A simulation approach to evaluate the cost efficiency of nonresponse follow-ups

Sara Westling

1 Introduction

1.1 Background

In surveys, nonresponse has become an almost unavoidable, though undesired, feature. Surveys with nonresponse suffer from many difficulties and possible deficiencies, practical in data collection and planning as well as in estimation and inference. The main concern in the presence of nonresponse is the obvious risk of bias, but the variance of point estimators will also be inflated. In addition, not only point estimators, but also variance estimators may be biased.

Although nonresponse is usually difficult to avoid entirely, survey administrators and researchers often try to minimize the nonresponse rate within the available budget, at the expense of other activities to reduce the total error. Many efforts are both expensive and time consuming, and may have little effect on the nonresponse error. One important category of methods to deal with nonresponse is *reduction efforts*, i.e. efforts that take place after the initiation of the data collection period and that aim at reducing the nonresponse rate. The efficiency of the reduction efforts must be measured by the nonresponse error reduction they result in (if any), but this reduction must also be weighed against the amount of resources that is required to achieve it. If the effect is small compared to the cost, the budget could be reallocated to other survey activities that improve quality or the survey could be conducted at a lower cost. In Tångdahl (2006), an approach to evaluate the cost efficiency of the nonresponse rate reduction efforts is proposed, based on a framework introduced in Tångdahl (2004) and in Tångdahl (2005). This approach, developed under an ideal situation where all required quantities

are assumed known, serves as a point of reference in this paper. Here the "ideal" procedure is adapted to some practical situations where the quantities required for the evaluation are unknown. Section 1 gives an introduction to the problem and lays out some notation and definitions. Section 2 contains a short summary of the approach suggested in Tångdahl (2006), and in sections 3 and 4 an adaptation of the approach to some specific situations that are more or less likely to occur in practice is proposed.

1.2 Notation and definitions

Consider a finite study population U of size N. The variable of interest is y with the value y_k for the kth element and we want to estimate the population total $t_{yU} = \sum_{k \in U} y_k = \sum_U y_k$.

To do this, select a probability sample s of fixed size n from U according to the design $p(\cdot)$ with positive first and second order inclusion probabilities π_k and π_{kl}. For simplicity, assume that direct element sampling without replacement is used. Data collection is initiated by an attempted contact with all elements in the sample. This results in a number of responding elements and nonresponding elements (e.g. non-contacts, refusals). During the data collection period, additional contact attempts are made with sample elements who have not yet responded or have not yet been categorized as definitive nonresponse. After a sufficiently long time period, usually at a predetermined time point in postal surveys or after a specified number of callbacks in interview surveys, say A, the data collection is terminated in the current survey setup. During the data collection period, successive response sets $r^{(1)} \subseteq \cdots \subseteq r^{(a)} \subseteq \cdots \subseteq r^{(A)} \subseteq s$, are generated, where $a = 1, \ldots, A$ denotes some point of time during the data collection period. The number of elements in $r^{(a)}$ is $m^{(a)}$. In interview surveys, it is more natural to define the response sets in terms of the number of callbacks. In the following we assume, for simplicity, that the survey is carried out by mail.

The response set $r^{(a)}$ is assumed to have been generated by the response distribution $RD^{(a)}$, influenced by the survey operations up to the time point a. Hence, for $a = 1, \ldots, A$, there is a corresponding sequence of response distributions. The response distribution $RD^{(a)}$ is such that elements respond independently of each other. The assumption of independence implies that $\Pr(k \& l \in r^{(a)}|s) = \theta^{(a)}_{k|s}\theta^{(a)}_{l|s}$ for every pair $k \neq l \in s$. We also require that the condition

$$\theta^{(a)}_{k|s} \geq \theta^{(a-1)}_{k|s} \tag{1}$$

is fulfilled for all k and all a.

An arbitrary point estimator for t_{yU} based on the response set $r^{(a)}$, generated by $RD^{(a)}$, is denoted $\hat{t}_{yc}^{(a)}$, while $\hat{t}_{ys}$ is the corresponding full response estimator, i.e. an estimator on the same form as $\hat{t}_{yc}^{(a)}$, but based on the total sample s.

Remark 1 *The superscript* (a) *will be used throughout to denote estimators and quantities based on response set* $r^{(a)}$.

There are many different methods available for handling nonresponse at the estimation stage, one being reweighting. In the following, we need not specify what estimator is currently being used in the survey. In situations where a specific estimator is needed to illustrate some technical points, the simple RHG estimator, studied in Tångdahl (2004), will be used. There, a slightly modified version of the RHG model described in Särndal, Swensson, and Wretman (1992), chapter 15, is defined. It is formulated as follows: assume that a partitioning of the realized sample can be made such that response probabilities are constant within groups s_h, $h = 1, \ldots, H$. It is also assumed that the response probabilities are positive for all elements and that elements respond independently. The partitioning need not be the same for different samples, but for a given sample, the grouping is always the same. In the present particular setting, we will assume that the same model, i.e. the same grouping, applies for any given time a, only the response probabilities within groups may change over time. The model can be stated formally as

$$\Pr(k \in r^{(a)}|s) = \theta_{k|s}^{(a)} = \theta_h^{(a)} > 0 \text{ for all } k \in s_h$$
$$\Pr(k\&l \in r^{(a)}|s) = \theta_{k|s}^{(a)}\theta_{l|s}^{(a)} \text{ for all } k \neq l \in s$$

for $h = 1, \ldots, H$ and for given time points $a = 1, \ldots, A$.

Let n_h be the size of s_h and let $r_h^{(a)}$ of size $m_h^{(a)}$ be the responding subset of s_h at time a. Conditioning on the response count vector $\mathbf{m}^{(a)}$, estimated first and second order response probabilities are

$$\hat{\theta}_{k|s}^{(a)} = \hat{\theta}_{hs}^{(a)} = \frac{m_h^{(a)}}{n_h} \text{ for all } k \in s_h$$

and

$$\hat{\theta}_{kl|s}^{(a)} = \begin{cases} \dfrac{m_h^{(a)}(m_h^{(a)}-1)}{n_h(n_h-1)} & \text{for } k \neq l \in s_h \\ \dfrac{m_h^{(a)}}{n_h}\dfrac{m_{h'}^{(a)}}{n_{h'}} & \text{for } k \in s_h, l \in s_{h'}; h \neq h' \end{cases}$$

Remark 2 *In the original presentation, the conditional response probabilities* $\theta_{k|s}^{(a)}$ *and* $\theta_{kl|s}^{(a)}$ *are defined under the assumption that the assumed model coincides with the true response distribution. In this presentation, these response probabilities are denoted* $\hat{\theta}_{k|s}^{(a)}$ *and* $\hat{\theta}_{kl|s}^{(a)}$ *to emphasize that this is an assumption made by the statistician and is not necessarily true.*

Details on the simple RHG estimator can be found in Särndal et al. (1992) and in Tångdahl (2004).

2 The evaluation approach in an ideal situation

To evaluate the cost efficiency of the nonresponse rate reduction efforts, the effect of the reduction efforts must be related to their costs. In Tångdahl (2006), this is done by using a *cost efficiency measure*, defined as a function of an error measure given at two arbitrary points of time during the data collection period, and the expected cost of the survey procedures at the same time points. Four alternative cost efficiency measures are proposed, each representing different error measures and different ways to relate this error to the expected costs.

The effect of the reduction efforts is defined as the change in the chosen error measure, which is suggested to be a combination of estimator bias and variance into one measure. The estimator bias and variance are defined in terms of the true, but unknown, response distribution, where sampled elements may have different individual response probabilities, between 0 and 1. By expressing the error, and also the cost of a survey procedure in terms of the true response distribution $RD^{(a)}$ at a single point of time, and using the defined sequence of response distributions, we can compare the nonresponse error at different points of time, and also define the cost efficiency of efforts between time a and time $a-1$.

Tångdahl (2006) presents an evaluation procedure that builds on this setup, and that can be used to evaluate the efficiency of the reduction efforts relative to their costs. An evaluation is initiated by calculating the

bias and variance of the point estimator at time A, i.e. under the current survey strategy. (In practice it may be necessary to estimate these quantities.) We may find that the bias is sufficiently small and that the precision requirements are met, in which case the evaluation is relevant to see if the nonresponse reduction efforts are cost efficient. Or we may find that the bias is unacceptably large and/or that the variance is too large. In this case an evaluation is motivated to find out which efforts, if any, do not have the desired effect on survey error, and to get input for alternative strategies to be considered. In any case, the procedure is developed under an ideal situation, i.e. it is assumed that all required population quantities and parameters are known. If that is the case, the following procedure for a full evaluation of the follow-up strategy is suggested. First choose which time points during the data collection that should be evaluated, and which pairs of time points are to be compared. It is suggested that the pairs are chosen as consecutive time points. For a complete evaluation of the data collection procedure, or rather, the strategy for nonresponse rate reduction, it is recommended that the evaluation is performed stepwise, starting with the last effort. Note that the evaluation is done pairwise, for each chosen pair of consecutive time points $a-1$ and a. The choice of time points to be evaluated is discussed in Tångdahl (2006).

The cost efficiency measures to consider in the suggested approach are

$$CE_{1,MSE}^{(a|a-1)} = \frac{MSE(\hat{t}_{yc}^{(a-1)})E(C_T^{(a-1)})}{MSE(\hat{t}_{yc}^{(a)})E(C_T^{(a)})} \tag{2}$$

$$CE_{1,BR}^{(a|a-1)} = \frac{BR^2(\hat{t}_{yc}^{(a-1)})E(C_T^{(a-1)})}{BR^2(\hat{t}_{yc}^{(a)})E(C_T^{(a)})} \tag{3}$$

$$CE_{2,MSE}^{(a|a-1)} = \frac{MSE(\hat{t}_{yc}^{(a)}) - MSE(\hat{t}_{yc}^{(a-1)})}{E(C_T^{(a)}) - E(C_T^{(a-1)})} \tag{4}$$

$$CE_{2,BR}^{(a|a-1)} = \frac{BR^2(\hat{t}_{yc}^{(a)}) - BR^2(\hat{t}_{yc}^{(a-1)})}{E(C_T^{(a)}) - E(C_T^{(a-1)})} \tag{5}$$

where $MSE(\hat{t}_{yc}^{(a)}) = V(\hat{t}_{yc}^{(a)}) + B^2(\hat{t}_{yc}^{(a)})$ and $BR^2(\hat{t}_{yc}^{(a)}) = B^2(\hat{t}_{yc}^{(a)})/V(\hat{t}_{yc}^{(a)})$ are the mean square error and squared bias ratio, respectively, while $C_T^{(a)}$ is the total cost of $\mathfrak{P}^{(a)}$, the survey procedure terminated at time a, which is a stochastic variable.

One of these measures is chosen. For every pair of time points, the following steps are then performed:

Step 1 The chosen cost efficiency measure is calculated. The decision rule is that the survey strategy $(\mathfrak{P}^{(a)}, \hat{t}_{yc}^{(a)})$ is cost efficient, relative to $(\mathfrak{P}^{(a-1)}, \hat{t}_{yc}^{(a-1)})$, if

$$CE_1^{(a|a-1)} > \delta,\ \delta > 0 \qquad \text{or} \qquad CE_2^{(a|a-1)} < \delta,\ \delta < 0$$

depending on which measure is chosen. The constant δ is chosen to reflect the importance of a quality improvement.

Step 2 The variance of the point estimator at time $a-1$ is calculated, to determine if the precision requirements are met under $\mathfrak{P}^{(a-1)}$.

Step 3 The magnitude of the absolute and relative bias at time $a-1$ is calculated. Even if one or more efforts are not found to be cost efficient and thus could be excluded, there may still be a large bias present. If so, the complete survey procedure must be reappraised.

Using the information from steps 1, 2 and 3 for all pairs of time points, an assessment of the efficiency of the reduction strategy as a whole can be made.

3 The evaluation approach in less than ideal situations

3.1 The problem

As stated in the introduction, the purpose of this paper is to suggest how to use the approach in Tångdahl (2006) in practical situations that may occur. The evaluation approach, described in section 2, is straightforward to apply given that all required quantities are known. In practice this is rarely, if ever, the case. In this paper, several more or less plausible situations, with varying amounts of information are considered. It is obvious that the less information we have, the more the evaluation must depend upon various assumptions.

We suggest how to deal with the uncertainty about the quantities used in steps 1-3 of the evaluation approach in each of the situations considered. These are described in section 3.2. Since steps 2 and 3 of the approach involve calculation or estimation of the variance and the bias, which are included in

step 1, focus will mainly be on step 1. Brief statements on how steps 2 and 3 should be performed are given.

Unfortunately, the nature of a situation with nonresponse is that assumptions about the nonresponse mechanism are unverifiable, since we lack data to test them on. In those cases, it is up to the survey administration, and the survey statistician in particular, to use their judgment and past experience of similar surveys to make different but realistic assumptions. The approach proposed here is simulation based. When extensive information is available, the CE measure, the variance and the bias can be estimated with arbitrary precision. In the cases with little information, the simulations must be based on several more or less plausible assumptions about unknown and unestimable quantities, performing an evaluation for each of those assumptions and comparing the conclusions drawn under the different assumptions. If the same or at least similar conclusions can be drawn, we can be fairly confident that they are correct. If, however, the results are inconclusive, resources to collect additional information must be found, so that we may narrow down the range of alternative assumptions.

An alternative to a simulation approach is to use the available information in each situation to find point estimators of the cost efficiency measure, the variance and the bias. Replacing the components of the cost efficiency measure with estimates requires that we find point estimators of the variance, the squared bias and the expected cost at time $a-1$ and a. As an unbiased estimator of the expected cost, the realized cost in the survey can be used. Since standard variance estimators may be biased, jackknifing can be used to estimate the point estimator variance. Jackknife for reweighting estimators under nonresponse has been treated in e.g. Kott (1998) and Kott (2001).

In more ideal situations, it may be possible to find an unbiased estimator of the bias, but in most cases this is impossible. Instead, auxiliary information must be used (if any is available) to make informed guesses about the bias. Even in cases where we can estimate the bias, the variance and the expected cost and insert these into the expression for the cost efficiency to arrive at an approximately unbiased point estimator for the cost efficiency, we must still try to provide some estimate of the uncertainty in the estimators. If it is possible to estimate the variance of the cost efficiency estimator, the bias estimator and the variance estimator, and also find their sampling distributions, confidence intervals can be constructed in each of steps 1-3. It turns out, however, that even in the most ideal situations this is difficult.

3.2 The situations

In practice, we may have access to different types of information, and it may be available at different levels. In particular, it may be that, at the evaluation stage, we have access to auxiliary information not used in the current estimation. Denote this additional information $\mathbf{z}_k$ for element k, to distinguish it from $\mathbf{x}_k$, the auxiliary information used in the current estimator. The vector $\mathbf{z}_k$ may be an updated version of $\mathbf{x}_k$, but may also contain one or more new variables, not used previously. We will assume that $\mathbf{z}$ has a stronger predictive ability than $\mathbf{x}$, when used in the estimation of t_{yU}. Another, even more favorable situation, is when the values of the study variable y_k become known at the evaluation stage, for all elements in the population. An example is *Omsättningsstatistiken* at Statistics Sweden, a monthly survey on turnover for businesses in the service sector. Estimation of monthly turnover is based on questionnaire data from the responding elements of a sample, while a register variable from which turnover can be calculated becomes available after each quarter.

Consider a survey that has been carried out regularly for some time, with few changes in the general setup. In this case, we may have a fair idea about the form of the true response distribution. Assume that the parameters of the response distribution are known from previous experience or that we have access to the information we need to estimate these parameters from the obtained sample. We denote this case *RD known*. Of course, if we have a fair idea about the form of the response distribution, that would be used at the estimation stage, in our current standard estimators. If that was possible, point and variance estimators would have negligible bias. What we assume here is that the information required is not available until the evaluation stage, so that less than perfect information must be used in the current estimator.

The case *RD unknown* denotes situations where the form of the true response distribution is unknown (which is the most common situation in practice), or where it is known but can not be fully identified with the available information (for example if we know that there are response groups defined by age by income classes, but income is unknown even at the evaluation stage).

Table 1 summarizes the different situations that are more or less likely to occur. We will focus on the situations 1a), 1c), 1d), 3c) and 3d), as marked in table 1. Case 1a) is unrealistic in practice, but will serve as a transition

		y_k known for		
		$k \in U$	$k \in s$	$k \in r^{(A)}$
RD known	$\mathbf{z}_k$ known for $k \in U$	**1a)**	2a)	3a)
	no additional information	1b)	2b)	3b)
RD unknown	$\mathbf{z}_k$ known for $k \in U$	**1c)**	2c)	**3c)**
	no additional information	**1d)**	2d)	**3d)**

Table 1: Different cases that may occur at the evaluation stage

between the ideal situation considered in section 2 and the less than ideal cases where RD is unknown. The other cases are chosen because in some respects, they represent "extreme" situations. The case where y_k is known for all $k \in s$ is most likely rare in practice and will not be covered here. The suggested methods for the case y_k known for $k \in r^{(A)}$ can also be applied when y_k is known for $k \in s$, the main difference being that we are in a better position to estimate unknown quantities. Although the situations in table 1 are different, they share some common features and conditions that must be fulfilled. Firstly, we will assume that data on the data collection process are collected, in particular that the inflow of each response is registered, i.e. we have a time stamp on incoming questionnaires or on interviews. This is necessary in order to determine which elements belong to the response set $r^{(a)}$, and which belong to the nonresponse set $s - r^{(a)}$, for any choice of $a \leq A$. Also, in interviewer assisted surveys, process data need to be collected so that all contact attempts can be identified. Secondly, we will assume that a cost model has been formulated, describing the data collection process in terms of the reduction efforts, and in factors that can be related to the error model, which in our case is a function of the (unknown) response distribution. Thirdly, detailed cost data must be collected, so that the costs of each reduction effort can be identified and separated from the other costs. To be able to use the methods suggested in the following, per element costs, not only total costs of each reduction effort must be known.

3.3 The next to ideal situation

In this section, we deal only with case 1a), the situation where y_k is known for all $k \in U$ and RD is known. Four different scenarios are identified in this case. The situations $\theta_{k|s}^{(a)}$ *is independent of* s and $\theta_{k|s}^{(a)}$ *depends on* s require slightly different approaches, and for each of these we may have $\theta_{k|s}^{(a)}$ *is known* or $\theta_{k|s}^{(a)}$ *must be estimated*.

3.3.1 RD known, $\theta_{k|s}^{(a)}$ independent of s and are known

In the ideal situation where y_k and the response probabilities $\theta_{k|s}^{(a)}$ ($=\theta_k^{(a)}$ for every s) are known for all $k \in U$ and for $a = 1, \ldots, A$, it is in principle possible to directly calculate the bias, variance, squared bias and expected total cost. This means that the cost efficiency measure can be calculated and steps 1-3 from section 2 directly applied . However, to do this, explicit expressions for the components of the cost efficiency measure, i.e. the variance, squared bias and expected cost, need to be derived. In many applications, this is prohibitively difficult or even impossible. In Appendix A, an explicit expression is derived for the cost efficiency measure for the simple case of a direct element simple random sampling design and the simple RHG estimator. In this simple case, the derivation is fairly straightforward, but for other designs and estimators, the expression can become unwieldy.

In situations when explicit expressions cannot be derived for the quantities required in the evaluation steps, particularly the components of the cost efficiency measure, the known values of the study variable and the response probabilities can be used to perform a Monte Carlo simulation to estimate these quantities as follows.

For $i = 1, \ldots, M$:

i. Draw a sample s_i using the same sampling design as in the original survey.

ii. Generate response sets $r_i^{(1)}, \ldots, r_i^{(a)}, \ldots, r_i^{(A)}$ using the known response probabilities.

iii. Calculate $\hat{t}_{yci}^{(1)}, \ldots, \hat{t}_{yci}^{(A)}$ and the costs $C_{Ti}^{(1)}, \ldots, C_{Ti}^{(A)}$ based on the response sets realized in step ii.

From the simulation, we get input for all three steps of the evaluation approach. Recall that the components of the cost efficiency measures are expected cost, bias and variance. The variance $V(\hat{t}_{yc}^{(a)})$ is estimated by

$$\hat{V}_{MC}(\hat{t}_{yc}^{(a)}) = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{t}_{yci}^{(a)} - \frac{1}{M}\sum_{i=1}^{M}\hat{t}_{yci}^{(a)}\right)^2 \tag{6}$$

and the squared bias $B^2(\hat{t}_{yc}^{(a)})$ can be estimated unbiasedly by

$$\widehat{B^2}_{MC} = \widehat{MSE}_{MC}(\hat{t}_{yc}^{(a)}) - \hat{V}_{MC}(\hat{t}_{yc}^{(a)}) \tag{7}$$

where

$$\widehat{MSE}_{MC}(\hat{t}_{yc}^{(a)}) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{t}_{yci}^{(a)} - t_{yU}\right)^2 \tag{8}$$

is an unbiased estimator of the mean square error $MSE(\hat{t}_{yc}^{(a)})$. An unbiased estimator of the expected cost is

$$\hat{E}_{MC}(C_T^{(a)}) = \frac{1}{M}\sum_{i=1}^{M} C_{Ti}^{(a)} \tag{9}$$

Remark 3 *For most cost models, it is likely that we could calculate the expected cost using the known response probabilities in an explicit expression instead of estimating it from the simulation. This will influence the precision of the estimator of the cost efficiency, but it has not been investigated which one of the two is the better.*

The bias, needed in step 3 of the evaluation approach, is estimated from the simulation using

$$\hat{B}_{MC} = \frac{1}{M}\sum_{i=1}^{M}\hat{t}_{yci}^{(a)} - t_{yU} \tag{10}$$

where $\frac{1}{M}\sum_{i=1}^{M}\hat{t}_{yci}^{(a)}$ is an unbiased estimator of $E_p E_{RD}(\hat{t}_{yc}^{(a)})$.

For step 1 of the evaluation approach, the estimates of variance, squared bias and expected cost, i.e. (6), (7) and (9), replace the corresponding true values in the formula for the CE measure, one of (2) to (5), which gives an estimate $\widehat{CE}_{MC}$. (In the cost efficiency measures involving MSE, formula

(8) can be used directly.) In steps 2 and 3, the simulation variance (6) and simulation bias (10) are used, respectively.

In the case when we can calculate directly from explicit expressions the cost efficiency measure in step 1, the variance in step 2 and the bias in step 3, there is no issue of inference and the decision rules from section 2 apply directly. If a simulation is used to approximate these quantities, the results are afflicted with some uncertainty, mainly determined by the number of iterations, M. M should be chosen to achieve desired precision in estimation of the variance, since more observations are usually required to estimate variances with satisfactory precision, compared to estimates of means, such as (10) above.

M can be determined by the following reasoning: Since $\hat{t}_{yci}^{(a)}$ are based on independent samples s_i, $i = 1, \ldots, M$ and are approximately normally distributed for large samples and response sets, we have approximately

$$\frac{(M-1)\hat{V}_{MC}(\hat{t}_{yc}^{(a)})}{V(\hat{t}_{yc}^{(a)})} \sim \chi^2_{(M-1)} \tag{11}$$

For large M, this can be approximated with the normal distribution. One approximation is

$$\chi^2_{(M-1)} \sim N(M-1, \sqrt{2(M-1)}) \tag{12}$$

so that

$$\frac{\frac{(M-1)\hat{V}_{MC}(\hat{t}_{yc}^{(a)})}{V(\hat{t}_{yc}^{(a)})} - (M-1)}{\sqrt{2(M-1)}} = Z \sim N(0,1) \tag{13}$$

Half the length of a $100(1-2\alpha)\%$ confidence interval for $V(\hat{t}_{yc}^{(a)})$, expressed as a percentage of $V(\hat{t}_{yc}^{(a)})$, is given by

$$\frac{z_{1-\alpha} - z_{\alpha}}{\sqrt{2(M-1)}} 100 \tag{14}$$

Say we require that (14) does not exceed ϵ, we can then solve for M and get

$$M - 1 \geq \frac{(z_{1-\alpha} - z_{\alpha})^2}{2\epsilon^2} 10000 \tag{15}$$

Since we can achieve arbitrary precision in estimates from the simulation by increasing the number of iterations, we can disregard this source of uncertainty in the analysis. It is then straightforward to follow the steps of the approach outlined in section 2.

3.3.2 RD known, $\theta_{k|s}^{(a)}$ independent of s but are unknown

In this case, assume that the form of the response distribution is known, but that the parameters of this distribution must be estimated. The approach proposed in the previous section can still be used, but since using estimates of the response probabilities in the calculations adds uncertainty, it must be adjusted.

Since the response probabilities are independent of s, data from the realized sample can be used to estimate them unbiasedly. How this should be done depends on the form of the response distribution. Here, the case of response groups is used as an example.

Assume that the response distribution is such that there are response groups U_g, defined in the population, with response probabilities

$$\Pr(k \in r^{(a)}) = \theta_k^{(a)} = \theta_g^{(a)} \text{ at time } a \text{ for } k \in U_g$$

and

$$\Pr(k \& l \in r^{(a)}) = \theta_k^{(a)}\theta_l^{(a)} \text{ at time } a \text{ for all } k \neq l \in U$$

Assuming that the event $n_g \leq 1$ has negligible probability, unbiased estimators of $\theta_g^{(a)}$, $g = 1, \ldots, G$, $a = 1, \ldots, A$, are given by

$$\hat{\theta}_g^{(a)} = \frac{m_g^{(a)}}{n_g} = \frac{1}{n_g}\sum\nolimits_{s_g} R_{k|s}^{(a)} \tag{16}$$

where $R_{k|s}^{(a)} = 1$ if $k \in r^{(a)}|s$, 0 otherwise, and $E_{RD}(R_{k|s}^{(a)}) = \theta_k^{(a)}$. The estimated response probabilities automatically fulfill condition (1). For other forms of the response distribution this may not be the case, so special care must be taken in forming estimators $\hat{\theta}_k^{(a)}$ for $a = 1, \ldots, A$.

The estimator $\hat{\theta}_g^{(a)}$ is unbiased since

$$\begin{aligned} E\left(\hat{\theta}_g^{(a)}\right) &= E_p\left(\frac{1}{n_g}E_{RD}\left(\sum\nolimits_{s_g} R_{k|s}^{(a)}|s\right)\right) \\ &= E_p\left(\frac{1}{n_g}\sum\nolimits_{s_g}\theta_g^{(a)}\right) = \theta_g^{(a)} \end{aligned}$$

If the estimated response probabilities, given by (16), are used as if they are the true ones, this would fail to take into account the additional uncertainty from using only estimates of the true response probabilities. Instead,

the variability of the estimated response probabilities should be incorporated into the simulation. If observations could be drawn from the distribution of $\hat{\theta}_g^{(a)}$ to use as input to the simulation, the variability of the output would reflect the variability of the estimated response probabilities. Now, the distribution of $\hat{\theta}_g^{(a)}$ is unknown but, conditional on s, we can estimate it as follows.

Define the mutually exclusive and exhaustive (on s_g) sets $r_g^{(1)}, r_g^{(2)} - r_g^{(1)}, \ldots, r_g^{(A)} - r_g^{(A-1)}, s_g - r_g^{(A)}$ for $g = 1, \ldots, G$. Conditional on s, the sets are of sizes $m_g^{(1)}, m_g^{(2)} - m_g^{(1)}, \ldots, m_g^{(A)} - m_g^{(A-1)}, n_g - m_g^{(A)}$. With n_g fixed, the vector of subset sizes $\mathbf{m}_g^* = (m_g^{(1)}, m_g^{(2)} - m_g^{(1)}, \ldots, n_g - m_g^{(A)})$, follows a multinomial distribution with parameters

$$(n_g, \boldsymbol{\theta}_g^*) = (n_g, \theta_g^{(1)}, \theta_g^{(2)} - \theta_g^{(1)}, \ldots, \theta_g^{(A)} - \theta_g^{(A-1)}, 1 - \theta_g^{(A)}) \tag{17}$$

where $\theta_g^{(a)} - \theta_g^{(a-1)}$ is the probability that an element in group g belongs to $r_g^{(a)} - r_g^{(a-1)}$. Since the vector of probabilities $\boldsymbol{\theta}_g^*$ is unknown, it is estimated with $\hat{\boldsymbol{\theta}}_g^* = (\hat{\theta}_g^{(1)}, \hat{\theta}_g^{(2)} - \hat{\theta}_g^{(1)}, \ldots, 1 - \hat{\theta}_g^{(A)})$ where $\hat{\theta}_g^{(a)}$ is given by (16). To generate observations from the estimated distribution of $\hat{\boldsymbol{\theta}}_g$, observations are first generated on $\mathbf{m}_g^*$ from the multinomial distribution with parameters $(n_g, \hat{\boldsymbol{\theta}}_g^*)$. The observations $\mathbf{m}_g^*$ are then transformed back into observations on $\hat{\boldsymbol{\theta}}_g$.

The simulation from the previous section is extended to incorporate the effect of using response probabilities generated from the estimated distribution. However, since the estimated distribution of $\hat{\theta}_g^{(a)}$ is conditional on s, so are the results from the simulation and the subsequent analysis. The idea is very similar to that of the parametric bootstrap, hence we rely on basic bootstrap results of asymptotic validity of the method. The parametric bootstrap is used when we know that the data are from a particular parametric class of distributions. In our case $\mathbf{m}_g^*$ follow the multinomial distribution. The parameters of the known distribution are replaced by maximum likelihood estimates and resampling is made from the resulting distribution. For details on the parametric bootstrap, see for example Efron and Tibshirani (1993).

Draw J repeated observations (sequences of response probabilities for $a = 1, \ldots, A$) from the estimated distribution of $\hat{\theta}_g^{(a)}$. For each generated sequence, either the CE-measure is calculated directly (if possible) or estimated in a simulation of size M as suggested in the previous case. That is, if explicit expressions can be derived for the CE measure, the variance, the

squared bias and expected cost, the iterations over i in the simulation below are replaced by a simple calculation of these quantities. If these quantities must be estimated, the total simulation will be of size $J \times M$. The steps of such a simulation are summarized in the following.

For $j = 1, \ldots, J$:

i. Generate response probabilities from the estimated conditional distribution of $\hat{\theta}_g^{(a)}$, for all G groups and all time points $a = 1, \ldots, A$.

   For $i = 1, \ldots, M$:

   a. Draw a sample s_i using the same sampling design as in the original sample.

   b. Generate response sets $r_i^{(1)}, \ldots, r_i^{(A)}$ using the response probabilities generated in step i.

   c. Calculate $\hat{t}_{yci}^{(1)}, \ldots, \hat{t}_{yci}^{(A)}$ and the realized costs $C_{Ti}^{(1)}, \ldots, C_{Ti}^{(A)}$ using the known form of the cost model.

ii. Calculate estimates of the variance, the squared bias and expected cost using formulas (6), (7) and (9), for all time points that are to be evaluated. Use these to calculate CE estimates for all the chosen pairs of time points.

The J iterations result in J estimates of the true CE, the variance and the bias. The distributions of the simulated values are estimates of the true distributions, conditional on s.

Conditional on a given generated sequence of response probabilities, the only variability in the estimators comes from using a limited number of iterations M. To determine M, the number of iterations required to estimate the variance, the bias and the total cost for each generated sequence of response probabilities with sufficient precision, the same reasoning as in the case *RD known, $\theta_{k|s}^{(a)}$ independent of s and are known* is used, i.e. M is determined by (15). As before, we can disregard this source of uncertainty when analyzing the simulation results since arbitrary precision can be achieved by increasing M.

The decision rules in steps 1-3 of the proposed evaluation approach should now be adjusted. Instead of comparing a single value of the cost efficiency, the bias and the variance with prespecified limits, we produce *uncertainty*

intervals[1] for these quantities, to be compared to the same limits. The output from the simulation is J estimates of the cost efficiency, the variance and the bias. For large J, uncertainty intervals can be calculated using the empirical distribution of the estimators. The basic idea is borrowed from the bootstrap. There are several different methods to produce bootstrap confidence intervals, and there is debate on which is the best, as there is no uniformly "best" method, it depends on the parameter being estimated. As a simple and straightforward approach, the limits of a $1-2\alpha$ uncertainty interval are defined by the α and $1-\alpha$ percentiles of the empirical cumulative distributions of $\widehat{CE}$, $\hat{B}$ and $\hat{V}$, respectively. The interval may be written as

$$\left[\hat{\Theta}_L, \hat{\Theta}_U\right] = \left[\hat{\Theta}_\alpha, \hat{\Theta}_{1-\alpha}\right] \tag{18}$$

where $\hat{\Theta}$ denotes either of $\widehat{CE}$, $\hat{B}$ and $\hat{V}$ and Θ_α is the $100 \cdot \alpha$th percentile of the distribution of the J simulation outcomes on $\hat{\Theta}$.

This corresponds to the method of percentile intervals in the bootstrap. It is preferred here since it avoids any normality assumptions which, for small response sets, may be grossly in error. For details on the bootstrap, see for example Efron and Tibshirani (1993) or Davison and Hinkley (1997).

For reliable and stable interval limits, J needs to be "sufficiently" large. Little is offered by the theory for the bootstrap. Anything from 50 replications (for estimation of means) up to 1000 (for estimation of percentiles) is recommended, depending on the type of parameter. In our application, it is difficult to give exact recommendations, so the choice of J must largely be determined by a compromise between the available computer power and the need for exact interval limits. Some guidance might be given by running the simulation a limited number of times to see if the results seem to point in any direction. If that is the case, we may achieve reasonably accurate results with a smaller number of iterations. If J is too small to provide probabilistic information, histograms of the simulation values may still provide an impression of the shape of the distributions the values are (indirectly) sampled from. As with real data, the simulated data may deserve a closer look, regardless of the size of J, as it may reveal unexpected patterns in the distributions.

Note that the proposed approach only takes into account the variability in the estimated response probabilities, given the realized sample, i.e.

[1] We use the term uncertainty interval rather than confidence interval, since what we propose is not a confidence interval in the strict, traditional, sense.

variability in $\hat{\theta}_g^{(a)}$ over the sampling design is ignored.

Since sampling randomly from the distribution of $\hat{\boldsymbol{\theta}}_g$ may cause the total number of iterations required ($J \times M$) to be prohibitively large, we may seek an alternative, more efficient way to sample from this distribution. In our specific application, we must generate observations independently for G groups, on A random variables that are dependent. Borrowing the idea of *Latin Hypercube sampling*, LHS, originally proposed by McKay, Beckman, and Conover (1979) in the context of computer simulations, this may be handled as follows.

i. For $a = 1$: The sample space $0 \leq m_g^{(1)} \leq n_g$ is partitioned into J_1 classes of (approximately) equal probability. Let the number of outcomes in class j_1 be n_{gj_1}. For every class j_1, draw one observation randomly with probability $1/n_{gj_1}$.

ii. For $a = 2$: For class j_1, the sample space for $m_g^{(2)} - m_g^{(1)}$ is $0 \leq m_g^{(2)} - m_g^{(1)} \leq n_g - m_{gj_1}^{(1)}$, where $m_{gj_1}^{(1)}$ is the generated number in class j_1 for group g and $a = 1$. Each sample space (one for each of the J_1 classes) is partitioned into J_2 classes of (approximately) equal probability. One observation is drawn randomly in each class j_2 with probability $1/n_{gj_2}$, where n_{gj_2} is the number of outcomes in class j_2.

iii. Repeat step ii for $a = 3, \ldots, A$.

Steps i–iii are repeated for all G groups. This produces sequences of response set sizes for the response sets $r_g^{(1)}, r_g^{(2)} - r_g^{(1)}, \ldots, s_g - r_g^{(A)}$, which are transformed into sequences of response probabilities $\hat{\theta}_g^{(1)}, \ldots, \hat{\theta}_g^{(A)}$. The generation of the sequences of response probabilities is done independently between groups. A sequence for group 1 is then combined at random without replacement with a sequence for group 2, and so on until sequences for all G groups are combined. Each of the $J_1 \times J_2 \times \cdots \times J_A = J$ generated sequences of response probabilities form an iteration in the Monte Carlo simulation. The resulting J estimates of the cost efficiency, the bias and the variance are no longer independent. Nonetheless, the distribution of values reflects the uncertainty in estimators due to uncertainty about the true response probabilities.

Estimators of means of functions of the input variables are shown to be unbiased under LHS. Although this has not been shown for our specific situation, we believe that it may work well for our purposes. Note, however, that

this approach can become computationally heavy, since we must calculate the probabilities of every possible outcome for each a, for each outcome of the preceding time point, $a-1$. The properties of the approach, in terms of unbiasedness of estimators and computational burden, and also on the choice of the number of classes for each variable (time point), must be further investigated before any recommendation on its use can be given.

3.3.3 *RD* known, $\theta_{k|s}^{(a)}$ dependent of s and are unknown

The difference between the case in the previous section and the one in the present section is that now the response probabilities depend on s. Direct calculation of the CE measure is no longer possible since the variance and expected values taken over all possible samples cannot be evaluated. However, under certain assumptions about the nature of the dependence of the response probabilities on s, we can still perform a simulation to estimate $V(\hat{t}_{yc}^{(a)})$, $B^2(\hat{t}_{yc}^{(a)})$ and $E(C_T^{(a)})$, but appropriate adjustments must be made to the estimation of the response probabilities. To capture the uncertainty in the CE measure resulting from using estimated response probabilities, a simulation similar to the one suggested in the previous case may be used. However, the simulation must be altered to take into account that the response probabilities now depend on s.

How to estimate the response probabilities depends on the (known) form of the response distribution. As an example, assume that the sequences of response distributions is such that a given sample s may be partitioned into response groups s_g, $g = 1, \ldots, G$ and that elements in the same group have the same response probability. The partitioning is allowed to vary between different samples, but for a given sample the partitioning is always the same. We have

$$\Pr(k \in r^{(a)}|s) = \theta_{k|s}^{(a)} = \theta_{g|s}^{(a)} \text{ at time } a \text{ for } k \in s_g \tag{19}$$

and

$$\Pr(k\&l \in r^{(a)}|s) = \theta_{k|s}^{(a)}\theta_{l|s}^{(a)} \text{ at time } a \text{ for } k \neq l \tag{20}$$

Assume also that, for a given sample s, the partitioning into groups s_g is known, but that the levels $\theta_{g|s}^{(a)}$ must be estimated for every a and g. The problem is that the data we have is from s, the realized sample, and the response probabilities need to be estimated for samples s_i, $i = 1, \ldots, M$, drawn in the simulation. Let the response groups in s_i be s_{ig_i}, $g_i = 1, \ldots, G_i$, with true but unknown response probabilities $\theta_{g_i|s_i}^{(a)}$ at time a for $k \in s_{ig_i}$.

The same partitioning can be applied to s, let these groups be s_{g_i} of size n_{g_i}. To estimate $\theta^{(a)}_{g_i|s_i}$, we propose to use

$$\hat{\theta}^{(a)}_{g_i|s_i} = \frac{1}{n_{g_i}} \sum\nolimits_{s_{g_i}} R^{(a)}_{k|s} = \frac{m^{(a)}_{g_i}}{n_{g_i}} \tag{21}$$

i.e. the response rates in s, in the groups g_i defined in s_i. The properties of this estimator depend on both samples s and s_i, generated by the design $p(\cdot)$. Since $E_{RD}(R^{(a)}_{k|s}) = \theta^{(a)}_{g|s}$, the estimator will be biased if $\theta^{(a)}_{g|s} \neq \theta^{(a)}_{g_i|s_i}$. If the dependence on s is not too strong, implying that the groups g and g_i and the levels $\theta^{(a)}_{g|s}$ and $\theta^{(a)}_{g_i|s_i}$ are not too different, the estimator (21) should work well enough for our purposes.

Remark 4 *To be able to use* (21) *we need to know, for every i, to which group g_i element $k \in s$ belongs. In the following, we assume this is the case.*

Using the estimated response probabilities, a simulation similar to the one in the previous case, *RD known, $\theta^{(a)}_{k|s}$ independent of s but are unknown*, can be performed. Due to the dependence on s, we must start by drawing a sample s_i so that we can calculate the estimated conditional response probabilities, given by (21). For each sample s_i, J observations from the estimated conditional distribution of $\hat{\theta}^{(a)}_{g_i|s_i}$ are generated to go into the simulation. If we are willing to assume that the dependence of $\theta^{(a)}_{k|s}$ on s is not too strong, the same method as in the previous case can be used to generate observations on $\hat{\theta}^{(a)}_{g_i|s_i}$, $a = 1, \ldots, A$.

Remark 5 *Since s is partitioned according to the grouping defined for s_i, there is a risk of empty groups, i.e. that $n_{g_i} = 0$ for one or more groups. If that happens, the estimates $\hat{\theta}^{(a)}_{g_i|s_i}$, $a = 1, \ldots, A$ cannot be calculated for those groups and the estimator $\hat{t}^{(a)}_{yc}$ will be biased for t_{yU}. In the proposed procedure, assume that the probability of the event*

$$(n_{g_i} = 0 \text{ for some } g_i = 1, \ldots, G_i \text{ given } s) \tag{22}$$

is negligible. This assumption is likely to hold when the groups g and g_i are not too different.

Under the assumption that the dependence on s is not too strong, the simulation approach from the previous case (*RD known, $\theta^{(a)}_{k|s}$ independent*

of s but are unknown) can be used. The subsequent analysis will also be the same, i.e. uncertainty intervals for CE, $V(\hat{t}_{yc}^{(a)})$ and $B(\hat{t}_{yc}^{(a)})$ are produced using (18).

To determine M, the same reasoning as in the previous cases is used. This gives a required precision for each j, i.e. conditional on the estimated response probabilities.

3.4 The less than ideal situations

A more plausible situation than that of the previous section is that the response distribution is unknown. It is also more common that the study variable values are known only for elements in the ultimate response set in the current survey, $r^{(A)}$. Our possibilities to evaluate the reduction efforts in these cases are even more limited and we must largely rely on available auxiliary information and/or unverifiable assumptions for the evaluation. Still, an analysis can be performed in much the same way as in section 3.3, but now under various reasonable assumptions about unknown factors going into the simulation.

3.4.1 y_k known for $k \in U$, RD unknown

In this case, since we have access to the study variable values y_k for all $k \in U$, these may be used in a simulation. However, since the true response probabilities are unknown, or even the form of the true response distribution, the simulation must be based on some assumed response distribution. The main source of uncertainty now comes from not knowing the true response distribution, not from estimating the parameters of this distribution. An evaluation can be done as in the previous cases, but the results of the evaluation (steps 1-3) now necessarily depend on the assumed response distribution. Since we want to know how robust our conclusions are, the simulation must be performed under a variety of different response distributions.

The different assumed (sequences of) response distributions should be plausible, but should also to some extent represent variation from "best case" to "worst case" scenarios. It is up to the statistician's experience, judgment and knowledge of the specific survey to formulate a set of such response distributions. In this process, all available information must be considered for use. For instance, if the auxiliary vector $\mathbf{z}$ is available for $k \in U$ (case 1c in table 1), the response probabilities can be defined as a function of

variables in this vector. Of course, since y is available for all $k \in U$, the response probabilities may alternatively be defined as a function of y, a case with potential for large bias in estimators of t_{yU}.

More or less complex models may be assumed for the true response distribution. However, the auxiliary information used to specify the model is most likely more important than the actual parametric form of the model, given the information used. Also, simple models are preferred over (unnecessarily) complex models. In the following, make the assumption that the response distribution is the same as in the example in the case *RD known, $\theta_{k|s}^{(a)}$ independent of s but are unknown*, i.e. that there are response groups defined in U with response probabilities $\theta_g^{(a)}$ for $k \in U_g$ at time a. The partitioning into groups then defines the properties of the estimator, given that the assumed response distribution coincides with the true one. Other reasonable parametric forms of the response distribution are logistic regression models, or other models suitable for binary data. The use of logistic regression is discussed in for example Alho (1990) and Laaksonen and Chambers (2006) and an early application is found in Ekholm and Laaksonen (1991).

A "best case" scenario is easily found by letting the model for the response distribution be close to, or even coincide with, the model used in the current point estimator. In that case, the bias will be small or negligible. (If the calibration estimator is used, the response distribution is not explicitly modeled, but the nonresponse bias is determined by the ability of the auxiliary vector $\mathbf{x}$ to predict the response probabilities and/or study variable values.) As a "worst case" scenario, the response probabilities may be specified as a function of y. Unless the auxiliary vector $\mathbf{x}$ has very strong predictive power, the point estimator can be severely biased. Intermediate cases include using different subsets of $\mathbf{z}$ to specify the assumed response distribution. The response probabilities are in any case estimated using (16), where g now represents the groups defined in the assumed true distribution. Note that the purpose in this case is not to find the "true" response distribution that we can use to evaluate the cost efficiency, but to make a variety of plausible assumptions about the unknown response distribution.

For each assumed sequence of response distributions, estimate the parameters of the distribution as if the sequence of *RD*'s is the true one and perform a simulation and subsequent analysis just as in section 3.3.1. The main uncertainty most likely comes from not knowing the true sequence of response distributions and not from not knowing the parameters of that distribution, should it be true. Thus, when comparing the conclusions under

the different assumed response distributions, a precision of a few percent, given the assumed RD, is not likely to be of any practical significance compared to, say, 0.5 percent. For all practical purposes, the precision obtained in the estimators from using $J \times M$ iterations as suggested in section 3.3.1, is likely more than enough. Since one $J \times M$ simulation must be performed for every assumed RD, the computing cost will increase by a factor equal to the number of different assumed RD's compared to the case when the true RD is known, and may be prohibitively large. To do the evaluation under the different alternatives, it should suffice to accept lower precision in estimators under each assumed sequence of response distributions than in the previous cases. If computing cost is an issue, an extreme option is to disregard the uncertainty in point estimates of the parameters of the response distribution and use the point estimates to do a simulation with only M iterations. The result from the simulation is then not uncertainty intervals for the cost efficiency measure, the bias and the variance, but point estimates of them.

The analysis of the simulation results is done separately for each assumed RD in the same way as in section 3.3. Evaluating the cost efficiency of the current data collection setup under the various assumed sequences of response distributions will provide a span of possible conclusions. The variation between the conclusions under the different response distributions will also show how sensitive the conclusions are to the underlying assumptions.

3.4.2 y_k known for $k \in r^{(A)}$, RD unknown, $\mathbf{z}_k$ known for $k \in U$

The situation where y_k is known only for $k \in r^{(A)}$ and the true RD is unknown is even less favorable than the previous one. In this case the study variable values cannot be used directly in the type of simulations suggested so far. However, it is possible to base the simulation on proxy values $\hat{y}$, estimated using the auxiliary vector $\mathbf{z}$ known for $k \in U$. Alternatively, the cost efficiency of efforts can be evaluated for an estimator of a parameter for one variable z, believed to be highly correlated with the study variable. If the patterns over the data collection period are similar for $\hat{y}$ and y or for z and y, an informed guess can be made about the cost efficiency for y based on the conclusions for $\hat{y}$ or z. When using z, the evaluation conditions are the same as in the previous case *y_k known for $k \in U$, RD unknown*, so we focus here on the use of a linear combination of the variables in $\mathbf{z}$ as proxy for y.

Assuming that the relationship between y and $\mathbf{z}$ in the population can be described by a linear regression $y_k = \boldsymbol{\gamma}'\mathbf{z}_k$, where $\boldsymbol{\gamma}$ is an unknown column

vector of regression coefficients, we can use theory from Särndal, Swensson, and Wretman (1992) to find a suitable estimator. The estimation of $\boldsymbol{\gamma}$ is discussed at the end of this section.

The true response distribution is unknown, but the same reasoning as in the previous case can be used here. Some plausible response distributions, spanning from "worst case" to "best case" scenarios, are formulated, and for each one of those a simulation and subsequent analysis is performed.

So, in this case, a separate simulation can be performed for each assumed sequence of response distributions just as in the case y_k *known for* $k \in U$, *RD unknown*, but now $\hat{y}$ is used instead of y.

To calculate the proxy values $\hat{y}_k$, the vector of regression coefficients, $\boldsymbol{\gamma}$ needs to be estimated. We do not think of $\boldsymbol{\gamma}$ as estimates of regression coefficients in a super population model, but rather as a characteristic describing the finite population point scatter $(y_k, z_{1k}, \ldots, z_{qk}) : k = 1, \ldots, N$. The parameter of interest is $\boldsymbol{\gamma} = \left(\sum_U \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_U \mathbf{z}_k y_k$, In the case of full response, $\boldsymbol{\gamma}$ could be estimated by

$$\hat{\boldsymbol{\gamma}}_s = \left(\sum_s \frac{\mathbf{z}_k \mathbf{z}_k'}{\pi_k}\right)^{-1} \sum_s \frac{\mathbf{z}_k y_k}{\pi_k} \tag{23}$$

Since y_k is only available for $k \in r^{(A)}$, we could try to estimate $\boldsymbol{\gamma}$ by

$$\hat{\boldsymbol{\gamma}}_r = \left(\sum_{r^{(A)}} \frac{\mathbf{z}_k \mathbf{z}_k'}{\pi_k}\right)^{-1} \sum_{r^{(A)}} \frac{\mathbf{z}_k y_k}{\pi_k} \tag{24}$$

If the response probabilities are defined for $k \in U$, i.e. are independent of s, the approximate expected value of $\hat{\boldsymbol{\gamma}}_r$ is

$$\begin{aligned} E\left(\hat{\boldsymbol{\gamma}}_r\right) &= E_p\left(\left(\sum_s \frac{\theta_{k|s}^{(A)} \mathbf{z}_k \mathbf{z}_k'}{\pi_k}\right)^{-1} \sum_s \frac{\theta_{k|s}^{(A)} \mathbf{z}_k y_k}{\pi_k} \middle| s\right) \\ &\approx \left(\sum_U \theta_{k|s}^{(A)} \mathbf{z}_k \mathbf{z}_k'\right)^{-1} \sum_U \theta_{k|s}^{(A)} \mathbf{z}_k y_k \end{aligned} \tag{25}$$

which, in general, is not equal to $\boldsymbol{\gamma}$. However, since the estimator is a ratio where the factors $\theta_{k|s}^{(A)}$ enter the formula in both the numerator and the denominator, we will expect that $E(\hat{\boldsymbol{\gamma}}_r)$ does not deviate too much from $\boldsymbol{\gamma}$ in most cases.

An alternative estimator of $\boldsymbol{\gamma}$ is based on the assumed response distribution. We get

$$\hat{\boldsymbol{\gamma}}_{r,alt} = \left(\sum_{r^{(A)}} \frac{\mathbf{z}_k \mathbf{z}_k'}{\hat{\theta}_k^{(A)} \pi_k} \right)^{-1} \sum_{r^{(A)}} \frac{\mathbf{z}_k y_k}{\hat{\theta}_k^{(A)} \pi_k} \tag{26}$$

where $\hat{\theta}_k^{(A)}$ is the estimated response probability for unit k under the assumed response distribution used in the simulation. This alternative gives different estimates of $\boldsymbol{\gamma}$ for each assumed response distribution. The approximate expected value of $\hat{\boldsymbol{\gamma}}_{r,alt}$ depends on how $\hat{\theta}_k^{(A)}$ are formed. If $E(\hat{\theta}_k^{(A)}) \approx \theta_k^{(A)}$, then $\hat{\boldsymbol{\gamma}}_{r,alt}$ is approximately unbiased. That case is very unlikely, and since we assume response distributions ranging from "best case" to "worst case" scenarios, the estimates of $\boldsymbol{\gamma}$ and consequently the proxy values $\hat{y}$ under different response distributions may differ considerably. In the analysis of the simulation results, it can then be difficult to separate the effects of the different assumed response distributions from the effect of using different proxy values $\hat{y}$. Thus, we recommend using $\hat{\boldsymbol{\gamma}}_r$ in each simulation, i.e. irrespective of the assumed response distribution. An additional argument in favor of $\hat{\boldsymbol{\gamma}}_r$ can be found in Särndal and Lundström (2005), chapter 9, in their discussion of near-unbiasedness of the calibration estimator. They state that if the study variable is perfectly linearly correlated with the auxiliary vector, then the bias of the calibration estimator is zero. Now, the calibration estimator, and its nearbias (approximate bias), can be expressed in terms of regression coefficients. It is stated that the nearbias of the calibration estimator is, in our notation, $(\sum_U \mathbf{z}_k)' (E(\hat{\boldsymbol{\gamma}}_r) - \boldsymbol{\gamma})$. [2] Zero bias then implies that $E(\hat{\boldsymbol{\gamma}}_r) \approx \boldsymbol{\gamma}$. The relationship between y and $\mathbf{z}$ will not be perfectly linear in practice, but if the predictive ability of $\mathbf{z}$ is strong, the estimator $\hat{\boldsymbol{\gamma}}_r$ should work well.

Remark 6 *In the cases where y_k is known for all $k \in s$ and RD is unknown, the estimator $\hat{\boldsymbol{\gamma}}_s$ can be used. Since it is (approximately) unbiased for $\boldsymbol{\gamma}$, that situation is slightly more favorable than when y_k is known only for $k \in r^{(A)}$.*

The values $\hat{y}_k = \hat{\boldsymbol{\gamma}}_r \mathbf{z}_k$ will tend to have less variability in the population than the set of values y_k, since all values $\hat{y}_k$ lie "on the regression surface". To introduce more natural variability in the data, we recommend adding a randomly selected residual. The residuals are selected among the set of observed values $\{e_k : k \in r^{(A)}\}$ where $e_k = y_k - \hat{\boldsymbol{\gamma}}_r \mathbf{z}_k$. This is in line with e.g.

[2] Note that the expression for the nearbias is based on the assumption that the response probabilities are independent of s and defined for all $k \in U$.

Särndal and Lundström (2005), where the approach is suggested for variance estimation purposes under regression imputation.

Let the residual selected (randomly) from $\{e_k : k \in r^{(A)}\}$ for element k be e_k^*. Using $\hat{y}_k = \hat{\boldsymbol{\gamma}}_r \mathbf{z}_k + e_k^*$ for all $k \in U$, a simulation study and subsequent analysis is performed as if the proxy values was the study variable, i.e. as when y_k is known for $k \in U$, for each of the assumed sequences of response distributions. However, since the inference now applies to $\hat{y}$ instead of y, the conclusions that can be drawn from the simulations must necessarily be different. They will depend on the population correlation between y and $\hat{y}$. If the correlation is high, the simulations will give at least an indication of the uncertainty in the conclusions (about y). A comparison of the patterns in point estimators and variance estimators from the successive response sets for both $\hat{y}$ and y may also give an indication of to what extent the conclusions about the proxy values apply also to y.

3.4.3 y_k known for $k \in r^{(A)}$, RD unknown, no additional information

This is the least favorable situation since we do not have any information other than what has already been used in our current point estimator. This situation is, unfortunately, perhaps the most common in practice. It is possible to use a simulation as suggested in the previous cases, but we will have no information to base the assumptions on. Of course, assumptions can be made about both the study variable and the response probabilities and simulations performed as suggested in the cases with more available information. This will give a span of possible conclusions, but since we cannot know how close our assumptions are to the truth, such an evaluation may be of very limited use.

In an attempt to follow the three steps of the suggested evaluation approach, one might be tempted to try the following unsatisfactory approach: To estimate the cost efficiency in step 1, use the "standard" variance estimator for $V(\hat{t}_{yc}^{(a)})$, estimate the bias using one of the methods proposed below, and estimate the expected cost by the realized cost in the survey. Insert this, taking the square of the bias estimate as an estimate of squared bias, into the expression for the CE measure. This gives a very crude estimate of the cost efficiency for use in step 1 of the evaluation. The variance and bias estimates are also used in steps 2 and 3 of the evaluation. Unfortunately, it is difficult to estimate the precision in the estimators, which is needed to draw

conclusions based on inference procedures. Also, since the estimators of the bias and the variance are most likely biased, using these point estimates can be grossly misleading.

Instead we will take on a different approach, focusing on estimation of the bias difference $B(\hat{t}_{yc}^{(a)}) - B(\hat{t}_{yc}^{(a-1)})$. It is generally acknowledged that point estimator bias is the main concern in surveys with nonresponse. Many different approaches to estimate the bias have been taken, most based on purely empirical studies. An overview of some common methods is given in Groves (2006). Methods presented therein include:

- Response rate comparisons across subgroups.
- Studying register variables for both respondents and nonrespondents.
- Comparisons with estimates from external sources.
- Nonresponse follow-up studies, e.g. comparisons across "waves".
- Comparisons of different types of estimators and with different auxiliary information.

Unfortunately, most of these methods require that additional auxiliary information is available for the evaluation, so in this particular case they simply cannot be used. However, it is not necessary to estimate the bias itself to be able to estimate the bias difference. Expressions for the bias difference are derived in Tångdahl (2004) for some commonly used estimators. A simple and unbiased estimator of the bias difference is the difference between two point estimators, i.e. $\hat{B}_{dif}^{(a-1,a)} = \hat{t}_{yc}^{(a)} - \hat{t}_{yc}^{(a-1)}$. Study of the changes in the levels of point estimates for $a = 1, \ldots, A$ can also be useful. This provides no information on the potential bias remaining after the current cut-off date for the data collection period, but gives some guidance to how the estimators would be affected if we cut down on the efforts to reduce nonresponse. An example of this is given in Japec (2005). Interpretation of the bias difference is discussed in Tångdahl (2004),Tångdahl (2005) and Japec (2005). To assess not only the effect on estimators, but the cost efficiency of reduction efforts, the bias difference must be related to the increase in costs.

Remark 7 *In the case when y_k is known for all $k \in s$ and RD is unknown, an unbiased estimator of the bias can be formed as $\hat{B}_{unb}^{(a)} = \hat{t}_{yc}^{(a)} - \hat{t}_{ys}$ where $\hat{t}_{ys}$ is the full response estimator corresponding to $\hat{t}_{yc}^{(a)}$. This puts us in a slightly better position than when y_k is known only for $k \in r^{(A)}$.*

All of the methods mentioned here are in general unsatisfactory. At best, they will give some insight into the change in point estimator bias between successive cut-off dates, but offer little information on the cost efficiency, the uncertainty in the evaluation results and the remaining nonresponse bias (if any). Another alternative, in case better support for decisions about the data collection is needed, is to set aside resources (time and money) to do a thorough study that will provide information on nonrespondents. The characteristics of the survey may dictate the possibilities, but alternatives include subsampling of nonrespondents, introduced by Hansen and Hurwitz (1946) and a sequential design proposed by Groves (1989). Successfully executed, subsampling of nonrespondents will give unbiased point and variance estimates, making it possible to estimate the bias of point estimators from successive response sets. The sequential design option involves collection of detailed information on the outcome of every effort to contact elements in the sample. By defining a simple error model and a cost model in terms of type of outcome (of the contact attempt or refusal conversion attempt) it is possible to estimate the change in bias, and to weigh this against the increase in costs. Note that the estimates in that approach are based on elements $k \in r^{(A)}$, which most likely are not representative of elements in the sample.

4 Concluding remarks

The problem of nonresponse is one of survey planning, resource allocation and estimation. It is the statisticians task to use all available information in the best way possible to resolve these issues. Uninformed pursuit of high response rates, whatever the cost, is not likely to mean that the money are well spent, so the allocation of resources in the survey should be based on informed decisions about the cost efficiency of different efforts. As a last resort, if nothing else is possible, we can at least make informed guesses based on a variety of realistic assumptions.

The cost efficiency evaluation approach proposed in Tångdahl (2006) can serve as a tool for the statistician in making such informed decisions. However, except in the most ideal situation, we are faced with several difficulties in performing the cost efficiency evaluation as proposed. The simulation approach suggested here is perhaps only a partial solution, since it, in most cases, must be based on unverifiable model assumptions. But it does provide

a quick and relatively easy way to try out different assumptions, even though it might be computationally intensive, and it also gives an estimate of the sensitivity of our conclusions to the model assumptions. Since the assumptions about the true response distribution must be made with the specific survey in mind, we cannot give absolute recommendations on what response distributions to use in the evaluation.

It should be noted that the typical survey situation is much more complex than this paper suggests. Only one study variable and one parameter is studied, while in practice there are usually many variables and many different types of parameters. Also, the estimation of parameters for domains is usually important, and response patterns may differ considerably between domains. What is cost efficient for one study variable, parameter or domain may not be cost efficient for another, so the results must be weighed against each other, depending on how important that particular parameter, variable or domain is. These issues have not been formalized in the approach, but rather we leave it up to the statistician to use his scrutiny and general knowledge of the survey to make such considerations before deciding on a strategy. Should one decide on a new strategy for nonresponse rate reduction, it should be tested in a formal (embedded) experiment first, to eliminate the risk of negative side effects that were not anticipated.

References

Alho, J. M. (1990). Adjusting for nonresponse bias using logistic regression. *Biometrika 77*, 617–624.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Efron, B. and R. J. Tibshirani (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ekholm, A. and S. Laaksonen (1991). Weighting via response modeling in the Finnish household budget survey. *Journal of Official Statistics 7*, 325–337.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

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

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

Japec, L. (2005). *Quality issues in interview surveys. Some contributions.* Ph. D. thesis, Stockholm University.

Kott, P. S. (1998). Using the delete-a-group jackknife variance estimator in NASS surveys. Technical report RD-98-01, Research Division, National Agricultural Statistics Service.

Kott, P. S. (2001). The delete-a-group jackknife. *Journal of Official Statistics 17*, 521–526.

Laaksonen, S. and R. Chambers (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics 22*, 81–95.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 239–245.

Särndal, C. E. and S. Lundström (2005). *Estimation in surveys with nonresponse*. New York: Wiley.

Särndal, C. E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Tångdahl, S. (2004). Nonresponse bias for some common estimators and its change over time in the data collection process. ESI Working Paper Series 2004:13, Örebro University.

Tångdahl, S. (2005). The variance of some common estimators and its components under nonresponse. ESI Working Paper Series 2005:9, Örebro University.

Tångdahl, S. (2006). On the evaluation of the cost efficiency of nonresponse rate reduction efforts – some general considerations. ESI Working Paper Series 2006:5, Örebro University.

A Explicit expression for the CE measure

When the response probabilities are independent of s (i.e. $\theta_{k|s}^{(a)} = \theta_k^{(a)}$ for every s), the RHG grouping is the same for all possible samples and the probability of empty groups is negligible, an explicit expression for the cost efficiency measure can be derived.

Remark 8 *The expression for the cost efficiency is valid whether the true response distribution is known or not, although its use requires that the true response distribution is known.*

Assume that simple random sampling is used to draw the sample, that the point estimator is

$$\hat{t}_{yc\pi^*}^{(a)} = \sum_{h=1}^{H} \frac{n_h}{m_h^{(a)}} \sum_{r_h^{(a)}} \check{y}_k = \frac{N}{n} \sum_{h=1}^{H} \frac{n_h}{m_h^{(a)}} \sum_{r_h^{(a)}} y_k \tag{A.1}$$

and that the cost model is the one given in Tångdahl (2006, Appendix A),

$$C_T^{(a)} = C_F + C_Q + C_{TR} + \sum_{j \leq a} C^{(j)} + C_P^{(a)} \tag{A.2}$$

where C_F are the fixed costs, $C_Q = nc_Q$, $C_{TR} = nc_{TR}$, $C_j = (n - m^{(j)})c_j$, $C_P^{(a)} = m^{(a)}c_P$ and c_Q, c_{TR}, c_j and c_P are the per element costs of the questionnaire, thank you/reminder card, reminder j and processing, respectively. These are specified in Appendix A in Tångdahl (2006).

The cost efficiency measures (2)–(5) are functions of the squared bias and variance of the point estimator, and the expected cost $E(C_T^{(a)})$, so we need to derive explicit expressions for each of those components.

Starting with the variance, we have

$$V(\hat{t}_{yc\pi^*}^{(a)}) = V_p E_{RD}(\hat{t}_{yc\pi^*}^{(a)}|s) + E_p V_{RD}(\hat{t}_{yc\pi^*}^{(a)}|s) \tag{A.3}$$

There are alternative ways to derive the variance, but in this particular case it is convenient to condition on the response homogeneity group sizes $\mathbf{n} =$

$(n_1, \ldots, n_h, \ldots, n_H)$. For the first term of (A.3) we have

$$V_p E_{RD}(\hat{t}^{(a)}_{yc\pi^*}|s) \approx V_p \left(\sum_{h=1}^{H} \frac{n_h}{\sum_{s_h} \theta_k^{(a)}} \sum_{s_h} \theta_k^{(a)} \check{y}_k \right)$$
$$= \left(\frac{N}{n}\right)^2 V_p \left(\sum_{h=1}^{H} \frac{n_h}{\sum_{s_h} \theta_k^{(a)}} \sum_{s_h} \theta_k^{(a)} y_k \right) \quad \text{(A.4)}$$

but for fixed $\mathbf{n}$, s can be treated as a stratified simple random sample, so

$$V_p E_{RD}(\hat{t}^{(a)}_{yc\pi^*}|s) \approx \left(\frac{N}{n}\right)^2 [V_{\mathbf{n}} E(\cdot|\mathbf{n}) + E_{\mathbf{n}} V(\cdot|\mathbf{n})] = \left(\frac{N}{n}\right)^2 [V_1 + V_2] \quad \text{(A.5)}$$

For notational convenience, let

$$K_h^{(a)} = \frac{\sum_{U_h} \theta_k^{(a)} y_k}{\sum_{U_h} \theta_k^{(a)}}$$

Then, using properties of the multivariate hypergeometric distribution, we get

$$\begin{aligned}
V_1 \approx & V_{\mathbf{n}} \left[\sum_{h=1}^{H} n_h K_h^{(a)} \right] = \sum_{h=1}^{H} (K_h^{(a)})^2 V(n_h) + \sum_{h \neq h'} K_h^{(a)} K_{h'}^{(a)} Cov(n_h, n_{h'}) \\
= & \sum_{h=1}^{H} (K_h^{(a)})^2 \frac{n}{N} \left(\frac{N-n}{N-1} \right) N_h \left(1 - \frac{N_h}{N} \right) \\
& - \sum_{h \neq h'} \sum K_h^{(a)} K_{h'}^{(a)} n \frac{N_h}{N} \frac{N_{h'}}{N} \left(\frac{N-n}{N-1} \right) \\
= & n \left(\frac{N-n}{N-1} \right) \left\{ \sum_{h=1}^{H} \frac{N_h}{N} (K_h^{(a)})^2 \left(1 - \frac{N_h}{N} \right) \right. \\
& \left. - \left[\left(\sum_{h=1}^{H} \frac{N_h}{N} K_h^{(a)} \right)^2 - \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 (K_h^{(a)})^2 \right] \right\} \\
= & n \left(\frac{N-n}{N-1} \right) \left\{ \sum_{h=1}^{H} \frac{N_h}{N} (K_h^{(a)})^2 - \left(\sum_{h=1}^{H} \frac{N_h}{N} K_h^{(a)} \right)^2 \right\} \quad \text{(A.6)}
\end{aligned}$$

Continuing with V_2, we have

$$\begin{aligned}
V_2 &= E_{\mathbf{n}} V\left(\sum_{h=1}^{H} \frac{n_h}{\sum_{s_h}\theta_k^{(a)}} \sum_{s_h}\theta_k^{(a)} y_k \middle| \mathbf{n}\right) = E_{\mathbf{n}}\left[\sum_{h=1}^{H} n_h^2 V\left(\frac{\sum_{s_h}\theta_k^{(a)} y_k}{\sum_{s_h}\theta_k^{(a)}} \middle| \mathbf{n}\right)\right] \\
&\approx E\left[\sum_{h=1}^{H} n_h^2 \frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{N_h - 1}\sum_{U_h}\left(\theta_k^{(a)} y_k - K_h^{(a)}\theta_k^{(a)}\right)^2\right] \\
&= \sum_{h=1}^{H}\left\{E(n_h) - \frac{E(n_h^2)}{N_h}\right\}\frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\frac{1}{N_h - 1}\sum_{U_h}\left(\theta_k^{(a)}\right)^2\left(y_k - K_h^{(a)}\right)^2
\end{aligned} \tag{A.7}$$

From the properties of the multivariate hypergeometric distribution, we know that

$$E(n_h) = n\frac{N_h}{N}$$

and

$$E(n_h^2) = V(n_h) + (E(n_h))^2 = n\frac{N_h}{N}\left(1 - \frac{N_h}{N}\right)\frac{N-n}{N-1} + \left(n\frac{N_h}{N}\right)^2$$

Inserting this into (A.7) and simplifying gives

$$\begin{aligned}
V_2 =& \sum_{h=1}^{H}\left\{n\frac{N_h}{N} - \frac{1}{N_h}\left[n\frac{N_h}{N}\left(1 - \frac{N_h}{N}\right)\frac{N-n}{N-1} + \left(n\frac{N_h}{N}\right)^2\right]\right\} \\
&\frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\frac{1}{N_h - 1}\sum_{U_h}\left(\theta_{k|s}^{(a)}\right)^2\left(y_k - K_h^{(a)}\right)^2 \\
=& \sum_{h=1}^{H}\frac{n}{N}\left\{N_h\left(1 - \frac{n}{N}\right) - \left(1 - \frac{N_h}{N}\right)\frac{N-n}{N-1}\right\} \\
&\frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\frac{1}{N_h - 1}\sum_{U_h}\left(\theta_k^{(a)}\right)^2\left(y_k - K_h^{(a)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{H} \frac{n}{N} \frac{N-n}{N-1} \left\{ \frac{N_h}{N}(N-1) - \left(1 - \frac{N_h}{N}\right) \right\} \\
&\qquad \frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2} \frac{1}{N_h - 1} \sum_{U_h} \left(\theta_k^{(a)}\right)^2 \left(y_k - K_h^{(a)}\right)^2 \\
&= \sum_{h=1}^{H} \frac{n}{N} \frac{N-n}{N-1} \frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2} \sum_{U_h} \left(\theta_k^{(a)}\right)^2 \left(y_k - K_h^{(a)}\right)^2 \qquad \text{(A.8)}
\end{aligned}
$$

For the second term of (A.3), we use the derivation given in Tångdahl (2005, eq. 14, page 12), where an approximate expression is given as an expected value over the sampling design. The general expression is

$$
E_p V_{RD}(\hat{t}_{yc\pi^*}^{(a)} | s) \approx E_p \left(\sum_{h=1}^{H} \left(\frac{n_h}{\hat{t}_{\theta h}^{(a)}} \right)^2 \sum_{s_h} \theta_{k|s}^{(a)} (1 - \theta_{k|s}^{(a)}) \left(\check{y}_k - \frac{\hat{t}_{y\theta h}^{(a)}}{\hat{t}_{\theta h}^{(a)}} \right)^2 \right)
$$

where $\hat{t}_{y\theta h}^{(a)} = \sum_{s_h} \theta_{k|s}^{(a)} \check{y}_k$ and $\hat{t}_{\theta h}^{(a)} = \sum_{s_h} \theta_{k|s}^{(a)}$. Evaluating the expected value under simple random sampling and under the assumption that the response probabilities are independent of s, we get

$$
\begin{aligned}
E_p V_{RD}(\hat{t}_{yc\pi^*}^{(a)} | s) &\approx \sum_{h=1}^{H} \left(\frac{\sum_{U_h} \pi_k}{\sum_{U_h} \pi_k \theta_k^{(a)}} \right)^2 \sum_{U_h} \pi_k \theta_k^{(a)} (1 - \theta_k^{(a)}) \left(\check{y}_k - \frac{\sum_{U_h} \theta_k^{(a)} y_k}{\sum_{U_h} \pi_k \theta_k^{(a)}} \right)^2 \\
&= \frac{N}{n} \sum_{h=1}^{H} \left(\frac{N_h}{\sum_{U_h} \theta_k^{(a)}} \right)^2 \sum_{U_h} \theta_k^{(a)} (1 - \theta_k^{(a)}) \left(y_k - \frac{\sum_{U_h} \theta_k^{(a)} y_k}{\sum_{U_h} \theta_k^{(a)}} \right)^2 \\
&= \frac{N}{n} \sum_{h=1}^{H} \frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2} \sum_{U_h} \theta_k^{(a)} (1 - \theta_k^{(a)}) \left(y_k - K_h^{(a)} \right)^2 \qquad \text{(A.9)}
\end{aligned}
$$

Remark 9 *To arrive at an expression analogous to* (A.6) *and* (A.8), $E_p V_{RD}(\hat{t}_{yc\pi^*}^{(a)} | s)$ *can be derived by conditioning on the response homogeneity group sizes* $\mathbf{n} = (n_1, \ldots, n_h, \ldots, n_H)$. *The resulting expression, slightly different from* (A.9), *is*

$$
E_p V_{RD}(\hat{t}_{yc\pi^*}^{(a)} | s) \approx \frac{N}{n} \sum_{h=1}^{H} \frac{D_h}{\left(\bar{\theta}_h^{(a)}\right)^2} \sum_{U_h} \theta_k^{(a)} (1 - \theta_k^{(a)}) \left(y_k - K_h^{(a)} \right)^2
$$

where $D_h = 1 + \frac{1}{n}\left(\frac{N}{N_h} - 1\right)\frac{N-n}{N-1}$

Inserting (A.6) and (A.8) into (A.5) and summing with (A.9) gives

$$V(\hat{t}^{(a)}_{yc\pi^*}) \approx \left(\frac{N}{n}\right)^2 n\frac{N-n}{N-1}\left[\sum_{h=1}^{H}\frac{N_h}{N}(K_h^{(a)})^2 - \left(\sum_{h=1}^{H}\frac{N_h}{N}K_h^{(a)}\right)^2 + \frac{1}{N}\sum_{h=1}^{H}\frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\sum_{U_h}\left(\theta_k^{(a)}\right)^2\left(y_k - K_h^{(a)}\right)^2\right] + \sum_{h=1}^{H}\frac{1}{\left(\bar{\theta}_h^{(a)}\right)^2}\sum_{U_h}\theta_k^{(a)}\left(1-\theta_k^{(a)}\right)\left(y_k - K_h^{(a)}\right)^2 \tag{A.10}$$

Further, the squared bias of $\hat{t}^{(a)}_{yc\pi^*}$ is

$$B^2(\hat{t}^{(a)}_{yc\pi^*}) = \left[E_p E_{RD}(\hat{t}^{(a)}_{yc\pi^*}) - t_{yU}\right]^2$$

and an approximate expression under simple random sampling is

$$\begin{aligned} B^2(\hat{t}^{(a)}_{yc\pi^*}) &\approx \left[E_p\left(\sum_{h=1}^{H}\frac{n_h}{\sum_{s_h}\theta_k^{(a)}}\sum_{s_h}\theta_k^{(a)}\breve{y}_k\right) - t_{yU}\right]^2 \\ &\approx \left[\sum_{h=1}^{H}\frac{N_h}{\sum_{U_h}\theta_k^{(a)}}\sum_{U_h}\theta_k^{(a)}y_k - t_{yU}\right]^2 \\ &= \left[\sum_{h=1}^{H}N_h K_h^{(a)} - t_{yU}\right]^2 \end{aligned} \tag{A.11}$$

Finally, the expected cost, i.e. the expected value of (A.2), is

$$\begin{aligned} E(C_T^{(a)}) &= E_p E_{RD}\left(C_F + C_Q + C_{TR} + \sum_{j\le a}C_j + C_P^{(a)}\right) \\ &= C_F + C_Q + C_{TR} + E_p E_{RD}\left(\sum_{j\le a}C_j + C_P^{(a)}\right) \end{aligned} \tag{A.12}$$

Since

$$\begin{aligned}\sum_{j\le a} C_j + C_P^{(a)} &= \sum_{j\le a}(n - m^{(j)})c_j + m^{(a)}c_P \\ &= \sum_{j\le a}(n - \textstyle\sum_s R_{k|s}^{(j)})c_j + \sum_s R_{k|s}^{(a)}c_P\end{aligned}$$

and $E_{RD}(R_{k|s}^{(a)}) = \theta_k^{(a)}$, we get

$$\begin{aligned}E(C_T^{(a)}) &= C_F + C_Q + C_{TR} \\ &\quad + E_p\left(\sum_{j\le a}\left(n - \textstyle\sum_s \theta_k^{(j)}\right)c_j + \sum_s \theta_k^{(a)}c_P\right) \\ &= C_F + C_Q + C_{TR} \\ &\quad + \sum_{j\le a}\left(n - \frac{n}{N}\textstyle\sum_U \theta_k^{(j)}\right)c_j + \frac{n}{N}\sum_U \theta_k^{(a)}c_P \qquad \text{(A.13)}\end{aligned}$$

As a final step, insert (A.10), (A.11) and (A.13) into the expression for the cost efficiency, one of equations (2)–(5).