International Statistical Review

International Statistical Review (2017), 85, 1, 143–163 doi:10.1111/insr.12159

Selecting Adaptive Survey Design Strata with Partial R-indicators

Barry Schouten[1] and Natalie Shlomo[2]

[1]*Department of Process Development, IT and Methodology, Statistics Netherlands, 24500, 2490 HA Den Haag, The Netherlands*
E-mail: jg.schouten@cbs.nl
[2]*Social Statistics, School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester M13 9PL, UK*
E-mail: natalie.shlomo@manchester.ac.uk

Summary

Recent survey literature shows an increasing interest in survey designs that adapt data collection to characteristics of the survey target population. Given a specified quality objective function, the designs attempt to find an optimal balance between quality and costs. Finding the optimal balance may not be straightforward as corresponding optimisation problems are often highly non-linear and non-convex. In this paper, we discuss how to choose strata in such designs and how to allocate these strata in a sequential design with two phases. We use partial R-indicators to build profiles of the data units where more or less attention is required in the data collection. In allocating cases, we look at two extremes: surveys that are run only once, or infrequent, and surveys that are run continuously. We demonstrate the impact of the sample size in a simulation study and provide an application to a real survey, the Dutch Crime Victimisation Survey.

Key words: Non-response; responsive survey design; representativeness; paradata.

1 Introduction

In the recent literature, there is an increased interest in survey data collection designs in which design features are adapted to characteristics of units in the survey target population (Groves and Heeringa, 2006 Wagner, 2008 and 2013, Särndal, 2011 and Schouten, Calinescu and Luiten, 2013b). These characteristics may come from the sampling frame, other linked administrative data or from paradata observations, and they form strata in which design features are differentiated. This paper is about the formation of such strata and the design of interventions given the strata.

Most literature about adapting survey design is restricted to non-response error and ignores other errors like measurement error. We will restrict ourselves also to non-response error in this paper in order not to make the stratification problem overly complex. However, there is a clear need for a more general approach (Calinescu, Schouten and Bhulai, 2012, Calinescu and Schouten, 2013) as the survey mode is one of the most prominent design features and is known to affect multiple errors simultaneously. We leave this to future research.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute. Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

The implementation of designs that differentiate design features, generally, follows a number of steps:

(1) Choose proxy measures for survey quality;
(2) Choose a set of candidate design features, for example, survey modes or incentives;
(3) Define cost constraints and other practical constraints;
(4) Link available frame data, administrative data and paradata;
(5) Form strata with the auxiliary variables for which design features can be varied;
(6) Estimate input parameters (e.g. contact and participation propensities, costs);
(7) Optimise the allocation of design features to the strata;
(8) Conduct, monitor and analyse data collection;
(9) In case of incidental deviation from anticipated quality or costs, return to step 7;
(10) In case of structural deviation from anticipated quality or costs, return to step 6;
(11) Adjust for nonresponse in the estimation.

See Groves & Heeringa (2006), Peytchev *et al.* (2010) and Schouten *et al.* (2013b) for general strategies towards the design of adaptive surveys. Most of the steps are, however, not specific to adaptive survey designs; rather, it is steps 5 to 7 where the adaptation comes in. In this paper, we consider these steps.

The actual implementation in practice depends on the setting and the type of survey. There is a wide range of labels for designs that vary design features over population units: responsive survey design, adaptive survey design, responsive data collection design, targeted survey design and tailored survey design. Their inventors come from different survey settings and as a result have slightly different viewpoints on how the designs should be constructed and what they need to achieve. The main differences in settings between surveys are as follows: (1) the length of the data collection period and the number of instances for intervention; (2) the application of refusal conversion methods; (3) the strength of prior knowledge from frame data, administrative data and paradata in previous waves of the same survey; (4) a focus on learning during data collection versus learning from wave to wave; and (5) a focus on both structural and incidental deviations versus a focus on just structural deviations. On the one end, there are the responsive survey designs as introduced by Groves & Heeringa (2006) where surveys have a long data collection with several instances for intervention, where refusal conversion is possible, where there is relatively little prior knowledge, where the focus is on learning during data collection and on both structural and incidental deviations. On the other end, there are adaptive survey designs as described by Schouten *et al.* (2013b) that refer to relatively short data collection periods with limited intervention and limited possibility to convert refusers, with strong prior knowledge, a focus on learning in between waves and on structural deviations only. In fact, any design phase of responsive survey design, that is, any period in between interventions, could be adaptive. Here, we discuss the selection of strata for a single intervention, that is, for a single adaptation of design features, and throughout the paper, we refer to such designs simply as adaptive survey designs.

If the focus is on non-response, ideally, the characteristics used for forming strata explain both the key survey variables and the propensity to respond. The formation of strata in adaptive survey designs is very similar to formation of strata in the post data collection adjustments for non-response. The reason for this similarity is very simple: Adaptive survey designs attempt to adjust for non-response by design rather than just post hoc in the estimation. Various authors have come up with explicit proxy measures for non-response error in the optimisation of adaptive survey designs: Schouten *et al.* (2009) propose to use representativeness indicators and the coefficient of variation of response propensities and Särndal (2011) and Lundquist & Särndal (2013) propose to use balance indicators. Other authors describe a less explicit approach in which sample units are prioritised based on response propensity models (e.g. Peytchev *et al.*,

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

2010, Wagner, 2013). However, all share a focus on reducing the variation in response propensities for a selected set of auxiliary variables. The obvious and legitimate question is whether such adjustment by design on a specified set of variables has any use when the same variables can also be employed afterwards in the estimation. In our opinion, adjustment by design as a supplement to adjustment afterwards is useful for two reasons: First, it is inefficient to have a highly unbalanced response; a large variation in adjustment weights is to be avoided and may inflate standard errors. Second, and more importantly, the adjustment by design originates from the rationale that stronger imbalance on relevant, auxiliary variables is, in the majority of cases, a signal of stronger imbalance on survey target variables. For a more elaborated discussion, see Schouten *et al.* (2014) and Särndal & Lundquist (2013). Schouten *et al.* (2014) provide theoretical and empirical evidence that, on average, a design with a more representative response has smaller non-response biases, even after post-survey weighting on the characteristics for which representativeness was assessed and evaluated.

Adaptive survey designs have some similarity to balanced sampling, for example, Deville & Tillé (2004), Grafström & Schelin (2014) and Hasler & Tillé (2014). However, adaptive survey designs attempt to balance response to a given sample, not the sample itself. In other words, adaptive survey designs optimise the allocation of treatments or design features given a sample but not the inclusion into the sample. For the same reason, the criticism that adaptive survey designs resemble quota sampling is false; the balance of response is assessed against a probability sample not against the population.

Schouten *et al.* (2012) state that partial R-indicators can be used as tools to monitor and analyse non-response and to improve survey response through adaptive survey design. The last claim has not been substantiated, however.

In this paper, we demonstrate how the partial R-indicators can be used to identify (and monitor) strata for adaptive survey designs and how to optimise interventions for the strata depending on the frequency and length of the survey data collection.

In order to be able to use the indicators for building population strata, it is imperative that they are accompanied by standard error approximations. In this paper, as an important by-product, we provide such approximations. At www.risq-project.eu code in SAS and R and a manual (De Heij *et al.*, 2015) are available for the computation of R-indicators, partial R-indicators and coefficients of variation. The code is extended with standard error approximations for all indicators and other features compared with the first version that was launched in 2010.

In Section 2, we briefly review the partial R-indicators, present bias and standard error properties and explain how the indicators can be used to build and evaluate profiles of non-respondents. In Section 3, we discuss the optimisation of an intervention. Next, we provide a simulation study in Section 4, where we evaluate the impact of the sample size on intervention decisions. In Section 5, we demonstrate the formation of strata and the optimisation of the design for a real data set, the Dutch Crime Victimisation Survey. We conclude with a discussion in Section 6.

2 Building Non-respondent Profiles

In this section, we discuss the use of partial R-indicators to identify strata for adaptive survey designs. We first revisit the various partial R-indicators. As for R-indicators, partial R-indicators have a bias and imprecision that depend on the sample size. We derive approximations to these biases and standard errors. Last, we discuss how the partial R-indicators can be used to form profiles. In the optimization of adaptive survey designs, we also employ the coefficient of variation (CV) of the estimated response propensities, which sets an upper bound to the absolute bias of response means.

2.1 R-indicators and Partial R-indicators Revisited

We use the notation and definition of response propensities as set out in Schouten *et al.* (2011) and Shlomo *et al.* (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

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

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: $\rho_X(x_i) = E(R_i = 1|X = x_i) = P(R_i = 1|X = x_i)$ and denote this response propensity by ρ_X. We assume that the values x_i are known for all sample units, that is, for both respondents and non-respondents. We also assume that we select the auxiliary variables in such a way that the missing at random (MAR) assumption holds as closely as possible.

We define the R-indicator as $R(\rho_X) = 1 - 2S(\rho_X)$. The estimation of the response propensities is typically based on a logistic regression model, and we denote the estimated response propensity by $\hat{\rho}_X(x_i)$ or the shortened $\hat{\rho}_X$. The estimator of the variance of the response propensities is

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

where $d_i = \pi_i^{-1}$ is the design weight and $\hat{\bar{\rho}}_X = \frac{1}{N}\sum_s d_i \hat{\rho}_X(x_i)$. We estimate the R-indicator as

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

The bias adjusted R-indicator as shown in Shlomo *et al.* (2012) is

$$\hat{R}_{\text{BIAS-ADJ}}(\hat{\rho}_X) = 1 - 2\sqrt{\left(1 + \frac{1}{n} - \frac{1}{N}\right)\hat{S}^2(\hat{\rho}_X) - \frac{1}{n}\sum\nolimits_{i\in s}\eta_i^{\mathrm{T}}\left[\sum\nolimits_{j\in s}\eta_j x_j^{\mathrm{T}}\right]^{-1}\eta_i} \tag{2}$$

where $\eta_i = \nabla h(x_i^T\hat{\beta})x_i$ and h is the link function of the logistic regression.

The standard error of the R-indicator is also presented in Shlomo *et al.* (2012).

The coefficient of variation, which standardises the variance of the response propensities, is calculated as the bias-adjusted standard error of the response propensities as shown in (2) divided by the average response propensity and estimated by $CV = \hat{S}_{BIAS-ADJ}(\hat{\rho}_X)/\hat{\bar{\rho}}$ where $\hat{S}^2_{\text{BIAS-ADJ}}(\hat{\rho}_X)$ is the term under the square root in (2). The estimate of the variance of the coefficient of variation is presented in De Heij *et al.* (2015). The coefficient of variation and its estimated standard error as well as all partial coefficients of variation and standard errors are included in the new versions of the SAS and R code on www.risq-project.eu (De Heij *et al.*, 2015).

The unconditional partial R-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 *et al.*, 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 $P_u(Z, \rho_X) = S_B(\rho_X|Z)$ and

$$S_B^2(\rho_X|Z) = \frac{1}{N-1}\sum\nolimits_{k=1}^{K} N_k(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \cong \sum\nolimits_{k=1}^{K}\frac{N_k}{N}(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \tag{3}$$

where $\bar{\rho}_{X,k}$ is the average of the response propensity in stratum k . This between variance is estimated by

$$\hat{S}_B^2(\hat{\rho}_X|Z) = \sum\nolimits_{k=1}^{K}\frac{\hat{N}_k}{N}\left(\hat{\bar{\rho}}_{X,k} - \hat{\bar{\rho}}_X\right)^2, \tag{4}$$

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

where $\hat{\overline{\rho}}_{X,k}$ is the design-weighted stratum mean of the estimated propensities and $\hat{N}_k$ is the estimated population size of stratum k.

At the category level $Z = k$, the unconditional partial R-indicator is defined as

$$P_u(Z,k,\rho_X) = S_B(\rho_X|Z=k)\frac{\left(\overline{\rho}_{X,k}-\overline{\rho}_X\right)}{\left|\overline{\rho}_{X,k}-\overline{\rho}_X\right|} = \sqrt{\frac{N_k}{N}}\left(\overline{\rho}_{X,k}-\overline{\rho}_X\right) \tag{5}$$

and is estimated by

$$\hat{P}_u(Z,k,\hat{\rho}_X) = \hat{S}_B(\hat{\rho}_X|Z=k) = \sqrt{\frac{\hat{N}_k}{N}}\left(\hat{\overline{\rho}}_{X,k}-\hat{\overline{\rho}}_X\right). \tag{6}$$

Conditional partial R-indicators measure the remaining variance because of variable Z within subgroups formed by all other remaining variables, denoted by X^- (Schouten *et al.*, 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 $P_c(Z,\rho_X) = S_W(\rho_X|X^-)$ and

$$S_w^2(\rho_X|X^-) = \frac{1}{N-1}\sum\nolimits_{j=1}^{J}\sum\nolimits_{i\in U_j}\left(\rho_X(x_i)-\overline{\rho}_{X,j}\right)^2 \tag{7}$$

and is estimated by

$$\hat{S}_w^2(\hat{\rho}_X|X^-) = \frac{1}{N-1}\sum\nolimits_{j=1}^{J}\sum\nolimits_{i\in s_j} d_i\left(\hat{\rho}_X(x_i)-\hat{\overline{\rho}}_{X,j}\right)^2. \tag{8}$$

At the categorical level of $Z = k$,we restrict the within variance to population units in stratum k and obtain

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

and estimated by

$$\hat{P}_c(Z,k,\hat{\rho}_X) = \sqrt{\frac{1}{N-1}\sum\nolimits_{j=1}^{J}\sum\nolimits_{i\in s_j} d_i\delta_{k,i}\left(\hat{\rho}_X(x_i),-\hat{\overline{\rho}}_{X,j}\right)^2} \tag{10}$$

In order to compare the partial R-indicator values and to select categories that show the strongest under representation, their values need to be accompanied by the standard errors. In earlier papers, however, the properties of partial R-indicators have only been simulated through resampling methods. In practical monitoring and analysis, resampling is too time-consuming and cumbersome. We, therefore, provide analytic bias and standard error approximations.

2.2 Bias Adjustment and Standard Error Approximations for Partial R-indicators

The strategy taken in the bias and standard error analytic approximations strongly resemble those for the overall R-indicator. For the sake of brevity, we, therefore, give only condensed derivations here. We refer to Shlomo *et al.* (2012) for an elaborate description.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

2.2.1 Bias adjustment

Empirical work has shown that the size-dependent bias affecting the R-indicator in (2) 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 for the partial R-indicators, defined as the decomposed variance components of the variance of the response propensities, their variance terms become more dominant compared with the bias terms. Therefore, for smaller sample sizes, we adopt a method of pro-rating the bias correction term of (2) 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 variances for each of the stratifying variables, we can implement a pro-rating of the bias correction term for (2) between these variance terms when sample sizes are small.

2.2.2 Standard error approximations for variable-level partial R-indicators

To obtain the variance estimates for the variable-level partial R-indicators, we observe that for the unconditional partial R-indicator $P_u(Z, \rho_X) = S_B(\rho_X|Z)$, we can obtain an estimate of the variance according to the methodology of obtaining the estimated variance of the overall R-indicator as set out in Shlomo *et al.* (2012) but with the change that the response propensities are modelled according to a stratification on the single variable Z. Similarly, for the conditional partial R-indicator $P_c(Z, \rho_X) = S_W(\rho_X|X^-)$, we can obtain an approximation of the variance according to the methodology of the overall R-indicator but with the change that the response propensities are modelled according to a stratification on X^-. This is an approximation because only main effects and second-order interactions are typically used to estimate response propensities in the logistic regression model as opposed to a complete cross-classification of auxiliary variables on X^- because many categories would have very small or zero sample sizes.

2.2.3 Standard error approximation for unconditional category-level partial R-indicators

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

The variance of the estimated unconditional category-level partial R-indicator, $\hat{P}_u(Z, k, \hat{\rho}_X)$ in (6) can be written as

$$Var\left(\hat{P}_u(Z,k,\hat{\rho}_x)\right) = \frac{\hat{N}_k}{N} Var\left(\hat{\bar{\rho}}_{x,k} - \hat{\bar{\rho}}_x\right) = \frac{\hat{N}_k}{N}\left[Var\left(\hat{\bar{\rho}}_{x,k}\right) + Var\left(\hat{\bar{\rho}}_x\right) - 2Cov\left(\hat{\bar{\rho}}_{x,k}, \hat{\bar{\rho}}_x\right)\right] \quad (11)$$

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. Because

$$\hat{\bar{\rho}}_X = \frac{\hat{N}_k}{N}\hat{\bar{\rho}}_{X,k} + \left(1 - \frac{\hat{N}_k}{N}\right)\hat{\bar{\rho}}_{X,k^C}$$

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

where

$$\hat{\bar{\rho}}_{X,k^C} = \sum_{i \in s} d_i \hat{\rho}_i \left(1 - \delta_i^k\right) / \left(N - \hat{N}_k\right),$$

we have that

$$Cov\left(\hat{\bar{\rho}}_{X,k}, \hat{\bar{\rho}}_X\right) = \frac{\hat{N}_k}{N} Var\left(\hat{\bar{\rho}}_{X,k}\right) \tag{12}$$

and from (11) and (12)

$$Var\left(\hat{P}_u(Z,k,\hat{\rho}_X)\right) = \frac{\hat{N}_k}{N}\left[\left(1 - \frac{\hat{N}_k}{N}\right)^2 Var\left(\hat{\bar{\rho}}_{X,k}\right) + \left(1 - \frac{\hat{N}_k}{N}\right)^2 Var\left(\hat{\bar{\rho}}_{X,k^C}\right)\right]. \tag{13}$$

We restrict ourselves to a first-order approximation and approximate $Var(\hat{\bar{\rho}}_{X,k})$in (13), where $\hat{\bar{\rho}}_{X,k}$ is the average response propensity under the response model of X^- for category k, by a standard design-based variance estimator of $\sum_{i \in s} d_i \hat{\varphi}_i$, where $\hat{\varphi}_i = \delta_i^k \hat{\rho}_i / \hat{N}_k$, and approximate $Var(\hat{\bar{\rho}}_{X,k^C})$ in (13), where $\hat{\bar{\rho}}_{X,k^c}$ is the average response propensity under the response model of X^- when not in category k, 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 (13).

2.2.4 Standard error approximation for conditional category-level partial R-indicators

For the conditional category-level partial R-indicator, $\hat{P}_c(Z,k,\hat{\rho}_X)$ in (10), we use similar methodology as the variance estimation of the R-indicator described in Shlomo *et al.* (2012), but we add in the stratification variable X^- indexed by $j = 1, 2, ..., J$. The estimate of the variance of the R-indicator was based on the decomposition of $\hat{S}^2(\hat{\rho}_X)$ into the part induced by the sampling design for a fixed value of $\hat{\boldsymbol{\beta}}$ and the part induced by the distribution of $\hat{\boldsymbol{\beta}}$ as obtained from the logistic regression response model. We take the latter to be $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{J}(\boldsymbol{\beta})^{-1} var\{\sum_s d_i [R_i - h(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i\}\mathbf{J}(\boldsymbol{\beta})^{-1}$ and $\mathbf{J}(\boldsymbol{\beta}) = E\{\mathbf{I}(\boldsymbol{\beta})\}$ is the expected information. The estimate of the variance for the conditional category-level partial R-indicator $\hat{P}_c(Z,k,\hat{\rho}_X)$ is given by

$$var\left(\hat{P}_c\left(Z,k,\hat{\rho}_X\right)\right) \approx var_s\left[\sum_{j=1}^{J} \sum\nolimits_{s_k} u_{ji}\right] + 4\mathbf{A}' \sum \mathbf{A} + var_{\hat{\beta}}\left\{tr\left[\mathbf{B}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}'\right)\right]\right\} \tag{14}$$

where u_{ji} replaces $d_i(\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})^2$, and the first term in (14) is treated as the standard design-based variance under a stratified sample design of a linear statistic. For the latter terms in (14), we replace the $\mathbf{A}$ and $\mathbf{B}$ from the derivations in Shlomo *et al.* (2012) under the stratified design: $\hat{\mathbf{A}} = N^{-1} \sum_{j=1}^{\mathrm{J}} \sum\nolimits_{S_k} d_i \delta_i^{\mathrm{k}} (\hat{\bar{\rho}} - \hat{\bar{\rho}}_{X=j})(\hat{z}_i - \hat{\bar{z}}_{x=j})$ and

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

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

2.3 Building Non-respondent Profiles from Auxiliary Variables

Schouten *et al.* (2012) argue that partial R-indicators can be used to improve survey design. However, this claim has not been substantiated in their paper. In this section, we show how to employ the partial R-indicators to form non-respondent strata. It must, however, be noted that the indicators are no pre-requisite to building non-respondents profiles; other statistics exist and can be applied to do the same. The utility of the partial R-indicators lies in four properties: (1) they can be computed at the variable level; (2) they form a suite with R-indicators and naturally allow for a top-down analysis; (3) they have analytical expressions of variance and hence can be used in statistical testing; and (4) they incorporate the size and impact on representativeness of subpopulations based on those variables included in the response models.

The construction of sensible non-respondent profiles and the efficacy of adaptive survey designs in reducing non-response error are fully dependent on the relevance of available auxiliary variables. Partial R-indicators, and any other proxy measure for that matter, are merely tools to transform and condense multi-dimensional information to useful and manageable dimensions. As for non-response adjustment in the estimation, adaptive survey design that are based on variables with a weak relation to survey target variables may be counterproductive and increase imprecision (Little & Vartivarian, 2005). The pre-selection of relevant auxiliary variables is, therefore, crucial, and the absence of such variables should warrant against adaptive survey designs. The utility of adaptive survey design, regardless of adjustment in the estimation, is proportional to the strength of the association between auxiliary variables and survey target variables.

There are two main differences between the pre-selection of auxiliary variables for adaptive survey design and for adjustment afterwards. The first difference is that adaptive survey design strata need to be formed before or during data collection, while non-response adjustment strata can be formed after data collection is completed. For some surveys, the actual publication date of their statistics may be several months after the completion of data collection so that there is sufficient time to update some of the auxiliary variables to the reference period of the survey. The second difference lies in the required properties of the auxiliary variables. Because non-response adjustment is performed when non-response is a given fact, auxiliary variables need to relate to key survey variables and to the specific realised non-response mechanism. However, because adaptive survey design is affecting the actual response propensities, auxiliary variables need to relate to key survey variables and to all likely non-response mechanisms linked to the candidate design features. Hence, for adaptive survey design, the set of auxiliary variables should be broader.

Importantly, in the selection of auxiliary variables, the variables should not be strongly collinear. Ideally, the variables should be more or less independent and cover different dimensions of the target population. In practice, it is inconvenient to first construct such variables from the available list of variables (e.g. as principal components or factors) as the resulting variables are difficult to interpret and to translate to effective data collection treatments. Hence, it is better to use a selection of meaningful variables, although they may correlate to some extent. Conditional partial R-indicators are designed to remove remaining collinearity, but they follow from the idea that collinearity is modest.

In this paper, we assume that survey key statistics are population means or totals, and we take the CV as the target proxy non-response measure. Given a pre-selected set of auxiliary variables, non-respondent profiles can be built by entering all pre-selected variables to a (logistic) regression model for response. The next step is to judge whether the resulting CV is acceptable or not, and, hence, whether it is necessary to inspect the variable-level and category-level partial R-indicators. It is important to remark here that it is the effect size that should determine

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

further inspection and not the standard error or significance; for large sample sizes, any partial R-indicator would be significantly different from zero. Adaptation of the design is needed only when the CV is above a specified threshold. Given that R-indicators and coefficients of variation are fully dependent on the choice of auxiliary variables, the choice of a threshold is not straightforward. There are two options: an internal threshold and an external threshold. An internal threshold is a threshold based on one or more earlier waves of the same survey of acceptable quality. An external threshold is a threshold based on one or more other surveys of acceptable quality. Regardless of the type of threshold, the CV threshold should be computed using exactly the same model, that is, the same auxiliary variables with the same categories, the same link function and the same main and interaction regression terms. At Statistics Netherlands, models can be kept constant because survey samples can be linked to the same administrative data, and paradata is standardised over surveys but this may not be true for other institutions. When a CV attains a value above the internal or external threshold providing evidence of a lack of representativity, then non-respondents profiles should be derived from the variables and categories within those variables that have large and significant unconditional and conditional partial R-indicators. In Section 3.3, we discuss how these partial R-indicators can be used in constructing adaptive design strata.

3 Designing an Adaptive Follow-up from Non-respondent Profiles

In this section, we discuss the optimisation of an intervention given a set of strata depending on the frequency and length of the survey data collection. In order to avoid an overly complex approach, we make some pragmatic assumptions: We assume that a survey designer is considering a single intervention in which the first phase is cheaper than the second phase. We assume, furthermore, that the designer anticipates that the second phase is really needed to improve accuracy of the key survey statistics. The last assumption is made because accuracy has two dimensions: bias and variance. Without such an assumption, one would have to check whether a smaller bias using phase 2 outweighs a smaller variance using phase 1, but with a (much) larger sample size.

3.1 Approaches to Allocate Cases for Follow-up

To date, four approaches to the optimisation of adaptive survey designs can be identified in the literature: (1) a trial-and-error approach (e.g. Laflamme and Karaganis, 2010, Luiten and Schouten, 2013); (2) a set of stopping rules (e.g. Lundquist and Särndal, 2013); (3) propensity-based prioritisation (e.g. Peytchev *et al.*, 2010, Wagner, 2013, Wagner and Hubbard, 2014); and and (4) a mathematical optimisation problem (e.g. Schouten *et al.*, 2013b). The approaches vary in their explicit focus on mathematical optimisation, their certainty to be effective, their ability to be linked to candidate data collection strategies, and their reliance on the accuracy of response propensity estimates.

The first approach is a trial-and-error approach. Different population subgroups receive different treatments that have proven to be effective from practical experience and historic survey data. The subgroup response propensities for each treatment are not explicitly modelled or estimated, and costs are only roughly kept at the available budget level. There is no explicit mathematical optimisation. As a result, a successful improvement of quality using the design is uncertain until it is fielded and analysed. Furthermore, the approach may be subjective and not easily reproducible by others. However, there is no dependence on models or estimated response propensities, and it has room to include expert knowledge.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

The second approach is a set of stopping rules to decide whether continued efforts are made to population subgroups. This approach comes closest to the responsive survey design paper by Groves & Heeringa (2006) that refers to phase capacity. Lundquist & Särndal (2013), for example, prolong efforts in subgroups until a lower limit, say 60%, is reached. This approach also does not make use of an explicit mathematical optimisation model, but stopping rules are constructed based on a quality objective function. As a result, this approach has some guarantee that quality is improved. It also allows to some extent for the inclusion of expert knowledge to choose the most effective strategies within subgroups, and it is only mildly sensitive to the accuracy of response propensities.

The third approach is prioritisation of sample units in data collection based on estimated response propensities. The response propensities are estimated during data collection, and the propensities of non-respondents are sorted. The lowest propensity cases have higher priority and receive more effort. This approach is not linked to a specific quality objective function, but it does aim at equalising response propensities. Similar to stopping rules, there is some guarantee that quality is improved. It is, however, more sensitive to accuracy of response propensities than the first two approaches. More importantly, it is harder to link effective data collection strategies as the sorted cases do not have an easy translation into characteristics.

The fourth approach is a fully mathematical formulation and optimisation. The probabilities that subgroups are assigned to treatments, so-called strategy allocation probabilities, form the set of decision variables. The quality objective function and cost and other constraints are explicitly written in terms of these decision variables and optimised using (non)linear programming. If all input parameters, for example, the response propensities, are estimated accurately, then the approach leads to optimal and predictable improvement of quality. However, the approach is sensitive to inaccurate input parameters. Furthermore, the optimization problems can be non-linear and non-convex, depending on the form of the objective function and cost constraints, which may become computationally intractable. The R-indicator and coefficient of variation are examples where the optimisation becomes non-linear.

A practically feasible and pragmatic approach may be in between a full trial-and-error and a full mathematical optimisation: it is robust but has some mathematical rigour, objectivism and structure and allows for quality-cost trade-offs. We call such an approach a structured trial-and-error approach. The stopping rules and propensity-based prioritisation come close to such an approach, but they are not explicitly linked to proxy non-response bias measures. Furthermore, the propensity-based prioritisation is sensitive to sampling variation and cannot easily be linked to effective treatments, and the stopping rules do not allow for an easy quality-cost trade-off. In Section 3.3, we present a structured trial-and-error approach based on partial R-indicator values. Before we do, we discuss a crucial aspect of a survey: the frequency and length of the data collection.

3.2 Types of Surveys

It makes a big difference in designing an adaptive survey design whether a survey has a finite or infinite horizon data collection period, whether it is run only once (or infrequently) or continuously and whether there is strong or weak prior knowledge about the response propensities. Adaptive survey designs are best suited for continuously running surveys with a long time horizon or surveys with strong prior knowledge about the effectiveness of treatments. In surveys with a long time horizon, budgets can be invested in trying different treatments to learn how the target population responds, and there is no immediate need to optimise treatment during data collection. Surveys with strong prior knowledge resemble surveys for which there has been a long period to learn how treatments work. At the opposite of the spectrum, in surveys with

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

weak prior information that are run once and for a short period, the only option is to learn and act during data collection. In this paper, we will consider both extremes.

Suppose a survey runs for m time periods, or it is requested that the design of a survey is left unchanged for m time periods. In practice, the survey may run for a longer time, but method effects, that is, a change of design, are constrained to be absent for this period. Suppose that for each time period statistics need to be produced. This publication frequency implies that there is room only to experiment with the design during time period 1 and room to optimise the design before the statistics for time period 1 are published. Once data collection starts for time period 2, the design needs to be fixed. Suppose further that, in each period, data collection is split into two phases. The first phase is conducted for the full sample, but the second phase can only be conducted for a proportion q of the non-respondents because of budget constraints. Now, say that a follow-up costs the same for each non-respondent. Then, over the full length of the survey, qm of a one time period full follow-up budget is available. In the first time period, a subsample of proportion p of the non-respondents may receive follow-up, where, generally, p will be larger than q, so that an investment is made in the first time period. For the remaining m-1 time periods, a budget of qm-p is left, which amounts to $(qm\text{-}p)/(m\text{-}1)$ per time period. For a continuous survey with a long time horizon, $(qm\text{-}p)/(m\text{-}1) \approx q$ and all non-respondents can be allocated for follow-up. For a one-time only survey, it must hold that $p = q$.

Apart from the length and budget of the survey, the choices of p and the individual subsampling probabilities depend on the strength of the prior knowledge about the phase 2 response propensities. The inclusion of such knowledge demands a Bayesian approach, which is beyond the scope of the present paper. In the following, we assume that only weak knowledge exists about the overall phase 2 response rate and costs.

3.3 A Structured Trial-and-Error Approach

We consider the two extreme settings: a one-time only survey and a continuous survey with a long time horizon. For the one-time only survey, we suggest the following steps, in which n_R is the size of the non-response after phase 1:

(1) After phase 1 of the one-time only survey derive the CV, R-indicator and partial R-indicators;
(2) Adapt to the non-response by
   (a) Inspect the variable-level partial R-indicators and select variables for which unconditional and conditional partial R-indicators are significantly different from zero;
   (b) Select all categories of those variables identified in (a) that have a significant negative unconditional value and a significant (positive) conditional value;
   (c) Form a stratification by crossing all categories and, possibly, collapse empty or small strata;
   (d) Compute the category-level unconditional partial R-indicator for the new stratification variable, and order the strata by their sign and their p-value with respect to the null-hypothesis that the partial R-indicator is equal to zero;
   (e) Select strata for follow-up based on their rank until pn_R cases are selected.

For the first wave of a continuous survey, the steps are as follows:

(1) After phase 2 of the continuous survey derive the CV, R-indicator and partial R-indicators;
(2) Adapt to the non-response by
   (a) Inspect the variable-level partial R-indicators and select variables for which unconditional and conditional partial R-indicators are significantly different from zero;

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

(b) Select all categories of those variables identified in (a) that have a significant positive unconditional value and a significant (positive) conditional value;
(c) Form a stratification by crossing all categories and, possibly, collapse empty or small strata;
(d) Compute the category-level unconditional partial R-indicator for the new stratification variable, and order the strata by their sign and their p-value with respect to the null-hypothesis that the partial R-indicator is equal to zero;
(e) Deselect strata until at most pn_R cases remain with the highest p-values.

In both settings, it is assumed that adaptation is needed. For the one-time only survey (setting 1), we have partial response, and there is a need to decide on strata to target for follow-up. It is assumed that the CV values do not satisfy the prescribed threshold. For the continuous survey (setting 2), we have a full response from the first phase of a survey, and there is a need to deselect strata for follow-up in the subsequent phase. It is assumed that the required budget to apply phase 2 to all non-respondents is too small.

The aforementioned approaches provide structure but still are essentially trial-and-error. There is no guarantee that the adaptation leads to an optimal allocation and better accuracy. In Sections 4, we analyse the first setting in a simulation study, and in Section 5, we analyse both settings in a real application.

4 A Simulation Study

For the simulation study, we use a data set from the 1995 Israel Census Sample of individuals aged 15 years and over. This was an equal probability sample of the Census population with a 1:5 sampling fraction. This data set is of size N=753 711 and will serve as our population for the simulation study. Population response propensities were calculated using a two-step process:

(1) Probabilities of response are defined according to variables: child indicator, income from earnings groups, age groups, sex, number of persons in household and three types of localities where the categories of the variables are described in Table 1. Some of the variables are at household level because typical social surveys will have the household as the sample unit, and certain variables impact on response, for example, if there are children in the household, but all individuals aged 16 years and over respond to the questionnaire. These variables define groups that are known to have differential response rates, and the probabilities of response are those found in social surveys in practice. Based on the probabilities of response, we generated a response indicator obtaining a value of 0 for a non-response and a value of 1 for a response according to a draw from the Bernoulli distribution.
(2) Using the response indicator as the dependent variable, we fit a logistic regression model on the population using the variables defined in step 1 as explanatory variables and where type of locality and size of household were interacted. The predictions from this model serve as the 'true' response propensities for our simulation study.

The overall response rate generated in the population data set was 69.2%. Table 1 presents the probabilities of response according to the variables in step 1 that were used to generate the population response propensities in step 2. High non-response rates in categories are likely to cause the subgroup in the population to be under-represented according to the partial R-indicators.

From the population, we drew three samples: 1:50 sample (sample size of 15 074), 1:100 sample (sample size of 7 537) and 1:200 sample (sample size of 3 769), using simple random sampling and generated a response indicator for each sample from a random draw of the

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

Table 1. *Percent response generated in the simulation population data set 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 (years)	15–17	77.4	Income group	Low	71.1
	18–21	65.2		2	67.8
	22–24	62.5		3	67.7
	25–34	64.6		4	67.5
	35–44	68.7		High	66.4
	45–54	72.2	Number of persons	1	68.5
	55–64	71.0	in household	2	66.4
	65–74	76.3		3	73.2
	75+	81.3		4	75.6
Type of locality	Type 1	66.7		5	68.2
	Type 2	70.7		6+	68.5
	Type 3	70.3			

Table 2. *R-indicators and coefficient of variation (with confidence intervals) for the three samples before and after targeted follow-up assuming 50% response conversion.*

		Original sample			With targeted follow-up non-response (50% response conversion)	
Sample	Response rate Original (%)	*R*-indicator	Coefficient of variation	Response rate Final (%)	R-indicator	Coefficient of variation
1:50	69.6	0.871	0.093	72.3	0.904	0.066
n=15 074		(0.857–0.886	(0.082–0.103)		(0.890–0.919)	(0.056–0.076)
1:100	69.0	0.854	0.105	71.9	0.886	0.079
n=7 537		(0.834–0.875)	(0.090–0.120)		(0.866–0.907)	(0.065–0.094)
1:200	70.1	0.843	0.112	72.8	0.871	0.088
n=3 769		(0.813–0.872)	(0.091–0.133)		(0.842–0.901)	(0.068–0.109)

Bernoulli distribution according to the propensity to respond as defined in the population in step 2. The response rates for each original sample prior to targeting of non-respondents and follow-up are in Table 2 in the second column, and the R-indicators and coefficient of variations are presented on the left side (columns 3 and 4).

Table 3 provides the variable-level partial R-indicators (unconditional and conditional, respectively) with '*' denoting significantly different from zero at the 5% significance level for the original samples on the left hand side of Table 2 (columns 2, 3 and 4). All variables are contributing to lack of representativity. There is generally a larger lack of representativity as the sample sizes get smaller. For the conditional partial R-indicators, which control for the effects of remaining variables, the within variation of the response propensities in categories of variables remains large. In other words, conditioning on other variables, response is still not representative with respect to the specified variable. In general, we see that the unconditional partial R-indicators are larger than the conditional partial R-indicators in the original sample for all variables. This suggests that the impact of each variable is reduced when controlling for other variables, and that the auxiliary variables show some collinearity.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

Table 3. *Variable-level partial R-indicators for the sample before and after targeted follow-up assuming 50% response conversion.*

	Original sample			With targeted follow-up (50% response conversion)		
Variable	1:50	1:100	1:200	1:50	1:10	1:20
			Unconditional variable partial R-indicator			
Persons in HH	0.032*	0.040*	0.051*	0.027*	0.034*	0.048*
Type of locality	0.011*	0.014*	0.020*	0.010*	0.011	0.019*
Age group	0.047*	0.054*	0.055*	0.033*	0.035*	0.039*
Children in HH	0.030*	0.033*	0.036*	0.014*	0.017*	0.021*
Income group	0.018*	0.031*	0.027*	0.011*	0.026*	0.021*
Sex	0.019*	0.012*	0.013	0.010*	0.018*	0.015*
			Conditional variable partial R-indicator			
Persons in HH	0.029*	0.033*	0.047*	0.029*	0.032*	0.046*
Type of locality	0.011*	0.013*	0.021*	0.009*	0.010*	0.020*
Age group	0.046*	0.050*	0.052*	0.037*	0.037*	0.041*
Children in HH	0.017*	0.017*	0.014	0.008*	0.009	0.004
Income group	0.005	0.022*	0.016*	0.007	0.022*	0.017*
Sex	0.017*	0.011*	0.010	0.009*	0.017*	0.016*

HH, household.
*Significance at the 5% significance level.

Table 4. *Category level (unconditional and conditional) partial R-indicators for the 1:50 original sample.*

Variable	Category	Uncond. partial	Cond. partial	Variable	Category	Uncond. partial	Cond. partial
Children in HH	None	−0.015*	0.012*	Locality type	Type 1	−0.010*	0.009*
	1+	0.026*	0.013*		Type 2	0.005*	0.004*
					Type 3	0.001	0.005
Age group (years)	15–17	0.020*	0.005*				
	18–21	−0.017*	0.021*	Sex	Male	−0.014*	0.013*
	22–24	−0.015*	0.013*		Female	0.013*	0.012*
	25–34	−0.016*	0.011*	Persons in HH	1	−0.007	0.012*
	35–44	−0.005	0.011*		2	−0.015*	0.008*
	45–54	0.005	0.007*		3	0.007	0.007*
	55–64	0.002	0.009*		4	0.025*	0.022*
	65–74	0.018*	0.020*		5	−0.003	0.008*
	75+	0.026*	0.026*		6+	−0.005	0.008*

HH, household.
*Significance at the 5% significance level.

We use the first setting of the structured trial-and-error approach in Section 3.3 to determine characteristics of non-respondents to target for follow-up. Based on the variable-level partial R-indicators, we inspect variables and choose those where the unconditional and conditional values are significantly different from zero as denoted by the '*' in the left-hand panel of Table 3. On the larger sample size 1:50, this check distinguishes the variables: number of persons in the household, type of locality, age group, child indicator and sex.

We next inspect the categories of these variables and determine which categories have a significant negative unconditional partial R-indicator (under-represented in the original sample) and a significant conditional value. For the 1:50 original sample (prior to the targeted follow-up of non-respondents), the category-level partial R-indicators are presented in Table 4, where '*' denotes significantly different from zero at the 5% significance level.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

As can be seen in Table 4, the categories that meet the requirements are as follows: males, persons aged between 18 and 34 years, persons living in two-person households, persons living in households without children and the first type of locality. We then form 32 strata defined by cross-classifying the following sets: {males, females}×{aged 18–34 years, other}×{two persons, other}×{no children, has children}×{locality type 1, other}. The unconditional categorical partial R-indicators are calculated for each of the new strata, and the strata were then sorted by their p-value with respect to the null-hypothesis that the partial R-indicator is equal to zero. For the 1:50 sample, the high and significant p-values on the under-represented strata were obtained for the following sets in order of significance: {males, aged 18–34 years, two persons, no children, type 1}; {males, aged 18–34 years, two persons, no children, not type 1}; {males, aged 18–34 years, not two persons, no children, type 1}; {males, aged 18–34 years, not two persons, no children, not type 1}. We repeated this process of identifying strata for the 1:100 and 1:200 samples. The number of non-respondents to target for follow-up in the identified strata are 838 (5.6%), 421 (5.6%) and 188 (5.0%) for the 1:50, 1:100 and 1:200 samples, respectively. We assume that after efforts to convert the non-respondents to respondents, we achieve a 50% response rate in the follow-up. For the simulation study, half of the non-respondents in the strata were randomly converted to respondents assuming that within each strata, the probabilities to respond are homogenous under the MAR assumption. This increased the response rates by approximately 2.7%, as can be seen in column 5 of Table 2. In addition, Table 2 presents the R-indicators and coefficients of variation after the targeted follow-up of non-response assuming a 50% conversion response rate (columns 6 and 7). There is evidence of a significant increase in the R-indicator and a decrease in the coefficient of variation after the non-response follow-up when comparing confidence intervals.

We return now to Table 3, containing the variable-level partial R-indicators, and focus on the right side of the table (columns 5, 6 and 7) after the 50% response conversion to the targeted follow-up. Based on the results, we generally see the same trend as the R-indicators with a reduction in the variable level partial R-indicators, although some collinearity has remained. In the 1:200 sample size, we see that the variable sex, which is a dichotomous variable, has gone from non-significant to significant, following the targeted response for both conditional and unconditional partial R-indicators. For the categorical level partial R-indicators (not shown here), there is an overall reduction following the targeted response, and many categories have become non-significant.

The conclusion from this simulation study is that even with a small increase in the overall response rate, albeit targeted at those non-respondents contributing to the lack of representativity, we are able to improve the representativeness of the data. This means that fewer adjustments will be needed to correct for non-response bias under post-stratification, leading to smaller variation in sampling weights. In each of the three samples drawn, post-stratification weights were calculated for individuals in the survey (ignoring household-level weighting) using standard weighting classes of cross-classified type of locality, age group and sex. Before the targeted follow-up, the CVs of the survey weights were 10.0%, 15.6% and 17.8% for the 1:50, 1:100 and 1:200 samples, respectively. Following the targeting of non-respondents and conversion, the CV's of the survey weights were 9.7%, 15.0% and 16.3%. The weighted estimate for the average income in Israeli Shekels (IS) for those employed went from 4 340 IS, 4 327 IS and 4 326 IS for the 1:50, 1:100 and 1:200 samples, respectively, before targeting non-respondents, to 4 342 IS, 4 358 IS and 4 354 IS following targeting of non-respondents. The true value in the population is 4 377 IS.

5 An Application to the Crime Victimisation Survey

In this section, we show an application to the 2011 Dutch Crime Victimisation Survey (CVS). Within the 2011 CVS, a large survey mode experiment was conducted that allows us to investigate various sequential mixed-mode adaptive survey designs. This experiment is described

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

and analysed in detail in Schouten *et al.* (2013a). In the construction of the adaptive survey designs, we adopt the two extreme settings: no learning period (a one-time only survey) and a long learning period (a continuous survey).

The design of the experiment was as follows: A sample of 8 800 persons was randomly assigned to one of four sequential mode strategies: Web followed by face-to-face, mail followed by face-to-face, telephone followed by face-to-face and face-to-face followed by face-to-face. The last strategy, face-to-face followed by face-to-face, is not a mixed-mode strategy but was added to evaluate time stability of CVS key variables. The experiment was designed to decompose mode effects into mode-specific selection and mode-specific measurement bias with face-to-face as the benchmark mode. In order to do so, both respondents and non-respondents to the first phase (Web, mail, telephone or face-to-face) received the second phase (face-to-face) in which the first key sections of the CVS questionnaire were repeated. At the first phase, persons were not aware of a second phase. In Schouten *et al.* (2013a), it was concluded that the mode of the first phase did not impact the size and composition of response to the second phase. Furthermore, the answers to the key CVS questions in the second phase could not be predicted by the mode of the first phase. We view the two phases, therefore, as a sequential mixed-mode design.

In the application, we consider two strategies: Web to face-to-face and mail to face-to-face. These two strategies come up naturally in mixed-mode designs where cheaper survey modes are tried first. We assume that there is insufficient budget to allocate all non-respondents to face-to-face, and we investigate which persons to allocate to the face-to-face second phase. In the following, we abbreviate face-to-face to F2F.

The evaluation of representativeness and the construction of strata for the second phase is performed using six socio-demographic registry variables: gender (male, female), age (15–25, 25–35, …, 65–75, 75+ years), urbanisation of residence (not, little, moderate, strong, very strong), income in Euro's (<3K, 310K, 10–15K, 15–20K, 20–30K, >30K), ethnicity (native, western non-native, non-western non-native) and registered landline phone number (yes, no). These variables have been linked to a wide range of survey data sets at Statistics Netherlands as they are related generally to survey variables and are used in weighting adjustments and publication tables. Particularly, gender, age, urbanisation and registration of a landline phone number relate strongly to key CVS variables: victimisation, perception of safety, judgement of police performance and perception of neighbourhood problems. Table 5 presents the response rate, R-indicator and coefficient of variation for various strategies given the specified auxiliary variables. The last row of Table 5 presents the values for the strategy with two F2F phases.

The response rate of this strategy is close to 70% and the R-indicator is around 0.80. In the following, we take the resulting CV of 0.160 as the target. We believe that two F2F phases

Table 5. *Response rate, R-indicator, coefficient of variation and costs for various strategies in the 2011 CVS experiment. Standard error approximations are given within brackets. Costs are given relative to the cost of one sample unit in Web.*

Strategy	Response rate	R-indicator	CV	Cost
Web	28.7%(1.0%)	0.806 (0.019)	0.368 (0.034)	1
Web → F2F	57.9%(1.1%)	0.829 (0.022)	0.168 (0.019)	22.3
Web setting 1	39.7%(1.0%)	0.808 (0.021)	0.267 (0.026)	9.1
Web setting 2	43.6%(1.1%)	0.846 (0.021)	0.206 (0.025)	13.2
Mail	49.0%(1.1%)	0.738 (0.020)	0.283 (0.020)	4.0
Mail→ F2F	66.0%(1.0%)	0.812 (0.021)	0.157 (0.016)	19.2
Mail setting 1	54.1%(1.1%)	0.855 (0.022)	0.159 (0.020)	8.5
Mail setting 2	59.5%(1.1%)	0.878 (0.022)	0.129 (0.019)	12.2
F2F → F2F	67.9%(1.0%)	0.801 (0.021)	0.160 (0.015)	41.5

CVS, Crime Victimisation Survey; CV, coefficient of variation; F2F, face-to-face.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

represent what can be achieved with reasonable effort; beyond this effort, the survey gets exceptionally expensive. We use the F2F → F2F CV as internal benchmark.

Rows 2, 3, 6 and 7 of Table 5 present the indicator values for Web only, Web to F2F, mail only and mail to F2F. The response rate for Web only is by far the lowest, but the R-indicator is similar to the benchmark strategy. However, resulting from the low response rate, the CV is much higher. For mail, the picture is somewhat reversed: the response rate is relatively high but the R-indicator is much lower. As a consequence, the CV is much higher but lower than that of the Web only strategy. When the F2F second phase is added, then the response rates increase considerably; for mail, it is now close to that of the benchmark strategy. The R-indicator and the CV become similar to that of the benchmark strategy.

We assume that the available budget is not sufficient to cover a second phase for all non-respondents in the first phase. The costs for approaching one CVS sample person through mail is approximately four times higher than through Web, and the costs for F2F are approximately 30 times higher. The last column of Table 5 gives the costs per strategy relative to the cost of assigning one sample unit to Web. The F2F to F2F strategy is approximately 45 times more expensive than Web only. For ease of demonstration, suppose that the available budget is 13.800. This implies there is budget to allocate 940 cases to F2F after a Web first phase and 720 cases after a mail first phase. The full F2F strategies for Web and mail cost, respectively, 49.1 and 42.3 per case, and are too expensive to assign all cases.

We adopted two extreme settings. The first setting is that of a one-time only survey or a low frequency survey in which there is no time to learn and to perform a full F2F phase 2. Under this setting, a decision to allocate non-respondents to F2F has to be based on the Web and mail phase 1 responses only. The second setting is that of a continuous survey in which budget can be invested to perform a pilot with a full F2F second phase. Under this setting, a decision to allocate non-respondents can be based using the responses to both phases. Table 5 includes the indicator values of the adaptive survey designs that are constructed under the two settings (rows 4 and 5 for Web and rows 8 and 9 for mail). We constructed the designs following the steps of Section 3.3. Table 6 presents the variable-level conditional and unconditional partial R-indicator under the different settings.

Table 6. *Variable-level unconditional and conditional partial R-indicators for various strategies in the 2011 CVS experiment.*

		Unconditional		Conditional	
		Phase 1	Phase 1 and 2	Phase 1	Phase 1 and 2
Gender	Mail	0.024 ***	0.014	0.040 *	0.024 ***
	Web	0.020 ***	0.003	0.001	0.007
Ethnicity	Mail	0.077 *	0.058 *	0.043 *	0.033 *
	Web	0.039 *	0.047 *	0.022 **	0.021 ***
Income	Mail	0.067 *	0.056 *	0.056 *	0.047 *
	Web	0.077 *	0.046 *	0.053 *	0.032 **
Urbanisation	Mail	0.026 ***	0.026 ***	0.014	0.015
	Web	0.015	0.053 *	0.014	0.034 *
Age	Mail	0.087 *	0.051 *	0.064 *	0.037 *
	Web	0.061 *	0.036 *	0.041 *	0.022 ***
Phone	Mail	0.038 *	0.027 **	0.016	0.011
	Web	0.029 *	0.046 *	0.016	0.026 **

CVS, Crime Victimisation Survey.
* p-value below 0.1%.
** p-value below 1%.
*** p-value below 5%.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

For the first setting, four categories turned up for both Web (income groups 10–15K and 15–20K, age group >75 years and non-western non-natives) and for mail (males, age groups 15–25 years and 25–35 years and non-western non-natives). From these categories, stratifications were formed, and strata with significant negative unconditional values were selected for follow-up; 594 cases for Web and 329 for mail. For the second setting, we found four categories for Web (income group >30K, natives, persons with a registered phone and persons living in little or non-urbanised areas) and five categories for mail (income group >30K, natives, persons with a registered phone and age groups 55–65 years and 65–75 years). From these categories again, stratifications were formed, and strata that did not have significant negative unconditional values were deselected for follow-up; leaving a total of 896 cases for Web and 601 cases for mail for follow-up. For both settings, the numbers of cases were within the budget levels.

Table 5 presents the resulting indicator values. For both settings, obviously, the response rates increases. Under the first setting for Web, the second phase does not improve the R-indicator, while for mail, the second phase leads to a large increase in the R-indicator. The CV for mail has become similar to the target from the F2F to F2F design, while for Web, it is still higher. Under the second setting, the R-indicator values increase for both Web and mail and are significantly higher than for strategy F2F to F2F. Because of the lower response rate, the CV for Web is still higher than the target but for mail it is significantly lower.

In the construction of the designs, we have concentrated on auxiliary variables. The important question is whether the various designs also affect the survey variables. Table 7 contains the design-weighted, non-response-adjusted response means for designs with Web and mail, respectively, of five survey variables observed in phase 2: the number of victimisations per 100 inhabitants, the percentage victimised over the last year, a five-point neighbourhood nuisance scale, the percentage of persons feeling unsafe at times and the percentage of persons being not satisfied with the police. The response means are computed using the phase 2 answer to the repeated question in F2F in order to avoid confounding with mode-specific measurement bias. The non-response adjustment was performed using linear weighting on the six auxiliary variables without interaction terms. The estimates of a full F2F phase 2 and the adaptive survey

Table 7. *Adjusted response means for five CVS survey variables and coefficient of variation for designs with a Web or mail first phase.*

Web	Phase 1	Phase 1 and 2	Setting 1	Setting 2
Coefficient of variation	0.368	0.168	0.267	0.206
# Victimisations per 100	29.5	31.8	27.4	30.5
% Victimised	8.9	11.0 **	9.0	10.5 *
Nuisance scale	1.4	1.4	1.4	1.4
% Unsafe	26.7	25.8	24.8	26.0
% Not satisfied police	45.8	47.0	46.2	46.2
Mail	Phase 1	Phase 1 and 2	Setting 1	Setting 2
Coefficient of variation	0.283	0.157	0.159	0.129
# Victimisations per 100	23.9	23.3	23.9	23.1
% Victimised	10.0	10.4	10.8	10.5
Nuisance scale	1.3	1.3	1.3	1.3
% Unsafe	28.4	25.5 **	26.7	26.0 *
% Not satisfied police	47.9	47.7	47.9	48.4

CVS, Crime Victimisation Survey.
*** p-value below 0.1% for test against phase 1 only.
** p-value below 1% for test against phase 1 only.
* p-value below 5% for test against phase 1 only.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

designs under the first and second settings are tested against the Web only and mail only designs under the null-hypothesis that they are equal. The test is conducted using bootstrap replication. For comparison, CVs are also shown.

The victimisation variable shows a significant difference against a Web only design at the 5% level, the feeling unsafe variable against a Mail only design at the 5% level and the other variables do not. Especially, the neighbourhood nuisance scale seems to be very robust against changes in design. There is some indication that decreases in the CV coincide with significant changes in the victimisation variables; the only design where the % victimised did not change significantly, the first setting for Web, still had a relatively high coefficient of variation. Remarkably, the number of reported victimisations in designs with a Web first phase is higher than those with a mail first phase, although percentages victimised are similar. We have no explanation for this other than differences in selection between the Web to F2F and mail to F2F designs.

The application confirms that building adaptive survey designs based on response to a first phase can be risky. For mail, the second phase allocation turned out right, and all indicators improved, but for Web, hardly any improvement was found; the F2F second phase helped raise response rates of some strata but was counterproductive on other strata. This risk reflects the lack of knowledge about the efficacy of the second phase, which is included in the scenario where both phases have been conducted first. The application shows that it may be fruitful to perform a first-pilot wave in which some investment is made in learning if and how a second phase improves response. After this wave, the design can be optimised for subsequent waves, and statistics for the first wave (and obviously future waves) can be based on the optimised design.

In the application, we performed a structured trial-and-error approach. Under the second setting where we have estimates for the response propensities for both phases, we could have performed an advanced optimisation following Schouten *et al.* (2013b). This would lead to a complex non-linear, non-convex optimisation problem when all variables and all variable interactions are included. If we would first construct stratifications for targeting sample cases, as we have performed here, the properties of the problem are the same but the dimensionality would be much lower, and perhaps, a brute force approach where many options are simply tried would become within reach. We have not tried this for the application, however.

6 Discussion

In this paper, we demonstrated how to use partial R-indicators in forming non-respondent profiles and strata for adaptive survey designs. We, furthermore, presented and demonstrated structured trial-and-error approaches to adaptive survey designs, and we identified two extreme scenarios: a one-time only survey and a continuous survey with a long time horizon. In the approaches, we adopted a pragmatic viewpoint in order to avoid complex optimisation. However, the crucial ingredient to the formation of strata and the efficacy of adaptive survey designs is the availability of auxiliary variables from frame data, administrative data or paradata that are relevant to the key survey variables.

We design and employ adaptive survey designs from the conviction that adjustment by design is profitable also when adjustment afterwards is applied. There are two motivations for this conviction: we want to reduce variation in adjustment weights by design, and we treat proxy measures of non-response error as process quality indicators. We believe, and there is now empirical evidence, that a larger variation in response propensities on known variables is indicative of even higher variation in response propensities on other variables.

In dividing the data collection into two phases, we assume that there is a strong conjecture that a second phase is needed to reduce the impact of non-response error. Although we do test indicators for their significant difference to a fully random non-response, we largely ignore the

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

trade-off between bias and variance that needs to be made when taking the mean square error as the ultimate quality measure.

To date, all approaches towards designing adaptive survey design are non-Bayesian and do not update response propensity distributions during data collection or from one wave to the other. A Bayesian approach is a promising alternative as uncertainty about adaptive survey design input parameters is included in a natural way; see, for example, Wagner & Hubbard (2014). However, such an approach requires a different framework (e.g. Schafer, 2013 for an application to data collection monitoring), and the most challenging element may be how to include and model multiple but diverse key survey variables. We advocate that a Bayesian approach is investigated but leave it to future research to explore this alternative.

Acknowledgments

The authors wish to thank the reviewers for their insightful comments, which greatly improved the manuscript.

References

Calinescu, M., Schouten, B. & Bhulai, S. (2012). Adaptive survey designs that minimize nonresponse and measurement risk. *Discussion paper 201224*, CBS, Den Haag.

Calinescu, M. & Schouten, B. (2013). Adaptive survey designs to minimize mode effects. A case study on the Dutch Labour Force Survey. *Discussion paper 201312*, CBS, Den Haag.

Deville, J. C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, **91**(4), 893–912.

Grafström, A. & Schelin, L. (2014). How to select representative samples? *Scand. J. Stat.*, **41**, 277–290.

Groves, R. M. & Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *J. R. Stat. Soc. Ser. A*, **169**, 439–457.

Hasler, C. & Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Comput. Statist. Data Anal.*, **74**, 81–94.

De Heij, V., Schouten, B. & Shlomo, N. (2015). *RISQ 2.1 manual. Tools in SAS and R for the computation of R-indicators and partial R-indicators*. Available at www.risq-project.eu.

Laflamme, F. & Karaganis, M. (2010). *Implementation of responsive collection design for CATI surveys at statistics.* Helsinki Finland.

Little, R. & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means *Surv. Methodol.*, **31**, 161–168.

Luiten, A. & Schouten, B. (2013). Adaptive fieldwork design to increase representative household survey response: A pilot study in the Survey of Consumer Satisfaction. *J. R. Stat. Soc. Ser. A*, **176**(1), 169–190.

Lundquist, P. & Särndal, C.E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *J. Off. sta.*, **29**(4), 557–582.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. & Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, **4**, 21–29.

Särndal, C. E. (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation. *J. Off. sta.*, **27**(1), 1–21.

Särndal, C. E. & Lundquist, P. (2013). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. Working paper. Statistics Sweden, Sweden.

Schafer, J. L. (2013). Bayesian penalized spline models for statistical process monitoring of survey paradata quality indicators, Chapter 13. In *Improving surveys with Paradata. Analytic Uses of Process Information*, Ed. Kreuter, F., pp. 311–340. New York, USA: Wiley.

Schouten, J. G., Bethlehem, J., Beulens, K., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N. & Skinner, C. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators. *Int. Stat. Rev.*, **80**(3), 382–399.

Schouten, B., Brakel, J.van der, Buelens, B., Laan, J.van der & Klausch, L.T. (2013a). Disentangling mode-specific selection and measurement bias in social surveys. *Soc. Sci. Res.*, **42**, 1555–1570.

Schouten, B., Calinescu, M. & Luiten, A. (2013b). Optimizing quality of response through adaptive survey designs. *Surv. Methodol.*, **39**(1), 29–58.

Schouten, J.G., Cobben, F. & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Surv. Methodol.*, **35**(1), 101–113.

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.

Schouten, B., Cobben, F., Lundquist, P. & Wagner, J. (2014). Theoretical and empirical support for adjustment of nonresponse by design? *Discussion paper 2014-15, Statistics Netherlands, Available at* www.cbs.nl.
Schouten, B., Shlomo, N. & Skinner, C.J. (2011). Indicators for monitoring and improving representativeness of response. *J. Off. sta.*, **27**(2), 231–253.
Shlomo, N., Skinner, C.J. & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *J. Stat. Plan. Inference*, **142**, 201–211.
Wagner, J. (2008). *Adaptive survey design to reduce nonresponse bias*, PhD thesis, University of Michigan, USA.
Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Surv. Res. Meth.*, **7**(1), 45–55.
Wagner, J. & Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *J. Surv. Stat. Methodol.*, **2**(3), 323–342.

[*Received August 2014, accepted October 2015*]

© 2015 The Authors. International Statistical Review © 2015 International Statistical Institute.