Statistics Netherlands

Discussion Paper

Theoretical and empirical support for adjustment of nonresponse by design

2014 | 15

Barry Schouten
Fannie Cobben,
Peter Lundquist,
James Wagner

12-05-2014

Theoretical and empirical support for adjustment of nonresponse by design

Barry Schouten, Fannie Cobben, Peter Lundquist and James Wagner

Summary: Recently, various indicators have been proposed as indirect measures of nonresponse error in surveys. They employ auxiliary variables to detect non-representative or unbalanced response. A class of survey designs known as adaptive survey designs maximizes these indicators by applying different treatments to different subgroups. The natural question is whether the decrease in nonresponse bias caused by adaptive survey designs could also be achieved by nonresponse adjustment methods. We discuss this question and provide theoretical and empirical considerations, supported by a range of household and business surveys. We find evidence that balancing response reduces bias more than adjustment does.

Keywords: Adaptive survey design; Adaptive treatment regime; Missing data mechanism; Poststratification; Survey nonresponse

1. Introduction

This paper is a follow-up to Schouten and Cobben (2012). In the current paper, we extend the theoretical and empirical evidence of Schouten and Cobben (2012) and further motivate balancing of survey response by design. The number of data sets is substantially increased and broadened by including surveys from Statistics Sweden and the University of Michigan. Furthermore, we perform an extensive simulation study and construct a rank test to find consistent patterns in rankings of representativeness indicators over different designs. In order to maintain a coherent paper, the paper overlaps in parts with Schouten and Cobben (2012). Throughout the paper, we use the terms representative and balanced response interchangeably. The terms are defined by Schouten, Cobben and Bethlehem (2009) and Särndal (2011), respectively. Representative response is defined as equal response propensities whereas balanced response is defined as equal nonresponse adjustment weights. Although they are slightly different, the two features are very similar in nature and indicators based on the two definitions usually rank designs the same. Balancing response refers to data collection efforts that lead to more representative or less unbalanced response.

The main research question to this paper is: Does balancing of data collection effort on auxiliary variables lead to less nonresponse bias on survey target variables, even after adjustment using the same auxiliary variables? If we choose survey design features differently, like the survey mode or the timing and number of interviewer calls, for say different age and income groups with the goal of balancing response, would the reduction of nonresponse bias be larger than achieved by nonresponse adjustment of

data collected under a uniform treatment design using age and income as weighting variables.

The context of this paper is survey data collection. However, the research question and discussion can be applied to any setting where different candidate treatments exist with varying amounts of missing data and where external data or covariates are available about population units.

To date, the confluence of declining survey budgets, the rise of web as a data collection mode, and the changing survey climate in which it is harder to obtain survey response, urge statistical institutes and market research companies to make difficult decisions regarding trade-offs between costs and accuracy of survey statistics. Survey organizations are forced to efficiently allocate resources to attract respondents. As a result in the survey methodology literature, two closely linked trends are emerging: looking in more depth and detail into the survey process itself, and tailoring and adapting treatment to the potential respondent. The first trend relates to the analysis of process data, termed "paradata" (Couper, 2000; Couper and Lyberg, 2005) in the survey literature. See Kreuter (2013) for an extensive overview of recent developments. The second trend refers to adaptive and responsive survey designs in which different population subgroups receive different treatments, see Groves and Heeringa (2006), Wagner (2008) and Schouten, Calinescu and Luiten (2013) as relevant examples of this trend. These adaptive or responsive designs may be viewed as an extension of traditional sampling designs, which allow for only two options – either in the sample or not. These new designs may select from multiple strategies for each unit. This paper is about the second trend and discusses the question whether it is worthwhile in terms of nonresponse bias reduction to adopt such designs.

Adaptive and responsive survey designs perform an adjustment by design as a supplement to adjustment afterwards in the estimation. The adjustment by design is conjectured to be 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.

Adaptive and responsive survey designs resemble adaptive treatment regimes in medical statistics. In survey data collection terminology, the treatments are called survey design features. The most prominent feature is the survey mode (web, mail, telephone, face-to-face) because of its large cost-quality differential, but many other design features may be varied during the course of a survey. Different population subgroups are identified in the sample based on covariates from registry data, sampling frames, and paradata that become available during data collection. An example of paradata is interviewer observations that relate to the main topics in the survey. In clinical trials, the outcomes of interest are unknown to the researcher in advance. Hence, decisions about continuing or altering treatment are based on proxy measures. This is analogous to the survey setting where the answers that respondents will give are not known before they respond. Recently, there have been developments

in research on proxy measures of the outcome variables in the survey setting. Indicators that have recently been proposed are representativeness indicators and the strongly related coefficient of variation of survey response propensities (Schouten, Cobben, Bethlehem 2009), balance indicators (Särndal 2011) and the fraction of missing information (Wagner 2010, Andridge and Little 2011). In this paper we focus on representativeness indicators, abbreviated to R-indicators, and the coefficient of variation which are very similar to balance indicators. The fraction of missing information is part of a very different class of indicators, see Wagner (2012), which we do not consider here.

Adapting a design to optimize R-indicators or the coefficient of variation comes down to reducing variation in response propensities on available auxiliary variables, i.e. to equalizing response propensities. One of the criticisms to the use of these indicators for adapting survey design is that any utility gained by adaptation of the design during data collection may be achieved as well, and more cheaply, by post-hoc adjustment using the same auxiliary information. Hence, reducing bias on survey target variables by balancing response on auxiliary information could also be done through adjustment methods afterwards. However, since the validity of a model for the association between auxiliary variables and survey target variables may change during data collection, it is not necessarily true that adjustment based on such a model performs the same with or without any adaptation of the design. This paper discusses the question whether balancing response does reduce bias regardless relative to post-hoc adjustments.

There is no easy way to answer the main research question, as in most cases nonresponse biases on survey target variables are unknown. We circumvent this complication by dividing available auxiliary variables into two groups: a group to be used in the assessment and improvement of indicators and a group to be used in the evaluation of nonresponse bias. We do this in two ways. First, we apply indicators to models estimated with increasing numbers of auxiliary variables as predictors and investigate whether patterns emerge, i.e. whether worse indicator values based on models with few predictors are associated with worse values on models with many predictors. Second, we perform nonresponse adjustment using increasing numbers of weighting variables and search for consistency in the biases remaining after each adjustment, i.e. whether larger biases on models with few predictors coincide with larger biases on models with many predictors. In both cases we search for patterns by means of a rank test. The auxiliary variables held out of the adjustment models function as surrogates of survey target variables, or “Pseudo-Y” variables.

It is important to stress that the research question of this paper is mostly an empirical question. One can easily construct examples where balancing response does not reduce nonresponse bias. If we do find evidence in survey data that balancing helps, it does not, therefore, imply that the indicators have the feature that they detect nonresponse bias on other variables. It merely means that lower quality survey data collection, in the majority of cases, tends to affect the full range of potential variables and that the indicators successfully signal this tendency. This issue has also been

debated by Särndal (2011). Nonetheless, we show that there are also theoretical considerations that support balancing response through a combination of adaptive survey design and post-hoc adjustment.

The strength of the empirical evidence depends on the variety of surveys that are studied and the nature of the auxiliary variables that are input to the indicators. We have selected a wide range of survey data sets from three different countries to find empirical support. We compare the representativeness of response for growing sets of auxiliary variables over different surveys, over different waves of a survey, during data collection and after different survey process steps like establishing contact and obtaining cooperation. In each comparison the auxiliary variables are fixed, but different variables are used over different comparisons and different data sets. Some of the data sets that we have selected contain a relatively rich set of auxiliary variables that were linked from registry data. Our study is somewhat similar to that of Peytcheva and Groves (2009), who investigated whether biases on demographic variables covary with specialized studies of biases on survey target variables. They found little evidence for such an association. Our study, however, uses a variety of data sources from actual surveys and shows that there is consistency in biases for auxiliary variables, even after adjustment. We do not extrapolate to survey target variables, but do discuss how such a consistency may extend to these variables as well.

In this paper, we focus on nonresponse bias, i.e. on external validity. We assume that sample sizes are large, so that precision is not an issue. The restriction to nonresponse bias may, however, be too naïve, especially when multiple survey modes or intensive refusal conversion procedures are considered. For such design features differences in measurement bias, i.e. in internal validity, may need to be accounted for. Calinescu, Schouten and Bhulai (2012) generalize adaptive survey designs to nonresponse and measurement bias. Since adjustment methods for measurement error are different from adjustment methods for nonresponse error, we leave a debate about balancing or adjusting measurement error to a future paper.

Even if our results provide a rationale that adaptive and responsive survey designs are meaningful extensions of traditional survey sampling designs, implementation of such designs in survey practice is not straightforward or easy. It implies a different conceptual framework. We leave it to other papers to make recommendations on how to implement such designs, e.g. Wagner (2013) and Luiten and Schouten (2013).

The paper is organized as follows. In section 2, we review the various indicators for representative response. Next, in section 3, we show that the indicators appear in the bias intervals for all standard estimators: the Horvitz-Thompson estimator, the generalized regression estimator, the inverse propensity weighting estimator and double robust estimators. In this section, we also show that if the indicators favour certain designs over others based on an arbitrary selection of variables, then the indicators would also favour these designs when indicators would be based on any other non-selected variable. Furthermore, biases on any other non-selected variable after post-hoc adjustment would also be larger for the designs that are favoured by the indicators. In section 4, we discuss a rank test to investigate whether indicator

preferences are random and perform an extensive simulation study to explore the properties of the test. In section 5, we apply the rank test to the various datasets collected in the Netherlands, Sweden and the USA. In section 6, we end with a summary and conclusion.

2. Indicators for representative or balanced response

In this section, we review indicators that have recently been proposed in the survey methodology literature as indirect measures of nonresponse error. We discuss briefly how they are used as objective functions measuring quality in adaptive survey designs. We refer to Lundquist and Särndal (2013a), Särndal and Lundström (2010), Särndal (2011), Schouten, Cobben and Bethlehem (2009), Schouten, Shlomo and Skinner (2011) and Shlomo, Skinner and Schouten (2012) for detailed accounts of the indicators and their statistical properties. We also refer to Wagner (2012) for a comparison and taxonomy of indicators.

We first introduce a notation and framework. Throughout the paper, it is not essential whether we would take a design-based-finite-population or model-based-superpopulation approach. All theoretical and empirical results could be translated in a straightforward way from one approach to the other. For ease of notation, however, we adopt a model-based-superpopulation approach. Let $\{(X_i, Y_i, R_i)\}_{1 \le i \le n}$ be an independent, identically distributed vector of covariates X_i, variables of interest Y_i and 0-1 response indicators R_i, where $R_i = 1$ indicates response. We assume that missingness applies to Y_i only and that X_i is always observed. For the moment we do not model the joint distribution of (X_i, Y_i, R_i). We denote the conditional distribution $P[R_i = 1 \mid X_i = x, Y_i = y]$ by $\rho_{X,Y}(x, y)$, which is usually termed the response propensity for x and y. Similarly, we define response propensities $\rho_X(x)$ and $\rho_Y(y)$. For convenience, we will shorten the notation for the response propensity for unit i, given its values on the covariates and variables of interest, to $\rho_i = \rho_{X,Y}(x_i, y_i)$. Let Z be some vector of variables formed out of the auxiliary variables and survey target variables. In the following, we will often not specify the value of Z but treat the response propensity $\rho_Z(Z)$ as a random variable. It is straightforward to show that the expectation $E\rho_Z(Z)$ is always equal to the (expected) response rate $P[R_i = 1]$, which we denote simply by μ. We use the notation μ here rather than the more common ρ to avoid confusion with the shorthand notation for the response propensity on the combined set of variables (X, Y). We will denote variances by S^2 and standard deviations by S.

The representativeness indicator or R-indicator for a variable Z is defined as the transformed standard deviation of the response propensity function ρ_Z

$$R(Z) = 1 - 2S(\rho_Z(Z)) . \quad (1)$$

The R-indicator takes values between 0 and 1, where a value close to 1 corresponds to representative response. It corresponds to the Euclidean distance to the response propensity function that is constant over all values of Z. Schouten, Cobben and Bethlehem (2009) introduce this indicator in a design-based context and propose an estimator using logistic regression on the R_i's. The estimator itself is usually referred to as the R-indicator. If one would use linear regression instead of logistic regression,

then the R-indicator is equal to the balance indicator BI_3, proposed by Särndal (2011). In practice, the choice of link function is, however, rarely influential, see Bethlehem (2012). The rationale behind the indicators is that an absence of variation implies that response is a random subsample of the full sample with respect to the predictors in the model.

Two indicators have a close similarity to the R-indicator and balance indicator. The first is the coefficient of variation of the response propensity function

$$CV(Z) = \frac{S(\rho_Z(Z))}{\mu}.$$

(2)

The coefficient was labelled as the maximal bias for Z by Schouten, Cobben and Bethlehem (2009), because it limits the standardized absolute bias of any function of Z restricted to respondents (where standardization is done by the standard deviation of Y). The indicator (2) can be estimated in a manner similar to (1) by dividing over the observed response rate. If the identity link function is used then the estimator has a close similarity to the coefficient of variation of the nonresponse adjustment weights proposed by Särndal and Lundström (2010), which they denote as H_3. Nonresponse adjustment weights can be viewed as smoothed inverse response propensities. A Taylor expansion of the coefficient of variation of inverse response propensities shows that for large samples it is proportional to the coefficient of variation of response propensities. The second indicator that links to the R-indicator and balance indicator is the standardized contrast

$$C(Z) = \frac{S(\rho_Z(Z))}{\mu(1-\mu)}, \qquad (3)$$

which is equal to the standardized difference in the expectations of Z for respondents and nonrespondents. Traditionally, the impact of nonresponse on the locations of distributions is decomposed as the product of the contrast and the nonresponse rate $1-\mu$ (e.g. Bethlehem, Cobben and Schouten 2011). This product equals the coefficient of variation (2). Also the contrast has a counterpart in the Lundquist and Särndal paper (2013a), where it is denoted by $dist_{r|nr}$.

Groves and Heeringa (2006), Wagner (2008), Lundquist and Särndal (2013a) and Schouten, Calinescu and Luiten (2013) propose to differentiate level and type of effort in surveys for different population subgroups in order to maximally reduce bias of estimators based on the survey response within the available survey budget. These designs are termed adaptive or responsive survey designs and resemble adaptive treatment regimes (Collins, Murphy, and Bierman 2004 and Murphy, Lynch, Oslin, McKay, Tenhave 2007) in other areas of statistics. The rationale is that different population subgroups may prefer or react differently to different treatments. The indicators in this section are proposed by some of the authors as quality objective functions in these optimal quality-cost trade-offs. They are applied to a number of

candidate designs and the design that has the best indicator value is favoured. Adaptation to the sampled units can be done prior to data collection based on previous waves of the same survey or similar surveys, or during data collection based on observations made on the sampled units. Schouten, Shlomo and Skinner (2011) propose partial R-indicators to identify population subgroups that should be targeted in order to reduce variation in response propensities. Essentially, the variance of response propensities is decomposed into between and within components and the subgroups that have the largest within variances are prioritized. The authors define unconditional partial R-indicators, denoted by $P_U(Z \mid X)$, and conditional partial R-indicators, denoted by $P_C(Z \mid X)$, where Z is an element of the auxiliary vector X. $P_U(Z \mid X)$ is defined as the between variance for Z of the response propensity function ρ_X. $P_C(Z \mid X)$ as the within variance attributable to Z given a stratification on X without Z, again of ρ_X. For exact definitions, we refer to Schouten, Shlomo and Skinner (2011). Lundquist and Särndal (2013b) define partial imbalance indicators in a comparable fashion.

We now return to the main research question of this paper. Can these indicators usefully be applied to improve survey design? We first note that the division of the indicators by the response rate in (2) and by the product of the response rate and the nonresponse rate in (3) implies that the indicators generally lead to different design preferences. Only if the response rate is equal for different designs, it is true that the indicators rank designs identically. Hence, although they may be interesting in their own right, the three indicators cannot be used simultaneously in design decisions. More importantly, however, the indicators are criticized for two main deficits. Recently, Beaumont and Haziza (2011) rightfully remarked that the early adaptive and responsive survey design papers restrict attention to bias and ignore variance. Also, in this paper we will focus mostly on bias, because we want to address the other alleged deficit. Although the indicators have subtle differences, they share one important feature: They can be estimated only for auxiliary variables X and not for the variables of interest Y, unless a model is formulated. This feature is the second deficit; balancing response on X may not be meaningful or useful because the missingness on these variables can be accounted for through an adjustment procedure and the real variables of interest remain unaffected. This discussion links strongly to the paper by Andridge and Little (2011) in which missingness is modelled as a function of $Y(X) + \lambda Y$, where $Y(X)$ is the projection of Y on X and λ is a moderating parameter. λ cannot be estimated but allows for a sensitivity analysis. Andridge and Little (2011) do this by computing the fraction of missing information (FMI) for different choices of λ. The FMI cannot be easily used in adaptive and responsive survey designs since it is specific to the Y and we might make different modeling assumptions across different Y's.

In the following sections we provide both a theoretical and empirical discussion on this key question for adaptive and responsive survey designs.

3. Components of nonresponse bias

In this section, we provide theoretical considerations that support a focus on improving indicator values through adaptive design. We, first, derive approximations for the bias of four estimators in the context of survey nonresponse: the (unweighted) response mean, the generalized regression estimator, the inverse propensity weighting estimator and the double robust estimator. We show how the biases relate to the indicators of the previous section. Second, we give a statistical argument why they may be seen as measures of the quality of the data collection process, i.e. as opposed to product quality measures.

3.1 Bias of unweighted and weighted response means

We start by deriving expressions for the bias of four estimators of the location of the distribution of a survey target variable that are used in the context of survey nonresponse. These estimators are the Horvitz-Thompson or expansion estimator (Horvitz and Thompson 1952), the generalized regression estimator (e.g. Bethlehem 1988), the inverse propensity weighting estimator (e.g. Hirano, Imbens and Ridder 2003) and the double robust estimator (e.g. Bang and Robins 2005). For an excellent discussion on the various estimators see Schafer and Kang (2008). We restrict ourselves to the estimation of the location of distributions of interest. See Brick and Jones (2008) for derivations of the bias of other properties of distributions. The double robust estimator derived its name from the simultaneous modelling of a survey target variable and nonresponse using auxiliary variables X ; if either one of the models is correctly specified then the estimator is unbiased, i.e. it is doubly robust. The Horvitz-Thompson estimator does not employ a model. Essentially, the generalized-regression estimator models the survey target variable and the inverse propensity weighting estimator models the nonresponse. In our simplified framework, the Horvitz-Thompson estimator amounts to the response mean (RM), since the probability of being sampled does not depend on X or Y .

Since the indicators of the previous section are not specific to any survey target variable Y , any population parameter of that survey variable or any estimator, it can be stated beforehand that the indicators cannot be ideal indicators in any specific setting. The indicators do allow for flexibility in choosing different covariates X for each estimand, but for specific variables, parameters and estimators, other indicators may be preferred. Nonetheless, we show that the variance of response propensities and the response rate are important components of the bias of the four estimators.

We focus on bias of estimators. This is too simple a focus in general. We want to evaluate whether it is possible to reduce bias of nonresponse-adjusted estimators on variables of interest by altering the design. Future work should include the impact of design choices on variances of these estimators, or alternatively on variances of posterior distributions when adopting a Bayesian viewpoint. Furthermore, we do not discuss the selection of auxiliary variables, but treat them as given and fixed. From a

bias point of view, it is irrelevant whether auxiliary variables are selected that do not relate to variables of interest nor to nonresponse, but for the precision of estimators it is a crucial choice that has been much debated in the literature, see e.g. Little and Vartivarian (2005) and Huber, Lechner and Wunsch (2013).

We assume an outcome model and a selection model of the following form

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{4a}$$

$$\rho_X(X_i) = h(\gamma + \delta X_i), \tag{4b}$$

where α, β, γ and δ are parameters, h is a link function and ε_i is an iid residual with zero expectation. The combined model (4a and b) is often referred to as a sample-selection model or Heckman model. The model specifications may not be correct and the auxiliary vectors in (4a) and (4b) may be taken to be different, but this will not be problematic in the following. Assume for the moment that the parameters α, β, γ and δ are known.

The response mean and inverse propensity weighting (IPW) estimators have the following form

$$\bar{y}_{RM} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}, \tag{5}$$

$$\bar{y}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\rho_X(X_i)} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{h(\gamma + \delta X_i)}, \tag{6}$$

and the generalized regression (GREG) and double robust (DR) estimators are

$$\bar{y}_{GREG} = \bar{y}_{RM} + \beta(\bar{x}_n - \bar{x}_{RM}), \tag{7}$$

$$\bar{y}_{DR} = \bar{y}_{IPW} + \beta(\bar{x}_n - \bar{x}_{IPW}). \tag{8}$$

In this section, we make frequent use of the fact that any correlation between two variables, say X_1 and X_2, can be bounded from below and above by the correlations with a third variable, say X_3,

$$\begin{aligned} cor(X_1, X_2) \in [&cor(X_1, X_3)cor(X_2, X_3) - \sqrt{1 - cor^2(X_1, X_3)}\sqrt{1 - cor^2(X_2, X_3)}, \\ &cor(X_1, X_3)cor(X_2, X_3) + \sqrt{1 - cor^2(X_1, X_3)}\sqrt{1 - cor^2(X_2, X_3)}] \end{aligned} \tag{9}$$

It follows directly that

$$\begin{aligned} &\mathrm{cov}(X_1, X_2) \in \\ &[S(X_1)S(X_2)(cor(X_1, X_3)cor(X_2, X_3) - \sqrt{1 - cor^2(X_1, X_3)}\sqrt{1 - cor^2(X_2, X_3)}), \\ &S(X_1)S(X_2)(cor(X_1, X_3)cor(X_2, X_3) + \sqrt{1 - cor^2(X_1, X_3)}\sqrt{1 - cor^2(X_2, X_3)})] \end{aligned} \tag{10}$$

For convenience, we use the following notation to describe the interval in (10)

$$\left[S(X_1)S(X_2),\ cor(X_1, X_3)cor(X_2, X_3),\ \sqrt{1-cor^2(X_1, X_3)}\sqrt{1-cor^2(X_2, X_3)} \right],$$

where the first entry is the scaling constant, the second entry the centre point of the interval and the third entry the term determining the width of the interval.

Approximations to the bias of the four estimators are given by

$$B(\bar{y}_{RM}) = \frac{\mathrm{cov}(Y, \rho)}{\mu}, \qquad B(\bar{y}_{IPW}) = \frac{\mathrm{cov}(Y, \frac{\rho}{\rho_X})}{\mu},$$

$$B(\bar{y}_{GREG}) = \frac{\mathrm{cov}(Y - \beta X, \rho)}{\mu}, \qquad B(\bar{y}_{DR}) = \frac{\mathrm{cov}(Y - \beta X, \frac{\rho}{\rho_X})}{\mu},$$

where $\rho = \rho_{X,Y}(X, Y)$ and $\rho_X = \rho_X(X)$.

Now, we use two random variables to assess bias intervals, $\rho_X(X)$ and $Y(X) = \beta X$. It is relatively straightforward to derive pairs of approximate bias intervals for the absolute bias of the response mean and IPW and GREG estimators

$$\left[\frac{S(Y)S(\rho)}{\mu}, |cor(Y, \rho_X)\ ||\ cor(\rho, \rho_X)|, \sqrt{1-cor^2(Y, \rho_X)}\sqrt{1-cor^2(\rho, \rho_X)} \right], \quad \text{(RM1)}$$

$$\left[\frac{S(Y)S(\rho)}{\mu}, |cor(Y, \beta X)\ ||\ cor(\rho, \beta X)|, \sqrt{1-cor^2(Y, \beta X)}\sqrt{1-cor^2(\rho, \beta X)} \right], \quad \text{(RM2)}$$

$$\left[\frac{S(Y)S(\rho)}{\mu}, 0, \sqrt{1-cor^2(Y, \rho_X)}\sqrt{1-cor^2(\rho, \rho_X)} \right], \quad \text{(IPW1)}$$

$$\left[\frac{S(Y)S(\rho)}{\mu}, 0, \sqrt{1-cor^2(\rho, \rho_X)}\sqrt{1-cor^2(Y, \beta X)} \right], \quad \text{(IPW2)}$$

$$\left[\frac{S(Y)S(\rho)}{\mu}, 0, \sqrt{1-cor^2(Y, \beta X)}\sqrt{1-cor^2(\rho, \rho_X)} \right] \quad \text{(GREG1)}$$

$$\left[\frac{S(Y)S(\rho)}{\mu}, 0, \sqrt{1-cor^2(Y, \beta X)}\sqrt{1-cor^2(\rho, \beta X)} \right]. \quad \text{(GREG2)}$$

In the derivation of the bias intervals, we used that $cov(\rho/\rho_X, \beta X) = 0$ and $cov(\rho_X, Y - \beta X) = 0$. Furthermore, we used that $\beta = \left(S^2(X)\right)^{-1} cov(Y, X)$ so that it follows that $S^2(Y - \beta X) = S^2(Y) - cov^2(Y, \beta X)/S^2(\beta X) = S^2(Y)(1 - cor^2(Y, \beta X))$. Finally, we used that from Taylor approximations it holds that approximately $S^2(\rho/\rho_X) = S^2(\rho) - cov^2(\rho, \rho_X)/S^2(\rho_X) = S^2(\rho)(1 - cor^2(\rho, \rho_X))$.

The two choices of auxiliary variables lead to the same bias interval for the DR estimator

$$\left[\frac{S(Y)S(\rho)}{\mu},0,\sqrt{1-cor^{2}(Y,\beta X)}\sqrt{1-cor^{2}(\rho,\rho_{X})}\right], \qquad \text{(DR1)}$$

which can be rewritten as $\left[S(Y),0,\sqrt{1-cor^{2}(Y,\beta X)}\sqrt{\frac{S^{2}(\rho)}{\mu^{2}}-\frac{S^{2}(\rho_{X})}{\mu^{2}}}\right]$.

Hence, the IPW, GREG and DR estimators transfer the centre point of the bias interval for the response mean to zero. A closer look at the width of the intervals reveals that (RM1) and (IPW1) are the same as well as (GREG2) and (RM2). Hence, the IPW and GREG estimators only shift these bias intervals. Interval (IPW2) is equal to (GREG1) and they are smaller than (IPW1) and (GREG2). Thus, taking intersections of the bias intervals leads to smaller intervals for both estimators with respect to the response mean. However, (DR1) has the same width as (IPW2) and as (GREG1), so that the DR estimator does not lead to a further reduction of the width of the bias interval.

There are two messages from the bias intervals. First, all intervals have a scaling constant that is proportional to the coefficient of variation of the response propensity ρ. Second, the width of the bias interval for the estimators depends on the squared difference of the coefficients of variation applied to ρ and ρ_X, respectively,

$$\sqrt{\frac{S^{2}(\rho)}{\mu^{2}}-\frac{S^{2}(\rho_{X})}{\mu^{2}}}=\sqrt{cv^{2}(\rho)-cv^{2}(\rho_{X})} \qquad (11)$$

Schouten (2007) proposes to use the width of interval (GREG2) to select auxiliary variables for the GREG estimator. The derivations here show that, when fully ignoring variance, it is superior to choose X such that $\sqrt{1-cor^{2}(Y,\beta X)}\sqrt{1-cor^{2}(\rho,\rho_{X})}$ is minimal. However, correlations need to be large in order to get small bias intervals.

In practice, the parameters α, β, γ and δ are unknown. Consistent estimators for the parameters γ and δ exist, provided the outcome model is valid. However, the parameters α and β cannot be estimated without a potential bias. If, instead of the true β, we would use the slope parameter for the respondents, say $\tilde{\beta}$, the bias intervals for the GREG and DR estimators change to

$$\left[\frac{S(Y-\tilde{\beta}X)S(\rho)}{\mu},|\,cor\,(Y-\tilde{\beta}X,\rho_{X})\,\|\,cor\,(\rho,\rho_{X})\,|,\sqrt{1-cor^{2}(Y-\tilde{\beta}X,\rho_{X})}\sqrt{1-cor^{2}(\rho,\rho_{X})}\right],$$

(GREG3)

$$\left[\frac{S(Y-\tilde{\beta}X)S(\rho)}{\mu},|\,cor\,(Y-\tilde{\beta}X,\beta X)\,\|\,cor\,(\rho,\beta X)\,|,\sqrt{1-cor^{2}(Y-\tilde{\beta}X,\beta X)}\sqrt{1-cor^{2}(\rho,\beta X)}\right],$$

(GREG4)

$$\left[\frac{S(Y-\tilde{\beta}X)S(\rho)}{\mu},0,\sqrt{1-cor^{2}(Y-\tilde{\beta}X,\rho_{X})}\sqrt{1-cor^{2}(\rho,\rho_{X})}\right], \qquad \text{(DR2)}$$

$$\left[\frac{S(Y-\tilde{\beta}X)S(\rho)}{\mu},0,\sqrt{1-cor^{2}(Y-\tilde{\beta}X,\beta X)}\sqrt{1-cor^{2}(\rho,\rho_{X})}\right]. \qquad \text{(DR3)}$$

The bias intervals for the GREG estimator are now not centered around zero, and the two intervals for the DR estimator are different. Without further assumptions it is not possible to state whether (GREG3) is smaller than (GREG4) or vice versa. The same applies to the (DR2) and (DR3) intervals, Since (DR3) is always smaller than (GREG4), the DR estimator always has intervals that are equal in width or smaller. If we do not consider a specific Y variable, then $\sqrt{1-cor^{2}(Y-\tilde{\beta}X,\rho_{X})}$ can be made arbitrarily large and the important term again is (11).

In the following section, we motivate why a larger $cv(\rho_{X})$ may also indicate a larger difference between the squared $cv(\rho)$ and $cv(\rho_{X})$.

3.2 The indicators as measures of process quality

In this section, we formalize the utility of the indicators as process quality indicators. More specifically, we formalize the intuition that a larger variation of response propensities for X corresponds to a larger variation of the true individual response probabilities. Doing so, we capitalize on the existence of an individual response probability.

In section 3.1, we view the vector X as fixed and given. All derivations and conclusions that we have made so far, do apply, however, to any arbitrary vector. Here, we view auxiliary variables themselves as being sampled from the population of all possible random variables.

Suppose a large population consists of G fully homogeneous and equally sized groups, labelled by $g=1,2,\ldots,G$. All units in group g behave exactly the same in every way, and they have the same response probability for any given survey design. The stratification into the groups itself is not observed, but we do observe categorical variables X_{k}, $k=1,2,\ldots,K$, that cluster groups into smaller numbers of groups.

Let us for simplicity look at a 0-1 indicator variable X. Assume that X was constructed by a simple random sample without replacement of size G_{X} from the set of G groups. Let s_{g} be the 0-1 indicator that group g was selected. We then have the following definition of X

$$X=\begin{cases}1 & \forall g, s_{g}=1,\\ 0 & \forall g, s_{g}=0,\end{cases} \qquad (12)$$

i.e. X is one for all selected groups g and zero otherwise. Since the groups have equal size, the probability that $X=1$ is equal to G_{X}/G.

Now, let ρ_g be the response probability of group g, so that the response propensity function $\rho_X(x)$ for X is defined as

$$\rho_X(x) = \begin{cases} \frac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g & \text{if} \quad x = 1, \\ \frac{1}{G - G_X}\sum_{g=1}^{G}(1 - s_g)\rho_g & \text{if} \quad x = 0. \end{cases} \tag{13}$$

In order to investigate the relation between the indicators based on X and those based on the full stratification with the G groups, we consider the expected mean and the expected variance of the response propensity function ρ_X.

The mean response propensity can be derived as

$$\bar{\rho}_X = \frac{G_X}{G}\frac{1}{G_X}\sum_{g=1}^{G} s_g \rho_g + (1 - \frac{G_X}{G})\frac{1}{G - G_X}\sum_{g=1}^{G}(1 - s_g)\rho_g = \frac{1}{G}\sum_{g=1}^{G}\rho_g = \bar{\rho}. \tag{14}$$

From (14) we can conclude that the mean response propensity $\bar{\rho}_X$ is always equal to the mean individual response probability $\bar{\rho}$. Clearly, the expected mean response propensity is then also equal to $\bar{\rho}$. Hence, regardless of the choice of X, the mean response propensity is the mean of the individual probabilities.

The variance of ρ_X, $S^2(\rho_X)$, is equal to

$$S^2(\rho_X) = \frac{G_X}{G}(\rho_X(1) - \bar{\rho})^2 + (1 - \frac{G_X}{G})(\rho_X(0) - \bar{\rho})^2. \tag{15}$$

The expectation of $\rho_X(x)$ is always equal to $\bar{\rho}$, and, hence, the expectation of (15) equals

$$ES^2(\rho_X) = \frac{G_X}{G}Var(\rho_X(1)) + (1 - \frac{G_X}{G})Var(\rho_X(0)), \tag{16}$$

where $Var(\rho_X(x))$ is the variance of ρ_X with respect to the subgroup sampling design. Since X is constructed using a simple random sample without replacement, $Var(\rho_X(x))$ is equal to

$$Var(\rho_X(x)) = \begin{cases} \frac{1}{G_X}(1 - \frac{G_X}{G})S^2(\rho) & \text{if} \quad x = 1, \\ \frac{1}{G - G_X}\frac{G_X}{G}S^2(\rho) & \text{if} \quad x = 0. \end{cases} \tag{17}$$

Combining (16) and (17) gives

$$ES^2(\rho_X) = \frac{1}{G}S^2(\rho), \tag{18}$$

so that the expected variance of response propensities is equal to the variance of the individual response probabilities times the population diversity constant $1/G$.

With similar arguments, it can be reasoned that if X is a categorical variable with C categories, then

$$ES^2(\rho_X) = \frac{C-1}{G} S^2(\rho). \tag{19}$$

So for all X , the expected variance of the response propensity function ρ_X is proportional to the variance of the underlying variance of individual response probabilities. This is a very useful finding as it implies that, if for some survey design the R-indicator is smaller or the coefficient of variation is larger than for another survey design, then also the variance of the individual response probabilities is expected to be larger. As a consequence, the expected variance of the propensity function resulting from any other random draw of subgroups g would be larger too. Although it would not be true that the variance of all propensity functions is larger, there may in fact be various variables that lead to a smaller variance, it must hold that for an randomly selected variable on average the variance is larger. This conclusion supports the intuition that for surveys with many target variables, one would prefer larger R-indicators or smaller coefficients of variation. It also shows that for single topic surveys, it may actually be the survey target variable itself that is one of the exceptional variables.

When it comes to the bias of the RM and IPW, GREG and DR estimators, introduced in the previous section, the square root of the difference of the squared $cv(\rho)$ and $cv(\rho_X)$ in (11) can be derived from (19) as

$$\sqrt{\frac{S^2(\rho)}{\mu^2} - \frac{S^2(\rho_X)}{\mu^2}} = \frac{S(\rho)}{\mu}\sqrt{\frac{G-C+1}{G}} = \frac{S(\rho_X)}{\mu}\sqrt{\frac{G-C+1}{C-1}}, \tag{20}$$

which is the omnipresent term in the bias intervals. From (20) we can conclude that when X is drawn at random from all possible 0-1 variables then a larger coefficient of variation corresponds to a larger (expected) maximal remaining bias after adjustment with X for any other arbitrary variable.

When 0-1 auxiliary variables X are drawn using unequal inclusion probabilities for the population subgroups, then their expected variances of response propensities are proportional to a weighted variance of the individual response probabilities. The weights are equal to the inverse inclusion probabilities, normalized to one. This important result implies that if we consider a subset of all 0-1 auxiliary variables, then a larger coefficient of variation on an arbitrary variable from that subset indicates a larger expected coefficient of variation and remaining bias for any other arbitrary variable from that subset. One such subset are all variables that relate to the survey topics in a specified way, including the survey variables.

If we could assume that the set of auxiliary variables X_k, $k = 1,2,\ldots,K$, consists of independent random draws of subgroups, then we could estimate G and $S^2(\rho)$. Estimation of these parameters is, however, beyond the scope of this paper. Of course, the discussion in this section is conceptual, auxiliary variables cannot be considered as independent, random draws of population subgroups. However, models for nonresponse are often criticized for the lack of relevant, explanatory variables; standard variables like age or gender may have proved to be indicative of homogeneity in the population, they were certainly not picked for the specific purpose of modelling response.

4. Testing the reduction of nonresponse bias

Even when some theoretical considerations, as laid out in the previous section, would suggest that improving indicator values through changes in the survey design would be likely to control nonresponse bias, it would make a much stronger case if empirical results support such an endeavour. For this reason, we explore a wide range of survey data. We evaluate the validity of the preferences of the indicators using a rank test. In the test, we randomly divide the set of auxiliary variables into two groups: an evaluation group and a validation group. We test the null hypothesis that indicator values and nonresponse biases for evaluation variables are not indicative of indicator values and nonresponse biases for validation variables. If the test rejects this hypothesis, then designs are preferred by multiple variables simultaneously.

We describe the details of the rank test in appendix B. We developed the rank test but it has a strong resemblance to multiple sample location rank tests described in the literature (e.g. Van der Vaart 1998). Here, we provide only the necessary background to the test, discuss its assumptions and perform a simulation study. In the study, we manipulate the amount of collinearity between auxiliary variables, the amount of variability in response propensities over designs and the sample size, and we investigate the impact of these parameters on the type 1 and 2 errors of the test.

4.1 A rank test for indicator preferences

Assume there are V comparative data sets, i.e. different surveys, different designs of the same survey or different levels of effort, labelled by $v = 1,2,\ldots,V$. In the following, we will simply refer to different designs. Let D_v be the number of different designs in comparative data set v and M_v be the number of auxiliary variables.

The test statistic that we apply is based on the number of pairwise inversions that result when ranking the different designs according to the indicators and to remaining bias. A small number of pairwise inversions implies the indicator shows a consistent picture when different variables are considered. For each indicator we rank the designs within a data set while increasing the number of auxiliary variables used in the model. The included auxiliary variables function as the evaluation variables and the omitted auxiliary variables as the validation variables. The evaluation set of variables grows with each step, while the validation set of variables shrinks. If we assume that indicators computed on evaluation variables are not predictive of indicators based on validation variables, then it holds that the different rankings are independent. We derive the probability distribution of the number of pairwise inversions under independence in appendix B. It is possible to test all comparative data sets simultaneously. We will do both – that is, test datasets individually and jointly.

4.2 Assumptions underlying to the rank test

There are two basic assumptions underlying the rank test: 1) the ranks do not produce any ties, and 2) auxiliary variables are independent. In this section, we discuss both assumptions.

The rank test assumes that the indicators and biases have a continuous measurement level and do not produce ties. The indicators and biases are continuous but they are random variables and are subject to imprecision. As a result, two indicator values and two biases may not be statistically different at a certain significance level. Given a particular significance level, the indicators and biases could produce ties. In general, the number of ties increases with decreasing sample sizes of the surveys under study. Ignoring the standard errors in the rankings leads to more noisy and, hence, less consistent patterns and to a more conservative rank test. In fact if sample sizes were very small, then the indicators would not be able to detect any signal at all. There is, however, no straightforward method to account for the standard errors of the indicators and biases in the rank test without making assumptions about the probability distributions of the indicators and biases over the variables. We, therefore, accept that the tests will be conservative and we select survey data sets that have modest to large sample sizes, which holds for many surveys in official statistics.

The second assumption is more fundamental as it links to the independence assumption in the rank test. It is not true that the auxiliary variables are independent, and any dependence between the auxiliary variables may lead to spurious consistency in the rankings of indicator values. If we would take two copies of the same variables, then the indicator values would be exactly the same and would be fully consistent. Dependencies between the auxiliary variables, therefore, lead to smaller numbers of pairwise inversions and to more rejections of the null hypothesis. The probability of false rejections of the null hypothesis is larger and the power is smaller than anticipated. There are three options to reduce or remove the impact of dependent variables. First, the auxiliary variables can be made orthogonal or uncorrelated by performing a principal components analysis. One may also use the factors that have sufficiently large eigenvalues in a factor model. These solutions are unattractive, as, in practice, data collection monitoring and strategies are based on identifiable subgroups in the population. The second option is to adjust indicators themselves for correlations between variables. A possible choice is to rank conditional partial R-indicators rather than R-indicators themselves. Conditional partial R-indicators, see Schouten, Shlomo and Skinner (2011), compute the within variance in response propensities attributable to one auxiliary variable. In doing so, they adjust for any collinearity with other auxiliary variables. This option is attractive as it employs the untransformed auxiliary variables. However, for the other indicators there is not (yet) a conditional counterpart, so that we cannot apply it to these indicators. Alternatively, we can apply the test to remaining nonresponse bias after adjustment for some estimator, say the GREG estimator, adding variables to the adjustment one by one. The third option is to choose auxiliary variables in such a way that they are diverse and show little collinearity.

In section 5, we will apply the rank tests to R-indicators, coefficients of variation, contrasts, conditional partial R-indicators and to biases of GREG estimator. First, in the next section we investigate the performance of the rank test in a simulation study.

4.3 A simulation study on the impact of collinearity, response variation and sample size

In the simulation study the sample size, the amount of collinearity and the variation of response probability variances over designs are varied. The aim of the study is an analysis of type 1 and type 2 errors of the rank test as a function of these three parameters. The simulation study consists of repeated draws of: 1) M auxiliary variables on a population with G homogeneous subgroups, 2) response probabilities for D designs on these subgroups drawn from Beta distributions in which a parameter θ is used to manipulate variation between designs, 3) samples of n units from the population, and 4) responses to each design for each sampled unit given the response probabilities. The parameter G regulates the collinearity; the smaller the number of subgroups, the more the auxiliary variables will covary because they are based on the same selection of subgroups. The parameter θ takes values in the interval [0,1], where $\theta = 0$ implies that response probabilities are drawn from the same distribution for all designs and $\theta = 1$ correspond to a maximal variation in drawing response probabilities over different designs. Parameter n regulates the sample size. One draw in the simulation thus consists of four steps, which are described in the following.

First, a set of M auxiliary variables is constructed. Each auxiliary variable is a simple 0-1 variable, which reflects the categorical nature of many of the auxiliary variables available on sampling frames and registry data. Each of the M auxiliary variables is constructed independently as follows:

1. Draw G_X uniformly from the set $\{1,2,3,\dots,G-1\}$
2. Draw a simple random sample without replacement of size G_X from the set of subgroups $\{1,2,3,\dots,G\}$
3. The candidate m th variable is constructed by assigning the value one for elements in the subgroups that are selected and zero otherwise.
4. If the candidate variable is not a linear combination of the $m-1$ previously constructed variables and the unit variable taking on the value one for all subgroups, then it is accepted as the m th variable, otherwise it is rejected.

Second, for each subgroup g and design d a response probability is drawn. This is done in two steps: each design first gets a beta distribution where the two shape parameters α_d and β_d are drawn independently over designs, and, second, from this beta distribution design response probabilities are drawn for the subgroups. In the random draws of α_d and β_d, one other parameter is involved that can be varied. The probabilities are drawn as follows:

1. For each design d, a variance σ_d^2 is drawn from a uniform distribution on the interval $[(\xi-\theta)/4, (\xi+\theta)/4]$. σ_d^2 is the expected variance of the response probabilities for the design. Parameter θ moderates the amount to which different designs can have different variances. Parameter ξ moderates the variance of response probabilities over subgroups, independent of the design. Both parameters must be elements of the interval [0,1]. Only if $\theta = 0$, it holds that all designs have equal response probability distributions (in expectation).
2. For each design d, an expectation μ_d is drawn from a uniform distribution on the interval $[0.5 - 0.5\sqrt{1-4\sigma_d^2}, 0.5 + 0.5\sqrt{1-4\sigma_d^2})$. μ_d is the (expected) response rate of the design. The range of the uniform distribution depends on σ_d^2 through $\mu_d(1-\mu_d) \leq \sigma_d^2$.
3. The expectation μ_d and variance σ_d^2 are transformed to $\alpha_d = \mu_d(\mu_d(1-\mu_d)/\sigma_d^2 - 1)$ and $\beta_d = (1-\mu_d)(\mu_d(1-\mu_d)/\sigma_d^2 - 1)$ and subgroup response probabilities are independent draws from the Beta(α_d, β_d) distribution.

The last two steps consist of drawing a sample of n units and their 0-1 response indicators for each of the designs. This is done by drawing first a simple random sample with replacement of size n from the G subgroups, linking the constructed auxiliary variables and response probabilities, and then drawing responses from Bernoulli distributions.

In each iteration of the simulation, the R-indicator, partial R-indicators, coefficient of variation, contrast and remaining biases after adjustment are computed as follows:

- R-indicators, coefficients of variation, contrasts are computed by adding one by one the auxiliary variables to the response model. A $d \times m$ matrix results with entry (i, j) the indicator value with the first j variables included in the model for design i.
- Partial R-indicators are computed by taking the unconditional partial R-indicator for the first variable and taking conditional partial R-indicators for the other variables that are added one by one. A $d \times m$ matrix results with entry (i, j) the indicator value for variable j with the first j variables included in the model for design i.
- GREG biases are computed by taking the response mean for the first variable and for each new variable taking remaining biases after GREG adjustment on the previous variables. A $d \times m$ matrix results with entry (i, j) the remaining bias for variable j calibrated on the first $j-1$ variables for response to design i.

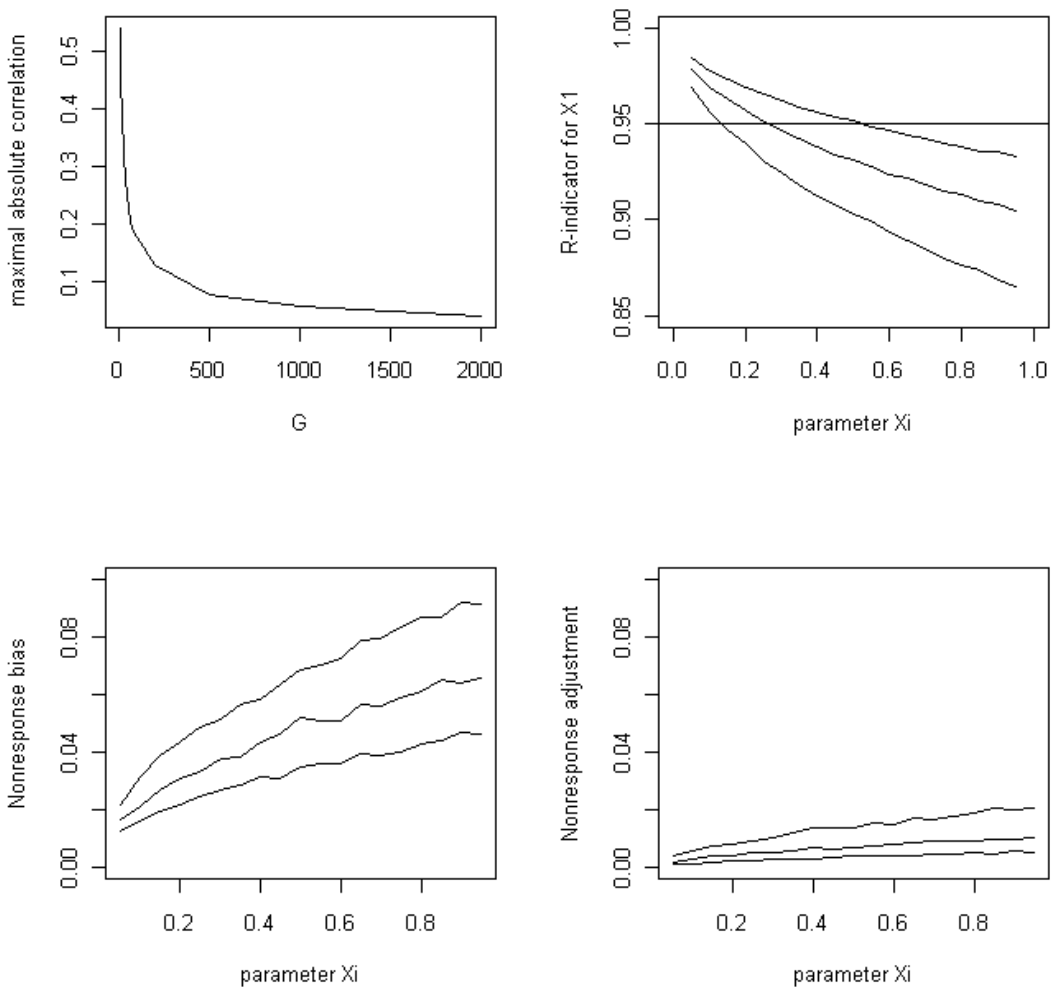

Figure 4.1: a) (top-left) maximal absolute correlation when six auxiliary variables are available as a function of population diversity G*, b) (top-right) R-indicator for one auxiliary variable as a function of* ξ *for* $G = 50$ *(bottom),* $G = 100$ *(middle) and* $G = 200$ *(top), c) (bottom-left) maximal absolute unadjusted nonresponse bias for six auxiliary variables for* $G = 50$ *(top),* $G = 100$ *(middle) and* $G = 200$ *(bottom), and d) (bottom-right) maximal absolute nonresponse adjustment for six auxiliary variables for* $G = 50$ *(top),* $G = 100$ *(middle) and* $G = 200$ *(bottom).*

The values of the different indicators are computed and stored when none of the categories of the auxiliary variables has zero respondents. If at least one category has no respondents, then the iteration is not further evaluated and used.

For each draw and for each indicator, the rank test is applied to the resulting matrix with indicator values and leads to a rejection or not. The type 2 error for an indicator is estimated by the proportion of draws that did not lead to a rejection for a specified significance level. The type 1 error for an indicator is estimated by the proportion of

draws that did lead to a rejection for a specified significance level under $\theta = 0$. Hence, each indicator is evaluated separately.

Before performing the simulation, we have to choose realistic values of the parameters G, ξ, φ and ψ. Although we will vary G in the study, we have to choose a range of values that fits to real survey data sets. We will vary θ based on realistic values for ξ and the sample size from $n = 2000$ to $n = 10000$. Figure 4.1 shows four plots that provide insight into the choice of the parameters. Figure 4.1a shows the maximal absolute correlation between the auxiliary variables when six auxiliary variables are available, $m = 6$, as a function of G. The maximal correlation shows a hyperbolic shape and decreases strongly for small values of G. For $G = 100$, the maximal correlation is approximately equal to 0.18, which is close to values we find for the datasets that we have investigated. For a realistic choice of ξ, we simulated the R-indicator for a single auxiliary variable for $G = 50$, $G = 100$ and $G = 200$. Figure 4.1b shows the decrease of the R-indicator as a function of ξ for the three values of G. Based on the R-indicators we typically find in survey datasets, we choose $\xi = 0.3$.

Figure 4.1c and d show, respectively, the maximal absolute bias of the response mean and the maximal absolute size of the adjustment produced by the GREG estimator for $G = 50$, $G = 100$ and $G = 200$. The biases and adjustments are based on a Monte Carlo simulation using 200 iterations, so that the plots still show some lack of smoothness. For figure 4.1c, the number of auxiliary variables was $m = 6$. For figure 4.1d, the number was $m = 9$ and the first three variables were used to perform the GREG estimators for the other six. The figures show that the larger the variance of response propensities and the larger the collinearity, the larger the bias and adjustments. However, as expected the remaining bias also increases with increasing ξ and decreasing G.

Tables 4.1 to 4.4 present the results of several simulations using different choices of n, G, θ, m and d. Tables 4.1 and 4.3 give estimated type 1 errors for $\theta = 0$, while tables 4.2 and 4.4 give estimated type 2 errors for $\theta = 0.3$. In all cases, we use a significance level of 5% to construct critical values. The estimated rejection and acceptance rates are based on Monte Carlo runs of 250 iterations. As a result, the estimated rates are still subject to some standard error. We are, however, mostly interested in patterns as a function of the parameters: The R-indicator, coefficient of variation and contrast almost always reject independence of rankings, regardless of the value of θ, and cannot be used to detect consistency of patterns over different designs. The partial R-indicator rejects less often than the R-indicator, coefficient of variation and contrast, but still has undesirable properties when the number of auxiliary variables m and the number of designs gets large. The bias of the GREG estimator as an indicator has good type 1 error properties except when collinearity is very weak, but it has a low statistical power. If collinearity is weak, then large samples are required to detect differences between designs. Large numbers of auxiliary variables are required to get acceptable power levels.

Before we performed the simulation study, we did not anticipate the low acceptance rates on the R-indicator, coefficient of variation and contrast. A closer look reveals that these indicators tend to move with small shocks; only a few auxiliary variables produce larger shifts in the indicator values. These shocks make the indicator patterns look consistent, but this is spurious because the majority of auxiliary variables give small changes. It is true for $\theta = 0$ that design preferences change from one variable to the other, while for $\theta = 0.3$ they point more often at the design with the largest variation in response propensities, but the indicators require adjusted critical levels for the rank test. The observed phenomenon applies to the partial R-indicator as well, even though this indicator adjusts for collinearity between auxiliary variables. Again, this result is surprising as the bias of the GREG estimator does not suffer from this problem. We believe that it is the result of modelling the same 0-1 response indicator while adding variables whereas the bias of the GREG estimator applies to a new variable every time a variable is added.

Table 4.1: Percentage of simulation runs that led to a rejection under $\theta = 0$ ($\xi = 0.3$, $m = 7$, $d = 6$).

	n=2000			*n=5000*			*n=10000*		
	G=50	*G=100*	*G=1000*	*G=50*	*G=100*	*G=1000*	*G=50*	*G=100*	*G=1000*
R	100%	100%	100%	100%	100%	100%	100%	100%	100%
CV	100%	100%	100%	100%	100%	100%	100%	100%	100%
Contrast	100%	100%	100%	100%	100%	100%	100%	100%	100%
Partial R	33%	31%	37%	37%	32%	35%	34%	32%	35%
NR bias	5%	15%	85%	9%	8%	51%	8%	6%	25%

Table 4.2: Percentage of simulation runs that did not lead to a rejection under $\theta = 0.3$ ($\xi = 0.3$, $m = 7$, $d = 6$).

	n=2000			*n=5000*			*n=10000*		
	G=50	*G=100*	*G=1000*	*G=50*	*G=100*	*G=1000*	*G=50*	*G=100*	*G=1000*
R	0%	0%	0%	0%	0%	0%	0%	0%	0%
CV	0%	0%	0%	0%	0%	0%	0%	0%	0%
Contrast	0%	0%	0%	0%	0%	0%	0%	0%	0%
Partial R	63%	58%	63%	57%	54%	60%	49%	52%	55%
NR bias	66%	57%	10%	71%	65%	32%	71%	67%	43%

Table 4.3: Percentage of simulation runs that led to a rejection under $\theta = 0$ ($\xi = 0.3$, $G = 100$, $n = 5000$).

	d=2			*d=6*			*d=10*		
	m=4	*m=8*	*m=12*	*m=4*	*m=8*	*m=12*	*m=4*	*m=8*	*m=12*
R	0%	48%	85%	92%	100%	100%	100%	100%	100%
CV	0%	58%	91%	99%	100%	100%	100%	100%	100%
Contrast	0%	60%	86%	93%	100%	100%	100%	100%	100%
Partial R	0%	8%	16%	26%	32%	42%	50%	54%	67%
NR bias	0%	1%	6%	8%	4%	6%	13%	8%	3%

Table 4.4: Percentage of simulation runs that did not lead to a rejection under $\theta = 0.3$ *(* $\xi = 0.3$, $G = 100$, $n = 5000$ *).*

	d=2			*d=6*			*d=10*		
	m=4	*m=8*	*m=12*	*m=4*	*m=8*	*m=12*	*m=4*	*m=8*	*m=12*
R	100%	51%	9%	5%	0%	0%	0%	0%	0%
CV	100%	45%	5%	2%	0%	0%	0%	0%	0%
Contrast	100%	46%	5%	3%	0%	0%	0%	0%	0%
Partial R	100%	89%	66%	72%	49%	39%	60%	29%	17%
NR bias	100%	91%	77%	77%	65%	52%	66%	52%	47%

Table 4.5: Percentage of simulation runs on 15 comparative datasets that did not lead to a rejection under $\theta = 0.0$ *and* $\theta = 0.3$ *(*$\xi = 0.3$, $G = 100$, $n = 5000$, $m = 7$, $d = 6$ *).*

	R	*CV*	*C*	*Partial R*	*NR bias*
$\theta = 0.0$	0%	0%	0%	0%	76%
$\theta = 0.3$	0%	0%	0%	0%	0%

Table 4.5 presents the estimated type 1 and type 2 errors, again for a significance level of 5%, when multiple comparative datasets are combined. We took 15 datasets and all datasets have seven auxiliary variables and six different designs for simplicity. As expected all indicators, except the indicator based on the remaining bias of the GREG estimator, always lead to a rejection of the null-hypothesis of independent rankings. The GREG bias indicator accepted 24% of the runs for $\theta = 0.0$ which is still much higher than the prescribed 5%.

Although the main purpose of the simulation study was to gain insight into the error properties of the rank test, we found that with only three parameters, G, θ and ξ, the simulation model was able to produce realistic correlations and indicator patterns. This finding may be a starting point for estimating the diversity of a population.

5. Application to survey data sets

We apply the rank test to a wide range of data sets collected by the Institute for Social Research of the University of Michigan, and the national statistical institutes of Sweden and The Netherlands. We present the various data sets and the linked auxiliary variables in appendix A. Table 5.1 gives the full names and labels of the surveys.

The p-values are estimated for the 14 datasets separately and jointly. The number of pairwise inversions is a discrete random variable and p-values are taken as the average of the probability distribution at *m-1* and *m,* when *m* is the observed number. Table 5.2 contains p-values for the rank test applied to the partial R-indicator (Pc) and bias of the GREG estimator (B). We also computed p-values for the other indicators, and as expected from the simulation study they are small. Furthermore, we computed adjusted p-values for Pc and B based on the simulation study; for these indicators p-values were robust for collinearity in the selected auxiliary variables.

Table 5.1: Overview of data sets and number of designs and number of auxiliary variables.

Label	*Survey*	D_v	M_v
HS	Dutch Health Survey	3	6
CVS	Dutch Crime Victimisation survey	3	4
HS-CVS	HS and CVS combined	6	4
LFS	Dutch Labour Force Survey	3	6
SCS	Dutch Survey of Consumer Sentiments	8	7
SCSASD	SCS adaptive design pilot study	2	5
STS-IND	Dutch Short Term Statistics survey Manufacturing Industry	7	6
STS-RET	Dutch Short Term Statistics survey Retail Industry	3	4
LISS	Longitudinal Internet panel for the Social Sciences	3	4
LCS	Swedish Living Conditions Survey	6	7
PPS	Swedish Party Preference Survey May	6	7
SCA	USA Survey of Consumer Attitudes	4	8
NSFG	USA National Survey of Family Growth	2	8
HRS	USA Health and Retirement Survey	2	9

For the partial R-indicator five out of the 14 datasets have an unadjusted p-value smaller than 0.05 and of these five values four are smaller than 0.01. For the nonresponse bias three values are smaller than 0.05 and these are also smaller than 0.01. The adjusted p-values for the different indicators are not always consistent. For instance, the HS-CVS, SCA and NSFG datasets lead to very different p-values. This is a somewhat puzzling result. There is no strong link to the type of comparison that is made in the dataset, see table A.2 of the appendix, but the number of datasets is too small to make strong conclusions here.

Table 5.3 contains the observed numbers of inversions and corresponding p-values when multiple datasets are combined into one overall test. Three combinations of datasets are combined: all nine datasets from Statistics Netherlands, all five datasets from Stat Sweden and ISR Michigan, and all 14 datasets. In all cases the p-values are smaller than 0.05, and with one exception they are much smaller. For the conditional partial R-indicator we know from the simulation that the rank test is too optimistic, so that results should be interpreted with some care. For the nonresponse bias the performance of the test is better and still p-values are generally (much) lower than 5%. The overall test, thus, indicates that the total observed numbers of inversions are much smaller than expected if design preferences per variable were random.

Table 5.2: p-values for various comparative datasets for the partial R-indicator and bias after adjustment.

	HS	*CVS*	*HS-CVS*	*LFS*	*SCS*	*SCS-ASD*	*LISS*	*STS IND*	*STS RET*	*LCS*	*PPS*	*SCA*	*NSFG*	*HRS*
Pc	0.32	0.97	0.96	0.82	0.00	0.06	0.03	0.72	0.50	0.00	0.00	0.81	0.01	0.36
B	0.18	0.12	0.00	0.82	0.00	0.06	0.00	0.72	0.88	0.07	0.10	0.35	0.77	0.36

Table 5.3: Expected numbers of inversions, observed numbers of inversions and p-values for combined datasets from Statistics Netherlands, from Stat Sweden and ISR Michigan and from all institutes.

	Number of inversions			*p-value*	
	Expected	*Pc*	*B*	*Pc*	*B*
Stat Netherlands	189.5	142	97	<0.001	<0.001
Stat Sweden/ISR	118.5	66	97	<0.001	0.02
All	308	208	194	<0.001	<0.001

Summarizing, while individual datasets do not point strongly at consistency in design preferences, their combination does clearly indicate that nonresponse affects multiple variables simultaneously, even when adjusting for collinearity. This is, in our view, a remarkable result. Our conclusion is based on 14 datasets with a specific selection of auxiliary variables. Nonetheless, given the wide range of surveys these results do provide empirical support for balancing response on auxiliary variables by design regardless of any adjustment based on these same variables afterwards.

Our simulation study indicated that it is not meaningful to track the R-indicator and coefficient of variation when adding variables one by one. Still it turned out that these indicators can be viewed as process quality indicators. The rankings of different surveys/designs/time points or processing steps of these indicators after adding all variables are very similar to the average rankings produced by the conditional partial R-indicator and nonresponse biases when adding variables one by one. Hence, when the value of the R-indicator is lower or the value of the coefficient of variation is higher, then the partial R-indicator and nonresponse biases tend to produce higher values for multiple variables.

6. Discussion

This paper contributes in five ways to the existing literature: 1) it derives bias intervals for commonly used estimators and shows how recently proposed proxy measures appear in these intervals, 2) it provides a motivation for viewing these proxy measures as indicators of process quality, 3) it provides a rank test for evaluating patterns of the proxy measures, 4) it describes a simulation study that produces realistic proxy measure values with only a few parameters, and 5) it gives empirical evidence that balancing of survey response may on average be fruitful.

With these contributions, we hope to have provided both a theoretical and an empirical rationale to balance survey response by design, regardless of adjustment afterwards. Such balancing allows for an efficient allocation of resources to different sample units, i.e. by minimising mean square error given a certain budget or vice versa minimize the required budget given restrictions on quality. Balancing by design does not mean that one should not apply adjustment afterwards nor that one should blindly assume that balancing response will always lead to a reduced bias, even after adjustment. First, it is easy to "trick" any indicator by simply subsampling response based on subgroup response propensities. This is what large web panels sometimes do; they take subsamples that resemble the overall population. It is clear that any subsampling cannot remove bias and will, generally, only increase variances (e.g. for a discussion, see Bethlehem 2014). Adaptive and responsive survey design are, however, not about subsampling; subsampling cases means removing them from the sample and, as a consequence, subsampling would not affect the value of proxy measures. These designs are about optimal allocation of resources. Adapting survey design focussing on the indicators that we considered in this paper, implies going for the subgroups that performed worst and attempting to equalize or balance response propensities. How to do this in practical settings is not at all straightforward. Second, a full balance of response will in most practical settings be infeasible and adjustment afterwards will always be necessary. Third, given that survey resources are finite, it is imperative that variances are taken into account when adapting survey designs. This is an important topic for future research. Fourth, our empirical evidence shows that only on average did nonresponse biases show consistency behaviour relative to the indicators; for some of the datasets this consistency was not apparent. For this reason we call on others to perform similar empirical studies, and, in practice, to monitor multiple indicators next to the response rate.

We have shown that the coefficient of variation of response propensities appears in bias intervals for all commonly used estimators for the location of population distributions, i.e. design-weighted response means, inverse propensity weighting, generalized regression estimators and double robust estimators. Although, this was not the purpose of this paper, this result can be used to construct variable selection strategies for each of those estimators.

Our theoretical result that less representative response on an arbitrary (auxiliary) variable translates to less representative response and more nonresponse bias after

adjustment on other arbitrary variables is intuitive. Still, as far as we know, there is no statistical literature that attempts to model the construction of random variables on a population itself. Such modelling is needed in order to make statements about the transfer of observed relations on one subset of variables to another subset of variables; in our case the relation between missingness and auxiliary information. It is obvious that the available auxiliary information in a survey cannot be viewed as a random selection from the universe of variables. However, observed correlations between available auxiliary variables and nonresponse are generally close to zero which one would indeed expect when sampling variables at random rather than selecting them based on their predictive power for missingness. In panels where there is usually a wide range of topics offered and in surveys that include a wide range of topics, there may be sufficient rationale to balance on available auxiliary information. In surveys with a small number of main survey variables, it is insufficient to do so because these variables may drive the missingness and may thus form the exception. In practice, therefore, it is important to protect against undue assumptions by sensitivity analyses and to monitor multiple statistics, including estimates, to protect against situations that do not fit the pattern described in this paper.

References

Andridge, R.R., Little, R.J.A. (2011), Proxy pattern-mixture analysis for survey nonresponse, Journal of Official Statistics, 27 (2), 153 – 180.

Bang, H., Robins, J.M. (2005), Doubly robust estimation in missing data and causal inference models, Biometrics, 61, 962 – 972.

Beaumont, J.F., Haziza, D. (2011), A theoretical framework for adaptive collection designs, Paper presented at 5th International Total Survey Error Workshop, June 21 – 23, Quebec, Canada.

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

Bethlehem, J. (2012), Using response probabilities for assessing representativity, Discussion paper 201212, CBS, Den Haag, available at www.cbs.nl.

Bethlehem, J. (2014), Solving the nonresponse problem with sample matching?, Discussion paper 201404, CBS, Den Haag, available at www.cbs.nl.

Bethlehem, J.G., Cobben, F., Schouten, J.G. (2011), Handbook of Nonresponse in Household Surveys, Handbook, Wiley Series in Survey Methodology, USA.

Beullens, K., Loosveldt, G. (2012), Should high response rates really be a primary objective?, Survey Practice, 5 (3), 1 – 5.

Bona, M. (2004), Combinatorics of Permutations, Discrete Mathematics and Its Applications Series, Chapmann and Hall, Boca Raton, USA.

Brick, J.M., Jones, M.E. (2008), Propensity to respond and nonresponse bias, METRON – International Journal of Statistics, LXVI (1), 51 – 73.

Calinescu, M., Schouten, B., Bhulai, S. (2012), Adaptive survey designs that minimize nonresponse and measurement risk, Discussion paper 201224, CBS, Den Haag, available at www.cbs.nl..

Collins, L.M., Murphy, S.A., and Bierman, K.L. (2004), A conceptual framework for adaptive preventive interventions, Prevention Science, 5 (3), 185 – 96.

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

Groves, R.M. and 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 (3), 439 – 457.

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

Hirano, K., Imbens, G., Ridder, G. (2003), Efficient estimation of average treatment effects using the estimated propensity score, Econometrica, 2003, 1161 – 1189.

Horvitz, D., Thompson, D. (1952), A generalization of sampling without replacement from a finite population, Journal of the American Statistical Association, 47, 663 – 685.

Huber, M., Lechner, M., Wunsch, C. (2013), The performance of estimators on the propensity score, Journal of Econometrics, 175, 1 – 21.

Kreuter, F. (2013), Improving surveys with paradata. Analytic use of process information, Edited book, Wiley Series in Survey Methodology, John Wiley & Sons.

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. (2013a), Aspects of responsive design with applications to the Swedish Living Conditions Survey, Journal of Official Statistics, 29 (4), 557 – 582.

Lundquist, P., Särndal, C.E. (2013b), Responsive design, Phase II – Features of the nonresponse and applications, R&D report 2013:X, Statistics Sweden, www.scb.se

Murphy, S.A. (2003), Optimal dynamic treatment regimes, Journal of the Royal Statistical Society: Series B, 65, 331 – 355.

Murphy, S. A., Lynch, K.G., Oslin, D., McKay, J.R., TenHave, T. (2007), Developing adaptive treatment strategies in substance abuse research, Drug and Alcohol Dependence, 88 (2), 24-30.

Olsen, K., Groves, R.M., (2012), An examination of within-person variation in response propensity over the data collection period, Journal of Official Statistics, 28 (1), 29 – 51.

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. and Lundström, S. (2010), Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias, Survey Methodology, 36 (2), 131 – 144.

Schafer, J.L., Kang, J. (2008), Average causal effects from nonrandomized studies: A practical guide and simulated example, Psychological Methods, 13 (4), 279 – 313.

Schouten, B. (2007), A selection strategy for weighting variables under a not-missing-at-random assumption, Journal of Official Statistics, 23 (1), 1 – 19.

Schouten, J.G., Bethlehem, J., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N., Skinner, C. (2012), Indicators for evaluating, comparing, monitoring and improving survey response, To appear in International Statistical Review.

Schouten, J.G., Calinescu, M., Luiten, A. (2013), Optimizing quality of response through adaptive survey designs, Survey.Methodology, forthcoming.

Schouten, B., Cobben, F. (2012), Does balancing of survey response reduce nonresponse bias?, Discussion paper 201226, CBS, Den Haag, available at www.cbs.nl.

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

Schouten, J.G., Shlomo, N., Skinner, C. (2011), Indicators for monitoring and improving representativeness of response, Journal of Official Statistics, 27(2), 231 – 253.

Shlomo, N., Skinner, C., Schouten, J.G. (2012), Estimation of an indicator of the representativeness of survey response, Journal of Statistical Planning and Inference, 142, 201 – 211.

Van der Vaart, A.W. (1998), Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, UK.

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

Wagner, J. (2010), The fraction of missing information as a tool for monitoring the quality of survey data, Public Opinion Quarterly, 74 (2), 223 – 243.

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, 45 – 55.

Zajonc, T. (2012), Bayesian inference for dynamic treatment regimes: mobility, equity and efficiency in student tracking, Journal of the American Statistical Association, 107, 80 – 92.

Appendix A: Survey documentation

The empirical illustration in the paper is based on a wide variety of survey data sets from three countries. We distinguish four types of comparisons: 1) a comparison of different surveys or different survey designs with the same target population, 2) a comparison of the same survey with repeated administrations at different time periods, 3) an evaluation of a survey during data collection (such as comparisons based on time or numbers of visits or calls), and 4) an evaluation of a survey after different processing steps, where "processing steps" refers to obtaining contact, obtaining participation, recruitment for a panel.

Table A.1 shows the various data sets with some important characteristics of each. The first column contains the label of the data set that we will use in the remainder of this section. The survey data sets and their auxiliary variables were taken as they are regularly produced by the statistical subject-matter departments. In the social survey data sets, the potential number of auxiliary variables is usually larger, as various additional registry data exist. The LFS, STS and SCSASD data sets contain auxiliary variables that are closely related to the survey topics. The SCSASD dataset requires additional explanation as it concerns an experimental study directed at improving representativeness and balance of survey response. The study was linked to the Survey of Consumer Sentiments (SCS) 2009. Based on historic SCS data for the years previous to 2009, an adaptive survey design was constructed that treated population subgroups differently based on their gender, ethnicity, income, age, type of household, their zip code percentage of non-native inhabitants and the urbanization of their area of residence. The adaptive survey design aimed at improving representativeness with respect to these subgroups while keeping the survey costs and survey response rate fixed. Details can be found in Luiten and Schouten (2013). The study ran parallel to the regular SCS in which all subgroups were treated the same and as usual. Four new registry variables were linked to the SCSASD dataset that became available after the survey data collection and that were not considered in the design of the experiment. These new variables are used as validation variables to compare the indicators for the control group and the experimental group.

Table A.1: Overview of data sets.

Label	*Survey*	*Designs*	*Auxiliary variables*
HS	Dutch Health Survey 2010	web CAPI web → CATI+CAPI	Employment status, ethnicity, age, urbanization, type of household and zip code house value
CVS	Dutch Crime Victimisation survey 2006	web CATI+CAPI web → CATI+CAPI	Ethnicity, age, urbanization and type of household
HS-CVS	HS and CVS combined	mode designs of HS and CVS	Ethnicity, age, urbanization and type of household
LFS	Dutch Labour	CAPI 2009	Employment status,

	Force Survey of 2009 and 2010	CAPI 2010 CATI+CAPI 2010	ethnicity, age, urbanization, type of household and zip code house value
SCS	Dutch Survey of Consumer Sentiments 2009	after varying numbers of phone calls	Gender, ethnicity, income, age, zip code percentage nonnative, urbanization and type of household
SCSASD	SCS adaptive design pilot study 2009	regular SCS (CATI) pilot study (web+mail → CATI)	Adaptive survey design strata based on gender, ethnicity, income, age, zip code percentage nonnative, urbanization and type of household, and new variables: owns car from company, type of business, number and size of jobs
STS	Dutch Short Term Statistics survey 2007	after 25 days of data collection after 30 days of data collection after 60 days of data collection	Business size, NACE, VAT of reference month in previous year, VAT of reference month
LISS	Longitudinal Internet panel for the Social Sciences 2007 - 2010	contact recruitment interview response recruitment interview willing to be panel member registered as a panel member active in panel after one year active in panel after two years active in panel after three years	Employment status, ethnicity, age, urbanization, type of household and zip code house value
LCS	Swedish Living Conditions Survey 2009	after different number of calls including follow up	Gender, age, ethnicity, education, civil status, property ownership, income, employment status, region, benefits
PPS	Swedish Party Preference Survey May 2012	after different numbers of calls	Gender, age, ethnicity, education, civil status, income, region
SCA	Survey of Consumer Attitudes 2011-	Four waves: Sep 2011, Jan 2012, Apr 2012 and Jun 2012	Census Region (4 regions), proportion in ZCTA (neighborhood) with income

	2012		$100,000+, proportion in ZCTA Hispanic, proportion in ZCTA black, proportion of households in ZIP code with a listed telephone number, proportion in ZCTA that is 25-34 years old, telephone number with address, median Home Value of households in ZCTA
NSFG	National Survey of Family Growth 2006-2010	Two time points during data collection (intermediate and final)	Indicator for access problems to the area segment, Census Region (1-4), Indicator for urban area within large MSA, sampling domain (1= <10% black, <10% Hispanic; 2= >10% black, <10% Hispanic; 3= <10% black, >10% Hispanic; 4= >10% black, >10% Hispanic), indicator for housing unit in structure with more than one unit, indicator for evidence of non-English speakers in neighborhood, proportion eligible for NSFG in ZCTA, indicator for neighborhood has safety concerns
HRS	Health and Retirement Survey 2006, 2008	Two waves 2006 and 2008	Age group cohort, indicator for less than high school education, indicator for part of a couple, gender, marital status, indicator for black, indicator for in a nursing home, indicator for 65 years or older in 2006, indicator for Hispanic

Table A.2 gives the type of comparison that is made for each of the data sets. The 13 datasets allow for all types of comparisons; four look at different surveys or designs of the same survey, three consider the same survey in different waves, four correspond

to different amounts of effort during data collection and two involve different data collection steps.

Table A.2: Overview of comparisons.

	Type of comparison			
	Different surveys/designs	*Survey in time*	*Data collection*	*Processing steps*
HS	x			
CVS	x			
HS-CVS	x			
LFS		x		
SCS			x	
SCSASD	x			
STS			x	
LISS				x
LCS			x	
PPS				x
SCA		x		
NSFG			x	
HRS		x		

Appendix B: Probability distribution of the rank test

This appendix describes the construction of the rank test that is used in the main document to test absence of patterns in indicator values and nonresponse biases. We first introduce notation and derive the probability distribution of the number of pairwise inversions or the number of discordant pairs in independent rankings. This distribution forms the basis for the rank test. In the main document we analyse the properties of the rank test in a simulation study.

B.1 The probability distribution of the number of pair wise inversions

Assume there are V comparative data sets, i.e. different surveys, different designs of the same survey or different levels of effort, labelled by $v = 1,2,\dots,V$. In the following, we will simply refer to different designs.

For illustration purposes, we use the following example with $V = 2$: data set 1 is the Labour Force Survey conducted through three different survey modes on independent samples and data set 2 is the Health Survey of five different calendar years. In data set 1, we have auxiliary variables age, type of household, employment status, urbanization of residence area, ethnicity and average house value at zip code. In data set 2, we have gender, age and income.

We introduce some notation. Let D_v be the number of different designs in comparative data set v. In the example, we have $D_1 = 3$ and $D_2 = 5$. Let $X_v = (X_{v,1}, X_{v,2}, \dots, X_{v,M_v})^T$ be the vector of auxiliary variables that is input to the comparison in data set v. The comparison thus consists of M_v auxiliary variables, which may again be different in each data set v. In the example they are $M_1 = 6$ and $M_2 = 3$. Assume that the auxiliary variables in X_v are independent.

The test statistic that we propose in the section B.2 is based on the number of pairwise inversions that result when ranking the different designs according to the indicators and to remaining bias. A small number of pairwise inversions implies clustered preferences, i.e. the indicator shows a consistent picture when different variables are considered.

With a pairwise inversion we mean the inversion of two consecutive designs. Two examples:

> 2,1,3,4 requires only one pairwise inversion to get to 1,2,3,4, namely (2,1).
>
> 1,3,4,2 requires two pairwise inversions to get to 1,2,3,4, namely (4,2) and (3,2).

In this section, we derive the probability distribution of the number of pairwise inversions between two independent rankings of the designs.

Without loss of generality, it can be said that the first ranking leads to a preferred order $1,2,3,\ldots,D$ of the designs. So design 1 performs best and design D performs worst. When the two rankings are independent, then the new sequence may be any permutation of the D designs. There are $D!$ ways in which the designs can be ordered and each sequence has an equal probability of being realized, i.e. $1/D!$. The maximal number of pair wise inversions is $D(D-1)/2$ and is attained when $1,2,3,\ldots,D$ is reversed to the sequence $D, D-1,\ldots,2,1$.

It is easy to show that the probability distribution of the number of pairwise inversions is symmetrical and has expectation $D(D-1)/4$. The number of sequences of design labels that requires d pairwise inversions has no easy closed form, but it can be shown (Bona 2004) that it is the coefficient of the term z^d in the polynomial

$$\prod_{l=1}^{D}\left(\sum_{k=0}^{l} z^k\right) = 1\times(1+z)\times(1+z+z^2)\times\ldots\times(1+z+z^2+\ldots+z^D). \qquad \text{(B1)}$$

The probability of d pairwise inversions is equal to this coefficient times $1/D!$. In the example, the maximal number of inversions for the first data set is three and for the second data set it is ten. The probability distributions are (1/6, 1/3, 1/3, 1/6) and (1/120, 1/30, 3/40, 1/8, 1/6, 11/60, 1/6, 1/8, 3/40, 1/30, 1/120).

B.2 A rank test for indicator preferences

For each indicator we rank the designs within a comparative data set based on increasing the number of auxiliary variables used in the model. We start by ranking the designs on X_1, then add X_2, and continue to add variables until the whole vector $X_v = (X_{v,1}, X_{v,2}, \ldots, X_{v,M_v})^T$ is included. Hence, the included auxiliary variables function as the evaluation variables and the omitted auxiliary variables as the validation variables. The evaluation set of variables grows with each step, while the validation set of variables shrinks.

If we assume that indicators computed on evaluation variables are not predictive of indicators based on validation variables, then it holds that the different rankings are independent. The total number of pair wise inversions needed to go from the first ranking to the last ranking is the sum of $M_v - 1$ independent numbers or pairwise inversions. We derived the probability distribution of this number in the previous section.

A small total number of pairwise inversions implies clustered preferences, i.e. the indicator shows a consistent picture when different variables are considered. Let $I_{v,m}$ be the number of pairwise inversions in data set v when adding variable m. Let μ_v be

the expected total number of pairwise inversions. Then the null hypothesis is $H_{v,0}: \mu_v \geq (M_v - 1)D_v(D_v - 1)/4$, and the test statistic is

$$T_v = \sum_{m=1}^{M_v - 1} I_{v,m} \ . \qquad \text{(B2)}$$

It is possible to test all comparative data sets simultaneously, i.e. summing the individual test statistics T_v in (B2) over all data sets. We will do both – that is, test datasets individually and jointly.

Essentially, the testing problem in this section is a multiple sample location problem. A standard test is the Kruskal-Wallis test, see e.g. Van der Vaart (1998). The Kruskal-Wallis is a chi-square test for large M . In our case, M is generally small, so that we cannot assume a chi-square distribution but would have to derive the exact probability distribution. The test we propose here uses a different statistic and is more intuitive.

Tables B.1 and B.2 contain the 5% and 1% quantiles, respectively, for different values of D_v and M_v. The rank test is implemented in R by the authors and the code is available upon request.

Table B.1: 5% quantiles for various values of D and M.

	D						
M	*2*	*3*	*4*	*5*	*6*	*7*	*8*
2	-	-	0	1	2	4	6
3	-	0	2	4	8	12	18
4	-	1	4	8	14	21	29
5	-	2	6	12	20	30	42
6	0	3	9	16	27	39	54
7	0	4	11	21	33	49	67
8	0	5	14	25	40	58	79
9	1	7	16	30	47	68	92
10	1	8	19	34	53	77	105

Table B.2: 1% quantiles for various values of D and M.

	D						
M	*2*	*3*	*4*	*5*	*6*	*7*	*8*
2	-	-	-	0	1	2	4
3	-	-	0	2	5	9	14
4	-	0	2	6	11	17	25
5	-	1	4	10	17	26	36
6	-	2	6	14	23	34	48
7	-	3	9	17	29	43	60
8	0	4	11	22	35	52	72
9	0	5	13	26	42	61	85
10	0	6	16	30	48	70	97

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
empty cell	Not applicable
2013–2014	2013 to 2014 inclusive
2013/2014	Average for 2013 to 2014 inclusive
2013/'14	Crop year, financial year, school year, etc., beginning in 2013 and ending in 2014
2011/'12–2013/'14	Crop year, financial year, etc., 2011/'12 to 2013/'14 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68

XXXX 20130X X-00