Survey Methodology, June 2009
Vol. 35, No. 1, pp. 101-113
Statistics Canada, Catalogue No. 12-001-X

Indicators for the representativeness of survey response

Barry Schouten, Fannie Cobben and Jelke Bethlehem [1]

Abstract

Many survey organisations focus on the response rate as being the quality indicator for the impact of non-response bias. As a consequence, they implement a variety of measures to reduce non-response or to maintain response at some acceptable level. However, response rates alone are not good indicators of non-response bias. In general, higher response rates do not imply smaller non-response bias. The literature gives many examples of this (*e.g.*, Groves and Peytcheva 2006, Keeter, Miller, Kohut, Groves and Presser 2000, Schouten 2004).

We introduce a number of concepts and an indicator to assess the similarity between the response and the sample of a survey. Such quality indicators, which we call R-indicators, may serve as counterparts to survey response rates and are primarily directed at evaluating the non-response bias. These indicators may facilitate analysis of survey response over time, between various fieldwork strategies or data collection modes. We apply the R-indicators to two practical examples.

Key Words: Quality; Non-response; Non-response reduction; Non-response adjustment.

1. Introduction

It is a well-developed finding in the survey methodological literature that response rates by themselves are poor indicators of non-response bias, see *e.g.*, Curtin, Presser and Singer (2000), Groves, Presser and Dipko (2004), Groves (2006), Groves and Peytcheva (2006), Keeter *et al.* (2000), Merkle and Edelman (2002), Heerwegh, Abts and Loosveldt (2007) and Schouten (2004). However, the field has yet to propose alternative indicators of non-response that may be less ambiguous as indicators of survey quality.

We propose an indicator, which we call an R-indicator ('R' for representativeness), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as "representative response". In the literature, there are many different interpretations of the 'representativeness' concept. See Kruskal and Mosteller (1979a, b and c) for a thorough investigation of the statistical and non-statistical literature. Rubin (1976) introduced the concept of ignorable non-response; the minimal conditions that allow for unbiased estimation of a statistic. Some authors explicitly define representativeness. Hájek (1981) links "representative" to the estimation of population parameters; the pair formed by an estimator and a missing-data mechanism are representative when, with probability one, the estimator is equal to the population parameter. Following Hajèk's definition, calibration estimators (*e.g.*, Särndal, Swensson and Wretman 2003) are representative for the auxiliary variables that are calibrated. Bertino (2006) defines a so-called univariate representativeness index for continuous random variables. This index is a distribution-free measure based on the Cramér – Von Mises statistic. Kohler (2007) defines what he calls an internal criterion for representativeness. His univariate criterion resembles the Z-statistic for population means.

We separate the concept of representativeness from the estimation of a specific population parameter but relate this concept to the impact on the overall composition of response. By separating indicators from a specific parameter, they can be used as tools for comparing different surveys and surveys over time, and for a comparison of different data collection strategies and modes. Also, the measure gives a multivariate perspective of the dissimilarity between sample and response.

The R-indicator that we propose employs estimated response probabilities. The estimation of response probabilities implies that the R-indicator itself is a random variable, and, consequently, has a precision and possibly a bias. The sample size of a survey, therefore, plays an important role in the assessment of the R-indicator as we will show. However, this dependence exists for any measure; small surveys simply do not allow for strong conclusions about the missing-data mechanism.

We show that the proposed R-indicator relates to Cramèr's V measure for the association between response and auxiliary variables. In fact, we view the R-indicator as a lack-of-association measure. The weaker the association the better, as this implies there is no evidence that non-response has affected the composition of the observed data.

In order to be able to use R-indicators as tools for monitoring and comparing survey quality in the future, they need to have the features of a measure. That is, we want an R-indicator to be interpretable, measurable, able to be normalized and also to satisfy the mathematical properties of a measure. Especially since the interpretation and normalization are not straightforward features.

1. Barry Schouten, Fannie Cobben and Jelke Bethlehem, Statistics Netherlands, Department of Methodology and Quality, PO Box 4000, 2370 JM Voorburg, The Netherlands. E-mail: bstn@cbs.nl.

We apply the R-indicator to two studies that were conducted at Statistics Netherlands in 2005 and 2006. The objectives of those studies were the comparison of different data collection strategies. The studies involved different data collection modes and different non-response follow-up strategies. For each of the studies, a detailed analysis was done and documented. These studies are, therefore, suited to an empirical validation of the R-indicator. We compare the values of the R-indicator to the conclusions in the analyses. We refer to Schouten and Cobben (2007) and Cobben and Schouten (2007) for more illustrations and empirical investigations.

In section 2, we start with a discussion of the concept of representative response. Next, in section 3, we define the mathematical notation for our R-indicator. Section 4 is devoted to the features of the R-indicator. Section 5 describes the application of the R-indicator to the field studies. Finally, section 6 contains a discussion.

2. The concept of representative response

We, first, discuss what it means when a survey respondent pool is representative of the sample. Next, we make the concept of representativeness mathematically rigorous by giving it a definition.

2.1 What does representative mean?

Literature warns us not to single-mindedly focus on response rates as an indicator of survey quality. This can easily be illustrated by an example from the 1998 Dutch survey POLS (short for Permanent Onderzoek Leefsituatie or Integrated Survey on Household Living Conditions in English).

Table 1 contains the one- and two-month POLS survey estimates for the proportion of the Dutch population that receives a form of social allowance and the proportion that has at least one parent that was born outside the Netherlands. Both variables are taken from registry data and are artificially treated as survey items by deleting their values for non-respondents The sample proportions are also given in Table 1. After one month, the response rate was 47.2%, while after the full two-month interview period, the rate was 59.7%. In the 1998 POLS, the first month was CAPI (Computer Assisted Personal Interview). Non-respondents after the first month were allocated to CATI (Computer Assisted Telephone Interview) when they had a listed, land-line phone. Otherwise, they were allocated once more to CAPI. Hence, the second interview month gave another 12.5% of response. However, from table 1 we can see that after the second month, the survey estimates have a larger bias than after the first month.

Table 1
Response means in POLS for the first month of interviews and the full two-month interview period

Variable	After 1 month	After 2 months	Sample
Receiving social allowance	10.5%	10.4%	12.1%
Non-native	12.9%	12.5%	15.0%
Response rate	47.2%	59.7%	100%

From the example, it seems clear that the increased effort led to a less representative response with respect to both auxiliary variables. But what do we mean by representative in general?

It turns out that the term "representative" is often used with hesitation in the statistical literature. Kruskal and Mosteller (1979a, b and c) make an extensive inventory of the use of the word "representative" in the literature and identify nine interpretations. A number of interpretations they have found are omnipresent in the statistical literature. The statistical interpretations that Kruskal and Mosteller named 'absence of selective forces', 'miniature of the population', and 'typical or ideal cases' relate to probability sampling, quota sampling and purposive sampling. In the next section, we will propose a definition that corresponds to the 'absence of selective forces' interpretation. First, we will explain why we make this choice.

The concept of representative response is also closely related to the missing-data mechanisms Missing-Completely-at-Random (MCAR), Missing-at-Random (MAR) and Not-Missing-at-Random (NMAR) that are often referred to in the literature, see Little and Rubin (2002). A missing-data mechanism is MCAR when the probability of response does not depend on the survey topic of interest. The mechanism is MAR if the response probability depends on observed data only, which is, hence, a weaker assumption than MCAR. If the probability depends on missing data also, then the mechanism is said to be NMAR. These mechanisms, in fact, find their origin in model-based statistical theory. Somewhat loosely interpreted with respect to a survey topic, MCAR means that respondents are on average the same as non-respondents, MAR means that within known subpopulations, respondents are on average the same as non-respondents, and NMAR implies that even within subpopulations, respondents are different. The addition of the survey topic is essential. Within one questionnaire, some survey items can be MCAR, while other items are MAR or NMAR. Furthermore, the MAR assumption for one survey item holds for a particular stratification of the population. A different item may need a different stratification.

Given that we wish to monitor and compare the response to different surveys in topic or time, it is not appealing to define a representative response as dependent on the survey topic itself nor as dependent on the estimator used. We focus instead on the quality of data collection and not on the estimation. This setting leads us to compare the response composition to that of the sample. Clearly, the survey topics influence the probability that households participate in the survey, but the influence cannot be measured or tested and, hence, from our perspective, this influence cannot be the input for assessing response quality. We propose to judge the composition of response by pre-defined sets of variables that are observed outside of the survey and can be employed for each survey under investigation. We want the respondent selection to be as close as possible to a 'simple random sample of the survey sample', *i.e.*, with as little relation as possible between response and characteristics that distinguish units from each other. The latter can be interpreted as having selective forces which are absent in the selection of respondents, or as MCAR with respect to all possible survey variables.

2.2 Definition of a representative response subset

Let $i = 1, 2, 3, \ldots, N$ be the unit labels for the population. By s_i we denote the 0-1-sample indicator, *i.e.*, when unit i is sampled, it takes the value 1 and 0 otherwise. By r_i we denote the 0-1-response indicator for unit i. If unit i is sampled and did respond then $r_i = 1$. It is 0 otherwise. The sample size is n. Finally, π_i denotes the first-order inclusion probability of unit i.

The key to our definitions lies in the individual response propensities. Let ρ_i be the probability that unit i responds when it is sampled.

The interpretation of a response propensity is not straightforward by itself. We follow a model-assisted approach, *i.e.*, the only randomness is in the sample and response indicators. A response probability is a feature of a labelled and identifiable unit, a biased coin that the unit carries in a pocket, so to speak, and is, therefore, inseparable from that unit. With a little effort, however, all concepts can be translated into a model-based context.

First, we give a strong definition.

Definition (*strong*): *A response subset is representative with respect to the sample if the response propensities ρ_i are the same for all units in the population*

$$\rho_i = P[r_i = 1 \mid s_i = 1] = \rho, \quad \forall i, \tag{1}$$

and if the response of a unit is independent of the response of all other units.

If a missing-data mechanism would satisfy the strong definition, then the mechanism would correspond to Missing-Completely-at-Random (MCAR) with respect to all possible survey questions. Although the definition is appealing, the validity of it can never be tested in practice. We have no replicates of the response of one single unit. We, therefore, also construct a weak definition that can be tested in practice.

Definition (*weak*): *A response subset is representative of a categorical variable X with H categories if the average response propensity over the categories is constant*

$$\bar{\rho}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} \rho_{hk} = \rho, \quad \text{for} \quad h = 1, 2, \ldots, H, \tag{2}$$

where N_h is the population size of category h, ρ_{hk} is the response propensity of unit k in class h and summation is over all units in this category.

The weak definition corresponds to a missing-data mechanism that is MCAR with respect to X, as MCAR states that we cannot distinguish respondents from non-respondents based on knowledge of X.

3. R-indicators

In the previous section, we defined strong and weak representative response. Both definitions make use of individual response probabilities that are unknown in practice. First, we start with a population R-indicator. From there on, we base the same R-indicator on a sample and on estimated response propensities.

3.1 Population R-indicators

We first consider the hypothetical situation where the individual response propensities are known. Clearly, in that case we can even test the strong definition and we simply want to measure the amount of variation in the response propensities; the more variation, the less representative in the strong sense. Let $\rho = (\rho_1, \rho_2, \ldots, \rho_N)'$ be a vector of response propensities, let $\mathbf{1} = (1, 1, \ldots, 1)'$ be the N-vector of ones, and let $\rho_0 = \mathbf{1} \times \bar{\rho}$ be the vector consisting of the average population propensity.

Any distance function d in $[0, 1]^N$ would suffice in order to measure the deviation from a strong representative response by calculating $d(\rho, \rho_0)$. Note that the height of the overall response does not play a role. The Euclidean distance is a straightforward distance function. When applied to a distance between ρ and ρ_0, this measure is proportional to the standard deviation of the response probabilities

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

It is not difficult to show that

$$S(\rho) \leq \sqrt{\bar{\rho}(1-\bar{\rho})} \leq \frac{1}{2}. \qquad (4)$$

We want the R-indicator to take values on the interval [0, 1] with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness. We propose the R-indicator R, which is defined by

$$R(\rho) = 1 - 2S(\rho). \qquad (5)$$

Note that the minimum value of (5) depends on the response rate, see figure 1. For $\bar{\rho} = 1/2$, it has a minimum value of 0. For $\bar{\rho} = 0$ and $\bar{\rho} = 1$, clearly no variation is possible and the minimum value is 1. Paradoxically, the lower bound increases when the response rate decreases from $1/2$ to 0. For a low response rate, there is less room for individual response propensities to have a large variation.

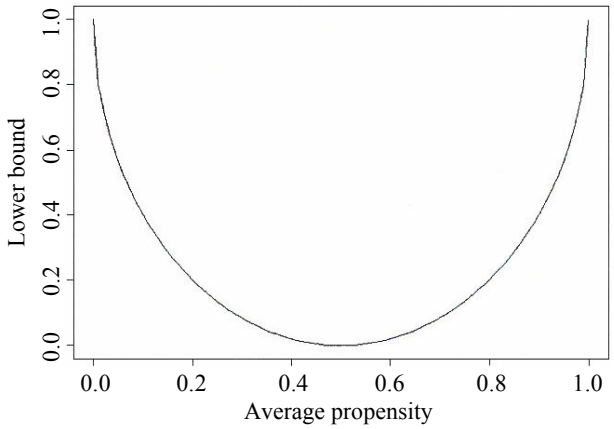

Figure 1 Minimum value of R-indicator (5) as a function of the average response propensity

One may view R as a lack of association measure. When $R(\rho) = 1$ there is no relation between any survey item and the missing-data mechanism. We show that R in fact has a close relation to the well-known χ^2-statistic that is often used to test independence and goodness-of-fit.

Suppose that the response propensities are only different for classes h defined by a categorical variable X. Let $\bar{\rho}_h$ and f_h be, respectively, the response propensity and the population function of class h, *i.e.*,

$$f_h = \frac{N_h}{N}, \quad \text{for} \quad h = 1, 2, \ldots, H. \qquad (6)$$

Hence, for all i with $X_i = h$ the response propensity is $\rho_i = \bar{\rho}_h$.

Since the variance of the response propensities is the sum of the 'between' and 'within' variances over classes h, and the within variances are assumed to be zero, it holds that

$$S^2(\tilde{\rho}) = \frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2$$

$$= \frac{N}{N-1}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2 \approx \sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2. \qquad (7)$$

The χ^2-statistic measures the distance between observed and expected proportions. However, it is only a true distance function in the mathematical sense for fixed marginal distributions f_h and $\bar{\rho}$. We can apply the χ^2-statistic to X in order to 'measure' the distance between the true response behaviour and the response behaviour that is expected when response is independent of X. In other words, we measure the deviation from weak representativeness with respect to X.

We can rewrite the χ^2-statistic to get

$$\chi^2 = \sum_{h=1}^{H} \frac{(N_h\bar{\rho}_h - N_h\bar{\rho})^2}{N_h\bar{\rho}}$$

$$+ \sum_{h=1}^{H} \frac{(N_h(1-\bar{\rho}_h) - N_h(1-\bar{\rho}))^2}{N_h(1-\bar{\rho})}$$

$$= \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{\bar{\rho}} + \sum_{h=1}^{H} \frac{Nf_h(\bar{\rho}_h - \bar{\rho})^2}{(1-\bar{\rho})}$$

$$= \frac{N}{\bar{\rho}(1-\bar{\rho})}\sum_{h=1}^{H} f_h(\bar{\rho}_h - \bar{\rho})^2$$

$$= \frac{N-1}{\bar{\rho}(1-\bar{\rho})} S^2(\tilde{\rho}). \qquad (8)$$

An association measure that transforms the χ^2-statistic to the [0, 1] interval, see *e.g.*, Agresti (2002), is Cramèr's V

$$V = \sqrt{\frac{\chi^2}{N(\min\{C, R\} - 1)}}, \qquad (9)$$

where C and R are, respectively, the number of columns and rows in the underlying contingency table. Cramèr's V attains a value 0 if observed proportions exactly match expected proportions and its maximum is 1. In our case, the denominator equals N since the response indicator has only two categories: response and non-response. As a consequence, (9) changes into

$$V = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{N-1}{N\bar{\rho}(1-\bar{\rho})}}\, S(\tilde{\rho}). \qquad (10)$$

From (10) we can see that for large N, Cramèr's V is approximately equal to the standard deviation of the response propensities standardized by the maximal standard deviation $\sqrt{\bar{\rho}(1-\bar{\rho})}$ for a fixed average response propensity $\bar{\rho}$.

3.2 Response-based R-indicators

In section 3.1, we assumed that we know the individual response propensities. Of course, in practice these propensities are unknown. Furthermore, in a survey, we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R. An obvious way to do this is to use response-based estimators for the individual response propensities and the average response propensity.

We let $\hat{\rho}_i$ denote an estimator for ρ_i which uses all or a subset of the available auxiliary variables. Methods that support such estimation are, for instance, logistic or probit regression models (Agresti 2002) and CHAID classification trees (Kass 1980). By $\hat{\bar{\rho}}$ we denote the weighted sample average of the estimated response propensities, *i.e.*,

$$\hat{\bar{\rho}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_i \frac{s_i}{\pi_i}, \qquad (11)$$

where we use the inclusion weights.

We replace R by the estimators $\hat{R}$

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\bar{\rho}})^2}. \qquad (12)$$

Note that in (12) there are in fact two estimation steps based on different probability mechanisms. The response propensities themselves are estimated and the variation in the propensities is estimated. We return to the consequences of the two estimation steps in section 4.

3.3 Example

We apply the proposed R-indicators to the survey data from the 1998 POLS that we described in section 2.1. Recall that the survey was a combination of face-to-face and telephone interviewing in which the first month was CAPI only. The sample size was close to 40,000 and the response rate was approximately 60%. We linked the fieldwork administration to the sample and deduced whether each contact attempt resulted in a response. This way, we can monitor the pattern of the R-indicator during the fieldwork period.

For the estimation of response rates we used a logistic regression model with region, ethnic background and age as independent variables. Region was a classification with 16 categories, the 12 provinces and the four largest cities – Amsterdam, Rotterdam, The Hague and Utrecht – as separate categories. Ethnic background has seven categories: native, Moroccan, Turkish, Surinam, Dutch Antilles, other non-western non-native and other western non-native. The classification is based on the country of birth of the parents of the selected person. The age variable has three categories: 0 – 34 years, 35 – 54 years, and 55 years and older.

In figure 2, $\hat{R}$ is plotted against the response rate for the first six contact attempts in POLS. The leftmost value corresponds to the respondent pool after one attempt was made. For each additional attempt, the response rate increases but the indicator shows a drop in representativeness. This result confirms the findings in Schouten (2004).

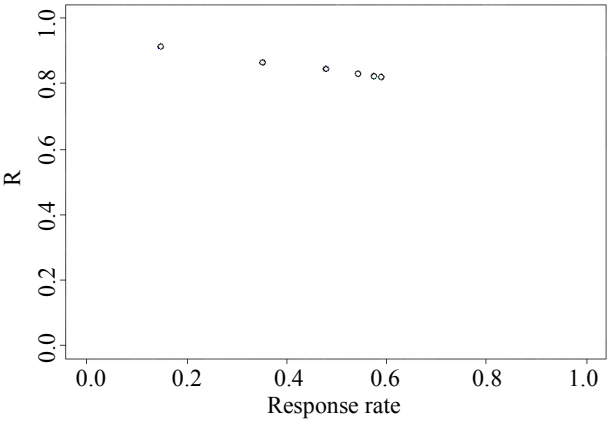

Figure 2 R-indicator for first six contact attempts in POLS 1998

4. Features of R-indicators

In section 3, we propose a candidate indicator for representativeness. However, other indicators can be constructed. There are many association measures or fit indexes, *e.g.*, Goodman and Kruskal (1979), Bentler (1990) and Marsh, Balla and McDonald (1988). Association measures have a strong relation to R-indicators. Essentially, R-indicators attempt to measure in a multivariate setting the lack of association. In this section, we discuss the desired features of R-indicators. We show that the proposed R-indicator R allows for a straightforward upper bound on the non-response bias.

4.1 Features in general

We want R-indicators to be based on a distance function or metric in the mathematical sense. The triangle inequality property of a distance function allows for a partial ordering of the variation in response propensities which enables interpretation. A distance function can easily be derived from any mathematical norm. In section 3, we chose to use the Euclidean norm as this norm is commonly used. The Euclidean norm led us to an R-indicator that uses the standard deviation of response propensities. Other norms, like the supremum norm, would lead us to alternative distance functions. In section 4.3, however, we show that the Euclidean norm based R-indicators have interesting normalization features.

We must make a subtle distinction between R-indicators and distance functions. Distance functions are symmetric while an R-indicator measures a deviation with respect to a specific point, namely the situation where all response propensities are equal. If we change the vector of individual propensities, then this point is in most cases shifted. However, if we fix the average response propensity, then the distance function facilitates interpretation.

Apart from a relation to a distance function, we want to be able to measure, interpret and normalize the R-indicators. In section 3.2, we already derived response-based estimators for 'population' R-indicators that are not measurable when response propensities are unknown and all we have is the response to a survey. Hence, we made R-indicators measurable by switching to estimators. The other two features are discussed separately in the next two sections.

4.2 Interpretation

The second feature of R-indicators is the ease with which we can interpret their values and the concept they are measuring. We moved to an estimator for an R-indicator that is based on the samples of surveys and on estimators of individual response probabilities. Both have far-reaching consequences for the interpretation and comparison of the R-indicator.

Since the R-indicator is an estimator itself, it is also a random variable. This means that it depends on the sample, *i.e.*, it is potentially biased and has a certain accuracy. But what is it estimating?

Let us first assume that the sample size is arbitrarily large so that precision does not play a role and also suppose the selection of a model for response propensities is no issue. In other words, we are able to fit any model for any fixed set of auxiliary variables.

There is a strong relation between the R-indicator and the availability and use of auxiliary variables. In section 2, we defined strong and weak representativeness. Even in the case where we are able to fit any model, we are not able to estimate response propensities beyond the 'resolution' of the available auxiliary variables. Hence, we can only draw conclusions about weak representativeness with respect to the set of auxiliary variables. This implies that whenever an R-indicator is used, it is necessary to complement its value by the set of covariates that served as a grid to estimate individual response propensities. If the R-indicator is used for comparative purposes, then those sets must be the same. We must add that it is not necessary for all auxiliary variables to be used for the estimation of propensities, since they may not add any explanatory power to the model. However, the same sets should be available. The R-indicator then measures a deviation from weak representativeness.

The R-indicator does not capture differences in response probabilities within subgroups of the population other than the subgroups defined by the classes of X. If we let $h = 1, 2, \ldots, H$ again denote strata defined by X, N_h be the size of stratum h, and $\bar{\rho}_h$ be the population average of the response probabilities in stratum h, then it is not difficult to show that $\hat{R}$ is a consistent estimator of

$$R_X(\rho) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{h=1}^{H} N_h(\bar{\rho}_h - \bar{\rho})^2}, \qquad (13)$$

when standard models like logistic regression or linear regression are used to estimate the response probabilities. Of course, (13) and (5) may be different.

In practice, the sample size is not arbitrarily large. The sample size affects both estimation steps; the estimation of response propensities and the estimation of the R-indicator using a sample.

If we knew the individual response propensities, then the sample-based estimation of the R-indicator would only lead to variance and not to bias. We would be able to estimate the population R-indicator without bias. Hence, for small sample sizes, the estimators would have a small precision which could be accounted for by using confidence intervals instead of merely point estimators.

The implications for the estimation of response probabilities are, however, different because of model selection and model fit. There are two alternatives. Either one imposes a model to estimate propensities fixing the covariates beforehand, or one lets the model be dependent on the significant contribution of covariates with respect to some predefined level. In the first case, again no bias is introduced but the standard error may be affected by over fitting. In the second case, the model for the estimation of response propensities depends on the size of the sample; the larger the sample, the more interactions that are accepted as significant. Although it is standard statistical practice to fit models based on a significance level, model selection may introduce bias and variance to the estimation of any R-indicator. This can be easily understood by going to the extreme of a sample of, say, size 10. For such a small sample, no interaction between response behaviour and auxiliary characteristics will be accepted, leaving an empty model and an estimated R-indicator of 1. Small samples simply do not allow for the estimation of response propensities. In general, a smaller sample size will, thus, lead to a more optimistic view on representativeness.

We should make a further subtle distinction. It is possible that, for one survey, a lot of interactions contribute to the prediction of response propensities but each one contributes very little, while in another survey there is only one but it strongly contributes a single interaction. None of the small contributions may be significant, but together they are as

strong as the one large contribution that is significant. Hence, we would be more optimistic in the first example even if sample sizes would be comparable.

These observations show that one should always use an R-indicator with some care. It cannot be viewed as separate from the auxiliary variables that were used to compute it. Furthermore, the sample size has an impact on both bias and precision.

4.3 Normalization

The third important feature is the normalization of an R-indicator. We want to be able to attach bounds to an R-indicator so that the scale of an R-indicator, and, hence, changes in the R-indicator get a meaning. Clearly, the interpretation issues that we raised in the previous section also affect the normalization of the R-indicator. Therefore, in this section we assume the ideal situation where we can estimate response propensities without bias. This assumption holds for large surveys. We discuss the normalization of the R-indicator $\hat{R}$.

4.3.1 Maximal absolute bias and maximal root mean square error

We show that for any survey item Y, the R-indicator can be used to set upper bounds to the non-response bias and to the root mean square error (RMSE) of adjusted response means. We use these bounds of the R-indicator to show the impact under worst-case scenarios.

Let Y be some variable that is measured in a survey and let $\hat{\bar{y}}_{\text{HT}}$ be the Horvitz-Thompson estimator for the population mean based on the survey response. It can be shown (*e.g.*, Bethlehem 1988, Särndal and Lundström 2005) that its bias $B(\hat{\bar{y}}_{\text{HT}})$ is approximately equal to

$$B(\hat{\bar{y}}_{\text{HT}}) = \frac{C(y,\, \rho)}{\bar{\rho}}, \tag{14}$$

with $C(y,\, \rho) = 1/N \sum_{i=1}^{N} (y_i - \bar{y})\,(\rho_i - \bar{\rho})$ the population covariance between the survey items and the response probabilities. For a close approximation of the variance $s^2(\hat{\bar{y}}_{\text{HT}})$ of $\hat{\bar{y}}_{\text{HT}}$ we refer to Bethlehem (1988).

A normalization of R is found by the Cauchy-Schwarz inequality. This inequality states that the covariance between any two variables is bounded in absolute sense by the product of the standard deviations of the two variables. We can translate this to bounds for the bias (14) of $\hat{\bar{y}}_{\text{HT}}$

$$\begin{aligned} |B(\hat{\bar{y}}_{\text{HT}})| &\leq \frac{S(\rho)\,S(y)}{\bar{\rho}} = \frac{(1 - R(\rho))\,S(y)}{2\bar{\rho}} \\ &= B_m(\rho,\, y). \end{aligned} \tag{15}$$

Clearly, we do not know the upper bound $B_m(\rho,\, y)$ in (15) but we can estimate it using the sample and the estimated response probabilities. We denote the estimator by $\hat{B}_m(\hat{\rho},\, y)$.

In a similar way, we can set a bound to the root mean square error (RMSE) of $\hat{\bar{y}}_{\text{HT}}$. It holds approximately that

$$\begin{aligned} \text{RMSE}(\hat{\bar{y}}_{\text{HT}}) &= \sqrt{B^2(\hat{\bar{y}}_{\text{HT}}) + s^2(\hat{\bar{y}}_{\text{HT}})} \\ &\leq \sqrt{B_m^2(\rho,\, y) + s^2(\hat{\bar{y}}_{\text{HT}})} \\ &= E_m(\rho,\, y). \end{aligned} \tag{16}$$

Again, we do not know $E_m(\rho,\, y)$. Instead, we use the sample-based estimator that employs the estimated response probabilities, denoted by $\hat{E}_m(\hat{\rho},\, y)$.

The bounds $\hat{B}_m(\hat{\rho},\, y)$ and $\hat{E}_m(\hat{\rho},\, y)$ are different for each survey item y. For comparison purposes it is, therefore, convenient to define a hypothetical survey item. We suppose that $\hat{S}(y) = 0.5$. The corresponding bounds we denote by $\hat{B}_m(\hat{\rho})$ and $\hat{E}_m(\hat{\rho})$. They are equal to

$$\hat{B}_m(\hat{\rho}) = \frac{(1 - \hat{R}(\hat{\rho}))}{4\hat{\bar{\rho}}} \tag{17}$$

$$\hat{E}_m(\hat{\rho}) = \sqrt{\hat{B}_m^2(\hat{\rho}) + \hat{s}^2(\hat{\bar{y}}_{\text{HT}})}. \tag{18}$$

We compute (17) and (18) in all studies described in section 5. We have to note that (17) and (18) are again random variables that have a certain precision and that are potentially biased.

4.3.2 Response-representativeness functions

In the previous section, we used the R-indicator to set upper bounds to the non-response bias and to the root mean square error of the (adjusted) response mean. Conversely, we may set a lower bound to the R-indicator by demanding that either the absolute non-response bias or the root mean square error is smaller than some prescribed value. Such a lower bound may be chosen as one of the ingredients of quality restrictions put upon the survey data by a user of the survey. If a user does not want the non-response bias or root mean square to exceed a certain value, then the R-indicator must be bigger than the corresponding bound.

Clearly, lower bounds to the R-indicator depend on the survey item. Therefore, again we restrict ourselves a hypothetical survey item for which $\hat{S}(y) = 0.5$.

It is not difficult to show from (17) that if we demand that

$$\hat{B}_m(\hat{\rho}) \leq \gamma, \tag{19}$$

then it must hold that

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\gamma = r_1(\gamma, \hat{\rho}). \quad (20)$$

Analogously, using (18) and demanding that

$$\hat{E}_m(\hat{\rho}) \leq \gamma, \quad (21)$$

we arrive at

$$\hat{R} \geq 1 - 4\hat{\bar{\rho}}\sqrt{\gamma^2 - \hat{s}^2(\hat{\bar{y}}_{\mathrm{HT}})} = r_2(\gamma, \hat{\rho}). \quad (22)$$

In (20) and (22) we let $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ denote lower limits to the R-indicator. In the following section, we refer to $r_1(\gamma, \hat{\rho})$ and $r_2(\gamma, \hat{\rho})$ as response-representativeness functions. We compute them for the studies in section 5.

4.3.3 Example

We again illustrate the normalization with the same example used in sections 2.1 and 3.3. Figure 3 contains the response-representativeness function $r_1(\gamma, \hat{\rho})$ and the observed R-indicators $\hat{R}$ for the six contact attempts in POLS 1998. Three values of γ are chosen, $\gamma = 0.1$; $\gamma = 0.075$ and $\gamma = 0.05$.

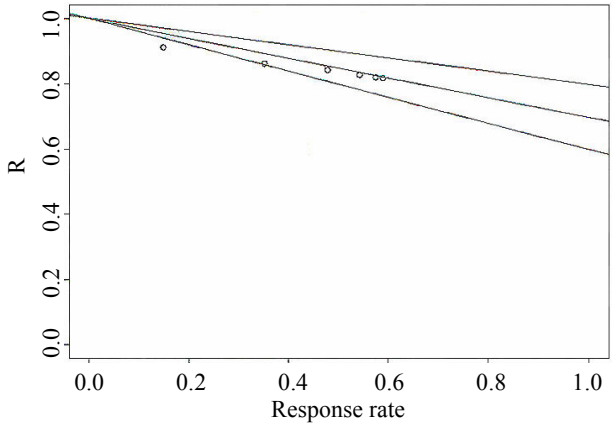

Figure 3 Lower bounds for R-indicator $\hat{R}$ for the first six contact attempts of POLS 1998. Lower bounds are based on $\gamma = 0.1$, $\gamma = 0.075$ and $\gamma = 0.05$

Figure 3 indicates that after the second contact attempt, the values of the R-indicator exceed the lower bound corresponding to the 10%-level. After four attempts, the R-indicator is close to the 7.5%-level. However, the values never exceed the other lower bound that is based on the 5%-level.

In figure 4, the maximal absolute bias $\hat{B}_m(\hat{\rho})$ is plotted against the response rate of the six contact attempts. After the third contact attempt, the R-indicator has converged on a value around 8%.

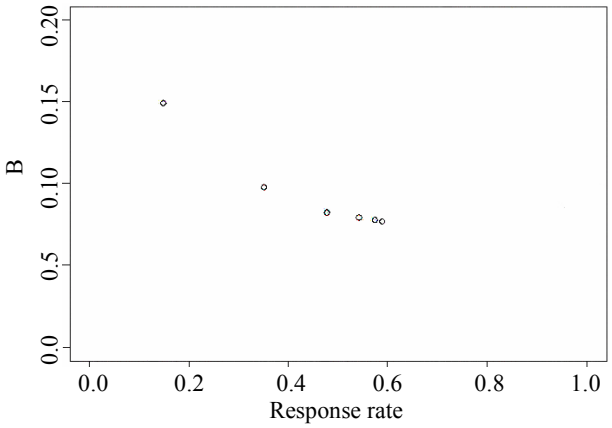

Figure 4 Maximal absolute bias for the first six contact attempts of POLS 1998

5. Application of the R-indicator

In this section, we apply the R-indicator to two studies that investigate different non-response follow-up strategies and different combinations of data collection modes. The first study involves the Dutch Labour Force Survey (LFS). The study is an investigation of both the call-back approach (Hansen and Hurwitz 1946) and the basic-question approach (Kersten and Bethlehem 1984). The second study deals with mixed-mode data collection designs applied to the Dutch Safety Monitor survey.

In sections 5.2 and 5.3 we take a closer look at the studies in connection with the representativeness of their different fieldwork strategies. First, in section 5.1 we describe how we approximate standard errors.

5.1 Standard error and confidence interval

If we want to compare the values of the R-indicator for different surveys or data collection strategies, we need to estimate their standard errors.

The R-indicator $\hat{R}$ involves the sample standard deviation of the estimated response probabilities. This means that there are two random processes involved. The first process is the sampling of the population. The second process is the response mechanism of the sampled units. If the true response probabilities were known, then drawing a sample would still introduce uncertainty about the population R-indicator and, hence, lead to a certain loss of precision. However, since we do not know the true response probabilities, these probabilities are estimated using the sample. This introduces additional precision loss.

An analytical derivation of the standard error of $\hat{R}$ is not straightforward due to the estimation of the response probabilities. In this paper, we are resigned to naïve numerical

approximations of the standard error. We estimate the standard error of the R-indicator by non-parametric bootstrapping (Efron and Tibshirani 1993). The non-parametric bootstrap estimates the standard error of the R-indicator by drawing a number $b = 1, 2, \ldots, B$ of so-called bootstrap samples. These are samples drawn independently and with replacement from the original dataset, of the same size n as the original dataset. The R-indicator is calculated for every bootstrap sample b. We thus obtain B replications of the R-indicator; $\hat{R}_b^{\mathrm{BT}}$, $b = 1, 2, \ldots, B$. The standard error for the empirical distribution of these B replications is an estimate for the standard error of the R-indicator, that is

$$s_R^{\mathrm{BT}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{R}_b^{\mathrm{BT}} - \hat{\bar{R}}^{\mathrm{BT}})^2} \qquad (23)$$

where $\hat{\bar{R}}^{\mathrm{BT}} = 1/B \sum_{b=1}^{B} \hat{R}_b^{\mathrm{BT}}$ is the average estimated R-indicator.

In the approximations, we take $B = 200$ for all studies. We experimented with larger numbers of B of up to $B = 500$, but found that in all cases, the estimate of the standard error had converged by $B = 200$.

We determine $100(1-\alpha)\%$ confidence intervals by assuming a normal approximation of the distribution of $\hat{R}$ employing the estimated standard errors using (23)

$$CI_\alpha^{\mathrm{BT}} = (\hat{R} \pm \xi_{1-\alpha} \times s_R^{\mathrm{BT}}) \qquad (24)$$

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

5.2 Labour Force Survey; follow-up study 2005

From July to December 2005, Statistics Netherlands conducted a large-scale follow-up of non-respondents in the Dutch Labour Force Survey (LFS). In the study, two samples of non-respondents in the LFS were approached once more using either a call-back approach (Hansen and Hurwitz 1946) or a basic-question approach (Kersten and Bethlehem 1984). The samples consisted of LFS households that refused, were not processed or were not contacted in the LFS for the months July – October. In the design of the follow-up study, we used the recommendations in the studies by Stoop (2005) and Voogt (2004).

The main characteristics of the call-back and basic-question approaches applied to the LFS are given in Table 2. For more details, we refer to Schouten (2007) and Cobben and Schouten (2007). The call-back approach employed the original household questionnaire in CAPI, while the basic-question approach used short questionnaires in a mixed-mode setting. The mixed-mode design involved web, paper and CATI. CATI was used for all households with a listed phone number. Households without a listed phone number received an advance letter, a paper questionnaire and a login to a secure website containing the web questionnaire. Respondents were left the choice to fill in either the paper or web questionnaire.

Table 2
Characteristics of the two approaches in the follow-up study

Call-back approach	Basic-question approach
• LFS questionnaire to be answered by all members of the household in CAPI	• A strongly condensed questionnaire with key questions of the LFS which takes between 1 and 3 minutes to answer or fill in
• 28 interviewers geographically selected from historically best-performing interviewers	• Mixed-mode data collection design using web, paper and CATI
• Interviewer was different from interviewer that received non-response	• The questionnaire was to be answered by one person per household following the next birthday method
• Interviewers received additional training in doorstep interaction	• The timing is one week after the household is processed as a non-response
• Extended fieldwork period of two months	
• Interviewer could offer incentives	
• Interviewers could receive a bonus	
• A paper summary of the characteristics of the non-responding household was sent to the interviewer	
• Allocation of address one week after non-response	

The sample size of the LFS pilot was $n = 18{,}074$ households, of which 11,275 households responded. The non-responding households were stratified according to the cause of non-response. Households that were not processed or contacted, and households that refused were eligible for a follow-up. It was considered to be unethical to follow-up households that did not respond due to other causes like illness. In total, 6,171 households were eligible. From these households, two simple random samples were drawn of size 775. In the analyses, the non-sampled eligible households were left out. The sampled eligible households received a weight accordingly. The 11,275 LFS respondents and the 628 ineligible households all received a weight of one. This implies that the inclusion probabilities are unequal for this example.

Schouten (2007) compared the LFS respondents to the converted and persistent non-respondents in the call-back approach using a large set of demographic and socio-economic characteristics. He used logistic regression models to predict the type of response. He concluded that the

converted non-respondents in the call-back approach are different from the LFS respondents with respect to the selected auxiliary variables. Furthermore, he found no evidence that the converted non-respondents were different from persistent non-respondents with respect to the same characteristics. These findings have led to the conclusion that the combined response of the LFS and call-back approach is more representative with respect to the selected auxiliary variables.

The additional response in the basic question approach was analyzed by Cobben and Schouten (2007) using the same set of auxiliary variables and employing the same logistic regression models. For this follow-up, the findings were different for households with and without a listed phone number. When restricted to listed households, they found the same results as for the call-back approach; the response becomes more representative after the addition of the listed basic-question respondents. However, for the overall population, *i.e.*, including the unlisted households, the inverse was found. The basic-question approach gives 'more of the same' and, hence, sharpens the contrast between respondents and non-respondents. Combining LFS response with basic-question response leads to a less representative composition. In the logistic regression models by Cobben and Schouten (2007) the 0-1 indicators for having a listed phone number and having a paid job gave a significant contribution.

Cobben and Schouten (2007) and Schouten (2007) used the set of auxiliary variables listed in Table 3. The auxiliary variables were linked to the sample from various registers and administrative data. The variables in logistic regression models for response probabilities were selected when the variables gave a significant contribution at the 5% level. Otherwise, they were excluded.

Table 3
The auxiliary variables in the studies by Schouten (2007) and Cobben and Schouten (2007). The household core is the head of the household and his or her partner if present

Variable
Household has a listed phone number
Region of the country in 4 classes
Province and 4 largest cities
Average age in 6 classes
Ethnic group in 4 classes
Degree of urbanization in 5 classes
Household type in 6 classes
Gender
Average house value at zip code level in 11 classes
At least one member of household core is self-employed
At least one member of household core has a subscription to the CWI
At least one member of household core receives social allowance
At least one member of household core has a paid job
At least one member of household core receives disability allowance

Table 4 shows the weighted sample size, response rate, $\hat{R}$, $\mathrm{CI}^{\mathrm{BT}}_{0.05}$, $\hat{B}_m$ and $\hat{E}_m$ for the response to the LFS, the response of the LFS combined with the call-back response and the response of the LFS combined with the basic-question response. The standard errors are relatively large with respect to the studies in subsequent sections due to the weighting. There is an increase in $\hat{R}$ when the call-back respondents are added to the LFS respondents. As both the response rate and the R-indicator increase, the maximal absolute bias $\hat{B}_m$ decreases. The confidence intervals $\mathrm{CI}^{\mathrm{BT}}_{0.05}$ for the LFS response and the combined LFS and call-back response overlap. However, the one-sided null hypothesis H_0: $R_{\mathrm{LFS}} - R_{\mathrm{LFS+CB}} \geq 0$ is rejected at the 5%-level.

Table 4
Weighted sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, LFS plus call-back, and LFS plus basic-question for the extended set of auxiliary variables

Response	n	Rate	$\hat{R}$	$\mathrm{CI}^{\mathrm{BT}}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
LFS	18,074	62.2%	80.1%	(77.5-82.7)	8.0%	8.0%
LFS + call-back	18,074	76.9%	85.1%	(82.4-87.8)	4.8%	4.9%
LFS + basic-question	18,074	75.6%	78.0%	(75.6-80.4)	7.3%	7.3%

In Table 4, there is a decrease in $\hat{R}$ when we compare the LFS response to the combined response with the basic-question approach. This decrease is not significant. $\hat{B}_m$ slightly decreases. In Table 5, this comparison is restricted to households with a listed phone number. The R-indicator in general is much higher than for all the households. Because the sample size is now smaller, the estimated standard errors are larger as is reflected in the width of the confidence interval. $\hat{B}_m$ is decreased. For the combined response in the LFS and the basic-question approach, we see an increase of $\hat{R}$ but again this increase is not significant. $\hat{B}_m$ decreases.

Table 5
Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for LFS, and LFS plus basic-question restricted to households with listed phone numbers and for the extended set of auxiliary variables

Response	n	Rate	$\hat{R}$	$\mathrm{CI}^{\mathrm{BT}}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
LFS	10,135	68.5%	86.3%	(83.1-89.5)	5.0%	5.1%
LFS + basic-question	10,135	83.0%	87.5%	(84.3-90.7)	3.8%	3.8%

We find in the example of the LFS follow-up that the R-indicators confirm the conclusions for the call-back approach and the basic question approach. Furthermore, the increase in the R-indicator that follows by adding the call-back response is significant at the 5% level.

5.3 Safety Monitor; pilot mixed-mode 2006

In 2006, Statistics Netherlands conducted a pilot on the Safety Monitor to investigate mixed-mode data collection strategies. See Cobben, Janssen, Berkel and Brakel (2007) for details. The regular Safety Monitor surveys individuals of 15 years and older in the Netherlands about issues that relate to safety and police performance. The Safety Monitor is a mixed-mode survey. Persons with a listed phone number are approached by CATI. Persons that cannot be reached by telephone are approached by CAPI. In the 2006 pilot, the possibility of using the Internet as one of the modes in a mixed-mode strategy was evaluated. Persons in the pilot were first approached with a web survey. Non-respondents to the web survey were re-approached by CATI when they had a listed phone number and by CAPI otherwise. In Table 6 we give the response rates for the normal survey, the pilot response to the web only, and the response to the pilot as a whole. The response to the web survey alone is low. Only 30% of the persons filled in the web questionnaire. This implied that close to 70% of the sampled units were re-allocated to either CAPI or CATI. This resulted in an additional response of approximately 35%. The overall response rate is slightly lower than that of the normal survey.

Fouwels, Janssen and Wetzels (2006) performed a univariate analysis of response compositions. They argue that the response rate is lower for the pilot but that this decrease is quite stable over various demographic sub-groups. They observe a univariate decline in response rate for the auxiliary variables age, ethnic group, degree of urbanization and type of household. However, they do find indications that the response becomes less representative when the comparison is restricted to the web respondents only. This holds, not surprisingly, especially for the age of the sampled persons.

Table 6 contains the sample size, response rate, $\hat{R}$, $\mathrm{CI}^{\mathrm{BT}}_{0.05}$, $\hat{B}_m$ and $\hat{E}_m$ for three groups: the regular survey, the pilot survey restricted to web and the pilot survey as a whole. The auxiliary variables age, ethnic group, degree of urbanization and type of household were linked from registers and were selected in the logistic model for the response probabilities. Table 6 shows that the R-indicator for the web response is lower than that of the regular survey. The corresponding p-value is close to 5%. As a consequence of both a low response rate and a low R-indicator, the maximal absolute bias $\hat{B}_m$ is more than twice as high as for the regular survey. However, for the pilot as a whole, both the R-indicator and $\hat{B}_m$ are close to the values of the regular survey. Due to the smaller sample size of the pilot, the estimated standard errors are larger than in the regular survey.

Table 6
Sample size, response rate, R-indicator, confidence interval, maximal bias and maximal RMSE for response for the regular Safety Monitor, the pilot with web only and the pilot with web and CAPI/CATI follow-up

Response	n	Rate	$\hat{R}$	$\mathrm{CI}^{\mathrm{BT}}_{0.05}$	$\hat{B}_m$	$\hat{E}_m$
Regular	30,139	68.9%	81.4%	(80.3-82.4)	6.8%	6.8%
Pilot - web	3,615	30.2%	77.8%	(75.1-80.5)	18.3%	18.4%
Pilot - web plus	3,615	64.7%	81.2%	(78.3-84.0)	7.3%	7.4%

The findings in Table 6 do not contradict those of Fouwels *et al.* (2006). We also find that the web response in the pilot has a less balanced composition, whereas the composition of the full pilot response is not markedly worse than that of the Safety Monitor itself.

6. Discussion

We have three main objectives in this paper: a mathematically rigorous definition and perception of representative response, the construction of a potential indicator for representativeness, and the empirical illustrations of the indicator's use. As we saw, the proposed indicator is an example of what we call R-indicators, where 'R' stands for representativeness. With the empirical illustration, we want to find support for the idea that such R-indicators are valuable tools in the comparison of different surveys and data collection strategies. R-indicators are useful if they confirm findings in elaborate analyses of studies that involve multiple surveys in time or on a topic.

The R-indicator in this paper is promising because it can easily be computed and allows for interpretation and normalization when response propensities can be estimated without error. The application to real survey data shows that the R-indicator confirms earlier analyses of the non-response composition. Other R-indicators can, of course, simply be constructed by choosing different distance functions between vectors of response propensities. The R-indicator and graphical displays showed in this paper can be computed using most standard statistical software packages.

The computation of R-indicators is sample-based and employs models for individual response propensities. Hence, R-indicators are random variables themselves and there are two estimation steps that influence their bias and variance. However, it is mostly the modelling of response propensities that has important implications. The restriction to the sample for the estimation of R-indicators implies that those indicators are less precise, but this restriction does not introduce a bias asymptotically. Model selection and model fit usually are performed by choosing a significance level and adding only those interactions to the model that give a significant contribution. The latter means that the size of the

sample and the availability of auxiliary variables play an important role in the estimation of response propensities. Bias may be introduced by the model selection strategy. There are various obvious approaches for dealing with the dependence on the size of the sample. One may not do a model selection but fix a stratification beforehand. That way, bias is avoided but standard errors are not controlled and may be considerable. One may also let empirical validation be the input to develop 'best practices' for R-indicators.

We applied the proposed R-indicator to two studies that were conducted at Statistics Netherlands in recent years, and that were thoroughly investigated by other authors. The increase or decrease in the R-indicator conforms to the more detailed analyses done by these authors. We, therefore, conclude that R-indicators can be valuable tools. However, more empirical evidence is clearly needed.

The application of the R-indicator showed that there is no clear relation between response rate and representativeness of response. Larger response rates do not necessarily lead to a more balanced response. Not surprisingly, we do find that higher response rates reduce the risk of non-response bias. The higher the response rate, the smaller the maximal absolute bias of survey items.

Application to the selected studies showed that standard errors do decrease with increasing sample size as expected, but they are still relatively large for modest sample sizes. For example, for a sample size of 3,600, we found a standard error of approximately 1.3%. Hence, if we assume a normal distribution, then the 95% confidence interval has an approximate width of 5.4%. The sample size of the LFS is about 30,000 units. The standard error is approximately 0.5% and the corresponding 95% confidence interval is approximately 2% wide. The standard errors are larger than we expected.

This paper contains a first empirical study of an R-indicator and its standard error. Much more theoretical and empirical research is necessary to fully understand R-indicators and their properties. First, we did not consider survey items at all. Clearly, it is imperative that we do this in the future. However, as we already argued, R-indicators are dependent on the set of auxiliary variables. It can, therefore, be conjectured that, as for non-response adjustment methods, the extent to which R-indicators predict non-response bias of survey items is dependent on the missing-data mechanism. In a missing-data mechanism that is strongly non-ignorable, R-indicators will not do a good job. However, without knowledge about the missing-data mechanism, no other indicator would either. For this reason, we constructed the notion of maximal absolute bias, as this gives a limit to non-response bias under the worst-case scenario. A second topic of future research is a theoretical derivation of the standard error of the R-indicator used in this paper. The non-parametric bootstrap errors only give naïve approximations. However, if we want R-indicators to play a more active role in the comparison of different strategies, then we need (approximate) closed forms. Third, we will need to investigate the relation between the selection and number of auxiliary variables and the standard errors of the R-indicator.

Acknowledgements

The authors would like to thank Bob Groves, Björn Janssen, Geert Loosveldt, and the associate editor and two referees for their useful comments and suggestions.

References

Agresti, A. (2002). Categorical data analysis. *Wiley Series in Probability and Statistics*. New York: John Wiley & Sons, Inc., NY, USA.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 2, 238-246.

Bertino, S. (2006). A measure of representativeness of a sample for inferential purposes. *International Statistical Review*, 74, 149-159.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 3, 251-260.

Cobben, F., Janssen, B., Berkel, K. van and Brakel, J. van den (2007). Statistical inference in a mixed-mode data collection setting. Paper presented at ISI 2007, August 23-29, 2007, Lisbon, Portugal.

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

Curtin, R., Presser, S. and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly*, 64, 413-428.

Efron, B., and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fouwels, S., Janssen, B. and Wetzels, W. (2006). Experiment mixed-mode waarneming bij de VMR. Technical paper SOO-2007-H53, CBS, Heerlen.

Goodman, L.A., and Kruskal, W.H. (1979). Measures of association for cross-classifications. Springer-Verlag, Berlijn, Duitsland.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 5, 646-675.

Groves, R.M., and Peytcheva, E. (2006). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Paper presented at 17th International Workshop on Household Survey Nonresponse, August 28-30, Omaha, NE, USA.

Groves, R.M., Presser, S. and Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68, 2-31.

Hájek, J. (1981). Sampling from finite populations. New York: Marcel Dekker, USA.

Hansen, M.H., and Hurwitz, W.H. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Heerwegh, D., Abts, K. and Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods*, 1, 1, 3-10.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 2, 119-127.

Keeter, S., Miller, C., Kohut, A., Groves, R.M. and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64, 125-148.

Kersten, H.M.P., and Bethlehem, J.G. (1984). Exploring an reducing the nonresponse bias by asking the basic question. *Statistical Journal of the United Nations*, ECE 2, 369-380.

Kohler, U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods*, 1, 2, 55-67.

Kruskal, W., and Mosteller, F. (1979a). Representative sampling I: Non-scientific literature. *International Statistical Review*, 47, 13-24.

Kruskal, W., and Mosteller, F. (1979b). Representative sampling II: Scientific literature excluding statistics. *International Statistical Review*, 47, 111-123.

Kruskal, W., and Mosteller, F. (1979c). Representative sampling III: Current statistical literature. *International Statistical Review*, 47, 245-265.

Little, R.J.A., and Rubin, D.B. (2002). Statistical analysis with missing data. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc., NY, USA.

Marsh, H.W., Balla, J.R. and McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 3, 391-410.

Merkle, D.M., and Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A Little). New York: John Wiley & Sons, Inc., 243-258.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

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

Särndal, C., Swensson, B. and Wretman, J. (2003). Model-assisted survey sampling. Springer Series in Statistics, Springer, New York.

Schouten, B. (2004). Adjustment for bias in the Integrated Survey on Household Living Conditions (POLS) 1998. Discussion paper 04001, CBS, Voorburg, available at website http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2004/default.htm.

Schouten, B., and Cobben, F. (2007). R-indicators for the comparison of different fieldwork strategies and data collection modes, Discussion paper 07002, CBS, Voorburg. Available at website http://www.cbs.nl/nl-NL/menu/methoden/research/discussionpapers/archief/2007/default.htm.

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

Voogt, R. (2004). I am not interested: Nonresponse bias, response bias and stimulus effects in election research. PhD dissertation, University of Amsterdam, Amsterdam.