Selecting Adaptive Survey Design Strata with Partial R-indicators

Barry Schouten [1] and Natalie Shlomo[2]

[1]Statistics Netherlands and Utrecht University,

Statistics Netherlands
Department of Process Development, IT and Methodology
PO Box 24500, 2490HA Den Haag, The Netherlands
Tel +31 70 3374905
e-mail: jg.schouten@cbs.nl

[2]University of Manchester, (corresponding author)

Social Statistics, School of Social Sciences
University of Manchester
Humanities Bridgeford Street
Manchester M20 2PD, United Kingdom
Tel: +44 161 275 0269
e-mail: natalie.shlomo@manchester.ac.uk,

Abstract

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 optimization 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.

Keywords: Nonresponse; Responsive survey design; Representativeness.

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 2013). 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 nonresponse error and ignores other errors like measurement error. We will restrict ourselves also to nonresponse 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.

The implementation of designs that differentiate design features is marked by the following steps:

1. Choose proxy measures for survey quality;
2. Choose a set of candidate design features, e.g. 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. Optimize 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.

For an elaboration of these steps see Groves and Heeringa (2006), Peytchev et al (2010) and Schouten, Calinescu and Luiten (2013). 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: 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 and 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, Calinescu and Luiten (2013) 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, i.e. any period in between interventions, could be adaptive. Here, we discuss the selection of strata for a single intervention, i.e. 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 nonresponse, 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 nonresponse. The reason for this similarity is very simple: Adaptive survey designs attempt to adjust for nonresponse by design rather than just post-hoc in the estimation. Various authors have come up with explicit proxy measures for nonresponse error in the optimization of adaptive survey designs: Schouten, Cobben and Bethlehem (2009) propose to use representativeness indicators and the coefficient of variation of response propensities and Särndal (2011) and Lundquist and Särndal (2013) propose to use balance indicators. Other authors describe a less explicit approach in which sample units are prioritized based on response propensity models (e.g. Peytchev et al 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 a signal of even stronger imbalance on survey target variables. For a more elaborated discussion see Schouten, Cobben, Lundquist and Wagner (2013) and Särndal and Lundquist (2013). Schouten et al (2013) provide theoretical and empirical evidence that, on average, a design with a more representative response has smaller nonresponse 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, e.g. Deville and Tillé (2004), Grafström and Schelin (2014) and Hasler and Tillé (2014), However, adaptive survey designs attempt to balance response to a given sample, not the sample itself. In other words, adaptive survey designs optimize 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 design 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 nonresponse and to improve survey response through adaptive survey design. The last claim has not been substantiated, however. In this paper, we answer three research questions: 1) Can partial R-indicators be used to identify (and monitor) strata for

adaptive survey designs?, 2) If so, how to optimize intervention for the strata?, and 3) How to account for the frequency and length of the survey data collection in the optimization?

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, Schouten and Shlomo 2014) are available for the computation of R-indicators and partial-R-indicators. The code is extended with standard error approximations for all indicators and other features compared to 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 nonrespondents. In Section 3, we discuss the optimization 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 optimization of the design for a real dataset, the Dutch Crime Victimisation Survey. We conclude with a discussion in Section 6.

2. Building Nonrespondent Profiles

In this section, we discuss the first research question: Can partial R-indicators be used 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 they 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 Partial R-indicators Revisited

We use the notation and definition of response propensities as set out in Schouten, Shlomo and Skinner (2011) 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: $\rho_X(x_i) = E(R_i = 1 | X = x_i) = P(R_i = 1 | X = x_i)$ and denote this response propensity by ρ_X. We assume that the values x_i are known for all sample units, i.e. for both respondents and non-respondents.

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. The estimator of the variance of the response propensities is

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

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

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

The bias adjusted R-indicator as shown in Shlomo, Skinner and Schouten (2012) is:

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

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

The standard error of the R-indicator is also presented in Shlomo, Skinner and Schouten (2012).

The coefficient of variation is calculated as the bias-adjusted standard error of the response propensities as shown in (2) divided by the average response rate and estimated by: $CV = \hat{S}_{BIAS-ADJ}(\hat{\rho}_X)/\hat{\bar{\rho}}$. The estimate of the variance of the coefficient of variation is presented in De Heij, Schouten, Shlomo, 2014. The coefficient of variation and its estimated standard error are included in the new versions of the SAS and R code on www.risq-project.eu (De Heij, Schouten, Shlomo, 2014).

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 $P_u(Z,\rho_X) = S_B(\rho_X \mid Z)$ and

$$S_B^2(\rho_X \mid Z) = \frac{1}{N-1}\sum_{k=1}^{K} N_k(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \cong \sum_{k=1}^{K}\frac{N_k}{N}(\bar{\rho}_{X,k} - \bar{\rho}_X)^2 \qquad (3)$$

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

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

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

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

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

and is estimated by

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

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 $P_c(Z,\rho_X) = S_W(\rho_X \mid X^-)$ and

$$S_w^2(\rho_X \mid X^-) = \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in U_j}(\rho_X(x_i) - \bar{\rho}_{X,j})^2 \tag{7}$$

and is estimated by

$$\hat{S}_w^2(\hat{\rho}_X \mid X^-) = \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in s_j} d_i(\hat{\rho}_X(x_i) - \hat{\bar{\rho}}_{X,j})^2 \,. \qquad (8)$$

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

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

and estimated by

$$\hat{P}_c(Z,k,\hat{\rho}_X) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{J}\sum_{i\in s_j} d_i\delta_{k,i}(\hat{\rho}_X(x_i),-\hat{\bar{\rho}}_{X,j})^2} \qquad (10)$$

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, Skinner and Schouten (2012) for an elaborate description.

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 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 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 variance for each of the stratifying variables, we can implement a pro-rating of the bias correction term for (2) between the complimentary between and within variances for small sample sizes.

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 $P_u(Z,\rho_X)=S_B(\rho_X \mid Z)$ 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, Skinner and Schouten (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 $P_c(Z,\rho_X)=S_W(\rho_X \mid X^-)$ we can obtain an estimate 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 approximation is due to the fact that 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.

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

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 $k = 1,2,\ldots,K$.

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. 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

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

and from (11) and (12)

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

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 (13).

2.2.4 Standard error approximation for conditional category-level partial R-indicator:

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, Schouten, and Skinner (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 $\boldsymbol{\Sigma} = \mathbf{J}(\boldsymbol{\beta})^{-1}\,\mathrm{var}\{\sum_s d_i[R_i - h(\mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i\}\mathbf{J}(\boldsymbol{\beta})^{-1}$ 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:

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

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 **A** and **B** from the derivations in Shlomo, Skinner and Schouten (2012) under the stratified design: $\hat{\mathrm{A}} = N^{-1}\sum_{j=1}^{J}\sum_{s_k} d_i\delta_i^k(\hat{\rho}_i - \hat{\bar{\rho}}_{X=j})(\hat{z}_i - \hat{\bar{z}}_{X=j})$

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

2.3 Building nonrespondents profiles from auxiliary variables

Schouten et al (2012) argue that partial R-indicators can be used to improve survey design. This claim has not been substantiated, however, in their paper. In this section, we show how to employ the indicators to form nonrespondent strata. It must, however, be noted that the indicators of course are no prerequisite to building nonrespondents profiles; other statistics exist and can be applied to do the same. The utility of the indicators lies in three 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, , and 3) they incorporate the size and impact of subpopulations.

The construction of sensible nonrespondent profiles and the efficacy of adaptive survey designs in reducing nonresponse 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 such multi-dimensional information to useful and manageable dimensions. As for nonresponse 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 and Vartivarian 2005). The pre-selection of relevant auxiliary variables is, therefore, crucial, and the absence of such variables should warrant against adaptive survey designs.

Schouten, Cobben, Lundquist and Wagner (2013) show that when auxiliary variables are a random selection from the universe of all variables on a population, then larger values of the R-indicator and coefficient of variation for one of the candidate designs imply larger expected values of the indicators for any other randomly drawn variable on that design. They further show that under the same condition a larger coefficient of variation implies a larger expected remaining nonresponse bias for commonly used estimators that employ the selected variables to adjust bias of an arbitrary other variable. Finally, they show that when the random selection of variables is from a subset of the universe of all variables, for example those variables that correlate above a given threshold with survey target variables, then these two results still hold for any other variable selected from that subset. Although, we cannot assume that auxiliary variables originate from a random selection process, these conclusions do provide guidance on the importance of auxiliary variables: 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 nonresponse 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 link additional administrative data or population tables and to update some of the variables to the reference period of the survey. Consequently, the set of auxiliary variables for nonresponse adjustment may be much larger and the variables may be updated. The second difference lies in the required properties of the auxiliary variables. Since nonresponse adjustment is done when nonresponse is a given fact, auxiliary variables need to relate to key survey variables and to the specific realized nonresponse mechanism. However, since adaptive survey design is affecting the actual response propensities, auxiliary variables need to relate to key survey variables and to all likely nonresponse mechanisms linked to the candidate design features. Hence, for adaptive survey design, the set of auxiliary variables should be made wider.

Importantly, in the selection of auxiliary variables, it should be avoided that the variables are 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 survey key statistics are population means or totals and we take the coefficient of variation (CV) as the target proxy nonresponse measure. Given a pre-

selected set of auxiliary variables, nonrespondent 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 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, it is, however, not straightforward how to choose such a threshold. 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. When a CV attains a value above the internal or external threshold, then nonrespondents profiles should be derived from the variables and categories within those variables that have large and significant unconditional and conditional values. In section 3.3, we discuss how these categories can be used in constructing adaptive design strata.

3. Designing an adaptive follow-up from nonrespondent profiles

In this section, we discuss the other two research questions: How to optimize intervention given a set of strata?, and how to account for the frequency and length of the survey data collection in the optimization? 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 only with a (much) larger sample size.

3.1 Approaches to allocate cases for follow-up

To date, four approaches to the optimization 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 prioritization (e.g. Peytchev et al 2010, Wagner 2013, Wagner and Hubbard 2013), and 4) a mathematical optimization problem (e.g. Schouten, Calinescu and Luiten 2013). The approaches vary in their explicit focus on mathematical optimization, their certainty to be effective, their ability to be linked to candidate data collection strategies, their reproducibility, 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 optimization. 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. Luiten and Schouten (2013) describe such an approach that did lead to improved (partial) R-indicator values for the Survey of Consumer Sentiments without exceeding budget levels.

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 and Heeringa (2006) that refers to phase capacity. The approach is implemented and studied by Lundquist and Särndal (2013). Lundquist and Särndal estimate response propensities for subgroups and prolong efforts in subgroups until a lower limit, say 60%, is reached. This approach also does not make use of an explicit mathematical optimization 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 prioritization of sample units in data collection based on estimated response propensities. The response propensities are estimated during data collection and the propensities of nonrespondents 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 equalizing response propensities; like the 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. See for example Wagner and Hubbard (2013) for a discussion.

The fourth approach is a fully mathematical formulation and optimization as is presented by Schouten, Calinescu and Luiten (2013). 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 optimized using (non)linear programming. If all input parameters, e.g. 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 nonlinear 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 optimization becomes nonlinear.

A practically feasible and pragmatic approach may be in between a full trial-and-error and a full mathematical optimization: it is robust but has some mathematical rigor, 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 prioritization come close to such an approach, but they are not explicitly linked to proxy nonresponse bias measures. Furthermore, the propensity-based prioritization 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, budget can be invested in trying different treatments to learn how the target population responds and there is no immediate need to optimize treatment during data collection. Surveys with strong prior knowledge resemble surveys for which there has been a long time to learn how treatments work. At the opposite of the spectrum, in surveys with 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, i.e. 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 optimize 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 nonrespondents due to budget constraints. Now, say that a follow-up costs the same for each nonrespondent. 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 nonrespondents 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 $(qm-p)/(m-1)$ per time period. For a continuous survey with a long time horizon, $(qm-p)/(m-1) \approx q$ and p can be taken equal to 1. 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 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 scenarios: a one-time only survey and a continuous survey with a long-time horizon. For the first scenario, we suggest the following steps:

1. After phase 1 derive the CV, R-indicator and partial R-indicators;
2. Adapt to the nonresponse by:
    a. Inspect the variable-level partial R-indicators and select variables for which unconditional and conditional values are significantly different from zero;
    b. Select all categories of those variables that have a significant negative unconditional value and a significant 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 *p*-value;
e. Select strata for follow-up based on their rank until *p* cases are selected.

For the second scenario, the steps are:

1. After phase 2 derive the CV, R-indicator and partial R-indicators;
2. Adapt to the nonresponse by:
    a. Inspect the variable-level partial R-indicators and select variables for which unconditional and conditional values are significantly different from zero;
    b. Select all categories of those variables that have a significant positive unconditional value and a significant 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 *p*-value;
    e. Select strata for follow-up based on their rank until *p* cases are selected.

In both scenarios it is assumed that adaptation is needed. Under scenario 1, it is assumed that the indicator values do not satisfy the prescribed threshold. Under scenario 2, it is assumed that the required budget to apply phase 2 to all nonrespondents is too small. The above 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 and 5, we analyse the scenarios in a simulation study and a real application..

4. A simulation study

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 4.1 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 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/nonresponse indicator according to the propensity to respond as defined in the population. The response rates for each original

sample are in Table 1 in the second column and the R-indicators and coefficient of variations are presented on the left side (columns 3 and 4) of Table 4.2.

Table 4.1: 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	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 in Household	1	68.5
	55-64	71.0		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 4.3 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 (columns 2,3 and 4). All variables are contributing to lack of representativity. There is generally more impact on the 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, there remains a lack of representative response for the specific 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.

Table 4.2: R-indicators and Coefficient of Variation (with confidence intervals) for the three sample before and after targeted follow-up assuming 50% response

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

We use the first scenario of the structured trial-and-error approach in Section 3.3 to determine characteristics of individuals to target for follow-up on non-respondents. Based on the variable level partial indicators, we inspect those variables where the unconditional and conditional values are significantly different from zero as denoted by the '*' in the left-hand panel of Table 4.3. 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 4.4 where '*' denotes significantly different from zero at the 5% significance level.

Table 4.3: Variable level Partial R-indicators ('' denotes significance at the 5% significant level) for the sample before and after targeted follow-up assuming 50% response*

Variable	*Original Sample*			*With Targeted Follow-up (50% Response rate)*		
	1:50	1:100	1:200	1:50	1:100	1:200
Unconditional Variable Partial R-indicators						
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-indicators						
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*

As can be seen in Table 4.4, the categories that meet the requirements are: males, persons aged between 18 and 34, persons living in 2-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, other}×{2 persons, other}×{no children, has children}×{locality type 1, other}. The unconditional categorical partial R-indicators were calculated for each of the new strata and the strata were then sorted by their p-value. 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, 18-34, 2 persons, no children, type 1}; {males, 18-34, 2 persons, no children, not type 1}; {males, 18-34, not 2 persons, no children, type 1}; {males, 18-34, not 2 persons, no children, not type 1}. The number of non-respondents to target for follow-up in these four 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 four strata were randomly converted to respondents. This increased the response rates by approximately 2.7%, as can be seen in column 5 of Table 4.2. In addition, Table 4.2 presents the R-indicators and Coefficients of Variation after the targeted follow-up of non-response assuming a 50% response rate (columns 6 and 7). There is a clear and significant increase in the R-indicator and a decrease in the Coefficient of Variation after the non-response follow-up assuming a 50% response rate.

We turn now to Table 4.3, containing the variable level partial R-indicators, and focus on the right side of the table (columns 5, 6 and 7) for the 50% responding targeted follow-up on the non-response. 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 smaller 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 that have become non-significant, following the targeted response.

Table 4.4: Category level (Unconditional and Conditional) Partial R-indicators ('' denotes significance at the 5% significance level) 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*
Age Group	15-17	0.020*	0.005*		Type 3	0.001	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*

The conclusion from this simulation study is that even with a small increase of response rate, albeit targeted to those non-respondents contributing to the lack of representativity, we are able to improve the representativeness of the data. This means that less adjustments are needed to correct for non-response bias, leading to smaller variation in sampling weights and increased efficiency.

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 and analysed in detail in Schouten et al (2013). In the construction of the adaptive survey designs, we adopt two scenarios: no learning period and a long learning period.

The design of the experiment was as follows: A sample of 8800 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 nonrespondents 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 (2013) 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 nonrespondents to face-to-face and we investigate what 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 done using six socio-demographic registry variables: gender (male, female), age (15-25, 25-35, …, 65-75, 75+), employment (yes, no), urbanization of residence (not, little,

moderate, strong, very strong), income in Euro's (<3K, 3-5K, 5-10K, 10-15K, …, 25-30K, >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 datasets at Statistics Netherlands as they relate generally to survey variables and are used in weighting adjustments and publication tables. Particularly, gender, age, urbanization and registration of a landline phone number relate strongly to key CVS variables: victimization, perception of safety, judgment of police performance and perception of neighbourhood problems. Table 5.1 presents the response rate, R-indicator and coefficient of variation for various strategies given the specified auxiliary variables. The last row of table 5.1 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 coefficient of variation of 0.160 as the target. We believe that two F2F phases represent what can be achieved with reasonable effort; beyond this effort the survey gets exceptionally expensive. We use the F2F → F2F coefficient of variation as internal benchmark.

Rows 2, 3, 6 and 7 of table 5.1 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 coefficient of variation is much higher. For mail, the picture is somewhat reversed: the response rate is relatively high but the R-indicator is very low. As a consequence, again, the coefficient of variation is much higher and even higher 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 coefficient of variation become similar to that of the benchmark strategy.

Table 5.1: 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 in 1000's of Web sample unit costs.

Strategy	*Response rate*	*R-indicator*	*CV*	*Cost*
Web	28.7% (1.0%)	0.806 (0.019)	0.368 (0.034)	2.2
Web → F2F	57.9% (1.1%)	0.829 (0.022)	0.168 (0.019)	49.1
Web scenario 1	39.7% (1.0%)	0.808 (0.021)	0.267 (0.026)	20.0
Web scenario 2	43.6% (1.1%)	0.846 (0.021)	0.206 (0.025)	29.1
Mail	49.0% (1.1%)	0.738 (0.020)	0.283 (0.020)	8.8
Mail → F2F	66.0% (1.0%)	0.812 (0.021)	0.157 (0.016)	42.3
Mail scenario 1	54.1% (1.1%)	0.855 (0.022)	0.159 (0.020)	18.7
Mail scenario 2	59.5% (1.1%)	0.878 (0.022)	0.129 (0.019)	26.8
F2F → F2F	67.9% (1.0%)	0.801 (0.021)	0.160 (0.015)	91.3

We assume that the available budget is not sufficient to cover a second phase for all nonrespondents 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 5.1 gives the costs per strategy in thousands of sample unit costs for one Web approach. The F2F to F2F strategy is approximately 45 times more expensive than Web only. For ease of demonstration, suppose that the available budget is one third of the expensive F2F to F2F strategy, i.e. 30.4. 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, and are too expensive.

*Table 5.2: Variable-level unconditional and conditional partial R-indicators for various strategies in the 2011 CVS experiment. (p-value: * = below 0.1%, † = below 1% , # = below 5%).*

		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 †
Urbanization	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 †

We adopted two extreme scenarios. The first scenario 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 scenario, a decision to allocate nonrespondents to F2F has to be based on the Web and mail phase 1 responses only. The second scenario is that of a continuous survey in which budget can be invested to perform a pilot with a full F2F second phase. Under this scenario, a decision to allocate nonrespondents can be based using the responses to both phases. Table 5.1 includes the indicator values of the adaptive survey designs that are constructed under the two scenarios (rows 4 and 5 for Web and rows 8 and 9 for mail). We constructed the designs following the steps of section 3.3. For the first scenario, 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 scenario, we found four categories for Web (income group >30K, natives, persons with a registered phone and persons living in little or non- urbanized 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 for Web and 601 for mail for follow-up. For both scenarios the numbers of case were within the budget levels.

Table 5.1 presents the resulting indicator values. For both scenarios, obviously, the response rates increases. Under scenario 1, for Web the second phase does not improve the R-indicator while for mail the second phase leads to an enormous increase in the R-indicator. The coefficient of variation for mail has become similar to the target from the F2F to F2F design while for Web it is still higher. Under scenario 2, the R-indicator increases for both Web and mail and are significantly higher than for strategy F2F to F2F. Because of the lower response rate the coefficient of variation for Web is still higher than the target but for mail it is significantly lower.

In the construction of the designs, we have concentrated ourselves on auxiliary variables. The important question is whether the various designs also affect the survey variables. Table 5.3 contains the design-weighted but unadjusted response means for designs with Web and mail, respectively, of five survey variables observed in phase 2: the number of victimizations per 100 inhabitants, the percentage victimized 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 estimates of a full F2F phase 2 and the adaptive survey designs under scenarios 1 and 2 are tested against the Web only and mail only designs. For comparison also coefficients of variation are shown. The victimization variables show significant differences against a Web only or a mail only design at the 5% level, 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 coefficient of variation coincide with significant changes in the victimization variables; the only design where it did not change significantly, scenario 1 for Web, still had a relatively high coefficient of variation. Remarkably, the number of reported victimizations in designs with a Web first phase is a lot higher than those with a mail first phase, although percentages victimized are similar.

What can we learn from this application? 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 optimized for subsequent waves and statistics for the first wave (and obviously future waves) can be based on the optimized design.

In the application we performed a structured trial-and-error approach. Under scenario 2 where we have estimates for the response propensities for both phases, we could have done an advanced optimization following Schouten, Calinescu and Luiten (2013). This would lead to a complex non-linear, non-convex optimization problem when all variables and all variable interactions are included. If we would first construct stratifications for targeting sample cases, as we have done here, the properties of the problem are the same but the dimensionality would be much lower and perhaps a brute force approach where a lot of options are simply tried would become within reach. We have not tried this for the application, however.

*Table 5.3: Unadjusted response means for five CVS survey variables and coefficient of variation for designs with a Web or mail first phase. (p-value for test against phase 1 response only: * = below 0.1%, † = below 1% , # = below 5%).*

Web	*Phase 1*	*Phase 1 and 2*	*Scenario 1*	*Scenario 2*
Coefficient of variation	0.368	0.168	0.267	0.206
# victimizations per 100	26.6	30.3	26.6	30.1
% victimized	8.1	10.7 †	8.9	10.8 †
Nuisance scale	1.3	1.3	1.4	1.4
% unsafe	25.5	25.4	25.6	26.9
% not satisfied police	45.3	47.1	46.9	47.0

Mail	*Phase 1*	*Phase 1 and 2*	*Scenario 1*	*Scenario 2*
Coefficient of variation	0.283	0.157	0.159	0.129
# victimizations per 100	17.6	22.3 #	23.4 †	22.4 #
% victimized	8.8	10.0 #	10.6 #	10.3 #
Nuisance scale	1.2	1.3	1.3	1.3
% unsafe	27.1	25.4	26.4	26.2
% not satisfied police	47.8	47.5	48.0	48.1

6. Discussion

In this paper, we demonstrated how to use partial R-indicators in forming nonrespondent profiles and strata for adaptive survey designs. We, furthermore, presented and

demonstrated structured trial-and-error approaches to design 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 optimization. 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 nonresponse error as process quality indicators. We believe, and there now is 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 nonresponse error. Although we do test indicators for their significant difference to a fully random nonresponse, we largely ignore the 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 e.g. Wagner and Hubbard (2013). 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.

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?. *Scandinavian Journal of Statistics*, 41, 277 – 290.

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

Groves, R.M., Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72, 167 – 189.

Hasler, C., Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, 74, 81 – 94.

De Heij, V., Schouten, B., Shlomo, N. (2014). RISQ 2.0 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 Canada. *Paper presented at Q2010*, 3 – 6 May, Helsinki, Finland.

Little, R., Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 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. *Journal of Royal Statistical Society, Series A*, 176 (1), 169 – 190.

Lundquist, P., Särndal, C.E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey, .*Journal of Official Statistics*, 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.

Peytcheva, E., Groves, R.M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, 25, 193 – 201.

Särndal, C.E. (2011). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 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, 311 – 340, In *Improving surveys with paradata. Analytic uses of process informatio*n, ed. F. Kreuter, Wiley, New Yor, USA.

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. *International Statistical Review*, 80 (3), 382 – 399.

Schouten, B., Brakel, J. van den, Buelens, B., Laan, J. van der, Klausch, L.T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42, 1555 – 1570.

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

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

Schouten, B., Cobben, F., Lundquist, P., Wagner, J. (2013). Does balancing survey response reduce nonresponse bias?. *Paper presented at 68th AAPOR conference*, May 16 – 19, Boston, USA.

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.

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.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, PhD thesis, University of Michigan, USA.

Wagner, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias, *Public Opinion Quarterly*, 76 (3), 555 – 575.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys, *Survey Research Methods*, 7 (1), 45 – 55.

Wagner, J., Hubbard, F. (2013). Using propensity models during data collection for responsive designs: Issues with estimation. *Paper presented at 68th AAPOR conference*, May 16-19, Boston, USA.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., Kruger Ndiaye, S. (2013). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics,* 28 (4), 477 – 499.