Extending Hansen and Hurwitz's Approach for Nonresponse in Sample Survey

A. Demnati
Independent Research, Ottawa, Canada, Abdellatif_Demnati@msn.com

Abstract

Collecting information from sampled units over the Internet or by mail is much more cost-efficient than conducting interviews. These methods make self-enumeration an attractive data collection method for surveys and censuses. However, self-enumeration data collection can produce low response rates compared to interviews. To increase response rates, nonrespondents are subject to follow-up treatments, which influence the resulting probability of response. Because response occurrence is intrinsically conditional, we primarily record response occurrence in discrete intervals, and we characterize the probability of response by a discrete time hazard. This approach facilitates examining when a response is most likely to occur and how the probability of responding varies. Because response rates are presumed to be low, a widely used approach is to consider a second-phase of data collection, where only sub-sampled nonrespondents are followed-up. However, in practice, data collection from self-enumeration and from follow-ups are done in parallel, which makes sub-sampling from nonrespondents difficult to apply. In this case, excluding late self-enumeration responses – not obtained from the follow-up subsample after follow-up has been started – is recommended in the literature to avoid a nonresponse bias. Finally, we study the estimator of the finite population total that use all observed responses. Simulation results on the performance of the proposed estimators are also presented.

Key Words: Event history analysis, Mixed-mode surveys, Partially classified responses, Two-phase data collection.

1. Introduction

Survey or census studies start with a collection of distinct units of interest known as the finite population. There are multiple random variables attached to each unit, as each unit holds their own individual characteristics and aptitudes. Each particular study targets a small subset of these random variables. Measurements on these variables of interest are intended to be collected during the data collection stage from each selected unit, and involve a questionnaire used to collect the data from the respondents. The methodology behind estimating finite population parameters, based on observation of randomly selected units, is well described by survey literature. See Cochran (1977), or, Särndal *et al*. (1992) as an example. Consider the estimation of $Y=\sum_k y_k$ the total of a characteristic of interest y in a finite population P of size N, which is a recurrent parameter of interest in survey sampling, where $\sum_k$ represents summation over the finite population of units, and $\boldsymbol{y} = (y_1,..., y_N)^T$ is the vector of values of y. Under complete response and for a general sampling design with known positive inclusion probabilities $\pi_k(\wp \mid P)$, π_k for short, a customary design unbiased estimator of the total Y is given by the Horvitz-Thompson (HT) estimator

$$\hat{Y}_{HT} = \sum_k d_k(\wp) y_k \,. \tag{1.1}$$

where $d_k(\wp) = d_k(\wp \mid P) = 1_k(\wp \mid P)/\pi_k(\wp \mid P)$ denote the design weights associated with the random sample $\wp$ selected from the finite population, $1_k(\Omega \mid \Omega^*) = 1(k \in \Omega \mid k \in \Omega^*)$ is the set Ω membership indicator variable for unit k given that unit k belongs to the set Ω^*, $1(\mathit{condition})$ is the truth function, i.e., $1(\mathit{condition}) = 1$ if the *condition* is true, $1(\mathit{condition}) = 0$ if not, $\pi_k(\Omega \mid \Omega^*) = E_\Omega\{1_k(\Omega \mid k \in \Omega^*)\}$ is the set Ω inclusion probability for unit k given $k \in \Omega^*$, and E_Ω denotes expectation with respect to the inclusion mechanism.

Mixing modes of data collection offers the possibility of offsetting the disadvantages of one mode with the advantages of another. For example, recognizing that the Internet, unlike mail, offers the ability to move data capture and edit closer to the respondent, many statistical agencies are now offering electronic questionnaires as a voluntary option to: a) improve both response rates and quality of statistical processes; and, b) reduce survey costs. This potential increase in survey quality – in combination with the fact that collecting information from sampled units over the Internet or by mail is much more cost-effective than conducting interviews – makes mixed-mode self-enumeration an attractive data-collection method for surveys and censuses. Although there are benefits associated with mixed-mode self-enumeration surveys – in particular, Internet-based data collection – as well as an expected wider application of this approach in future, mixed-mode self-enumeration surveys bring particular difficulties to surveys and censuses. Observed values of typical variable of interest y might depend on the variable y_m associated with mode m of data collection, $m = 1,..., M$, where M is the number of modes of data collection under consideration for a given study. In principle, each unit k of the finite population P has all responses, i.e., a response $y_{m;k}$ that would have resulted if it had chosen mode m. Since each unit receives or chooses only one mode, only one response is observed. If the variable of interest is defined uniquely and independently from each mode, then $y_{m;k}$ represents the value the unit k believes is the correct answer to y, resulting from the medium of mode m in which the question is presented to the unit.

One of the main objectives of the mixed-mode of data collection is to influence the unit to get its cooperation, regardless of its preference for data-collection mode. If the mixed-mode of data collection can increase the overall response rates, we will be pleased to quantify and examine the contribution of each mode on the response probability. In reality, self-enumeration can produce low response rates in comparison to interviews. To gain non-respondents' cooperation and therefore maximize survey quality, each non-respondent is assigned to a follow-up strategy, where each strategy consists of a mixed-mode of predefined follow-up treatments (Schouten *et al*. 2013). Different costs are associated with different follow-up treatments. For example, face-to-face follow-up is more expensive than telephone follow-up. Currently, in some business surveys, to reduce the global cost of data collection, follow-up for nonresponse is performed on only a portion of non-respondents. These units are often identified in a deterministic way; for example, based on their expected contribution to the estimate. In addition, since a significant number of units are never followed up for nonresponse, the final response rate can be very low. So given the presence of nonresponse in almost all studies, the HT estimator given by (1.1) is rarely used. Suppose that the response probability, $\xi_k = \xi_{\mathrm{I}_{\max};k} = \pi_k(P^{(r)}_{\mathrm{I}_{\max}} \mid P)$, after $\mathrm{I}_{\max}$ time periods of data collection, is known for every unit in the population, where $\mathrm{I}_{\max}$ denotes the survey limited length of duration of data collection, and $P^{(r)}_{\mathrm{I}_{\max}}$ is the subpopulation of respondents after $\mathrm{I}_{\max}$ time periods of data collection. In this case, a design-response unbiased estimator of the population total Y is given by

$$\breve{Y}_{\xi} = \sum_k d_k(\wp)(r_k / \xi_k) y_k , \tag{1.2}$$

where $r_k = r_{\mathrm{I}_{\max};k} = 1_k(P^{(r)}_{\mathrm{I}_{\max}} \mid P)$ is the response indicator for unit k. Since the probability of response ξ_k is unknown, estimated response probability $\hat{\xi}_k$ is used to get

$$\hat{Y}_{\hat{\xi}} = \sum_k d_k(\wp)(r_k / \hat{\xi}_k) y_k . \tag{1.3}$$

As noted by Rosenbaum (1987) and others, estimator $\hat{Y}_{\hat{\xi}}$ using the estimated response probability can be more efficient than estimator $\breve{Y}_{\xi}$ using the true response probability.

The use of response probability to adjust for nonresponse supposes: a) a known relationship between the response mechanism and a set of auxiliary information; b) the availability of the auxiliary information; and, c) the response mechanism is missing at random given the available auxiliary information (Little and Rubin 2002). However, if the response mechanism is related to an unavailable information then the distribution of the variable of interest in the set of sampled respondents, after adjusting for both the sampling selection mechanism and the assumed response mechanism, do not reflect the distribution in the population, and this in turn may result in biased estimates. A population P after I time periods of data collection under self-enumeration (without follow-up) can be seen as composed of two subpopulations: the subpopulation $P^{(r)}_{self|\mathrm{I}}$ of size $N^{(r)}_{self|\mathrm{I}}$ of respondents under self-enumeration, and the subpopulation $P^{(m)}_{self|\mathrm{I}}$ of size $N^{(m)}_{self|\mathrm{I}}$ of nonrespondents under self-enumeration, so that $N = N^{(r)}_{self|\mathrm{I}} + N^{(m)}_{self|\mathrm{I}}$. A census of all N units during I time periods of data collection under self-enumeration yields information on $N^{(r)}_{self|\mathrm{I}}$ units, leaving $N^{(m)}_{self|\mathrm{I}}$ units missing. Hansen and Hurwitz (1946) considered the problem of nonresponse in mail surveys, and proposed to take a subsample $\wp_{f|\mathrm{I}}$ from $\wp^{(m)}_{self|\mathrm{I}}$ to get estimate for $P^{(m)}_{self|\mathrm{I}}$. Only sub-sampled unit are followed-up extensively during $\mathrm{I}_{\max} - \mathrm{I}$ time periods of data collection to get their cooperation and therefore optimize survey quality and cost. Primarily, an inconvenience for the Hansen and Hurwitz (1946) approach is that the set $\wp^{(m)}_{self|\mathrm{I}}$ may vary from one time period to another. Another inconvenience is that there would not be enough time for the extra subsampling and follow-ups after I time periods of data collection period, if both I is large and the statistical figures need to be published (Swensson 2007). In reality, data collections from self-enumeration without follow-up and from follow-up in addition to self-enumeration are somehow done in parallel, which makes sub-sampling from self-enumeration nonrespondents difficult to apply in some applications. Hansen and Hurwitz (1946) recommended excluding late self-enumeration responses not obtained from the follow-up subsample after follow-up has been started to avoid a nonresponse bias. However, in practice, such observations are included in every study. Hansen and Hurwitz (1946) discussed the issue as follow:

"*All schedules arriving before the deadline constitute the mail response and the field follow-up sampling ratio must be applied to all on the mailing list that did not respond before that date. The relatively few schedules arriving after that date, unless they are designed for interview, must be excluded from the sample, in order to avoid a bias of nonresponse of the type which we are trying to eliminate. The cut-off date of course should be held off until the mail response is substantially completed in order to take full advantage of the economies of the mail questionnaire. However, once a sample is designed for field follow-up and the respondent is actually interviewed in the field, the mail questionnaires returned (other than designed for field follow-up) must be discarded*".

Given the above issues, one question should first be of particular interest to statistical agencies: how should both the response probabilities, under mixed-mode of data collection and follow-ups, and the influence of the follow-up treatments on the resulting probability of response be modelled? Other relevant questions include the following: if one factor of the mixed-mode is improved, what will be the effect on the performance of the response mechanism? How can we estimate the response probability due to a particular mixed-mode factor of interest with the presence of the other mixed-mode factor? And finally, how do we extend the Hansen Hurwitz' approach to select the follow-up subsample at any time period of data collection and to make full use of all observed responses?

In an attempt to discuss these issues, our work below is organized as follows: in Section 2, the response probability is first characterized by a discrete-time hazard, followed with the use of regression analysis to investigate the effect of mixed-mode on the response probability, we extend the response model to cover more results such as refusal and ineligibility, and we present results of a small simulation study on the performances of the proposed estimators of the response model parameter and the finite population total under mixed-mode of data collection; and finally, in Section 3, two phases of data collection in the context of nonresponse is covered, and estimators of the response model parameter and the population total that use all observed values are derived. Simulation results on the proposed estimators are also presented.

2. Modeling Response Indicators as Discrete-time Hazard

In this Section, we give a brief account of the Demnati (2015, 2016) method for modeling response indicators as discrete-time survival. We first characterize the response probability by a discrete-time hazard, and investigate the effect of the mixed-mode on the response probability using regression analysis. Afterwards, we extend the response model to cover multiple results such as refusal and ineligible. Finally, we present simulation results on the performance of the estimators of the response model parameter and the population totals associated with mixed-mode surveys.

2.1 Discrete-time Hazard

Consider a homogeneous sample of units, each at risk of experiencing a single target event response. The target event is nonrepeatable. To record response occurrence in discrete intervals, we divided continuous time of the entire data collection period into a sequence of continuous time periods: 1, 2, and so on, and we let $\mathrm{I}_{\min}$ denote the minimum length of data collection period to obtain full responses. Suppose the survey limited length of duration of data collection is made up of $\mathrm{I}_{\max}$ time periods, with $\mathrm{I}_{\max} < \mathrm{I}_{\min}$. Let t represent the discrete random variable that indicates the time period i when the response occurs for a randomly selected unit from the sample. We assume that every unit in the sample lives through each successive discrete time period until the unit responds or is censored by the end of data collection. Then each unit k is observed until some period I_k, with $\mathrm{I}_k \le \mathrm{I}_{\max}$. Observation of the unit could be discontinued for two reasons: 1) the unit response; or, 2) the survey data collection period ends. In the first case, $t_k = \mathrm{I}_k$. In the second case, we only know that $t_k > \mathrm{I}_{\max}$. Units with $t_k > \mathrm{I}_{\max}$ are right-censored – when they respond is unknown. Note that t is defined only when the unit will eventually respond. Since censoring is planned and observation is terminated at the end of data collection, the censuring mechanism is noninformative (Lagakos 1979) in the sense that the act of censoring imparts no information about the response mechanism. Because response occurrence is intrinsically conditional, we characterize t by its conditional probability function – the distribution of the probability that a response will occur in each time period given that it has not already occurred in a previous time period – known as the discrete-time hazard function. Discrete-time hazard $h_{ki}(\boldsymbol{x}_{ki}, \boldsymbol{\beta})$, h_{ki} for short, is defined as the conditional probability that unit k will respond in time period i, given that the unit did not respond prior to i:

$$h_{ki} = \Pr(t_k = i \mid t_k \geq i) ,$$

where $\boldsymbol{x}_{ki}$ refers to both time-invariant and time-varying explanatory variables and $\boldsymbol{\beta}$ is the unknown $q_r \times 1$ vector parameter to be estimated. For unit with $t_k = i$, the probability of obtaining a response at time period i could be expressed in terms of the hazard as

$$\Pr(t_k = i) = h_{ki} \prod\nolimits_{j=1}^{i-1} (1 - h_{kj}) . \tag{2.1}$$

For units with $t_k > i$, the probability of obtaining a response can be expressed as

$$\Pr(t_k > i) = \prod\nolimits_{j=1}^{i} (1 - h_{kj}) . \tag{2.2}$$

We have

$$f(t_k) = \Pr(t_k = \mathrm{I}_k)^{\delta_k} \Pr(t_k > \mathrm{I}_k)^{1-\delta_k} , \tag{2.3}$$

where $\delta_k = 1$ if unit k is uncensored (responds) and $\delta_k = 0$ if unit k is censored. Substituting (2.1) and (2.2) into (2.3), yields

$$f(t_k) = \{h_{k\mathrm{I}_k} / (1 - h_{k\mathrm{I}_k})\}^{\delta_k} \prod\nolimits_{i=1}^{\mathrm{I}_k} (1 - h_{ki}) . \tag{2.4}$$

Expression (2.4) can be rewritten (Allison 1982) as

$$f(t_k) = \prod\nolimits_{i=1}^{\mathrm{I}_k} \{h_{ki} / (1 - h_{ki})\}^{r_{ki}} \prod\nolimits_{i=1}^{\mathrm{I}_k} (1 - h_{ki}) , \tag{2.5}$$

where r_{ki} is a sequence of response indicators defined for each unit k whose values are defined as $r_{ki} = 1$ if the unit does respond in period i and $r_{ki} = 0$ if the unit does not respond in period i. Taking the first derivatives of $\sum_k \log f(t_k)$ yields the estimating equation (EE)

$$\mathbf{S}(\boldsymbol{\beta}) = \sum\nolimits_k \mathbf{s}_k(\boldsymbol{\beta}) = \mathbf{0} , \tag{2.6}$$

where $\mathbf{s}_k(\boldsymbol{\beta}) = \partial \log f(t_k) / \partial \boldsymbol{\beta} = \sum_{i=1}^{\mathrm{I}_k} \dot{\boldsymbol{h}}_{ki} (r_{ki} - h_{ki}) \{h_{ki}(1 - h_{ki})\}^{-1}$, and $\dot{\boldsymbol{h}}_{ki} = \partial h_{ki} / \partial \boldsymbol{\beta}$. For the logistic regression model $\log(h_{ki} / (1 - h_{ki}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}$, $\dot{\boldsymbol{h}}_{ki} = \boldsymbol{x}_{ki} h_{ki} (1 - h_{ki})$, $\mathbf{s}_k(\boldsymbol{\beta}) = \sum_{i=1}^{\mathrm{I}_k} \boldsymbol{x}_{ki} (r_{ki} - h_{ki})$, and the matrix of second partial derivatives is

$$\frac{\partial \mathbf{S}^T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum\nolimits_k \sum\nolimits_{i=1}^{\mathrm{I}_k} \boldsymbol{x}_{ki} h_{ki} (1 - h_{ki}) \boldsymbol{x}_{ki}^T \equiv -\mathbf{J}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) .$$

Adjusting (2.6) for sampling unequal probabilities, we get the weighted EE

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \sum_k d_k(\wp)\mathbf{s}_k(\boldsymbol{\beta}) = \mathbf{0} \; . \qquad (2.7)$$

Starting with a guessed value, $\boldsymbol{\beta}_0$, then for $b = 1,2,...$ updates are made using

$$\boldsymbol{\beta}_b = \boldsymbol{\beta}_{b-1} + \{\hat{\mathbf{J}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}_{b-1})\}^{-1}\hat{\mathbf{S}}(\boldsymbol{\beta}_{b-1}) \, ,$$

where $\hat{\mathbf{J}}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\partial\hat{\mathbf{S}}^T(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. The solution obtained by a Newton-Raphson-type iterative method gives the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

The marginal probability of obtaining a response after $\mathrm{I}_{\max}$ time periods is given by

$$\xi_k = 1 - \prod_{i=1}^{\mathrm{I}_{\max}}(1 - h_{ki}) = \sum_{i=1}^{\mathrm{I}_{\max}} \Pr(t_k = i) \, . \qquad (2.8)$$

It is easily seen from (2.8) that ξ_k increases (or stays the same) as the level of effort increases, where the level of effort is seen in terms of follow-up treatments and data-collection period. This suggests that costs and benefits of increasing the level of effort should be explored given that, in some circumstances, there are a number of follow-up treatments made with a high percentage of cost, expanded to get values from a few non-respondents.

2.2 Influence of Follow-up on Response Probability

We expressed the inverse link function of the hazard rate as a function of explanatory variables $\boldsymbol{x}_{ki}$ and a vector parameter $\boldsymbol{\beta}$ to be estimated. For units under self-enumeration data collection without follow-up, the inverse link form of the hazard-rate is expressed as

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}) \, , \qquad (2.9)$$

for known function $\eta(.)$, where $\boldsymbol{x}_{ki}^{(0)}$ is the vector of explanatory variables for self-enumeration, $\boldsymbol{\beta}^{(0)}$ is the associated unknown vector parameter to be estimated, $\boldsymbol{x}_{ki} = \boldsymbol{x}_{ki}^{(0)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ and $g(.)$ is a link function – although the link function is generally used to transform (or to link) the conditional mean to the linear predictor $\boldsymbol{x}_{ki}^T\boldsymbol{\beta}$. For example, $g(a) = a$ with $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T\boldsymbol{\beta}$ gives a linear regression model and $g(a) = \exp(a)/\{1 + \exp(a)\}$ with $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^T\boldsymbol{\beta}$ gives a logistic regression model for binary responses r_{ki}.

Additional influences on response probability can be investigated by adding further predictors to the initial discrete-time hazard model. For instance, the following model differs from the model in (2.9) by the inclusion of the time-variant follow-up predictor $\gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)}$, the influence of which is captured by the parameter $\boldsymbol{\beta}^{(1)}$:

$$g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}^{(0)}, \boldsymbol{\beta}^{(0)}; \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)}, \boldsymbol{\beta}^{(1)}) \, , \qquad (2.10)$$

where the value of $\gamma_{ki}^{((1)}$ is set to 1 if the first follow-up treatment is started, or set to 0 if this is not the case, with $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T})^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T})^T$. Note that (2.10) can be used to define different slopes and intercepts, in which case the parameter $\boldsymbol{\beta}^{(1)}$ reflects the changes in the intercepts and in the slopes associated with changing from self-enumeration only to self-enumeration followed by the first follow-up treatment. For example, in the specification $\eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta}) = \boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} + \boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)}$, $i = 1,\dots,\mathrm{I}$, with $\boldsymbol{x}_{ki}^{(0)T}\boldsymbol{\beta}^{(0)} = \alpha_{0i}^{(0)} + x_{ki}\alpha_{1i}^{(0)}$ and $\boldsymbol{x}_{ki}^{(1)T}\boldsymbol{\beta}^{(1)} = \gamma_{ki}^{(1)}(\alpha_{0i}^{(1)} + x_{ki}\alpha_{1i}^{(1)})$, the regression parameters $\alpha_{0i}^{(0)}$, $\alpha_{1i}^{(0)}$ and the values x_{ki} represent respectively the intercept, the slope and the predictor associated with self-enumeration data collection in time period i. We have $\boldsymbol{x}_{ki}^{(0)} = (D_{k1}^{(0)},\dots,D_{kI}^{(0)}, x_{k1}^{(0)},\dots,x_{kI}^{(0)})^T$ and $\boldsymbol{\beta}^{(0)} = (\alpha_{01}^{(0)},\dots,\alpha_{0I}^{(0)},\alpha_{11}^{(0)},\dots,\alpha_{1I}^{(0)})^T$, where $D_{ki}^{(0)} = 1$, $x_{ki}^{(0)} = x_{ki}$, $D_{kj}^{(0)} = 0$ and $x_{kj}^{(0)} = 0$ for $j \neq i$. The vector predictor follow-up is given by $\boldsymbol{x}_{ki}^{(1)} = (D_{k1}^{(1)},\dots,D_{kI}^{(1)}, x_{k1}^{(1)},\dots,x_{kI}^{(1)})^T$ and the changes due to the follow-up in the intercepts and slopes are reflected by vector parameter $\boldsymbol{\beta}^{(1)} = (\alpha_{01}^{(1)},\dots,\alpha_{0I}^{(1)},\alpha_{11}^{(1)},\dots,\alpha_{1I}^{(1)})^T$, where $D_{ki}^{(1)} = 1$, $x_{ki}^{(1)} = x_{ki}$, $D_{kj}^{(1)} = 0$ and $x_{kj}^{(1)} = 0$ for $j \neq i$. To increase response rates, non-respondents are subject to intensive multiple follow-ups by telephone or other treatments to encourage them to participate. A follow-up treatment can take the form of mailed reminders, emailed reminders, telephone calls or in-person interviews. The follow-up process through treatments is conducted using data collection calendars with a specific strategy for each sampled unit. In the case of $1+\mathrm{T}$ follow-up treatment, the inverse link form of the hazard-rate can be expressed as $g^{-1}(h_{ki}) = \eta(\boldsymbol{x}_{ki}, \boldsymbol{\beta})$, where $\boldsymbol{x}_{ki} = (\boldsymbol{x}_{ki}^{(0)T}, \gamma_{ki}^{(1)}\boldsymbol{x}_{ki}^{(1)T},\dots,\gamma_{ki}^{(\mathrm{T})}\boldsymbol{x}_{ki}^{(\mathrm{T})T})^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\beta}^{(1)T},\dots,\boldsymbol{\beta}^{(\mathrm{T})T})^T$.

Consider the simple example of $\mathrm{T} = 1$ as it is the case in business surveys, where T_1 consists of intensive follow up and T_0 consists of sending the questionnaire, and suppose for simplicity the case in which the response outcome is instant. After collecting the response from self-enumeration without follow-up respondents, follow-up is performed in a deterministic way – non-respondents with $u_k \geq c_u$ are assigned to treatment T_1, where c_u is a predetermined constant and u is an auxiliary variable with values available for all sampled units. Suppose all units under T_1 responded, while the other units have still not responded. We have $\xi_k = h_{k1} + (1 - h_{k1})1 = 1$ for unit k with $u_k \geq c_u$ and $\xi_k = h_{k1} + (1 - h_{k1})0 = h_{k1}$ for units with $u_k < c_u$. This highlights the significant effect of follow-up on the probabilities of response.

2.3 Modeling Multiple Results as Discrete-time Hazards

To distinguish different results during data collection such as refusal, ineligibility, and mode of data collection, we develop a discrete-time model for multiple kinds of results or events, by extending the Bernoulli model to the multinomial model (see, for example, Prentice *et al.* 1978, Allison 1982, or Lancaster 1990). Assume that there are E specific results or events and the $(E+1)^{th}$ category of no results, where $E \geq 1$. Define a vector of event indicator variables as $r_{ki}^{(e)} = 1$ if outcome e occurs from unit k at time period i, and $r_{ki}^{(e)} = 0$ if not, where $e \in \{1,...,E\}$, $E_r(r_{ki}^{(e)}) = h_{ki}^{(e)}$, $Var_r(r_{ki}^{(e)}) = h_{ki}^{(e)}(1-h_{ki}^{(e)})$, and for $e \neq e'$ $Cov_r(r_{ki}^{(e)}, r_{ki}^{(e')}) = -h_{ki}^{(e)} h_{ki}^{(e')}$. The combined discrete-time hazard is

$$h_{ki} = \sum_{e=1}^{E} h_{ki}^{(e)} = \Pr(t_k = i \mid t_k \geq i) .$$

For units with $t_k = i$, the probability of obtaining result e at time period i could be expressed in terms of the hazard as

$$\Pr(t_k = i, r_{ki}^{(e)} = 1) = h_{ki}^{(e)} \prod_{j=1}^{i-1} (1 - h_{kj}) = h_{ki}^{(e)} \prod_{j=1}^{i-1} h_{kj}^{(E+1)} . \qquad (2.11)$$

For units with $t_k > i$, the probability of obtaining a response can be expressed as

$$\Pr(t_k > i) = \prod_{j=1}^{i} (1 - h_{kj}) = \prod_{j=1}^{i} h_{kj}^{(E+1)} . \qquad (2.12)$$

Substituting (2.11) and (2.12) into (2.3), and using the sequence of response indicators $r_{ki}^{(e)}$ yields

$$f(t_k) = \prod_{i=1}^{\mathrm{I}_k} \prod_{e=1}^{E} \{h_{ki}^{(e)} / h_{ki}^{(E+1)}\}^{r_{ki}^{(e)}} \prod_{i=1}^{\mathrm{I}_k} h_{ki}^{(E+1)}) . \qquad (2.13)$$

which reduces to (2.4) when $E = 1$. The marginal probability of obtaining result e after $\mathrm{I}_{\max}$ time periods is given by

$$\xi_k^{(e)} = \sum_{i=1}^{\mathrm{I}_{\max}} \Pr(t_k = i, r_{ki}^{(e)} = 1) . \qquad (2.14)$$

We consider the $(E+1)^{th}$ category of no results as an omitted or reference category. For the multinomial logistic regression model, logits of the first E events are constructed with the reference category in the denominator

$$\log(h_{ki}^{(e)} / h_{ki}^{(E+1)}) = \boldsymbol{x}_{ki}^T \boldsymbol{\beta}_e , \; e = 1,...,E ,$$

where $\boldsymbol{x}_{ki}$ is the $q^{(1)} \times 1$ vector of explanatory variables, $\boldsymbol{\beta}_e$ is the $q^{(1)} \times 1$ unknown vector of parameter associated with result e, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T,...,\boldsymbol{\beta}_E^T)^T$ is the $q_r = q^{(1)} E \times 1$ unknown vector parameter to be estimated. It follows that the $E+1$ conditional probabilities given the vector of explanatory variables are

$$h_{ki}^{(E+1)} = \{1 + \sum_{e=1}^{E} \exp(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}_e)\}^{-1} ,$$

and for $e = 1,...,E$,

$$h_{ki}^{(e)} = h_{ki}^{(E+1)} \exp(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}_e) .$$

Thus, the log likelihood function for multinomial logistic regression models is:

$$\ell(\boldsymbol{\beta}) = \sum_k \log f(t_k) = \sum_k \sum_{i=1}^{\mathrm{I}_k} \sum_{e=1}^{E} r_{ki}^{(e)} \boldsymbol{x}_{ki}^T \boldsymbol{\beta}_e - \sum_k \sum_{i=1}^{\mathrm{I}_k} \log\{1 + \sum_{e=1}^{E} \exp(\boldsymbol{x}_{ki}^T \boldsymbol{\beta}_e)\} ,$$

where $f(t_k)$ is given by (2.13). Taking the first derivatives, we get

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_e} = \textstyle\sum_k \mathbf{s}_e(\boldsymbol{\beta}) \, , \; e = 1,\dots,E \, , \tag{2.16}$$

where $\mathbf{s}_e(\boldsymbol{\beta}) = \sum_{i=1}^{\mathrm{I}_k} \boldsymbol{x}_{ki}(r_{ki}^{(e)} - h_{ki}^{(e)})$. Taking the second derivatives, we get the matrix of second partial derivatives for the multinomial logistic regression model

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_e \partial \boldsymbol{\beta}_c^T} = \begin{cases} -\sum_k \sum_{i=1}^{\mathrm{I}_k} \boldsymbol{x}_{ki} h_{ki}^{(e)}(1 - h_{ki}^{(e)}) \boldsymbol{x}_{ki}^T & e = c \\ \sum_k \sum_{i=1}^{\mathrm{I}_k} \boldsymbol{x}_{ki} h_{ki}^{(e)} h_{ki}^{(c)} \boldsymbol{x}_{ki}^T & e \neq c. \end{cases}$$

Adjusting (2.16) for sampling unequal probabilities, we get the weighted EE

$$\hat{\mathbf{S}}(\boldsymbol{\beta}) = \textstyle\sum_k d_k(\wp) \mathbf{s}_k(\boldsymbol{\beta}) = \mathbf{0} \, , \tag{2.17}$$

where $\mathbf{s}_k(\boldsymbol{\beta}) = (\mathbf{s}_{1;k}^T(\boldsymbol{\beta}),\dots,\mathbf{s}_{E;k}^T(\boldsymbol{\beta}))^T$.

2.4 Simulation Study

We conducted a small simulation study to illustrate the performances of the estimator of the a) response model parameter; as well as, b) finite population total related to the mixed-mode of data collection. To show the simulation results, we need to present first data generation model. Next we discuss steps required for design pre-specification such as sampling scheme, estimator used for design pre-specification and its associated variance, expected survey global cost, and resources allocation within stages of the survey design. Finally, we present each parameter of interest which leads us into simulation results.

2.4.1 Finite Population Values

We generate values for each unit k of a finite population P of size $N = 5000$ independently from the model

$$\begin{pmatrix} y_k \\ u_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_u \end{pmatrix}, \begin{bmatrix} \sigma_{y;y} & \sigma_{y;u} \\ \sigma_{u;y} & \sigma_{u;u} \end{bmatrix} \right),$$

where y is the variable of interest, u is the auxiliary variable, $\mu_y = \mu_u = 50$, $\sigma_{i;j} = \sigma_i \sigma_j \rho_{i;j}$, $\sigma_y = \sigma_u = 10$, and $\rho_{y;u} = .8$. Then we generated values of the variable $y_{m;k}$ associated with mode $m = M$ ($Mail$) and mode $m = I$ ($Internet$) of data collection from the conditional distributions

$$y_{m;k} \mid y_k \sim N\left(\mu_{m|y}, \sigma_{m|y;m|y}\right) \text{ for } m \in \{M, I\} \, ,$$

with $\mu_{m|y} = \mu_m + \sigma_{m;y} \sigma_{y;y}^{-1}(y_k - \mu_y)$, $\sigma_{m|y;m|y} = \sigma_{m;m} - \sigma_{m;y} \sigma_{y;y}^{-1} \sigma_{y;m}$, $\mu_M = 47$, $\mu_I = 55$, $\sigma_M = \sigma_I = 10$, $\rho_{M;y} = \rho(y_M, y) = .3$ and $\rho_{I;y} = \rho(y_I, y) = .7$.

2.4.2 Sampling Scheme

We use Poisson sampling with selection probability parameterized as $\pi_k = \{lb + ub \times \exp(\boldsymbol{v}_{\wp;k}^T \boldsymbol{\lambda}^{(\wp)})\}/\{1 + \exp(\boldsymbol{v}_{\wp;k}^T \boldsymbol{\lambda}^{(\wp)})\}$, where lb and ub are respectively the lower and upper bounds with $0 \le lb < ub \le 1$. Unlike the customary probability of selection $\pi_{c;k} = np_k$ with p_k a measure of size, this expression for π_k fulfill the two criterions: a) $0 < lb \le \pi_k \le ub \le 1$; and, b) $\pi_k \to ub$ when $n \to N$, where n denotes the expected sample size. We set $(lb, ub) = (0,1)$, $\boldsymbol{v}_{\wp;k} = (1, u_k)^T$, and $\boldsymbol{\lambda}^{(\wp)} = (\lambda_0^{(\wp)}, \lambda_1^{(\wp)})^T$.

2.4.3 Response Model

In this first simulation, it is assumed that in case of nonresponse each sampled unit is subject to the follow-up during data collection. We set $\mathrm{I}_{\max} = 10$ and $\mathrm{I} = 3$, and we generated the follow-up entry time period e_k for unit k from the Discrete Uniform (DU) distribution, $e_k \sim DU[\mathrm{I}+1, \mathrm{I}_{\max} - 1]$. Self-enumeration responses for sampled unit k are generated using $r_{ki} \sim M_{E+1}(1, h_{M;ki}, h_{I;ki}, h_{O;ki})$, with $E = 2$, $h_{O;ki} = 1 - h_{M;ki} - h_{I;ki}$, $x_{ki}^T \boldsymbol{\beta} = \beta_{m;1} + u_k \beta_{m;2} + i\beta_{m;3} + 0\beta_{m;4}$ for $m \in \{M, I\}$ and $i = 1, \ldots, \mathrm{I}_{\max}$. For unit k under follow-up data collection in addition to self-enumeration we used $x_{ki}^T \boldsymbol{\beta} = \beta_{m;1} + u_k \beta_{m;2} + i\beta_{m;3} + (i - e_k + 1)\beta_{m;4}$ $i = e_k, \ldots, \mathrm{I}_{\max}$. Table 2 displays values of the response model parameter $\boldsymbol{\beta}$.

2.4.4 Estimator for Design Pre-specification and Associated Variance

For design pre-specification, we used $\breve{U}_\xi = \sum_k d_k(\wp)(r_k / \xi_k) u_k$ as estimator the known total $U = \sum_k u_k$. We decomposed the variance of $\breve{U}_\xi$ as

$$Var(\breve{U}_\xi) = E_\wp Var_r(\breve{U}_\xi) + Var_\wp E_r(\breve{U}_\xi) \equiv \mathrm{V}_r + \mathrm{V}_\wp .$$

Under independent response mechanism, the first component $\mathrm{V}_r = E_\wp Var_r(\breve{U}_\xi)$ is given by

$$\mathrm{V}_r = -\sum_k u_k^2 / \pi_k + \sum_k u_k^2 / (\xi_k \pi_k) .$$

Under Poisson sampling, the second component $\mathrm{V}_\wp = Var_\wp E_r(\breve{U}_\xi)$ is given by

$$\mathrm{V}_\wp = -\sum_k u_k^2 + \sum_k u_k^2 / \pi_k .$$

The sum $\mathrm{V}_r + \mathrm{V}_\wp$ constitutes the variance of $\breve{U}_\xi$, which may be written in the form

$$Var(\breve{U}_\xi) = v_0 + \sum_k v_k / \pi_k ,$$

where $v_0 = -\sum_k u_k^2$, and $v_k = u_k^2 / \xi_k$.

2.4.5 Expected Survey Global Cost

We decompose the expected total survey cost as $\overline{C}=\overline{C}^{(\wp)}+\overline{C}^{(f)}+\overline{C}^{(dc)}$. The sampling component $\overline{C}^{(\wp)}$ is given by $C^{(\wp)}=\sum_k \pi_k c_k^{(\wp)}$, where $c_k^{(\wp)}$ denotes the sampling cost for unit k. The follow-up component $\overline{C}^{(f)}$ is given by $\overline{C}^{(f)}=\sum_k \pi_k(1-{}^{(self)}\xi_{e_k-1;k})c_k^{(f)}$, where ${}^{(self)}\xi_{e_k-1;k}=\Pr(t_{self;k}<e_k)$, t_{self} represents the discrete random variable that indicates the time period i when the response occurs under self-enumeration without follow-up of data collection, and $c_k^{(f)}$ denotes the cost associate with follow-up for unit k. Finally, the data collection component $\overline{C}^{(dc)}$ is given by $\overline{C}^{(dc)}=\sum_k \pi_k \sum_{m\in\{M,I\}} \xi_k^{(m)} c_{m;k}^{(dc)}$, where $\xi_k^{(m)}$ is given by (2.14) and $c_{m;k}^{(dc)}$ denotes the data collection cost associated with mode m for unit k. We may write the total survey expected cost as

$$\overline{C}=\sum_k \pi_k c_k ,$$

with
$$c_k=c_k^{(\wp)}+(1-{}^{(self)}\xi_{e_k-1;k})c_k^{(f)}+\sum_{m\in\{M,I\}} \xi_k^{(m)} c_{m;k}^{(dc)} .$$

We set $c_k^{(\wp)}=1$, $c_k^{(f)}=10$, $c_{M;k}^{(dc)}=5$, and $c_{I;k}^{(dc)}=1$.

2.4.6 Resource Allocation within Stages of the Survey Design

To create a design, we determine the sample selection probability parameter $\boldsymbol{\lambda}^{(\wp)}=(\lambda_0^{(\wp)},\lambda_1^{(\wp)})^T$ by minimizing the expected cost subject to constraint on the coefficient of variation, $\sqrt{Var(\breve{U}_\xi)}/\breve{U}_\xi=.05$. Here, the expected sample size is implicitly defined through the vector parameter $\boldsymbol{\lambda}^{(\wp)}$. Table 1 displays the expected coefficient of variation in percentage, the expected sample size, and the expected total cost. Table 1 also displays the expected cost ratios in percentage for sampling, follow-up and data collection. Finally for more information, Table 1 displays the estimates of the sampling model parameter $\boldsymbol{\lambda}^{(\wp)}$.

2.4.7 Parameters of Interest and Simulation Results

We maintained the population values $(y_k,u_k,y_{M;k},y_{I;k},e_k)$ fixed for $k=1,...,N$, and we selected $A=1000$ Poisson samples from the generated population using $\boldsymbol{\lambda}^{(\wp)}=(-3.724,.0243)^T$ obtained from the last two columns of Table 1. Table 2 displays statistics on the realized sample size and number of respondents by mode of data collection. Our first vector parameter of interest is the response model parameter $\boldsymbol{\beta}=(\boldsymbol{\beta}_M^T,\boldsymbol{\beta}_I^T)^T$, with $\boldsymbol{\beta}_m=(\beta_{m;1},\beta_{m;2},\beta_{m;3},\beta_{m;4})^T$ for $m\in\{M,I\}$. Let $\hat{\theta}$ denote an estimator of the parameter of interest θ. We calculated $\hat{\theta}$ from each repetition a ($a=1,...,A$), and its average $\overline{\hat{\theta}}=A^{-1}\sum_{a=1}^{A}\hat{\theta}_a$, where $\hat{\theta}_a$ is the value of $\hat{\theta}$ for the a^{th} sample. The simulated bias and relative bias of $\hat{\theta}$ are calculated as $B(\hat{\theta})=(\overline{\hat{\theta}}-\theta)$, and $RB(\hat{\theta})=B(\hat{\theta})/\theta$. We calculated $\overline{\hat{\theta}}$ and $B(\hat{\theta})$ for response model parameters and those values are reported in Table 3. Table 3 clearly demonstrates that the bias is small for each response model parameter. We also considered the estimation of the finite population total: $\boldsymbol{\theta} = (\sum_k y_k, \sum_k y_{m;k}; m \in \{M, I\})^T$. We used two sets of weights: the first set of weights uses estimated response model parameter, $(d_k(\wp)(r_k / \hat{\xi}_k), d_k(\wp)(r_k^{(m)} / \hat{\xi}_k^{(m)}); m \in \{M, I\})^T$; while the second set of weights uses the true response model parameter $(d_k(\wp)(r_k / \xi_k), d_k(\wp)(r_k^{(m)} / \xi_k^{(m)}); m \in \{M, I\})^T$. The mean square error (MSE) of $\hat{\theta}$ is calculated as $MSE(\hat{\theta}) = A^{-1} \sum_{a=1}^{A} (\hat{\theta}_a - \theta)^2$. We calculated $RB(\hat{\theta})$ and MSE ratios for each estimator $\hat{\theta}$ with $\hat{Y}_{\hat{\xi}} = \sum_k d_k(\wp)(\wp)(r_k / \hat{\xi}_k) y_k$ and those values are reported in Table 4. Table 4 clearly indicates that all relative biases are small. The estimator using an estimated response model parameter is more efficient than an estimator using the true response model parameter. For comparison, Table 4 also provides results for calibrated estimators to the population total $U = \sum_k u_k$ of the auxiliary variable, which indicate that calibration-to-population total is highly efficient. Here the calibration adjustment factors are respectively: $\hat{g}_k = \{\sum_k d_k(\wp)(r_k / \hat{\xi}_k) u_k\}^{-1} \sum_k u_k$, $\hat{g}_{m;k} = \{\sum_k d_k(\wp)(r_k^{(m)} / \hat{\xi}_k^{(m)}) u_k\}^{-1} \sum_k u_k$, $g_k = \{\sum_k d_k(\wp)(r_k / \xi_k) u_k\}^{-1} \sum_k u_k$, and $g_{m;k} = \{\sum_k d_k(\wp)(r_k^{(m)} / \xi_k^{(m)}) u_k\}^{-1} \sum_k u_k$ for $m \in \{M, I\}$.

3. Two-phase Data Collection

In this Section, we first define the ingredients associated with two-phase data collection for follow-up, and then we estimate the conditional probability that a unit belongs to the subpopulation of self-enumeration respondents given that the unit responded under follow-up in addition to self-enumeration. Secondly, we estimate the following domain sizes: the size of unidentified self-enumeration respondents and the size of extra respondents due to the follow-up activities. Next, we derive the proposed estimator of the population total that use all observed data. Finally, we present results of a small simulation study on the performance of the proposed estimator.

3.1 Subsampling for Nonresponse Follow-up

After subsampling from the set $\wp_{self|\mathrm{I}}^{(m)}$ of nonrespondents at the I^{th} time period of data collection ($1 \le \mathrm{I} < \mathrm{I}_{\max}$), we have two subsamples, i.e., the follow-up subsample $\wp_{f|\mathrm{I}}$, and its complement $\wp_{\mathrm{s}|\mathrm{I}}$ with respect to $\wp_{self|\mathrm{I}}^{(m)}$, so that $\wp_{self|\mathrm{I}}^{(m)} = \wp_{f|\mathrm{I}} \cup \wp_{\mathrm{s}|\mathrm{I}}$ with $\wp_{f|\mathrm{I}} \cap \wp_{\mathrm{s}|\mathrm{I}} = \varnothing$. The subsample $\wp_{\mathrm{s}|\mathrm{I}}$ permits estimation of the net probability of response from self-enumeration without follow-up during period $[\mathrm{I}+1, \mathrm{I}_{\max}]$, while the subsample $\wp_{f|\mathrm{I}}$ permits estimation of the crude probability of response using follow-up in the presence of self-enumeration during period $[\mathrm{I}+1,\mathrm{I}_{\max}]$. After completed $\mathrm{I}_{\max}$ time periods of data collection, the set of all respondents is given by $\wp^{(r)} = \wp^{(r)}_{self|\mathrm{I}} \cup \wp^{(r)}_{\mathrm{s}|\mathrm{I}} \cup \wp^{(r)}_{f|\mathrm{I}}$ with $\wp^{(r)}_{self|\mathrm{I}} \cap \wp^{(r)}_{\mathrm{s}|\mathrm{I}} \cap \wp^{(r)}_{f|\mathrm{I}} = \varnothing$, where $\wp^{(r)}_{\mathrm{s}|\mathrm{I}}$ is the set of respondents from $\wp_{\mathrm{s}|\mathrm{I}}$, and $\wp^{(r)}_{f|\mathrm{I}}$ is the set of respondents from $\wp_{f|\mathrm{I}}$. Let's define the subpopulation of self-enumeration respondents membership indicator $l_k^{(Self)}$ for unit k as $l_k^{(Self)} = 1$ if unit k responds under self-enumeration without follow-up, and $l_k^{(Self)} = 0$ if not. When $l_k^{(Self)} = 0$, the unit is either a non-respondent or an extra-respondent due to the follow-up activity. For convenience, we set the follow-up entry time period e_k for unit k under self-enumeration data collection without any follow-up (unit $k \in \wp^{(r)}_{self|\mathrm{I}} \cup \wp_{\mathrm{s}|\mathrm{I}}$) to $e_k = \mathrm{I}_{\max} + 1$, while the follow-up entry time for units in the follow-up subsample (unit $k \in \wp_{f|\mathrm{I}}$) is between $\mathrm{I} \le e_k < \mathrm{I}_{\max}$. The indicator $l_k^{(Self)}$ is equal to 1 for each unit in $\wp^{(r)}_{self|\mathrm{I}} \cup \wp^{(r)}_{\mathrm{s}|\mathrm{I}}$ since $e_k = \mathrm{I}_{\max} + 1$; i.e., the only random process involved is just self-enumeration without follow-up. The indicator $l_k^{(Self)}$ is also equal to 1 for respondent from $\wp^{(r)}_{f|\mathrm{I}}$ with $\mathrm{I}_k < e_k$ (i.e., units that responded before follow-up has been started). However, respondents in the set $\wp^{(r)}_{f|\mathrm{I}}$ with $\mathrm{I}_k \ge e_k$ are arising from a mixture of two distributions: 1) late self-enumeration respondents; or, 2) extra-respondents due to the follow-up activity. If $k \in \wp^{(r)}_{f|\mathrm{I}}$ with $\mathrm{I}_k \ge e_k$, then $l_k^{(Self)}$ is unknown. When $l_k^{(Self)}$ is unknown we replace it by $\hat{\tau}_k = \pi_k(P^{(r)}_{self} \mid \wp^{(r)}_{\mathrm{I}_{\max}}, \mathbf{D}_k, \hat{\boldsymbol{\beta}})$, where $P^{(r)}_{self}$ is the set of self-enumeration respondents, and $\hat{\tau}_k$ is an estimate of τ_k, the conditional expectation of $l_k^{(Self)}$ given the observed data $\mathbf{D}_k$ and the response model parameter $\boldsymbol{\beta}$. Note that when $l_k^{(Self)}$ is known $\hat{\tau}_k = l_k^{(Self)}$. The conditional probability that unit k belong to the subpopulation of self-enumeration respondents given that the unit responded at period I_k with $\mathrm{I}_k \ge e_k$ is given by

$$\tau_k = \frac{{}^{(UnSelf)}N^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self;k} = \mathrm{I}_k)}{{}^{(UnSelf)}N^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self;k} = \mathrm{I}_k) + {}^{(f)}N^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self+f;k} = \mathrm{I}_k)},$$

where t_a represents the discrete random variable that indicates the time period i when the response occurs during time period $[1,\mathrm{I}_{\max}]$ under random process a, with $a \in \{self, self+f\}$, $\Pr(t_{a;k} = i)$ is given by (2.1) with t replaced by t_a, ${}^{(UnSelf)}N^{(r)}_{(i)}$ is the number of “Unidentified” self-enumeration respondents at time period i, ${}^{(f)}N^{(r)}_{(i)}$ is the number of “extra-respondents” due to the follow-up activities at time period i, with ${}^{(Self)}N^{(r)}_{(i)} + {}^{(f)}N^{(r)}_{(i)} = {}^{(Self+f)}N^{(r)}_{(i)}$, ${}^{(Self)}N^{(r)}_{(i)} = {}^{(IdSelf)}N^{(r)}_{(i)} + {}^{(UnSelf)}N^{(r)}_{(i)}$, and ${}^{(IdSelf)}N^{(r)}_{(i)}$ is the number of “Identified” self-enumeration respondents at time period i. When $l_k^{(Self)}$ is known, then the estimator of the conditional probability that unit k belong to the subpopulation of self-enumeration respondents given the observed data is $\hat{\tau}_k = l_k^{(Self)}$; and when $l_k^{(Self)}$ is unknown, our estimator is given by

$$\hat{\tau}_k = \frac{{}^{(UnSelf)}\hat{N}^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self;k} = \mathrm{I}_k \mid \hat{\boldsymbol{\beta}})}{{}^{(UnSelf)}\hat{N}^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self;k} = \mathrm{I}_k \mid \hat{\boldsymbol{\beta}}) + {}^{(f)}\hat{N}^{(r)}_{(\mathrm{I}_k)} \Pr(t_{self+f;k} = \mathrm{I}_k \mid \hat{\boldsymbol{\beta}})}, \tag{3.1}$$

with ${}^{(f)}\hat{N}^{(r)}_{(i)} = {}^{(self+f)}\hat{N}^{(r)}_{(i)} - {}^{(self)}\hat{N}^{(r)}_{(i)}$ and ${}^{(UnSelf)}\hat{N}^{(r)}_{(i)} = {}^{(self)}\hat{N}^{(r)}_{(i)} - {}^{(IdSelf)}\hat{N}^{(r)}_{(i)}$. It remains to derive estimates for ${}^{(UnSelf)}N^{(r)}_{(\mathrm{I}_k)}$ and ${}^{(f)}N^{(r)}_{(\mathrm{I}_k)}$.

3.2 Derivation of Elementary Statistics

From the follow-up subsample we can compute first ${}^{(Self)}\hat{N}^{(r)}_{(i)} = \sum_k d_k(\wp_{f|\mathrm{I}}) P^{(*)}(t_{self;k} = i \mid \hat{\boldsymbol{\beta}})$, and ${}^{(self+f)}\hat{N}^{(r)}_{(i)} = \sum_k d_k(\wp_{f|\mathrm{I}}) P^{(*)}(t_{self+f;k} = i \mid \hat{\boldsymbol{\beta}})$ as estimators of ${}^{(self)}N^{(r)}_{(i)}$, and ${}^{(self+f)}N^{(r)}_{(i)}$ respectively, where $d_k(\wp_{f|\mathrm{I}}) = d_k(\wp_{f|\mathrm{I}} \mid P) = d_k(\wp) d_k(\wp_{f|\mathrm{I}} \mid \wp^{(m)}_{self|\mathrm{I}})$ is the follow-up subsample design weight, $P^{(*)}(t_{a;k} = i \mid \boldsymbol{\beta}) = h_{ki} \prod_{j=\mathrm{I}+1}^{i-1} (1 - h_{kj})$ for $i > \mathrm{I}$, ${}^{(Self+f)}N^{(r)}_{(i)} = {}^{(Self)}N^{(r)}_{(i)} + {}^{(f)}N^{(r)}_{(i)}$ denotes the sum of ${}^{(Self)}N^{(r)}_{(i)}$, the number of self-enumeration respondents, and ${}^{(f)}N^{(r)}_{(i)}$, the number of extra-respondents due to the follow-up at time period i, $i = \mathrm{I}+1, \ldots, \mathrm{I}_{\max}$. From both estimators, one may derive ${}^{(f)}\hat{N}^{(r)}_{(i)}$ using ${}^{(f)}\hat{N}^{(r)}_{(i)} = {}^{(Self+f)}\hat{N}^{(r)}_{(i)} - {}^{(Self)}\hat{N}^{(r)}_{(i)}$. Then, we can compute ${}^{(IdSelf)}\hat{N}^{(r)}_{(i)} = \sum_k d_k(\wp_{f|\mathrm{I}}) 1(i < e_k) P^{(*)}(t_{self;k} = i \mid \hat{\boldsymbol{\beta}})$ as estimator of ${}^{(IdSelff)}N^{(r)}_{(i)}$, where ${}^{(IdSelf)}N^{(r)}_{(i)}$ denotes the number of identified self-enumeration respondents at time period i, $i = \mathrm{I}+1, \ldots, \mathrm{I}_{\max}$. From ${}^{(Self)}\hat{N}^{(r)}_{(i)}$ and ${}^{(IdSelf)}\hat{N}^{(r)}_{(i)}$ we may derive ab estimator of ${}^{(UnSelf)}N^{(r)}_{(i)}$, the number of unidentified self-enumeration respondents, using ${}^{(UnSelf)}\hat{N}^{(r)}_{(i)} = {}^{(Self)}\hat{N}^{(r)}_{(i)} - {}^{(IdSelf)}\hat{N}^{(r)}_{(i)}$.

3.3 Constrained Least Squares Estimator

The size N of the finite population P is assumed to be fixed during data collection period, while the vector size $\boldsymbol{\theta}_{(i)} = ({}^{(IdSelf)}N^{(r)}_{(i)}, {}^{(UnSelf)}N^{(r)}_{(i)}, {}^{(f)}N^{(r)}_{(i)})^T$ of the vector domain $({}^{(IdSelf)}P^{(r)}_{(i)}, {}^{(UnSelf)}P^{(r)}_{(i)}, {}^{(f)}P^{(r)}_{(i)})^T$ may varies from one time period to another. Others elementary statistics can be derived using the follow-up subsample as well as the main sample. For example from the main sample, we may compute $\sum_k d_k(\wp) \Pr(t_{self;k} = i \mid \hat{\boldsymbol{\beta}})$ and $\sum_k d_k(\wp) \Pr(t_{self;k} > i \mid \hat{\boldsymbol{\beta}})$ as estimator of ${}^{(Self)}N^{(r)}_{(i)}$ and ${}^{(Self)}\hat{N}^{(m)}_{i}$ respectively. In the linear regression with observed vector $\hat{\boldsymbol{\theta}}_{(i)}$ of elementary statistics, vector of regression parameters $\boldsymbol{\theta}_{(i)}$, known design matrix $\mathbf{M}_{(i)}$ of 0's and 1's, and vector residual errors $\boldsymbol{\varepsilon}_{(i)}$, we have $\hat{\boldsymbol{\theta}}_{(i)} = \mathbf{M}_{(i)} \boldsymbol{\theta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}^T_{(1)}, \ldots, \boldsymbol{\theta}^T_{(\mathrm{I}_{\max})}, N^{(m)}_{\mathrm{I}_{\max}})^T$ be the $(3\mathrm{I}_{\max} + 1) \times 1$ vector parameter, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}^T_{(1)}, \ldots, \hat{\boldsymbol{\theta}}^T_{(\mathrm{I}_{\max})}, \hat{N}^{(m)}_{(\mathrm{I}_{\max})})^T$ be the observed vector of elementary statistics, the known design matrix $\mathbf{M}$ of 0's and 1's, and vector residual errors be $\boldsymbol{\varepsilon}$. Then we have $\hat{\boldsymbol{\theta}} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. By minimizing the objective function $\mathrm{O}(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \mathbf{M}\boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}} - \mathbf{M}\boldsymbol{\theta})$, it follows that the resulting least squares estimator $\hat{\boldsymbol{\theta}}_{ls}$ is $\hat{\boldsymbol{\theta}}_{ls} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\hat{\boldsymbol{\theta}}$. The situation where exact information is available relative to q_l linear combinations of elements of the vector parameter $\boldsymbol{\theta}$ can be stated in the form of the linear equality restrictions:

$$\mathbf{L}\boldsymbol{\theta} = \boldsymbol{l}\,, \tag{3.2}$$

where $\mathbf{L}$ is a $q_l \times (3\mathrm{I}_{\max} + 1)$ known design matrix that expresses the structure of the information on the elements of the $\boldsymbol{\theta}$ vector, and $\boldsymbol{l}$ is a $q_l \times 1$ vector of constants. For example if the population size N is known, then $q_l = 1$, the $1 \times (3\mathrm{I}_{\max} + 1)$ vector $\mathbf{L}$ is given by $\mathbf{L} = [1 \;\; \dots \;\; 1] \equiv \mathbf{1}^T_{3\mathrm{I}_{\max}+1}$ and the 1×1 scalar value l is given by $l = N$, where $\mathbf{1}_p$ is the $p \times 1$ vector of 1's. Under (3.2), we use instead a constrained least squares estimator of $\boldsymbol{\theta}$, which is basically the least squares with some components of $\boldsymbol{\theta}$ restricted to q_l combinations. It can be: a) constructed by minimizing $\mathrm{O}(\boldsymbol{\theta})$ subject to the constraint given by (3.2); and, b) solved using the Lagrange multiplier procedure. The constrained least squares estimator $\breve{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is given by

$$\breve{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{ls} + \mathbf{A}(\boldsymbol{l} - \mathbf{L}\hat{\boldsymbol{\theta}}_{ls})\,,$$

where $\mathbf{A} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{L}^T\{\mathbf{L}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{L}^T\}^{-1}$.

3.4 Estimator of the Population Total

If $l_k^{(Self)}$ is known for every sampled unit and all subsampled units respond, then an estimator of Y would be

$$\breve{Y}^{(C)} = \textstyle\sum_k d_k(\wp) r_k l_k^{(Self)} y_k + \sum_k d_k(\wp_{f|\mathrm{I}}) r_k (1 - l_k^{(Self)}) y_k \equiv \sum_k d_k^{(C)} y_k\,, \tag{3.3}$$

with

$$d_k^{(C)} = d_k(\wp) r_k l_k^{(Self)} + d_k(\wp_{f|\mathrm{I}}) r_k (1 - l_k^{(Self)})\,, \tag{3.4}$$

where the superscript C in $\breve{Y}^{(C)}$ and in $d_k^{(C)}$ stands for complete response in the follow-up subsample. Given $l_k^{(Self)}$, the first part of the right term of (3.3), $\sum_k d_k(\wp) r_k l_k^{(Self)} y_k$, is design-unbiased estimator of the total of the subpopulation of self-enumeration respondents, and the second part, $\sum_k d_k(\wp_{f|\mathrm{I}}) r_k (1 - l_k^{(Self)}) y_k$, is design-unbiased estimator of the total of the rest of the population. Hence for any I, the estimator given by (3.3) is a conditional unbiased estimator of the total of the finite population:

$$E_{\wp}^{(2|1)}(\hat{Y}^{(C)}) = \textstyle\sum_k d_k(\wp) y_k = \hat{Y}_{HT}\,,$$

or

$$E_{\wp}^{(2|1)}(d_k^{(C)}) = d_k(\wp)\,,$$

where $E_{\wp}^{(2|1)}$ denotes expectation with respect to subsampling for nonresponse. We may write $d_k^{(C)}$ as

$$d_k^{(C)} = d_k(\wp) r_k \{ {}^{(Id)}l_k^{(Self)} + (1 - {}^{(Id)}l_k^{(Self)}) l_k^{(Self)} \} + d_k(\wp_{f|\mathrm{I}}) r_k (1 - l_k^{(Self)}) ,$$

where ${}^{(Id)}l_k^{(Self)} = 1$ if unit k is identified as a self-enumeration respondent (i.e., $k \in \wp_{self|\mathrm{I}}^{(r)} \cup \wp_{s|\mathrm{I}}^{(r)}$ or $k \in \wp_{f|\mathrm{I}}^{(r)}$ with $\mathrm{I}_k < e_k$), and ${}^{(Id)}l_k^{(Self)} = 0$ if not. When ${}^{(Id)}l_k^{(Self)} = 0$ unit k might or might not be a self-enumeration respondent. Note that ${}^{(Id)}l_k^{(Self)} l_k^{(Self)} = {}^{(Id)}l_k^{(Self)}$.

Unfortunately neither $l_k^{(Self)}$ is known for each respondent nor the complete response is likely to occur in the follow-up subsample. When $l_k^{(Self)}$ is unknown we replace it by its estimate and the resulting weight under complete response in the follow-up subsample is given by

$$\hat{d}_k^{(C)} = d_k(\wp) r_k \{ {}^{(Id)}l_k^{(Self)} + (1 - {}^{(Id)}l_k^{(Self)}) \hat{\tau}_k \} + d_k(\wp_{f|\mathrm{I}}) r_k (1 - \hat{\tau}_k) , \tag{3.5}$$

where k might or might not be in the follow-up subsample.

Because nonresponse is likely to occur in the follow-up subsample, our final weight is defined as

$$\hat{d}_k = d_k(\wp) r_k \{ {}^{(Id)}l_k^{(Self)} + (1 - {}^{(Id)}l_k^{(Self)}) \hat{\tau}_k \hat{d}_{f;k}^{(r)} \} + d_k(\wp_{f|\mathrm{I}}) r_k (1 - \hat{\tau}_k) \hat{d}_{f;k}^{(r)} , \tag{3.6}$$

where $\hat{d}_{f;k}^{(r)}$ is an adjustment factor for nonresponse in the follow-up subsample. For example, one may use the following adjustment factor

$$\hat{d}_{f;k}^{(r)} = \hat{d}_k(\wp_{f|\mathrm{I}}^{(r)} \mid \wp_{f|\mathrm{I}}) = \frac{r_{f;k}}{\hat{\xi}_{f;k}} , \tag{3.7}$$

where $r_{f;k} = 1_k(\wp_{f|\mathrm{I}}^{(r)} \mid \wp_{f|\mathrm{I}}, \mathrm{I}_k \geq e_k)$ is the conditional response indicator for the k^{th} unit in the follow-up subsample, $\xi_{f;k} = \pi_k(\wp_{f|\mathrm{I}}^{(r)} \mid \wp_{f|\mathrm{I}}, \mathrm{I}_k \geq e_k)$ is the conditional response probability in the follow-up subsample, and $\hat{\xi}_{f;k} = \xi_{f;k}(\hat{\boldsymbol{\beta}})$. Substituting $\hat{d}_{f;k}^{(r)}$ given by (3.7) into $\hat{d}_k$ given by (3.6), we get

$$\hat{d}_k = d_k(\wp) \{ {}^{(Id)}l_k^{(Self)} r_k + (1 - {}^{(Id)}l_k^{(Self)}) \hat{\tau}_k r_{f;k} / \hat{\xi}_{f;k} \} + d_k(\wp_{f|\mathrm{I}}) (1 - \hat{\tau}_k) r_{f;k} / \hat{\xi}_{f;k} . \tag{3.8}$$

Noting that $r_k r_{f;k} = r_{f;k}$. So, our proposed estimator of $Y = \sum_k y_k$ is

$$\hat{Y} = \sum_k \hat{d}_k y_k , \tag{3.9}$$

where $\hat{d}_k$ is given by (3.8).

3.5. Simulation Study (continuation)

We subsampled from each $\wp_{self|\mathrm{I}}^{(m)}$, the set of nonrespondents at the I^{th} time period, three Bernoulli subsamples for follow-up with different sampling fraction f : $f = .25$, $f = .50$, and $f = .75$. Note that I has been set to $\mathrm{I} = 3$. Table 6 gives statistics on the observed subsamples, and Table 7 gives statistics on the observed cost. Only subsampled nonrespondents are subject to the follow-up in this second simulation. Therefore for convenience, we set the entry time period e_k for non subsampled unit to $e_k = \mathrm{I}_{\max} + 1$.

Our first vector parameter of interest is the response model parameter $\boldsymbol{\beta}$. Table 8 displays values of the response model parameters. We calculated $\bar{\hat{\theta}}$ and $B(\hat{\theta})$ for response model parameters and those values are reported in Table 8. Table 8 clearly demonstrates that the bias is small for each regression parameter. The second vector parameter is the number of respondents by mode and by time period. We calculated $\bar{\hat{\theta}}$ from the main sample for the number of respondents under self-enumeration by time period and those values are reported in Table 9. Table 9 demonstrates that the bias is small for each parameter. Table 10, 11, and 12 display estimates of the number of respondents under self-enumeration, follow-up in addition to self-enumeration, and the number of extra-respondent due to the follow-up using the follow-up subsamples with sampling fraction .25, .50, and .75 respectively. Tables 10, 11 and 12 demonstrate that the bias is small for each parameter. The final parameter of interest is the population total $Y = \sum_k y_k$. For comparison, we also computed $\hat{Y}_{\xi}$ given by (1.3) as estimator of the population total. We calculated $RB(\hat{\theta})$ and MSE ratios for each estimator $\hat{\theta}$ with $\hat{Y}$ given by (3.9), and those values are reported in Table 13. Table 13 clearly indicates that all relative biases are small for all estimates. The estimator using estimated true response model parameter is efficient when the subsampling fraction is .25, while the proposed estimator is more efficient that the estimator using estimated true response model parameter when the subsampling fraction is .50 or .75. Table 10 also provides results for calibrated estimators to the population total $U = \sum_k u_k$ of the auxiliary variable, which indicate again that calibration-to-population total is highly efficient. Note that the calibration factor for the proposed estimator is given by $g = \sum_k u_k / \sum_k \hat{d}_k u_k$, where $\hat{d}_k$ is given by (3.6).

Concluding Remarks

We used discrete-time hazard to the analysis of response indicators in surveys and censuses. The proposed approach facilitates the examination of the shape of the hazard function. Since inspection of the shape of the hazard function indicates when a response is most likely to occur, and how the probability varies over both time and follow-up treatments, the description of the shapes of the hazard function have an important role to play in survey quality and cost. We used regression analysis to investigate the effect of mixed-mode on the response probability. Estimator of response model parameter as well as estimator of the finite population total associated with the mixed-mode of data collection is given. We also studied the situation where all sampled units are subject to self-enumeration without follow-up, while a random subsample of nonrespondents is subject to follow-up activities in addition to self-enumeration. Then, we extended the Hansen-Hurwitz' approach for nonresponse in sample survey to make full use of all observed responses in the estimation of the finite population total. Our approach permits flexibility regarding the time period in the: a) selection of the follow-up subsample; and, b) entry time of each sub-sampled unit into the follow-up activity.

References

Allison, P. D. 1982. Discrete-time Methods for the Analysis of Event Histories."S. Leinhardt (ed.). *Sociological Methodology*. San Francisco: Jossey–Bass. pp. 61–98.

Cochran, W. 1977. *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Demnati, A. 2015. Linearization Variance Estimators for Mixed-mode Survey Data when Response Indicators are Modeled as Discrete-time Survival. In Proceedings of the Federal Committee on Statistical Methodology Research Conference: American Statistical Association, December 1, 2015. Washington D.C. USA.

Demnati, A. 2016. Responsive Design – Side Effect Reduction of Prior Information on Survey Design. In *Joint Statistical Meetings* Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. July 30 – August 4, Chicago, Illinois, USA

Hansen, M. H., and W. N. Hurwitz. 1946. The Problem of Nonresponse in Sample Survey. *Journal of the American Statistical Association*, 41, 517-529.

Lancaster, T. 1990. *The Econometric Analysis of Transition Data*, Cambridge. Cambridge University Press.

Lagakos, S.W. 1979. General Right Censoring and its Impact on the Analysis of Survival Data. *Biometrics*, 35, pp. 139–156.

Little, R. and D. Rubin. 2002. *Statistical Analysis with Missing Data*. Hoboken (NJ): John Wiley & Sons, Inc.

Prentice, R. L., J. D. Kalbfleisch, A. V. Peterson, N. Jr., Flournoy, V. T. Farewell, and N. E. Breslow. 1978. The Analysis of Failure Times in the Presence of Competing Risks. *Biometrics*, 34, 541-554.

Rosenbaum, P. R. 1987. "Model-based Direct Adjustment." *Journal of the American Statistical Association*, 82, pp. 387–394.

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

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology*, 39, 29-58.

Schouten, B. and M. Calinescu. 2010. Optimizing Quality of Response Through Adaptive Survey Designs. Statistics Netherlands.

Swensson, J. 2007. Measurement Bias Adjustment in the Swedish Farm Accidents Survey. International Conference on Establishment Survey.

Table 1: Resource Allocation

Expected			Cost Ratio			Sampling Model Parameter	
CV	Sample size	Cost	Sampling	Follow-up	Data collection	$\lambda_0^{(\wp)}$	$\lambda_1^{(\wp)}$
.05	388	2635	.15	.42	.43	-3.724	.0243

Table 2: Observed Statistics on the Sample Counts – Single-phase Data Collection

		Mean	Minimum	Maximum
Sample Size		388	325	444
Number of respondents	All	371	313	424
	By Mail	194	140	239
	By Internet	176	138	216

Table 3: Observed Statistics on the Sample Cost – Single-phase Data Collection

Cost of		Mean	Minimum	Maximum
	Sampling	387	333	442
	Follow-up	1102	770	1470
	Data collection	1146	915	1382
	By Mail	969	745	1210
	By Internet	177	133	213
	Total Cost	2635	2149	3217

Table 4: Bias for Response Model Parameter Estimate – Single-phase Data Collection

	Parameter θ							
	$\beta_{M;1}$	$\beta_{M;2}$	$\beta_{M;3}$	$\beta_{M;4}$	$\beta_{I;1}$	$\beta_{I;1}$	$\beta_{I;1}$	$\beta_{I;1}$
Value	-1.5	.003	-.2	.6	-2	.005	-.1	.5
Estimate	-1.5	.003	-.20	.61	-2.00	.005	-.01	.50
$B(\hat{\theta})$	-.0058	.0000	-.0008	.0066	-.0342	.0003	.0018	.0060

Table 5: Relative Bias and Mean Square Error Ratios – Single-phase Data Collection

		Relative bias and (MSE ratios)	
		Response Model parameter	
Parameter of interest	Weights	Estimated	True
$\theta = \sum_k y_k$	Design-Response	.0004 (1.00)	.0002 (1.05)
	Calibration*	.00009 (.02)	.00008 (.02)
$\theta_M = \sum_k y_{M;k}$	Design-Response	.0007 (.97)	.0013 (1.91)
	Calibration*	.00008 (.11)	.00041 (.13)
$\theta_I = \sum_k y_{I;k}$	Design-Response	.0008 (1.27)	-.0004 (2.86)
	Calibration*	.00017 (.08)	.00017 (.10)

*Calibration to the population total $U = \sum_k u_k$.

Table 6: Observed Means – Two-phase Data Collection

		$f = .25$	$f = .5$	$f = .75$
Subsample Size		42	84	125
Number of Respondents	All	337	348	359
	By Mail	176	182	188
	By Internet	161	166	171

Table 7: Observed Means on the Sample Cost – Two-phase Data Collection

Cost of	$f = .25$	$f = .5$	$f = .75$
Sampling	430	472	515
Follow-up	276	550	828
Data collection	1046	1082	1115
By Mail	884	915	944
By Internet	162	167	172
By Self	936	861	784
By Follow-up	110	221	330
Total Cost	1752	2104	2459

Table 8: Bias for Response Model Parameter Estimate – Two-phase Data Collection

		Parameter θ							
		$\beta_{M;1}$	$\beta_{M;2}$	$\beta_{M;3}$	$\beta_{M;4}$	$\beta_{I;1}$	$\beta_{I;1}$	$\beta_{I;1}$	$\beta_{I;1}$
	Value	-1.5	.003	-.2	.6	-2	.005	-.1	.5
$f = .25$	Estimate	-1.5	.003	-.20	.61	-1.98	.004	-.10	.51
	$B(\hat{\theta})$	-.0126	.0002	.0007	.0147	.0128	-.0003	-.0002	.0159
$f = .5$	Estimate	-1.52	.003	-.20	.6	-1.97	.004	-.10	.50
	$B(\hat{\theta})$	-.0281	-.0003	-.0012	.0155	.0285	-.0000	-.0000	.0101
$f = .75$	Estimate	-1.5	.003	-.2	.6	-1.99	.005	-.10	.50
	$B(\hat{\theta})$	-.0046	-.0000	.0004	.0084	.0099	-.0003	-.0003	.0086

Table 9: Number of Self-enumeration Respondents by Time Period - Estimation from the Main Sample

Time Period	Value	Estimate
1	1349	1350
2	877	875
3	591	590
4	412	412
5	296	296
6	218	218
7	164	165
8	126	127
9	98	100
10	78	79

Table 10: Number of Respondents by Time Period - Estimation from the Follow-up Subsample with f=.25

	Self		Self + Follow-up				Follow-up Only	
Time Period	Value	Estimate	Value	Estimate	Identified	Unidentified	Value	Estimate
1	1349							
2	877							
3	591							
4	412	413	432	434	374	60	20	21
5	296	297	355	359	209	149	59	62
6	218	219	312	316	110	207	94	97
7	164	165	276	280	50	230	112	114
8	126	127	239	241	13	228	113	113
9	98	100	196	196	0	196	97	96
10	78	79	146	146	0	146	70	67

**Table 11: Number of Respondents by Time Period -
Estimation from the Follow-up Subsample with f=.50**

	Self		Self + Follow-up				Follow-up Only	
Time Period	Value	Estimate	Value	Estimate	Identified	Unidentified	Value	Estimate
1	1349							
2	877							
3	591							
4	412	412	432	433	374	59	20	21
5	296	296	355	357	209	149	59	61
6	218	219	312	315	109	206	94	97
7	164	165	276	279	50	230	112	114
8	126	127	239	241	13	228	113	114
9	98	100	196	196	0	196	97	97
10	78	79	146	146	0	148	70	68

**Table 12: Number of Respondents by Time Period -
Estimation from the Follow-up Subsample with f=.75**

	Self		Self + Follow-up				Follow-up Only	
Time Period	Value	Estimate	Value	Estimate	Identified	Unidentified	Value	Estimate
1	1349							
2	877							
3	591							
4	412	413	432	433	375	59	20	20
5	296	297	355	357	210	147	59	60
6	218	219	312	314	109	205	94	95
7	164	165	276	278	49	229	112	112
8	126	127	239	240	13	227	113	112
9	98	100	196	196	0	196	97	96
10	78	80	148	147	0	147	70	68

Table 13: Relative Bias and Mean Square Error Ratios for Finite Population Total $\theta = \sum_k y_k$

		Relative bias and (MSE ratios)	
		Type of Estimator	
Subsampling Fraction	Weights	$\hat{Y}$	$\hat{Y}_{\hat{\xi}}$
$f = .25$	Design-Response	-.0004 (1.00)	.025 (.79)
	Calibration*	-.0004 (.02)	-.0002 (.01)
$f = .5$	Design-Response	.0005 (1.00)	.0402 (1.54)
	Calibration*	-.0003 (.02)	-.0001 (.02)
$f = .75$	Design-Response	.0004 (1.00)	.0590 (2.46)
	Calibration*	-.0003 (.02)	-.0001 (.02)

*Calibration to the population total $U = \sum_k u_k$.