مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۱۴، شماره‌ی ۱، بهار و تابستان ۱۳۹۶، صص ۷۷–۹۹

J. Statist. Res. Iran **14** (2017): 77–99
DOI: 10.18869/acadpub.jsri.14.1.77

An Efficient Method for Estimating Population Parameters Using Split Questionnaire Design

Saeideh Kamgar and Hamidreza Navvabpour*

Allameh Tabataba'i University

Received: 10/16/2016 Approved: 7/4/2017

Abstract. The effect of survey questionnaire length on precision of survey statistics has been discussed in several studies. It is generally concluded that the lengthy questionnaire leads to increase non-sampling errors, especially nonresponse rate. Split questionnaire method has been introduced as a solution to decrease the response burden and nonresponse rate, involves splitting the questionnaire into subquestionnaires and then administering these subquestionnaires to different subsets of the original sample. In this paper, we suggest a method for splitting long questionnaire and analyzing resulting data, using small area estimation. The general idea behind this approach is to construct some socio-demographic or geographic small areas to apply small area estimation to improve the efficiency of survey statistics. Our new approach is supported by a simulation study based on a real dataset of the 2011 Iran Income and Expenditure survey, in which we show our method provides more reliable statistics than existing methods.

Keywords. Empirical best linear unbiased prediction; matrix sampling; multiple imputation; response burden; small area estimation.

MSC 2010: 62D05; 62G09.

* Corresponding author

1 Introduction

Several studies in the Survey Research Methodology have focused on effects of a lengthy survey questionnaire on declining response rate and precision of survey statistics (Adams and Darwin, 1982; Roszkowski and Bean, 1990; Dillman, et al., 1993).

A split questionnaire survey design (Raghunathan and Grizzle, 1995) has been introduced as a solution to decrease the response burden arising from lengthy questionnaires. Accordingly, the questionnaire is split into subquestionnaires and each subquestionnaire is assigned to a group of sample units. Under a split questionnaire design, the procedure of subsample assigning is at random. Therefore, the resulting nonresponse is completely at random.

The split questionnaire method has two related aspects: (i) how to design the subquestionnaires and (ii) how to analyze the resulting data. A basic design for split questionnaire method has been introduced by Raghunathan and Grizzle (1995) as a generalization of multiple matrix sampling design (Shoemaker, 1973). The design is comprised of splitting questionnaire to some components and randomly assigning some components to a part of the sample units. Raghunathan and Grizzle (1995) also applied MI (Multiple Imputation) method (Rubin, 1987) to fill out item nonresponse caused by splitting questionnaire using Gibbs sampler technique (Gelfand, 2000). This strategy is still a common way to perform the split questionnaire approach (designing and analyzing steps). Meanwhile, a number of recent studies suggest different approaches to design and/or analyze the split questionnaire. The outline and applications of some of these approaches in split questionnaire method have been reviewed in depth by Gonzalez and Eltinge (2007). Here, we briefly introduce the most relevant ones.

Regarding the design of the split questionnaire, an algorithm has been developed by Thomas, et al. (2006) to design the split questionnaire based on an index which ranks how well each item is predicted by other terms of the subquestionnaires. More precisely, the idea is to administrate highly correlated items to different subquestionnaires, aimed to improve the quality of completed datasets by applying imputation technique to analyze the data. In addition, a fixed set of items from the original questionnaire is considered to be asked from all sample units. Later, Adigüzel and Wedel (2008) proposed a strategy to design optimal split questionnaire surveys by applying the Kullback-Leibler distance (Kullback and Leibler, 1951). In other words, their proposed criterion to find the optimal design is to minimize the

information loss compared to the original questionnaire.

Regarding the analysis of the split questionnaire, Chipperfield and Steel (2009) introduced an approach to find the optimal allocation for a split questionnaire design, subjected to constraints on fixed costs and variances. They also considered estimators in the simple case of two variables, including the BLUE (best linear unbiased estimator) to impute missing data. Later, Merkouris (2010) suggested an estimation method to improve the precision of survey statistics in matrix sampling. The method used a generalized regression procedure based on the principle of the best linear unbiased estimation. Chipperfield and Steel (2011) discussed the efficiency of the split questionnaire design by means of measuring the cost required to meet constraints on the accuracy of estimates. They concluded that the split questionnaire design could be an efficient and practical method to sample survey.
It is possible to split the long questionnaire in different ways with regards to the aim and the contents of the survey. One way is to split the original questionnaire to different subquestionnaires (without overlap) and assign them to different (disjoint) subsamples. This design is useful when the aim is to estimate the population mean or total. Another application of this design is to integrate a number of common (independent) surveys which are asked from different samples. In this design, the resulting nonresponse can be considered as unit nonresponse. Consequently, enough sample units would not be available for each subquestionnaire. Another way for splitting the questionnaire is to create the subquestionnaires that have some items in common. Applying this design is useful, when the goal is to study some associations of the combination of the response variables. In this design, the missing part caused by splitting the long questionnaire can be viewed as item nonresponse and the common methods for imputation such as MI technique could be applied to improve the quality of survey data.

Moreover, in many split questionnaire designs, a core part that contains a number of high priority questions from the original questionnaire is administered to all respondents, whereas other questions, called split items, are only administered to a subset of respondents.

In this paper, we suggest a method to design split questionnaire and to estimate population parameters, using SAE (Small Area Estimation) technique. The parameter of the interest is the population mean. So, the first design (subquestionnaires without overlap) is used to split the questionnaire. We construct some socio-demographic or geographic groups and consider them as small areas. The general idea is to apply SAE as a solution to

Table 1. The proposed approach for administering subquestionnaires to subsamples.

	Subsample (1)	Subsample (2)	...	Subsample (K)
subquestionnaire 1	✓	—	...	—
subquestionnaire 2	—	✓	...	—
⋮	⋮	⋮	⋮	⋮
subquestionnaire K	—	—	...	✓

analyze data from split questionnaires to improve the efficiency of survey statistics. The rest of the paper is organized as follows. In Section 2, we briefly describe the design of splitting long questionnaire, required to apply SAE technique. Section 3 is devoted to discuss on available approaches to analyze survey data coming from split questionnaire design. Section 4 provides a simulation study based on a real dataset to investigate the performance of our proposed approach. In Section 5, measures of comparison used in our study are introduced. Finally, results of this study will be presented in Section 6.

2 A Design for Splitting Long Questionnaire

In this section, we introduce an approach for splitting long questionnaires required to apply SAE technique. In this design, one or more grouping variables from the original questionnaire are used to construct (L) areas. The areas could be considered as socio-demographic or geographic areas. Some items of the questionnaire that are important or highly correlated with the other items are considered as covariates (X). The other items are considered as response variables (Y) and split to (K) parts (split items). Thus, the original questionnaire is divided into (K) subquestionnaires.

A simple random sample (original sample) is selected from the population and each subquestionnaire is randomly assigned to a part of the original sample (a subsample). The proposed approach for administering subquestionnaires to subsamples is demonstrated in Table 1. According to the design, there is no overlap among (K) subquestionnaires assigned to the (K) subsamples.

In addition to the original questionnaire, the grouping variable(s) can be selected from a census. In other words, two sources are available for the grouping variable(s): A census, or the original questionnaire. For the latter

source, the selected variable(s) is asked from the original sample through an initial interview before administering the survey (e.g., telephone interview) or during the survey process. Therefore, the information for grouping variable(s) could be obtained before (Case I) or after administering the survey (Case II). These two cases are described below.

Case I: Source of the grouping variable(s) is a previous census or an initial interview from the original sample. Thus, the information for grouping variable(s) is available before carrying out the survey. All sample units are categorized by the grouping variable(s) before measuring the study variables. Consequently, units are classified into homogeneous classes and each class is considered as an area. Sample units belonging to each area are randomly divided into (K) subsamples and each subsample answers to one of the (K) subquestionnaires. Note that, in each area, the number of subquestionnaires and the number of subsamples must be equal.

Case II: In this case, the grouping variable(s) is included in all subquestionnaires. Thus, the split items and the grouping variable(s) are asked from the original sample (all sample units), simultaneously. Consequently, the necessary information to classify the sample units will be available after administering the survey. The classification of the sample units is performed by the measured grouping variable(s) and each class is considered as an area.

According to the SAE method, it is necessary to use the appropriate auxiliary information collected from large surveys (e.g., census) to compute the SAE estimates. On the other hand, in split questionnaire method the original sample can be viewed as a source of auxiliary information for parameter estimation based on a subsample. These auxiliary variables, asked from the original sample (all sample units), are called core variables. It is possible to ask the core variables and split items, simultaneously or separately. Therefore, in split questionnaire method, the auxiliary information required for the SAE method is available from either a census or the original sample.

Let $n_{kl}^{'} \geqslant 0, k = 1, \ldots, K,\ l = 1, \ldots, L$ denote the subsample size within the lth area, responded to the kth subquestionnaire. Here, $n^{(l)},\ l = 1, \ldots, L$ are the sample size in the lth area, where $n = \sum_{l=1}^{L} n^{(l)}$ is the target sample size. Moreover, $n_k{}^{'} = \sum_{l=1}^{L} n_{kl}{}^{'},\ k = 1, \ldots, K$ are the number of sample units who responded to the kth subquestionnaire, where $n = \sum_{k=1}^{K} n_k{}^{'}$. In

Table 2. The number of units that responded to each subquestionnaire in each area in SAE approach.

	Area 1	Area 2	...	Area L	Total
subquestionnaire 1	$n_{11}^{'}$	$n_{12}^{'}$	...	$n_{1L}^{'}$	$n_{1}^{'}$
subquestionnaire 2	$n_{21}^{'}$	$n_{22}^{'}$	...	$n_{2L}^{'}$	$n_{2}^{'}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
subquestionnaire K	$n_{K1}^{'}$	$n_{K2}^{'}$	...	$n_{KL}^{'}$	$n_{K}^{'}$
Total	$n^{(1)}$	$n^{(2)}$	...	$n^{(L)}$	n

split questionnaire method, it is common to consider subsamples with equal sizes ($n_k^{'} = n^{'},\ k = 1, \ldots, K$). Hence, there are $n^{'} = n/K$ sample units who responded to each subquestionnaire. Table 2 shows the number of sample units that responded to each subquestionnaire within each area.

The pattern introduced in Table 2 could be considered for the both cases of grouping variable(s) sources. We note that as we described in Case I, the original sample is classified by the grouping variable(s) before administering the survey. Thus, some extra relations exist among the sample sizes in the pattern for Case I. In this case, $n_{kl}^{'} = n^{(l)}/K, k = 1, \ldots, K,\ l = 1, \ldots, L$ are the sample sizes within the lth area devoted to the kth subquestionnaire, where $n^{(l)}/K, l = 1, \ldots, L$ are integer values. We also stress that, in Case II, as the classification process is performed after administering the survey, some of the sub-sample sizes ($n_{kl}^{'}$) may be zero.

3 Population Characteristics Estimation

The analysis of resulting data from splitting long questionnaire is an important aspect of the method. In this paper, our goal is to estimate a linear function (mean) of the survey variables y ($\bar{Y}_U$). As already discussed in Section 1, the sample size in each area is not large enough to support direct estimates of appropriate precision based on the proposed split questionnaire design. Moreover, regarding the constructed areas, the estimation procedure could be considered as the overall estimation of a population comprised of small areas. Accordingly, the local direct estimates are not reliable due to insufficient sample units within each area. Therefore, SAE technique could

be applied to improve local estimates precision, that consequently leads to an efficient overall estimate (Militino et al., 2007).

SAE technique is an efficient method to solve the problem of low precision of characteristics estimates in areas with small sample sizes. Here, the area is referred to a geographic domain (e.g. a province, a county, etc.), or a socio-demographic group (e.g. age, sex, etc.). In the case of existing auxiliary information for each unit, one of the common models that has been used in the SAE method is nested error regression model with equal error variances (Battese, et al., 1988). The model, investigated in this paper, is a special case of unit level linear mixed model with a block diagonal covariance structure and it is applicable to estimate linear parameters.

First, let us define the notations related to the model. Under the assumption that population size in the lth area, $N^{(l)}$, is large, the model for the response variables (y) of the kth subquestionnaire can be written as:

$$y_{lj} = x_{lj}\beta + v_l + e_{lj}, \qquad l = 1, \ldots, L; \quad j = 1, \ldots, n_{kl}{}', \tag{1}$$

where:
$n_{kl}{}'$: sample size in the lth area responded to the kth subquestionnaire,
y_{lj}: response variable for the jth unit in the lth area,
x_{lj}: vector of auxiliary variables for the jth unit in the lth area,
β: vector of regression coefficients,
v_l: area random effect for the lth area, and
e_{lj}: random error term.
The area effects, $v_l,\ l = 1, \ldots, L$ are assumed to be independent $N(0, {\sigma_v}^2)$ random variables. The error terms, $e_{lj},\ l = 1, \ldots, L$ are considered as independent $N(0, {\sigma_e}^2)$ random variables and e_{lj}s are assumed to be mutually independent of v_ls. The empirical best linear unbiased predictor (EBLUP) (Rao and Molina, 2015) in the context of linear mixed model is a model-based prediction that can improve the efficiency of SAE. The EBLUP is given by:

$$\bar{Y}_{EBLUP,l} = \bar{X}_l{}'\tilde{\beta} + \hat{\gamma}_l(\bar{y}_l - \bar{x}_l{}'\tilde{\beta}), \qquad l = 1, \ldots, L, \tag{2}$$

where $\bar{X}_l$, $\bar{x}_l$ and $\bar{y}_l$ are the population auxiliary variables mean vector, sample auxiliary variables mean vector and sample response variable mean of the lth area, respectively. Note that $\bar{X}_l$ can be obtained from the previous census or from the original sample depends on the type of available data sources (refer to Section 2). For auxiliary variables which are only available from the original sample, the unknown population means of these variables can

be estimated using the information from the original sample. Furthermore $\tilde{\beta}$ and $\hat{\gamma}_l$ can be expressed as:

$$\tilde{\beta} = \left(\sum_{l=1}^{L} \sum_{j=1}^{n_{kl}{}'} (x_{lj} - \gamma_l \bar{x}_l \bar{x}_l') \right)^{-1} \sum_{l=1}^{L} \sum_{j=1}^{n_{kl}{}'} (x_{lj} y_{lj} - \bar{x}_l \bar{y}_l), \tag{3}$$

$$\hat{\gamma}_l = \frac{\hat{\sigma_v}^2}{\hat{\sigma_v}^2 + \hat{\sigma_e}^2 / n_{kl}{}'}, \tag{4}$$

where $\hat{\sigma_v}^2$ and $\hat{\sigma_e}^2$ are unbiased estimators of ${\sigma_v}^2$ and ${\sigma_e}^2$, respectively (Rao and Molina, 2015). The model (1) is used for each subquestionnaire to compute the EBLUP of population means for each area. To estimate population mean, $\bar{Y}_U$, the overall estimate is derived as a weighted aggregation of the area means (Militino et al., 2007). Therefore, the estimate of $\bar{Y}_U$ for the population of size N takes the following form:

$$\hat{\bar{Y}}_U = \sum_{l=1}^{L} \frac{N_l}{N} \bar{Y}_{EBLUP,l}. \tag{5}$$

4 A Simulation Study

This section describes a design-based Monte Carlo simulation study conducted with a real and generated datasets to investigate the efficiency of the proposed method compared with some alternative approaches, introduced in the literature. As the real dataset, we use a number of variables of the Iran Households Income and Expenditure Survey (HIES) conducted by the Statistical Center of Iran in 2014. In addition to the census variables, the dataset used in this study comes from the annual income and expenditure survey of N households. Moreover, the real dataset used for comparison is limited and it is impossible to check the effect of increasing the length of questionnaire on results based on all methods. This issue has been considered and checked with a long questionnaire. For this purpose, we generated 18 variables as response variables and repeated the comparison process for the new variables.

Different approaches can be used to estimate the population mean based on the data coming from the proposed design for splitting the questionnaire:

a) HT.AC: The simplest method, by computing the Horvitz-Thompson

estimates (a direct estimate) using the available case (AC) method (estimation based on only the available data from each subquestionnaire). In other words, the direct estimate is computed based on the data collected only from subsamples.

b) GREG.AC: A more efficient method by computing the generalized regression(GREG) estimates (based on a single level model) using the AC method, when the auxiliary information is available. Thus, the information obtained only from the subsamples is used to compute the GREG estimate of the population mean.

c) MI: Estimation based on imputed dataset is a common way to deal with nonresponse caused by splitting questionnaire (Raghunathan and Grizzle, 1995). Therefore, the MI method is used to construct (m) completed datasets and then, all completed dataset are used to compute the direct estimate. The MI method is a general imputation strategy that constructs different sets of completed dataset by replacing the missing values with probable values. In our study, the imputation method to complete dataset is the predictive mean matching (PMM) (Little, 1988). Briefly, the PMM method imputes a missing value by matching its predictive mean to a nearest neighbor among the predictive means of the observed values. The imputation model is linear regression model. The auxiliary variables, $X's$, are used to impute the missing part of each response variable Y.

d) SAE: Our proposed method (described in Section 3) to compute small area estimation (using a multilevel model) to enhance the precision of the estimates, when the auxiliary information is available.

In addition to these approaches, the HT (complete) and GREG (complete) estimates were obtained using complete dataset. The aim is to check how much efficiency is lost when the questionnaire is split comparing to the complete one.

Note that, in this study, we have access to the complete data. Thus, it enables us to investigate the performance of our proposed split questionnaire design. It would be useful to perform a case study to quantify the benefit of splitting the questionnaire in terms of reducing response burden. However, this is out of our scope due to the logistic issue. This subject has been adressed by Adigüzel and Wedel (2008) who construct a field study to

Table 3. The variables from the real dataset used in the study[a].

Variable	Description	Source	Size
X1	Age	Census	N=24070
X2	Household size	Census	N=24070
X3	Current education activity	Census	N=24070
X4	Net expenditure	Original sample	$n = 10\%N$
X5	Income	Original sample	$n = 10\%N$
Y1	*Medical care and health services* expenditure	Subsample I	${n_1}' = 50\%n$
Y2	*Housing* expenditure	Subsample I	${n_1}' = 50\%n$
Y3	*Food* expenditure	Subsample II	${n_2}' = 50\%n$
Y4	*Non-food* expenditure	Subsample II	${n_2}' = 50\%n$

[a]Statistical Centre of Iran

investigate the behavioral effects of providing respondents with split questionnaires.

4.1 Using a Real Dataset

As mentioned earlier, we use a number of variables of the Iran Households Income and Expenditure Survey (HIES) conducted by the Statistical Center of Iran in 2014.

For the purpose of this study, the real dataset is supposed to be a fixed finite population and the performance of our proposed method is studied over repeated samples, drawn from the population. In each replication out of R = 5,000, a new sample is drawn from the fixed finite population according to simple random sampling without replacement (SRSWOR) of size $n = 2407, 1200, 700$ and the population mean has been estimated from each sample.

Following our proposed approach, different sources for variables of the dataset have been considered based on Section 2. Table 3 shows the description of all variables included in the study. Some properties of households ($X1$, $X2$ and $X3$) are available for all units of the fixed population. Additionally, according to the original questionnaire, two important items ($X4$ and $X5$) are considered as the core part and asked from the original sample,

Table 4. The pattern of correlations between core variables and split items.

	Moderate level			High level
	Y1	Y2	Y3	Y4
X4	0.47	0.54	0.58	0.96
X5	0.19	0.50	0.35	0.60

while the rest contains the variables of interest which are assigned into two subquestionnaires. We focus on three variables ($Y1$, $Y2$ and $Y3$) from the survey to estimate the population means of these variables. Additionally, as the correlations between the core variables and the response variables ($Y1$, $Y2$ and $Y3$) are not sufficiently high, an extra response variable ($Y4$) is considered to provide the effect of high correlation on the precision of the estimates.

Half of the original sample is randomly selected to respond to the first subquestionnaire ($Y1$ and $Y2$) and called subsample I. The items in the second subquestionanire ($Y3$ and $Y4$) are asked from the other half of the original sample which is called subsample II. The aim here is to estimate the population means of variables included in each subquestionnaire.

Moreover, it is interesting to see whether the correlations between core variables and split items affect the results. Table 4 illustrates correlation coefficients between these variables. As this table shows, the correlations between core variables and split items are summarized in two levels; i) moderate level of correlation between $X4$, $X5$ and $Y1$, $Y2$ and $Y3$. ii) high level of correlation between $X4$, $X5$ and $Y4$.

As described previously, it is necessary to construct areas based on the grouping variable(s) to apply SAE technique. Here, age ($X1$) and household size ($X2$) are considered as the grouping variables. These two variables are available from the census (refer to Table 3) and consequently, follow the Case I category of the grouping variable(s), introduced in Section 2. The combination of these variables makes 35 areas. The levels of grouping variables are described as:
Age: "15-35", "35-45", "45-55", "55-65", "65 and above", and
Household size: "1", "2", "3", "4", "5", "6" and "More than 6".
Additionally, current education activity ($X3$), net expenditure ($X4$) and income ($X5$) are considered as covariates. Moreover, the number of subsample units belong to each area are unequal (see Figure 1).

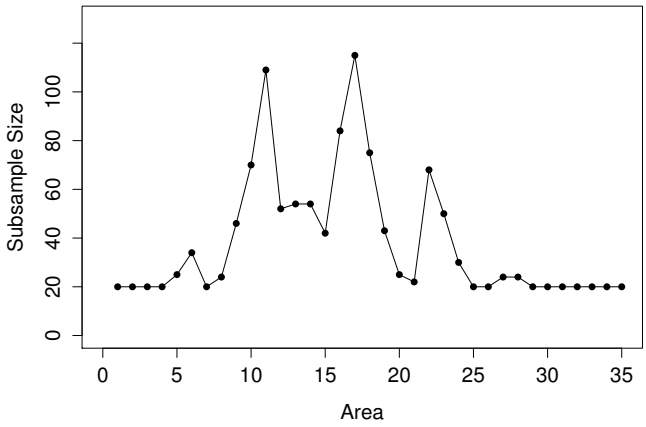


Figure 1. The subsample sizes within each constructed area

4.2 Using a Set of Generated Response Variables

As a long questionnaire, 18 dependent variables, Y, are generated through the following regression model based on independent variables, X.

$$Y = X\beta + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, \sigma^2).$$

Here, we have defined the values for β and σ, to generate Y. In order to investigate the effect of the correlation between X and Y on precision of estimates, three levels of correlation (R^2) have been considered: (I) high ($R^2 = 0.8$), (II) moderate ($R^2 = 0.4$), and (III) ($R^2 = 0.1$); have been considered in the generation process. The independent variables, X, are taken from the real dataset. Thus, a long questionnaire comprised of 18 generated response variables Y, which have been equally split into two subquestionnaires. The auxiliary variables used to generate the response variables and the corresponding generated variables are described in Table 5. In this setup, the areas are constructed, similar to the previous part.

Regarding the average value of correlation between X and Y, each subquestionnaire includes variables with three different levels of correlation. Table 6 shows the pattern of correlation between X and Y with respect to the subquestionnaire consisting of Y.

As before, 5000 Monte Carlo samples of size n (SRSWOR) are drawn

Table 5. A description of auxiliary variables[b] and generated response variables used in the study.

Variable	Description	Source	Size
$X1$	Age		
$X2$	Household size		
$X3$	SEX		
$X4$	Literacy	Census	N=24000
$X5$	Current education activity		
$X6$	Activity situation		
$X7$	Marriage situation		
$X8$	Net expenditure	Original sample	$n = 10\%N$
$X9$	Income		
$Y_{1.1}$	Generated variable 1.1		
$\vdots$	$\vdots$	Subsample I	$n_1{}' = 50\%n$
$Y_{1.9}$	Generated variable 1.9		
$Y_{2.1}$	Generated variable 2.1		
$\vdots$	$\vdots$	Subsample II	$n_2{}' = 50\%n$
$Y_{2.9}$	Generated variable 2.9		

[b]Statistical Centre of Iran

Table 6. The pattern of correlations between auxiliary variables X and Y, with respect to the subquestionnaire.

	High level	Moderate level	Low level
Correlation Value ($\approx$)	0.8	0.4	0.1
Subquestionnaire 1	$Y_{1.1}, Y_{1.2}, Y_{1.3}$	$Y_{1.4}, Y_{1.5}, Y_{1.6}$	$Y_{1.7}, Y_{1.8}, Y_{1.9}$
Subquestionnaire 2	$Y_{2.1}, Y_{2.2}, Y_{2.3}$	$Y_{2.4}, Y_{2.5}, Y_{2.6}$	$Y_{2.7}, Y_{2.8}, Y_{2.9}$

from the finite population to estimate the population mean, using different approaches.

5 Measures of Comparison

It would be useful to introduce some measures of comparison estimated over (R) Monte Carlo repeated samples. Let θ be the true value and $\hat{\theta}$ denote the corresponding estimate.

I. Estimated Relative Bias (ERBias) of $\hat{\theta}$:

$$ERBias(\hat{\theta}) = \frac{(\bar{\hat{\theta}} - \theta)}{\theta}, \tag{6}$$

where $\bar{\hat{\theta}} = \frac{1}{R}\sum_{i=1}^{R}\hat{\theta}_i$,

II. Estimated Root Mean Squared Error (ERMSE) of $\hat{\theta}$:

$$ERMSE(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\theta}_i - \theta)^2}, \tag{7}$$

III. Estimated Relative Root Mean Squared Error (ERRMSE) of $\hat{\theta}$:

$$ERRMSE(\hat{\theta}) = \frac{\sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\theta}_i - \theta)^2}}{\theta}. \tag{8}$$

6 Results of the Study

In this section, we evaluate the performances of estimators constructed under the AC, MI and SAE techniques, using i) real dataset, and ii) generated variables.

6.1 Results: Real Dataset

For the SAE method, we considered a set of auxiliary information ($X1$, $X2$, ..., $X5$). The variables $X1$, $X2$ and $X3$ are available for all units of the fixed population and the variables $X4$ and $X5$ surveyed from the original sample. These auxiliary information have been used to estimate the population mean

of the response variables ($Y1$, $Y2$, $Y3$, and $Y4$). To do this, the small area model (equation 1) was fitted to the data and the method of fitting constants (Searle, et al., 2009) was used to provide variance components estimators, $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$. The next step is to compute the EBLUP estimators using equation (2). Finally, we applied the equation (5) to obtain the overall estimate of the population mean.

As an alternative approach, the Horvitz-Thompson (HT) estimator was computed for the response variables ($Y1$, $Y2$, $Y3$ and $Y4$). This estimator is based on the data collected only from the subsamples (the AC method). Moreover, we used a single level model to compute the GREG estimate, while the auxiliary information is the same as for the SAE estimate.

For the next approach, we applied the MI method to impute the data collected from the subsamples. More precisely, the predictive mean matching was used to create (m=10) completed dataset by maximum of 100 iterations. The imputation model is linear regression model and the independent variables which have been used for the model are the same as the auxiliary variables for the SAE and GREG estimates. The next step is to combine the Horvitz-Thompson estimates, resulting from those (10) completed datasets by the formula $\bar{\theta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$, where K is the number of imputations (Rubin, 1987), to compute the population mean estimate.

As stated earlier in Section 4, we have access to the complete dataset. Thus, it provides an opportunely to investigate the influence of split questionnaire design on the precision of the estimates. For this purpose, the HT and GREG estimates were obtained based on complete dataset (dataset collected from the original sample), as well as the estimates for split questionnaires. These estimates are marked as HT (complete) and GREG (complete).

Table 7 shows the true values of $\theta(Yi),\ i = 1, ..., 4$ (population means of $Yi,\ i = 1, ..., 4$), the estimates of parameters and the corresponding ERMSEs using the HT.AC, GREG.AC, MI and SAE methods for the split questionnaire design, as well as the HT (complete) and GREG (complete), estimates obtained based on the original sample data.

Not surprisingly, the ERMSEs of HT (complete) and GREG (complete) estimates for $Y1$, $Y2$ and $Y3$ are smaller. Because, the simulation study is based on a dataset coming from a general survey, not a split questionnaire survey. Meanwhile, these values are not so far from the ERMSEs of the estimates computed from the split questionnaire design.

Additionally, among the different estimates from the split questionnaire design, the ERMSEs of the SAE estimates are smaller than others, specifi-

Table 7. The true value, estimates of parameters and the corresponding ERMSEs using the HT.AC, GREG.AC, MI and SAE methods using data coming from split questionnaire design and also, the HT.AC and GREG.AC methods for the original sample data (in 10,000 IRR*).

Variable Name	True Value	HT**	GREG**	HT.AC	GREG.AC	MI	SAE
Y1: *Medical care and health services* expenditure	1702	1698 (75)	1704 (76)	1696 (112)	1696 (104)	1694 (100)	1705 (100)
Y2: *Housing* expenditure	4216	4218 (80)	4220 (80)	4219 (117)	4214 (104)	4220 (108)	4224 (96)
Y3: *Food* expenditure	6024	6000 (73)	6026 (69)	6007 (103)	6021 (92)	5995 (96)	6020 (86)
Y4: *Non-food* expenditure	13732	13732 (244)	13747 (245)	13735 (350)	13710 (253)	13675 (257)	13735 (101)

*IRR stands for Iranian Rial.

**HT (complete) and GREG (complete), estimates obtained based on the original sample data.

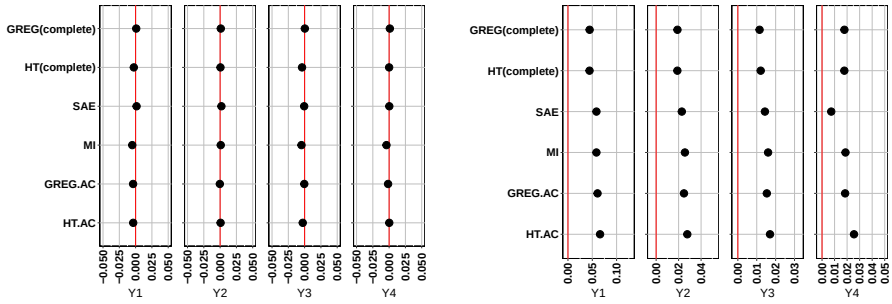


Figure 2. (a) The ERBias and (b) the ERRMSE of estimates for real dataset, with n=2407, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods

cally, for the response variable, $Y4$, in which the level of correlation is high.

The measures of comparisions, i.e., ERBias and ERRMSE (for sample size $n = 2407(10\%N)$) are computed for the HT.AC, GREG.AC, MI and SAE estimates and shown in Figure 2.

Figure 2(a) shows that for all response variables, the estimated relative biases of the SAE estimates are approximately zero and for other estimates, the values are considerably small.

Figure 2(b) indicates that the accuracy of the population mean estimates using the HT.AC, GREG.AC, MI and the SAE methods for the split questionnaire design are similar to those obtained from the original sample data.

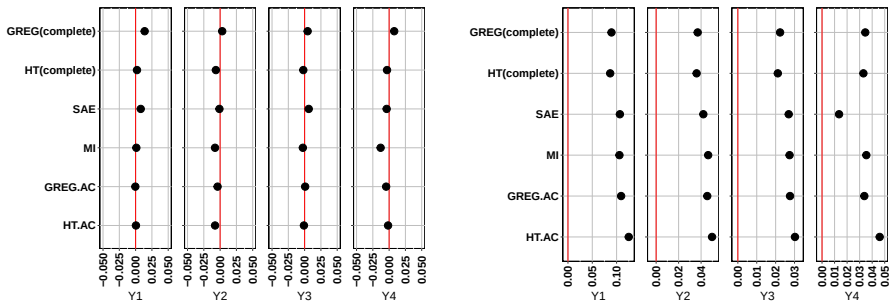
Focusing on the response variable, $Y4$, with high level of correlation, the SAE estimates are more accurate than those for the HT.AC, GREG.AC and MI estimates. For the response variables with moderate level of correlation, the accuracies of the SAE estimates are higher than or equal to the HT.AC, GREG.AC and MI estimates.

As a result, the correlation between the core part and split items has an important role on the efficiencies of the GREG.AC, MI and SAE estimates. Our result indicates that the SAE estimates for the response variables with high level of correlations are noticeably more accurate than the HT.AC, GREG.AC and MI estimates. For those response variables with moderate level of correlation, the accuracies of the SAE estimates are higher than or equal to the other estimates.

Moreover, we investigated the impact of the original sample size on the accuracy of the SAE, MI and AC estimates. For this purpose, we considered two more sample sizes $n = 700$ and $n = 1200$, which are roughly 3% and 5% of the population size, respectively. Figures 3 and 4 show the results of ERBias and ERRMSE for the HT.AC, GREG.AC, MI and SAE estimates for these two sample sizes. These figures demonstrate the same pattern as the case for $n = 2407$ (10% of the population size), i.e., the SAE approach provides better performance than the HT.AC, GREG.AC and MI approaches.

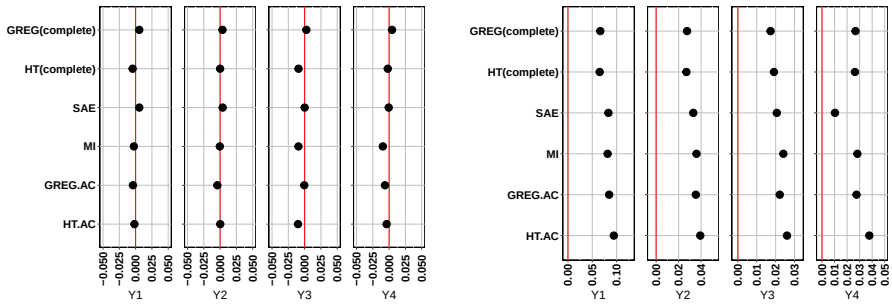
6.2 Results: Generated Variables

In this study, for all generated variables, the population means are calculated based on split questionnaires, using a) HT, b) GREG, c) MI and d) SAE. Additionally, the population means of all generated variables are obtained based on complete questionnaire. We have compared the ERBias and ERRSSE of estimates with $n = 2400(10\%N)$, with respect to the correlation



(a) (b)

Figure 3. (a) The ERBias and (b) the ERRMSE of estimates for real dataset, with n=700, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods



(a) (b)

Figure 4. (a) The ERBias and (b) the ERRMSE of estimates for real dataest, with n=1200, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods

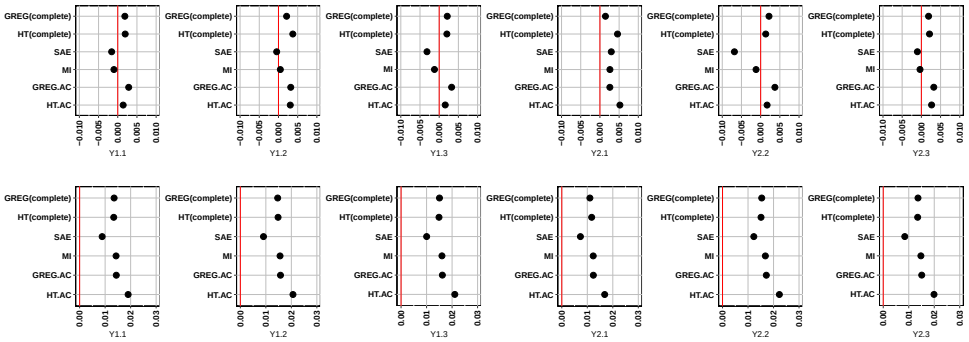


Figure 5. The ERBias and ERRMSE of estimates for generated variables with **high** correlation, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods

levels of variables described in Table 6.

Figure 5 shows that for all variables with high level of correlation, the SAE estimates are more precise than other methods. As it can bee seen in Figure 6, for those variables with moderate level of correlation, the SAE estimates have similar accuracies to other methods. Figure 7 demonstrates that for some variables with low level of correlation, the SAE estimates are less efficient than some methods. Although, the lack of efficiencies of SAE estimates for these variables are not high.

Following the results for both real and generated datasets, similar conclusion is reached. It was expected, because: i) the nonresponse burden caused by the length of a survey is not included in real and generated variables, ii) following the splitting strategy used in our study, there is no overlap between split questionnaires, so, a univariate model has been used for both multiple imputation and SAE methods. In this case, increasing the number of questions does not change the result.

7 Discussion

For the long questionnaire, split questionnaire design is a useful method to reduce nonresponse rate and other non-sampling errors. In this paper, focusing on the issue of insufficient sample size, caused from splitting the questionnaire, we proposed a strategy to split long questionnaire and estimate the

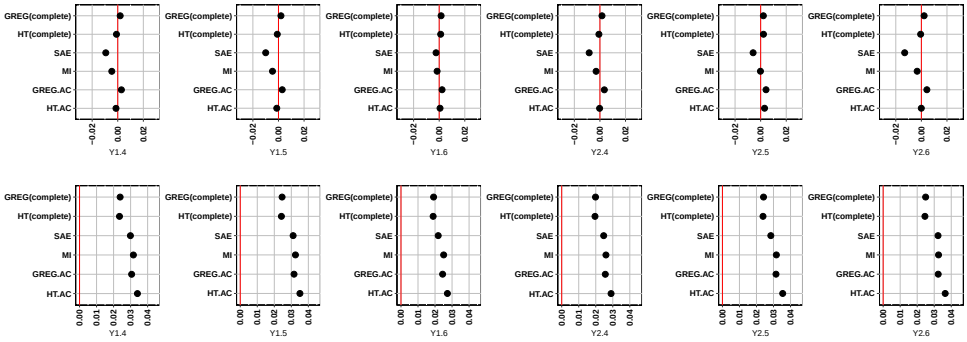


Figure 6. The ERBias and ERRMSE of estimates for generated variables with **moderate** correlation, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods

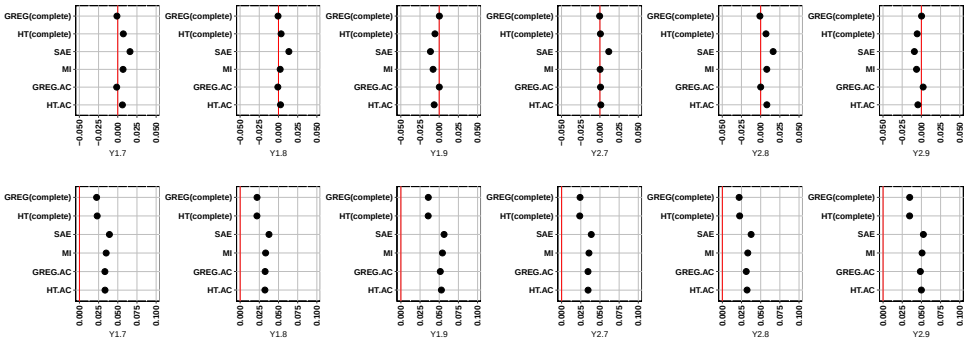


Figure 7. The ERBias and ERRMSE of estimates for generated variables with **low** correlation, based on the HT (complete), GREG (complete), HT.AC, GREG.AC, MI and SAE methods

population parameters based on data collected from subquestionnaires. The proposed method relies on the SAE method. Different cases were discussed with respect to the sources of the auxiliary variables. We investigated the efficiency of our proposed approach compared to the existing methods, e.g. the HT.AC, GREG.AC and MI approaches through a simulation study based on a real data. Moreover, the impact of sample sizes on the accuracy of all estimates was also studied. According to the results, the proposed approach in general achieves a higher accuracy for survey statistics than the existing methods, particularly, for those response variables which have high correlation with the core variables. Obviously, the SAE methods works more efficient if some proper auxiliary variables are applied as covariate. In our study, it is expected to get more accurate SAE estimate when more auxiliary variables are available from the registers or previous census, as the multiple imputation method only applies information from the original sample. This can be examine on a more comprehensive dataset in a future work.
In our study, we encounter an issue to obtain the GREG and SAE estimates, in split questionnaire design, where the information is available from three sources: i) the census, ii) the original sample and iii) the sub-samples. A problem arises in the case when the response variable is available on a subsample and we want to use core variables as covariates in a model. Clearly, the core variables are available only from the original sample and not the census; therefore, the GREG and SAE estimates cannot directly be calculated. As a solution, we have used the mean information of variables available from the original sample instead of census in the SAE method, while the variability of the estimates caused by this solution has been counted in the Monte Carlo estimation of MSE for SAE estimates.

We stress that the SAE model used in this paper is applicable for linear parameters. However, the proposed method could be extended to nonlinear parameter estimates using SAE technique. We also obtained the Monte Carlo (MC) estimation of MSE for all methods. Meanwhile, the precision of all estimates would be provided analytically in the future works. It should be noted that we have compared the performances of our approach with the existing methods in the literature, through a design-based simulation study. It would be beneficial to repeat this exercise using a model-based simulation. We will go over this issue in a future work.

In addition to the missing data caused by splitting the questionnaire, it is possible to have nonresponse in split questionnaires, due to some reasons, e.g. the sample unit cannot be located or does not participate in the survey or

not answer to some items of the corresponding subquestionnaire. In the case that the nonresponse occurs in split questionnaire design, a suggestion would be to consider an imputation method to complete the dataset. However, such solutions lead to decrease the accuracy of resulting estimates. The method presented in this work is applicable to estimate the population means when nonresponse happens in split questionnaires. In practice, the strategy to estimate the population parameters using SAE remains unchanged, while, the sample size reduces. A comparison of the loss of efficiency caused due to applying MI or SAE will be provided in a separate paper.

References

Adams, L.L.M. and Darwin, G. (1982). Solving the Quandary between Questionnaire Length and Response Rate in Educational Research, *Research in Higher Education*, **17**, 231-240.

Adigüzel, F. and Wedel, M. (2008). Split Questionnaire Design for Massive Surveys, *Journal of Marketing Research*, **45**, 608-617.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error-components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association*, **83**, 28-36.

Chipperfield, J.O. and Steel, D.G. (2009). Design and Estimation for Split Questionnaire Surveys, *Journal of Official statistics*, **25**, 227-244.

Chipperfield, J.O. and Steel, D.G. (2011). Efficiency of Split Questionnaire Surveys, *Journal of Statistical Planning and Inference*, **141**, 1925-1932.

Dillman, D.A., Sinclair, M.D. and Clark, J.R. (1993). Effects of Questionnaire Length, Respondent-friendly Design, and a Difficult Question on Response Rates for Occupant-addressed Census Mail Surveys, *Public Opinion Quarterly*, **57**, 289-304.

Gelfand, A.E. (2000). Gibbs Sampling, *Journal of the American statistical Association*, **95**, 1300-1304.

Gonzalez, J.M. and Eltinge, J.L. (2007). Multiple Matrix Sampling: A Review, Proceedings of the Section on Survey Research Methods, American Statistical Association, 3069-3075.

Kullback, S. and Leibler, R.A. (1951). On Information and Sufficiency, *The Annals of Mathematical Statistics*, **22**, 79-86.

Little, R.J.A. and Rubin, D.B. (2014). *Statistical Analysis with Missing Data*, John Wiley & Sons.

Little, R.J.A. (1988). Missing-data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, **6**, 287-296.

McCulloch, C.E. and Neuhaus, J.M. (2001). *Generalized Linear Mixed Models*, Wiley Online Library.

Merkouris, T. (2010). An Estimation Method for Matrix Survey Sampling, *Section on Survey Research Method*.

Militino, A.F., Ugarte, M.D. and Goicoa, T. (2007). A BLUP Synthetic Versus an EBLUP Estimator: An Empirical Study of a Small Area Estimation Problem, *Journal of Applied Statistics*, **34**, 153-165.

Molina, I. and Rao, J.N.K. (2010). Small Area Estimation of Poverty Indicators, *Canadian Journal of Statistics*, **38**, 369-385.

Raghunathan, T.E. and Grizzle, J.E. (1995). A Split Questionnaire Survey Design, *Journal of the American Statistical Association*, **90**, 54-63.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Roszkowski, M.J. and Bean, A.G. (1990). Believe It or Not! Longer Questionnaires Have Lower Response Rates, *Journal of Business and Psychology*, **4**, 495-509.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, CRC press.

Searle, S.R., Casella, G. and McCulloch, C.E. (2009). Variance Components, **391**, John Wiley & Sons.

Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling*, Ballinger.

Statistical Centre of Iran (2011). Iran Statistical Yearbook 1390, Islamic Republic of Iran, Management & Planning Organization, Statistical Centre of Iran, Tehran, **786**.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. and Johnson, C.L. (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey, *Survey Methodology*, **32**, 217.

Saeideh Kamgar
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *saeideh.kamgar@gmail.com*

Hamidreza Navvabpour
Department of Statistics,
Allameh Tabataba'i University,
Tehran, Iran.
email: *h.navvabpour@srtc.ac.ir*