Designing Adaptive Survey Designs with R-Indicators

Natalie Shlomo[1], Barry Schouten and Vincent de Heij[2]
[1]University of Manchester, e-mail: natalie.shlomo@manchester.ac.uk
[2]Statistics Netherlands, e-mail: jg.schouten@cbs.nl and vhey@cbs.nl

Abstract

Recent survey literature shows an increasing interest in the tailoring of survey design features. Survey designs that adapt data collection to characteristics of the survey target population derived through auxiliary data are termed adaptive (or responsive) survey designs. Given a specified quality objective function, the designs attempt to find an optimal balance between quality and costs. In this paper, we demonstrate how an adaptive survey design can be carried out using R-indicators. The R-indicators measure the degree to which respondents and non-respondents differ from each other (the contrast) and go beyond response rates alone. Through the analysis of R-indicators we can build profiles (characteristics) of the data units where more or less attention is required in the data collection. We demonstrate the effectiveness of targeted data collection in a simulation study.

Keywords: non-response analysis, partial R-indicators, survey design

1. Introduction

Recent literature has shown that the response or coverage rate is an insufficient quality indicator to measure the potential impact of non-response to a survey. There have been many studies that have shown that increased data collection efforts have led to a higher response rate but also to a larger non-response bias (Curtin, Presser and Singer (2000), Groves, Presser and Dipko (2004), Keeter et al. (2000) and Merkle and Edelman (2002)). For these surveys the contrast between response and non-response was increased by the increased effort.

The EU 7th Framework research project (Socioeconomic Sciences and the Humanities Part 8: FP7SSH20071) titled Representativity Indicators for Survey Quality (RISQ) was carried out between 2008 and 2010 and involved a consortium of European partners from the Netherlands, United Kingdom, Norway, Slovenia and Belgium. Representativity indicators (or R-Indicators) were developed to measure the extent to which a survey or register is representative of the population under investigation (Schouten, Cobben and Bethlehem 2009). Non-response research typically restricts itself to the investigation and identification of subpopulations that have low response rates. Implicitly, such research investigates the characteristics of households and enterprises for which a contrast can often be observed. These investigations can be conducted when demographic and socio-economic information is available that stems from other sources. The RISQ project

involved the translation of non-response analyses to quality indicators. The quality indicators (or R-Indicators) measure the degree to which respondents and non-respondents differ from each other. They can then be used to compare different surveys in time or topics, and, hence, to generalise findings, and for an assessment of quality that goes beyond the response rate alone.

Representativity is defined in terms of the response propensities of different sample units given their values on a specified set of auxiliary variables and is based on their variation. Response is said to be representative if all the response propensities in the sample are equal (and none are equal to zero). Our definitions of R-indicators will be most effective in capturing non-response bias in a survey estimate when the auxiliary variables are, in combination, strong predictors of the survey item(s) upon which the estimate is based and the model for estimating the response propensities is specified correctly. The R-indicators can be decomposed to produce partial R-indicators for measuring the impact of the specified variable/category on deviations from representative response. We make a distinction between unconditional and conditional partial R-indicators.

The partial R-indicators allow the building of profiles (characteristics) of the data units where more or less attention is required in the data collection in order to reduce the contrast between respondents and non-respondents. Monitoring and controlling data collection is known as adaptive (or responsive) survey designs. Adaptive survey designs aim to differentiate the field management and data collection with respect to known characteristics of the data units. By targeting the data collection and follow-up strategies, we ensure that efforts to increase response will be directed to those that are contributing the most to the non-response bias. By ensuring a more representative sample at the source, we aim to reduce the variation in final survey weights mainly due to non-response adjustments and thus produce more efficient estimators.

In this paper, we review the R-indicator and partial R-Indicators and present theoretical properties of these indicators in Sections 2 to enable significance testing. We then provide a simulation study in Section 3 where we show that even with a slight increase in the response rate, we obtain large gains in representativity when targeting the data collection in a responsive survey design. We conclude with a discussion in Section 4. At www.risq-project.eu code in SAS and R and a manual are available for the computation of R-indicators and partial-R-indicators. The code will be extended with standard errors approximations for all indicators and a number of other features.

2. Theoretical Properties of R- indicators and Partial R-Indicators

We use the notation and definition of response propensities as set out in Schouten, Cobben and Bethlehem (2009) and Shlomo, Skinner and Schouten (2012). We let U denote the set of units in the population and s the set of units in the sample. We define a response indicator variable R_i which takes the value 1 if unit i in the population responds and the value 0 otherwise. The *response propensity* is defined as the conditional expectation of R_i given the vector of values x_i of the vector X of auxiliary variables:

$\rho_X(x_i) = E(R_i = 1 \mid X = x_i) = P(R_i = 1 \mid X = x_i)$ and also denote this response propensity by ρ_X. We assume that the values x_i are known for all sample units, i.e. for both respondents and non-respondents, and can include both specified variables and survey fieldwork conditions.

2.1 Definition of R-indicator

We define the R-indicator as: $R(\rho_X) = 1 - 2S(\rho_X)$. The estimation of the propensities is typically based on a logistic regression model: $\log[\rho_X/(1-\rho_X)] = x'\beta$ where β is a vector of unknown parameters to be estimated, and *x* may involve the transformation of the original auxiliary variables (e.g. by including interaction terms) for the purpose of model specification. The estimator of the response propensity is: $\hat{\rho}_X = \frac{\exp(x'\hat{\beta})}{\exp(x'\hat{\beta})+1}$ where $\hat{\beta}$ is the estimator of β based on the model. The estimator of the variance of the response propensities: $\hat{S}^2(\hat{\rho}_X) = \frac{1}{N-1}\sum_s d_i(\hat{\rho}_X(x_i) - \bar{\hat{\rho}}_X)^2$ where $d_i = \pi_i^{-1}$ is the design weight or inclusion weight and $\bar{\hat{\rho}}_X = \frac{1}{N}\sum_s d_i \hat{\rho}_X(x_i)$. We estimate the R-indicator: $\hat{R}(\hat{\rho}_X) = 1 - 2\hat{S}(\hat{\rho}_X)$.

2.2 Theoretical Properties of the R-Indicator

As shown in Shlomo, Schouten and Skinner (2012), estimated R-indicators have a sample size dependent bias. When the sample size decreases, the bias increases. For this reason a bias adjustment was proposed for $\hat{R}(\hat{\rho}_X)$. When the sampling design is a simple random sample without replacement the bias-adjusted R-indicator has the form:

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

with $z_i = \nabla h(x_i^T\hat{\beta})x_i$ and *h* the link function in the model for response propensities, i.e. the logit function.

In addition, Shlomo, Schouten and Skinner (2012) also developed a variance calculation for the R-indicator $\hat{R}(\hat{\rho}_X)$. This was based on decomposing $\hat{S}_\rho^2$ into the part induced by the sampling design for a fixed value of $\hat{\boldsymbol{\beta}}$ and the part induced by the distribution of $\hat{\boldsymbol{\beta}}$. We take the latter to be $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{\Sigma} = \mathbf{J}(\boldsymbol{\beta})^{-1} \operatorname{var}\{\textstyle\sum_s d_i[R_i - h(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i\}\mathbf{J}(\boldsymbol{\beta})^{-1} \qquad (2)$$

and $\mathbf{J}(\boldsymbol{\beta}) = E\{\mathbf{I}(\boldsymbol{\beta})\}$ is the expected information rather than the observed information in (2). The decomposition can be written as:

$$\text{var}(\hat{S}^2_\rho) = E_{\hat{\boldsymbol{\beta}}}[\text{var}_s(\hat{S}^2_\rho)] + \text{var}_{\hat{\boldsymbol{\beta}}}[E_s(\hat{S}^2_\rho)], \tag{3}$$

where the subscript $\hat{\boldsymbol{\beta}}$ denotes the distribution induced by $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, which may be interpreted as arising from the response process. Following usual linearization arguments, denoting $\rho_i \equiv \rho(x_i)$ and given the consistency of $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ for standard sample designs, the first term can be written as:

$$E_{\hat{\boldsymbol{\beta}}}[\text{var}_s(\hat{S}^2_\rho)] \approx \text{var}_s[N^{-1}\sum_{i\in s} d_i(\rho_i - \bar{\rho}_U)^2]. \tag{4}$$

For the second component in (3), again using linearization arguments and approximating $\hat{\rho}_i \approx \rho_i + \mathbf{z}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ where $\mathbf{z}_i = \nabla h(\mathbf{x_i}'\boldsymbol{\beta})\mathbf{x}_i$, and assuming that $\hat{\boldsymbol{\beta}}$ is normally distributed so that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is uncorrelated with $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'$, we can write write:

$$\text{var}_{\hat{\boldsymbol{\beta}}}[E_s(\hat{S}^2_\rho)] \approx 4\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} + \text{var}_{\hat{\boldsymbol{\beta}}}\{tr[\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\}, \tag{5}$$

where $\mathbf{A} = N^{-1}\sum_{i\in U}(\rho_i - \bar{\rho}_U)(\mathbf{z}_i - \bar{\mathbf{z}}_U)$, $\mathbf{B} = N^{-1}\sum_{i\in U}(\mathbf{z}_i - \bar{\mathbf{z}}_U)(\mathbf{z}_i - \bar{\mathbf{z}}_U)'$ and $\boldsymbol{\Sigma}$ is defined in (2). The second term involves the fourth moments of $\hat{\boldsymbol{\beta}}$ which can also be expressed in terms of $\boldsymbol{\Sigma}$ since $\hat{\boldsymbol{\beta}}$ is assumed normally distributed.

The variance of $\hat{S}^2_\rho$ can be estimated by the sum of the estimated components of (3). We estimate the component in (4) by a standard design-based estimator of $\text{var}_s[\sum_{i\in s} d_i(\rho_i - \bar{\rho}_U)^2]$, where this is treated as the variance of a linear statistic $\text{var}_s[\sum_{i\in s} u_i]$ and u_i is replaced by $d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)^2$ in the expression for the variance estimator. We estimate the second component of the variance in (5) by estimating $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\Sigma}$. First, $\mathbf{z}_i$ may be estimated by $\hat{\mathbf{z}}_i = \nabla h(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i$. Then $\mathbf{A}$ may be estimated by $\hat{\mathbf{A}} = N^{-1}\sum_{i\in s} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_U)(\hat{\mathbf{z}}_i - \hat{\bar{\mathbf{z}}}_U)$, $\mathbf{B}$ may be estimated by $\hat{\mathbf{B}} = N^{-1}\sum_{i\in s} d_i(\hat{\mathbf{z}}_i - \hat{\bar{\mathbf{z}}}_U)(\hat{\mathbf{z}}_i - \hat{\bar{\mathbf{z}}}_U)'$, where $\hat{\bar{\mathbf{z}}}_U = N^{-1}\sum_s d_i\hat{\mathbf{z}}_i$, and $\boldsymbol{\Sigma}$ may be estimated by a standard estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$.

2.3 Definition of Partial R-Indicators

The unconditional partial indicators measure the distance to representative response for single auxiliary variables and are based on the between variance given a stratification with categories of Z (Schouten, Shlomo and Skinner (2011)). The variable Z may or may not be included in the covariates of the model X for estimating the response propensities. Given a stratification based on a categorical variable Z having categories $k = 1,2,\ldots,K$, the variable level unconditional partial R-indicator is defined as $P_u(Z, \rho_X) = S_B(\rho_X \mid Z)$ where

$$S^2_B(\rho_X \mid Z) = \frac{1}{N-1}\sum\nolimits_{k=1}^{K} N_k(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \cong \sum\nolimits_{k=1}^{K}\frac{N_k}{N}(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \tag{6}$$

where $\bar{\rho}_{X,k}$ is the average of the response propensity in stratum k. This between variance is estimated by: $\hat{S}_B^2(\hat{\rho}_X \mid Z) = \sum_{k=1}^{K} \frac{\hat{N}_k}{N}(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X)^2$ where $\hat{\bar{\rho}}_{X,k} = \frac{1}{N_k}\sum_{s_k} d_i \rho_X(x_i)$, $\hat{N}_k = \sum_{s_k} d_i$ is the estimated population size of stratum k and s_k is the set of sample units in the stratum. At the category level $Z=k$, the unconditional partial R-indicator is defined as:

$$P_u(Z,k,\rho_X) = S_B(\rho_X \mid Z=k)\frac{\left(\bar{\rho}_{X,k} - \bar{\rho}_X\right)}{\left|\bar{\rho}_{X,k} - \bar{\rho}_X\right|} = \sqrt{\frac{N_k}{N}}\left(\bar{\rho}_{X,k} - \bar{\rho}_X\right) \tag{7}$$

and is estimated by: $\hat{S}_B(\hat{\rho}_X \mid Z=k) = \sqrt{\frac{\hat{N}_k}{N}}(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X)$.

Conditional partial R- indicators measure the remaining variance due to variable Z within sub-groups formed by all other remaining variables, denoted by X^- (Schouten, Shlomo and Skinner (2011)). In contrast to the unconditional partial R- indicator, the variable Z must be included in the model for estimating response propensities. Let δ_k be the 0-1 dummy variable that is equal to 1 if $Z=k$ and 0 otherwise. Given a stratification based on all categorical variables except Z, denoted by X^- and indexed by j, $j=1...J$, the conditional partial R-indicator is based on the within variance and is defined as: $P_c(Z,\rho_X) = S_W(\rho_X \mid X^-)$ where:

$$S_w^2(\rho_X \mid X^-) = \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}(\rho_X(x_i) - \bar{\rho}_{X,j})^2 \tag{8}$$

and is estimated by $\hat{S}_w^2(\hat{\rho}_X \mid X^-) = \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in s_j} d_i(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,j})^2$. At the categorical level of $Z=k$, we restrict the within variance to population units in stratum k and obtain:

$$P_c(Z,k,\rho_X) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}\delta_{k,i}(\rho_X(x_i), -\bar{\rho}_{X,j})^2} \tag{9}$$

and estimated by: $\hat{P}_c(Z,k,\hat{\rho}_X) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in s_j} d_i\delta_{k,i}(\hat{\rho}_X(x_i), -\hat{\bar{\rho}}_{X,j})^2}$.

2.4 Theoretical Properties of Partial R-indicators

Empirical work has shown that the size dependent bias affecting the R-indicator in (1) has little impact on the variable level partial R-indicators when sample sizes are large and no impact on the categorical level partial R-indicators. The main reason for this is that the variance of the partial R-indicators becomes the dominant property which needs to be accounted for. Therefore, for smaller sample sizes, we adopt a method of pro-rating the bias correction term in (1) between the decomposed variance components defining the variable level partial R-indicators as follows: The variable level unconditional partial R-indicator $P_u(Z,\rho_X)$ is the between variance given the stratifying variable Z. The variable level conditional partial R-indicator $P_c(Z,\rho_X)$ is the within variance given the stratifying variable X^- (all auxiliary variables except Z). By calculating the complementary between and within variance for each of the stratifying variables, we can implement a

pro-rating of the bias correction term in (1) between the complimentary between and within variances.

To obtain the variance estimates for the variable level partial R-indicators, we observe that for the unconditional partial R-indicator $P_u(Z,\rho_X) = S_B(\rho_X \mid Z)$ we can obtain an estimate of the variance as shown in (3) when the response propensities are modelled based on a stratification on the single variable *Z*. Similarly, for the conditional partial R-indicator $P_c(Z,\rho_X) = S_W(\rho_X \mid X^-)$ we can obtain an approximation of the variance as shown in (3) when the response propensities are modelled by a stratification on X^-. The approximation is due to the fact that main effects and second order interactions are typically used to estimate response propensities in the logistic regression models as opposed to a complete cross-classification.

To obtain the variance estimates for the categorical level partial R-indicators, we denote X^- the auxiliary variables taking values $j = 1,2,...,J$ and Z a categorical variable for which the partial indicator is calculated with categories $k = 1,2,\ldots,K$.

Standard error of unconditional category-level partial R-indicator:
The variance of the estimated unconditional category-level partial R-indicator: $\hat{P}_u(Z,k,\hat{\rho}_X)$ can be written as:

$$Var\,(\hat{P}_u(Z,k,\hat{\rho}_X)) = \frac{\hat{N}_k}{N} Var\left(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X\right) = \frac{\hat{N}_k}{N}\left[Var\,(\hat{\bar{\rho}}_{X,k}) + Var\,(\hat{\bar{\rho}}_X) - 2Cov\,(\hat{\bar{\rho}}_{X,k}, \hat{\bar{\rho}}_X)\right]$$

assuming that N_k is the number of units with $Z=k$ and is known, $\hat{\bar{\rho}}_{X,k} = \sum_{i\in s} d_i \hat{\rho}_i \delta_i^k / \hat{N}_k$ where $\delta_i^k = 1$ if $Z = k$ and $\delta_i^k = 0$ otherwise, and $\hat{\bar{\rho}}_X = \sum_{i\in s} d_i \hat{\rho}_i / N$. In general N_k may not be known and we may need to estimate it by the sample-based estimator $\hat{N}_k = \sum_{s_k} d_i$. This will introduce a small additional loss of precision. Since $\hat{\bar{\rho}}_X = \frac{\hat{N}_k}{N}\hat{\bar{\rho}}_{X,k} + \left(1 - \frac{\hat{N}_k}{N}\right)\hat{\bar{\rho}}_{X,k^c}$ where

$\hat{\bar{\rho}}_{X,k^c} = \sum_{i\in s} d_i \hat{\rho}_i (1-\delta_i^k)/(N - \hat{N}_k)$, we have that: $Cov\,(\hat{\bar{\rho}}_{X,k}, \hat{\bar{\rho}}_X) = \frac{\hat{N}_k}{N} Var\,(\hat{\bar{\rho}}_{X,k})$ and therefore:

$$Var\,(\hat{P}_u(Z,k,\hat{\rho}_X)) = \frac{\hat{N}_k}{N}\left[\left(1 - \frac{\hat{N}_k}{N}\right)^2 Var\,(\hat{\bar{\rho}}_{X,k}) + \left(1 - \frac{\hat{N}_k}{N}\right)^2 Var\,(\hat{\bar{\rho}}_{X,k^c})\right] \qquad (10)$$

We restrict ourselves to a first-order approximation and approximate $Var(\hat{\bar{\rho}}_{X,k})$ by a standard design based variance estimator $\sum_{i\in s} d_i \hat{\phi}_i$, where $\hat{\phi}_i = \delta_i^k \hat{\rho}_i / \hat{N}_k$ and approximate $Var(\hat{\bar{\rho}}_{X,k^c})$ by a standard design based variance estimator $\sum_{i\in s} d_i \hat{\upsilon}_i$, where

$\hat{\upsilon}_i = (1-\delta_i^k)\hat{\rho}_i/(N-\hat{N}_k)$. The standard error is obtained by taking the square root of the expression in (10).

Standard error of conditional category-level partial R-indicator:

For the conditional category-level partial R-indicator $\hat{P}_c(Z,k,\hat{\rho}_X)$ we use the same methodology for the variance estimation of the R-indicator as shown in Section 2.2 but we add in the stratification variable X^- as follows: The first term in (4) may be estimated by a standard design-based estimator of $\text{var}_s[\sum_{j=1}^{J}\sum_{i\in s_k} d_i(\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})^2]$ where this is treated as the variance under a stratified sample design of a linear statistic $\text{var}_s[\sum_{j=1}^{J}\sum_{i\in s_k} u_{ji}]$ and u_{ji} is replaced by $d_i(\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})^2$. For the second term in (5), we replace the **A** and **B** under a stratified design using the estimates: $\hat{\mathbf{A}} = N^{-1}\sum_{j=1}^{J}\sum_{s_k} d_i\delta_i^k(\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})(\hat{z}_i - \hat{\bar{z}}_{X=j})$ and

$\hat{\mathbf{B}} = N^{-1}\sum_{j=1}^{J}\sum_{s_k} d_i\delta_i^k(\hat{z}_i - \hat{\bar{z}}_{X=j})(\hat{z}_i - \hat{\bar{z}}_{X=j})'$.

3. Simulation Study for a Responsive Design

For the simulation study, we use a dataset from the 1995 Israel Census Sample of Individuals aged 15 and over (N=753,711). Population response propensities were calculated using a 2-step process:

1. Probabilities of response were defined according to variables: child indicator, income from earnings groups, age groups, sex, number of persons in household and three types of localities. These variables define groups that are known to have differential response rates in practice. Based on the probabilities, we generated a response indicator.
2. Using the response indicator as the dependent variable, we fitted a logistic regression model on the population using the above explanatory variables where type of locality and size of household were interacted. The predictions from this model served as the 'true' response propensities for our simulation study.

The overall response rate generated in the population dataset was 69.2%. Table 1 presents the differential response rates according to the variables in the model that generated the population response propensities. High non-response rates in categories are likely to cause the sub-group in the population to be under-represented according to the partial R-indicators. From the population, we drew a 1:100 sample (sample size of 7,537) using simple random sampling and generated a response/nonresponse indicator according to the propensity to respond as defined in the population. The R-indicator was 0.859 with a confidence interval between 0.838 and 0.880.

Tables 2 and 3 provide the partial R-indicators and their confidence intervals for the original sample. The unconditional and conditional partial R-indicators can be used

during data collection to identify population subgroups that are candidates for follow-up. Doing so, a responsive survey design is created. We will explain how this was done for this particular study.

Table 1: Percent response generated in the simulation population dataset according to auxiliary variables

Variable	Category	Percent Response	Variable	Category	Percent Response
Children in Household	None	68.1	Sex	Male	68.4
	1+	74.8		Female	71.0
Age group	15-17	77.4	Income Group	Low	69.0
	18-21	65.2		2	63.5
	22-24	62.5		3	68.9
	25-34	64.6		4	73.3
	35-44	68.7		5	63.2
	45-54	72.2		6	68.7
	55-64	71.0		7	70.6
	65-74	76.3		8	61.5
	75+	81.3		9	68.5
Number of Persons in Household	1	68.5		10	57.9
	2	66.4		11	68.4
	3	73.2		12	71.8
	4	75.6		13	67.3
	5	68.2		14	73.0
	6+	68.5		High	70.3
Type of Locality	Type 1	66.7			
	Type 2	70.7			
	Type 3	70.3			

According to the original sample in Table 2, we see that the impact on representativity is occurring at all variables. For all variables, the 95% confidence intervals (using normal approximations) for both the unconditional and conditional partial R-indicators do not include zero. One exception is the conditional partial R-indicator for children present which is close to zero. Age group and size of household variables give the strongest contributions. This can be seen by the larger unconditional partial indicators for these variables which denotes that between grouping of categories of age (or household size), the average propensity to respond variation is larger. For the conditional partial indicator, controlling for the effects of remaining variables, the within variation of the response propensities in categories of age group (or household size) is still larger. In other words, conditioning on the other variables, there remains a lack of representative response. In general, we see that the unconditional partial indicators are larger than the conditional partial indicators in the original sample for all variables. This suggests that the impact of each variable is reduced when controlling for the other variables and that the auxiliary variables show some multicollinearity.

Table 2: Variable level partial R-Indicators for the original and targeted sample

	Unconditional Partial Indicator							Conditional Partial Indicator						
	Original Sample			Targeted Increase of 1% Response			Percent	Original Sample			Targeted Increase of 1% Response			Percent
	Estimate	CI_LB	CI_UB	Estimate	CI_LB	CI_UB	Reduction	Estimate	CI_LB	CI_UB	Estimate	CI_LB	CI_UB	Reduction
Persons in HH	0.043	0.032	0.053	0.036	0.025	0.046	16.2	0.034	0.023	0.044	0.030	0.020	0.041	10.4
Type of Locality	0.013	0.002	0.024	0.006	-0.009	0.020	57.4	0.015	0.005	0.025	0.008	-0.002	0.019	45.0
Age Group	0.053	0.042	0.063	0.045	0.035	0.055	14.3	0.047	0.036	0.058	0.043	0.032	0.053	9.4
Children Present	0.030	0.020	0.040	0.025	0.015	0.035	16.3	0.010	-0.001	0.020	0.011	0.000	0.021	-10.2
Income Groups	0.031	0.020	0.041	0.025	0.013	0.036	19.7	0.018	0.008	0.028	0.017	0.007	0.027	6.6
Sex	0.020	0.010	0.031	0.012	0.001	0.022	42.9	0.018	0.007	0.028	0.010	-0.001	0.020	44.6

We now turn to Table 3 for the categorical level of partial indicators for the original sample. For the unconditional partial R-indicator, we can see the categories of the variables that are under-represented by the 'minus' signs. These roughly coincide with the response rates in the population as seen in Table 1. For the conditional partial R-indicator, values are always positive because the subgroups may be underrepresented within some of the categories of the other variables and overrepresented for other categories. We see that for some variables there are categories that have confidence intervals which overlap with zero, indicating that they do not significantly contribute to non-representative response. This holds true for many of the income and age group categories.

From these results we built a profile of individuals that were (artificially) targeted in a responsive design data collection. We targeted 64 individuals for follow up according to the profile: male, living in the first type of locality in a household size of 1 or 2, had no children, between the ages of 18 and 34 and in income groups from level 2 to 11. This profile was built by selecting all categories for which unconditional partial R-indicators were negative and where both the unconditional and conditional values were significantly different from zero. For our simulation study, we then assumed that all 64 individuals responded and we compared the R-indicators and partial R-indicators after the targeted responsive data collection. Under this scenario, the response rate increased slightly from 69.8% to 70.7%, an increase of 0.9%. The R-indicator for the targeted sample was 0.884 (compared to 0.859) with a confidence interval between 0.863 and 0.905. The overall R-indicator increased by 3% in spite of the very small increase in response albeit targeted to a particular subset. The difference however was not significant.

The partial R-indicators from the targeted responsive design sample also appear in Tables 2 and 3 for comparison to the original sample. We also calculated the percent reduction in the partial R-indicators. We see that many more categories of variables now have confidence intervals that overlap with zero, indicating that following the targeted response they do not significantly contribute to non-representative response. Whilst we see in the tables a clear reduction in the partial R-indicator across all variables and their categories, the differences however were not significant.

Table 3: Categorical level partial R-Indicators for the original and targeted sample

	Unconditional Partial Indicator							Conditional Partial Indicator						
	Original Sample			Targeted Increase of 1% Response			Percent	Original Sample			Targeted Increase of 1% Response			Percent
	Estimate	CI_LB	CI_UB	Estimate	CI_LB	CI_UB	Reduction	Estimate	CI_LB	CI_UB	Estimate	CI_LB	CI_UB	Reduction
Persons in HH														
1	-0.006	-0.019	0.008	-0.002	-0.016	0.012	62.7	0.011	0.002	0.019	0.008	-0.001	0.016	29.9
2	-0.026	-0.035	-0.018	-0.021	-0.030	-0.013	18.7	0.015	0.009	0.022	0.013	0.007	0.020	11.9
3	0.022	0.009	0.034	0.018	0.005	0.031	17.6	0.017	0.010	0.025	0.015	0.008	0.023	11.6
4	0.023	0.010	0.036	0.020	0.007	0.032	15.6	0.018	0.011	0.026	0.017	0.009	0.024	9.3
5	-0.008	-0.021	0.006	-0.010	-0.023	0.004	-30.7	0.012	0.004	0.020	0.012	0.004	0.021	-7.8
6+	0.006	-0.010	0.021	0.004	-0.011	0.019	33.9	0.004	-0.002	0.010	0.004	-0.003	0.010	16.7
Type of Locality														
Type 1	-0.011	-0.021	0.000	0.003	-0.009	0.014	125.5	0.011	0.002	0.019	0.002	-0.009	0.014	77.1
Type 2	0.004	-0.003	0.011	-0.003	-0.010	0.004	181.6	0.005	0.000	0.009	0.004	-0.002	0.011	6.4
Type 3	0.006	-0.008	0.020	0.004	-0.010	0.017	43.8	0.009	0.001	0.018	0.006	-0.002	0.015	31.9
Age Group														
15-17	0.021	0.005	0.037	0.019	0.003	0.035	10.8	0.006	0.001	0.012	0.006	0.001	0.012	-1.6
18-21	-0.013	-0.026	0.000	-0.015	-0.028	-0.002	-14.7	0.019	0.011	0.028	0.019	0.011	0.027	3.1
22-24	-0.016	-0.029	-0.003	-0.014	-0.028	-0.001	11.7	0.015	0.006	0.023	0.013	0.004	0.022	10.3
25-34	-0.024	-0.035	-0.014	-0.012	-0.024	-0.001	49.6	0.014	0.008	0.020	0.010	0.006	0.013	31.9
35-44	-0.009	-0.021	0.002	-0.013	-0.024	-0.001	-40.7	0.012	0.006	0.018	0.014	0.007	0.022	-19.5
45-54	0.010	-0.003	0.024	0.007	-0.006	0.020	30.4	0.009	0.004	0.014	0.008	0.003	0.012	15.6
55-64	0.005	-0.009	0.019	0.002	-0.012	0.016	53.1	0.011	0.006	0.016	0.009	0.005	0.013	15.5
65-74	0.024	0.008	0.039	0.021	0.006	0.037	10.5	0.024	0.016	0.032	0.021	0.014	0.029	11.3
75+	0.023	0.006	0.039	0.021	0.004	0.037	8.8	0.022	0.014	0.029	0.019	0.011	0.027	11.6
Child Indicator														
Yes	0.026	0.014	0.038	0.022	0.010	0.033	16.5	0.007	0.001	0.013	0.008	0.002	0.014	-8.3
No	-0.015	-0.022	-0.008	-0.013	-0.019	-0.006	16.7	0.007	0.001	0.013	0.007	0.002	0.013	-10.4
Income Groups														
Low	0.009	-0.009	0.027	0.008	-0.010	0.026	7.7	0.007	-0.002	0.016	0.007	-0.003	0.016	3.0
2	-0.002	-0.017	0.013	-0.002	-0.017	0.013	-6.2	0.002	-0.020	0.024	0.001	-0.051	0.053	58.8
3	-0.010	-0.022	0.001	-0.006	-0.018	0.006	38.2	0.003	-0.009	0.015	0.002	-0.015	0.019	45.2
4	0.011	-0.009	0.031	0.010	-0.009	0.030	4.6	0.009	0.001	0.016	0.009	0.001	0.016	0.0
5	-0.008	-0.022	0.006	-0.006	-0.020	0.009	27.8	0.003	-0.011	0.018	0.003	-0.014	0.019	19.4
6	-0.006	-0.018	0.005	-0.005	-0.016	0.007	28.6	0.006	-0.003	0.015	0.005	-0.004	0.015	3.6
7	-0.008	-0.022	0.006	-0.008	-0.022	0.006	0.0	0.006	-0.006	0.018	0.006	-0.005	0.018	-8.6
8	-0.009	-0.023	0.005	-0.007	-0.021	0.007	24.7	0.006	-0.006	0.018	0.005	-0.009	0.018	25.0
9	-0.004	-0.017	0.009	-0.004	-0.017	0.009	-5.1	0.002	-0.019	0.023	0.003	-0.015	0.020	-35.0
10	-0.004	-0.018	0.011	-0.004	-0.019	0.011	-17.1	0.002	-0.016	0.021	0.003	-0.012	0.019	-37.5
11	-0.003	-0.018	0.011	-0.004	-0.018	0.011	-8.8	0.001	-0.037	0.038	0.001	-0.035	0.037	0.0
12	0.007	-0.008	0.022	0.005	-0.010	0.020	27.1	0.008	-0.002	0.017	0.007	-0.003	0.016	12.0
13	0.004	-0.012	0.020	0.002	-0.014	0.018	33.3	0.002	-0.015	0.020	0.001	-0.052	0.053	72.7
14	0.016	0.006	0.027	0.012	0.001	0.022	29.3	0.003	-0.005	0.012	0.002	-0.007	0.012	27.3
High	-0.001	-0.016	0.014	0.000	-0.015	0.015	122.2	0.001	-0.023	0.025	0.001	-0.032	0.034	28.6
Sex														
Male	-0.015	-0.023	-0.007	-0.008	-0.017	0.000	43.2	0.013	0.006	0.020	0.007	0.000	0.014	44.9
Female	0.014	0.006	0.022	0.008	0.000	0.016	43.0	0.012	0.006	0.019	0.007	0.007	0.014	44.7

4. Discussion

In this paper, we have presented the theoretical properties of the R-indicator and their partial R-indicators and carried out a simulation study for an adaptive survey design. We have shown that even for a very small increase in response targeted specifically to those individuals whose characteristics were shown to be impacting on representativity, we were able to reduce non-response bias at the source as seen by the reduction in the R-indicator and their partial R-indicators. The reductions however were not shown to be significant since estimated confidence intervals were large. Further work will aim at increasing the precision of the confidence intervals. With the results of the simulation study demonstrating that R-indicators are effective in planning a responsive survey design, we aim to carry out testing on real data for a cross-sectional survey as well as follow-up strategies for a longitudinal survey.

References

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.

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

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

Merkle, D.M. and Edelman, M. (2002), Non-response in exit polls: a comprehensive analysis. In *Survey Non-response* (eds. Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R..J.A.), John Wiley & Sons, 243-258.

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

Shlomo, N., Skinner, C.J. and Schouten, B. (2012). Estimation of an Indicator of the Representativeness of Survey Response. *Journal of Statistical Planning and Inference* 142, 201-211.

Schouten, B., Shlomo, N. and Skinner, C.J. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics,* Vol. 27, No. 2, 231-253.