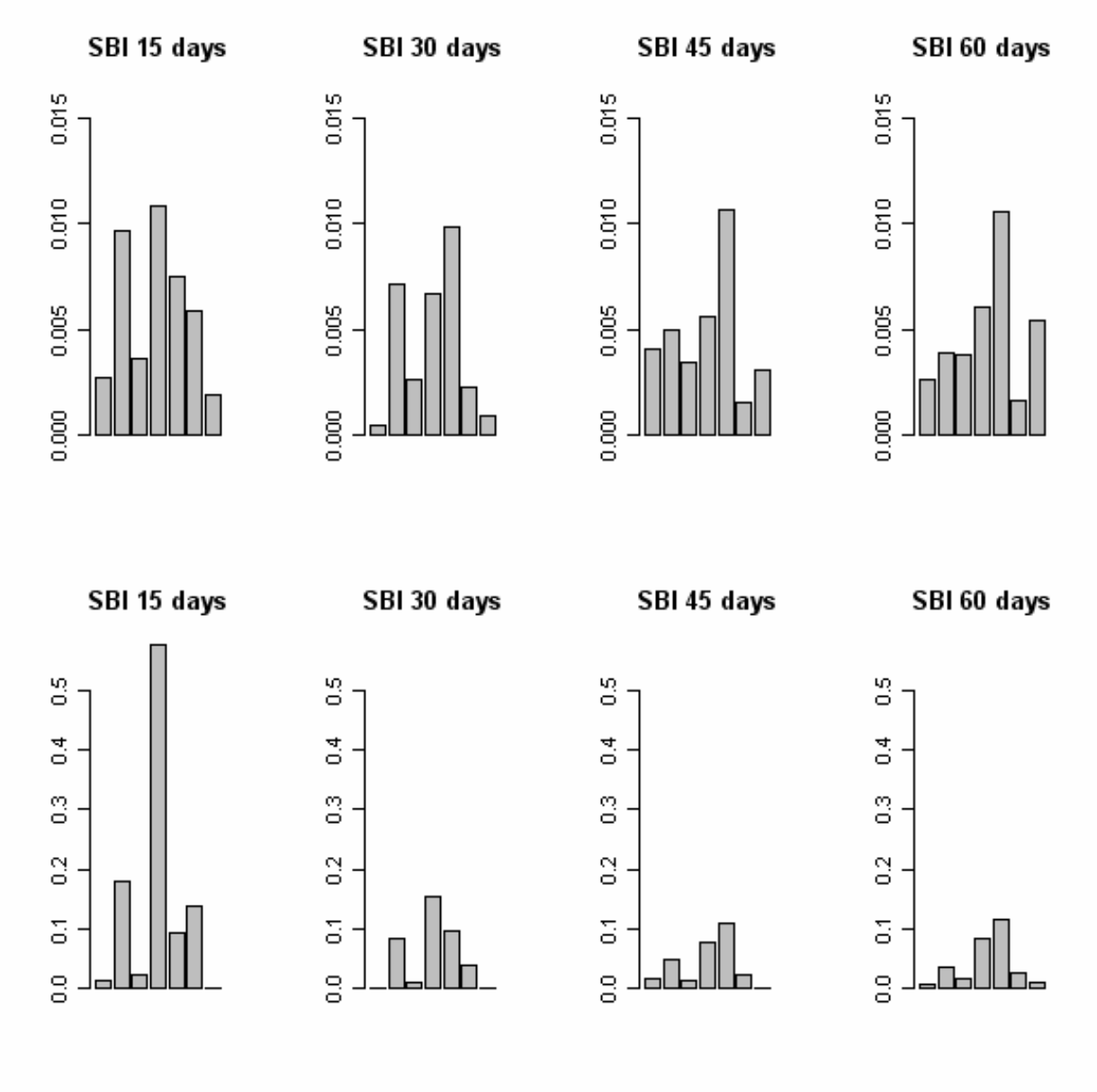


We have to restrict the assessment of representative response to available auxiliary variables. Only if we would dispose of a "super"vector $\aleph$, containing all relevant variables for explaining response behaviour, we would be able to apply the variation in response propensities $S(\rho_{\aleph})$, the maximal absolute bias $B_m(\rho_{\aleph})$ and the maximal absolute contrast

$C_m(\rho_{\aleph})$ to all possible survey items. For a specific choice of X, it implies that $B_m(\rho_X)$ and $C_m(\rho_X)$ may underestimate the true maximal bias and maximal contrast. As a consequence these measures need to be used with some care.

The interpretation of the R-indicator and the partial R-indicators is straightforward. They are based on response propensities which have a clear interpretation. It measures the (transformed) standard deviation of those propensities which is a measure that is commonly used in many statistical settings and its components, the between and within variance. The more diverse the response behaviour is, the larger the standard deviation.

Figure 3.5: For t=15, 30, 45, 60 conditional partial indicators and differences in maximal absolute bias for STS retail for Z = business type (SBI) and X = business size x VAT.



We note that lower and upper limits for each of the measures is dependent on the average response propensity $\bar{\rho}$. Suppose that we would fix $\bar{\rho}$, then

$$R(\rho_X) \geq 1 - 2\bar{\rho}(1-\bar{\rho})$$

$$B_m(\rho_X) \leq \sqrt{\frac{1-\overline{\rho}}{\overline{\rho}}}$$

$$C_m(\rho_X) \leq \frac{1}{\sqrt{\overline{\rho}(1-\overline{\rho})}}.$$

The lower limit for the R-indicator is smallest when $\overline{\rho} = 0{,}5$. The upper limits for the maximal bias and contrast are unbounded and increase when $\overline{\rho}$ gets smaller. The maximal bias and contrast are standardised by $S(y)$ in order to remove dependence on particular Y's. If we would not standardise, then the maximal bias and maximal contrast are bounded by, respectively, $1-\overline{\rho}$ and 1. However, for general Y both are unbounded as clearly bias and contrasts can be arbitrarily large for quantitative variables. It must be remarked that when $S(\rho_X)$ decreases and $\overline{\rho}$ increases, all measures will lead to the same conclusion.
The dependence of lower and upper limits on the average response propensity, i.e. the expected response rate, does not hamper the interpretation of the R-indicator itself. The interpretation in terms of the definition is still clear; it measures the amount of variation. When the average response propensity is closer to 0 or 1, then less variation is possible and response behaviour is becoming more and more similar. The dependence does play an important role in the normalizability of indicators.

The objective of RISQ is the development of measures that can be used irrespective of the set of survey variables, the population parameters or statistics that one is interested in, and the models that are used to explain response behaviour. The R-indicator corresponds to that goal. It does not depend on survey items, estimators or models. The partial R-indicators that are derived from the R-indicator correspond to variations in response propensities within and between subpopulations. In that sense they appeal directly to the practice in data collection departments as data collection strategies are usually improved based on groups that have a relatively low response propensity.

It would be natural to compare R-indicator values of different surveys based on bounds for the variation of response propensities. If we take a simplified view, then we may assume that nature selects response propensities according to a normal distribution. We may then conclude that approximately 100α% of the response propensities lays within $\xi_{1-0.5\alpha}S(\rho_X)$ from the average response propensity $\overline{\rho}$ (with $\xi_{1-0.5\alpha}$ representing the $100(1-\alpha/2)\%$ quantile from the standard normal distribution). If the R-indicator increases then the response propensities become more concentrated around the response rate, and as a consequence there is less reason for data collectioners to focus on specific subpopulations. Also it implies that in general the set of survey items and statistics that are potentially affected by nonresponse becomes smaller. If we do not restrict ourselves to a specific survey or a set of surveys, then obviously we have the full range of survey items and the full range of population parameters. Without such limitation it does not make sense to look at maximal bias or maximal contrast as these measures derive from the bias of response means. The response-representativity curve RR_1 was designed for this purpose.

However, if one considers a single survey or a restricted set of single surveys, it may be desirable to tailor the comparison and to take contrast and bias into account. Many household surveys aim at population means of categorical survey variables. Many business surveys aim at the population sum of quantitative variables. For these surveys one would be more

interested in the consequences for bias and contrast. The two other curves RR_2 and RR_3 can be used when one is especially interested in bias or contrast.

We make the following recommendations based on various analyses:

- 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 should be adjoined by a confidence interval.
- The number of selected auxiliary variables has only a mild effect on the size of confidence intervals for R-indicators.
- The inclusion of response-unrelated 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 restricting analysis to auxiliary variables for which it is known from the literature that they relate to response behaviour.
- 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 or the contrast of survey variables, and nor does the response rate. The maximal absolute bias combines the response rate and the R-indicator and is designed to make comparisons of non-response bias under worst case scenarios. The maximal absolute contrast does the same for the contrast 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. However, now the sets may also 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. Finally, maximal contrast or bias may be used rather than the R-indicator itself.
- In the comparison of different surveys, partial R-indicators are supplementary to R-indicators. Models for the estimation of response propensities are simple and employ general auxiliary variables only.
- In the comparison of a survey in time, partial R-indicators are again supplementary to R-indicators. Models for the estimation of response propensities may be more complex, e.g. define multiple model equations or levels, and may employ paradata additionally to auxiliary variables.
- In the monitoring of data collection, partial R-indicators assist in identifying groups that are underrepresented and may support decisions in responsive designs (Groves and Heeringa 2006, Mohl and Laflamme 2007, Wagner and Raghunathan 2007) or a change in future survey designs. Propensities may be modelled for different non-response types and data collection steps that produce missing data. Models may employ paradata additionally to auxiliary variables.
- In improving representativity of response it must always be the objective to increase the response rate and to decrease the variation in response propensities.

5. Future research

Future research within RISQ is dedicated to the elaboration and evaluation of partial R-indicators (paper scheduled for June 2009), the bias-correction of population-based R-

indicators (paper scheduled for July 2009), and the use of both types of indicators during data collection (papers scheduled for July and December 2009) with two pilots planned in October-December 2009. For monitoring response representativeness during survey data collection we will employ more advanced models that distinguish different causes for non-response and include fieldwork paradata. Papers are available at www.R-indicator.eu.

Acknowledgements: We thank the members of the RISQ team, Koen Beullens, Geert Loosveldt, Katja Rutar, Øyvin Kleven, Chris Skinner, Natalie Shlomo, and Li-Chun Zhang for their valuable input and comments.

References

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

Little, R.A., Rubin, D. (2002), Statistical analysis with missing data, Wiley Series in Probability and Statistics, Wiley: New Yorl, USA.

Madow, W.G., Olkin, I. (1983), Incomplete data in sample surveys, Proceedings of a Symposium, Academic Press, New York, USA.

Mohl, C., Laflamme, F (2007), Research and responsive design options for survey data collection at Statistics Canada, Proceedings of ASA Joint Statistical Meeting, Section 293, July 29 – August 2, Salt Lake City, USA.

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

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N., Skinner, C. (2009), How to use R-indicators?, RISQ deliverable, www.R-indicator.eu .

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J., Zhang, L.C. (2009a), Statistical properties of representativity indicators, RISQ deliverable, www.R-indicator.eu .

Shlomo, N., Skinner, N., Schouten, B., Carolina, N., Morren, M. (2009b), Partial indicators for representative response, RISQ deliverable, available at www.r-indicator.eu.

Shlomo, N., Skinner, C., Schouten, B., Heij, V. de, Bethlehem, J., Ouwehand, P. (2009c), Indicators for representative response based on population totals, RISQ deliverable, available at www.r-indicator.eu.

Wagner, J., Raghunathan, T. (2007), Bayesian approaches to sequential selection of survey design protocols, Proceedings of ASA Joint Statistical Meeting, Section 501, July 29 – August 2, Salt Lake City, USA.