Randomly Split Zones for Samples of Size One as Reserve Replicates and Random Replacements for Nonrespondents

A.C. Singh and C. Ye
Survey and Data Sciences Division, American Institutes for Research, Washington, DC 20007
asingh@air.org; cye@air.org

Abstract

Low response may render a probability sample behave like a nonprobability sample. Achieving high weighted response rate after a small nonresponse follow-up survey may be misleading due to instability in the resulting estimator. Release of many reserve replicate samples helps in reaching the target sample size but relies heavily on correct specification of the nonresponse model so that units from response-prone domains are appropriately weighted. Use of ad hoc substitution by similar units to offset nonresponse is subject to selection bias due to lack of correct selection probabilities. As an alternative, a random replacement strategy for unbiased estimation with appropriate selection probabilities along with a nonresponse model is proposed based on the idea of reserve samples of size one which can be viewed as follow-ups for nonresponding units. It is a take-off from the random group method of Rao, Hartley, and Cochran (RHC, 1962) for probability-proportional-to-size (PPS) sampling where each stratum is randomly split into groups, and then a single unit is drawn within each group. In the proposed method, each stratum is partitioned further into zones formed after sorting for the purpose of implicit stratification so that values of nonresponse predictors used as sorting variables are well distributed over zones. The number of zones is about half the allocated sample size. Each zone is randomly split into groups as in RHC within which replicate samples of size one are selected in order to obtain a responding unit. This way responding units from almost all zones are obtained and then weighted estimates from all responding groups are combined after adjustments for nonresponding groups as well as zones. The nonresponse adjustment is made through a one-step calibration for nonresponse and post-stratification as the usual two-step approach is not applicable because in addition to the information about model covariates for the rejected units, the first step for nonresponse adjustment requires selection probabilities for each given sequence of nonresponding units before obtaining a responding unit within a group, and these probabilities are not known. Due to relatively well distribution of responding units over the range of covariate values, the calibration for nonresponse is expected to provide robust estimation with respect to nonresponse bias even if the model is misspecified. The unit level response rate remains low and is not altered by the new design, but the notion of a group response rate becomes meaningful which can be made high by choosing suitably the number of replicate release within the data collection time frame and the budget allowed. Simulation results are presented to illustrate the nonresponse bias reduction property of the proposed estimator and the robustness of its mean squared error under misspecified models.

Key Words: Random group method; Reserve Sample Replicates; Unit vs. Group Response Rates; Weighted Response rate; Nonresponse follow-up surveys.

1. Introduction

High nonresponse is quite common in many surveys (especially with telephone and mail) and there is the growing concern among survey practitioners that a probability sample may behave like a nonprobability sample. This problem can be mitigated only marginally using innovations in questionnaire design, interview protocol, and use of incentives. In practice, there are three major approaches to the nonresponse problem listed below along with their limitations:

(i) Use of a Nonresponse Follow-Up Survey (NRFUS) to increase the weighted response rate but it may be misleading because of the high variability of sampling weights resulting from the fraction of the follow-up subsample being small due to budgetary constraints. This, in turn, makes the estimator quite unstable.
(ii) Release of many reserve replicate samples helps in reaching the target sample size but it puts a lot of burden on model-based adjustment for nonresponse bias because the respondents may be concentrated more in response-prone domains and not well dispersed over the range of values of auxiliary variables used in the model.
(iii) Use of ad hoc substitution by similar units to offset nonresponse is subject to selection bias because the choice of units for substitution is not based on any random mechanism designed for unbiased estimation.

In view of the above concerns about ways to reach the goal of meeting the target number of completes, there is clearly a need for an alternative to the traditional method of inflating the released sample size to compensate for ineligibility and nonresponse which is typically followed up by release of reserve replicates when faced with a lower number of completes than expected. NRFUS is also generally not a viable option due to cost constraints. A natural option is to develop ways in which substitution for nonrespondents by similar units can be justified. Clearly, there is need to substratify strata into zones (or deep strata) by good anticipated nonresponse predictor variables in addition to the variables used for explicit stratification so that each zone is represented in the sample of completes. These zones can be created by using the additional nonresponse predictors as sorting variables for implicit stratification in systematic sampling. Both explicit and implicit stratification variables are deemed to be correlated with the outcome as well as response indicator variables. Within each zone, we can use rejective sampling by repeated random draws with replacement to obtain a desired sample size of distinct respondents. Here we reject nonrespondents in favor of a respondent in the sense that since we don't know in advance the subpopulation of respondents for sampling, we sample from the larger known population (of respondents and nonrespondents) and resort to rejection as and when necessary. However, in practice, draw by draw selection to find an allocated number of responding units in each zone would be an onerous task and impractical due to time and budgetary constraints in data collection. Moreover, this will not be conducive for unbiased variance estimation in general. Alternatively, the rejective sampling strategy can be relatively easily implemented with samples of size one where replicates correspond to reserve releases. This is where the method of random groups comes in; see Rao, Hartley, and Cochran (1962, RHC for short) and Cochran (1977, pp. 266).

The RHC method was originally developed for providing a simplified PPS selection in which primary sampling units (PSUs) in a stratum are split into random groups of about equal size. The number of groups corresponds to the desired number of sampled PSUs, and one PSU is drawn at random from each group. It is also useful for replacing retiring PSUs with new ones in rotating partially overlapping panel surveys. By analogy between a retiring unit and a nonresponding unit, RHC can be adapted to replace nonresponding units with responding units. The purpose of this paper is to generalize RHC under the full sample case (i.e., no nonresponding PSU) to the case of a respondent subsample (which may come from single stage design with no PSUs) in order to obtain an unbiased estimate under a quasi-randomization model where nonrespondents are replaced at random by respondents. This problem arose in the context of education surveys where schools are typically stratified by school type, urbanicity, enrollment size, and percentage of Native Americans; and each stratum is further implicitly stratified by sorting variables such as % white non-Hispanic, and % eligible for free or reduced lunch among others. Here, in the first phase, schools are PSUs which are selected using PPS with student enrollment as size measures, and can be nonrespondents. The second phase units are students or teachers within selected schools. In some education surveys, nonresponding schools are substituted by neighboring schools in the sorted list within each stratum, and the corresponding selection probabilities are adjusted in an ad hoc manner by the new enrollment sizes. The substitution and the associated weight adjustment do not have any theoretical justification but do reflect the selection probabilities had the substituted unit been drawn in the first place. This ad hoc substitution may not be serious if the nonresponse rate is low but in recent times, surveys are experiencing high nonresponse.

The proposed method termed randomly split zones (RSZ) for samples of size one partitions each stratum into approximately equal sized zones via implicit stratification where the number of zones is set equal to half the allocated stratum sample size. Random groups of about equal size are then created within zones (or deep strata) and one unit is drawn at random from each group along with replacements if necessary. In Section 2, a brief review of the RHC method is presented along with a motivation for the proposed RSZ method. Section 3 contains a description of RSZ followed by Section 4 on point and variance estimates when response probabilities are assumed to be known and under the more realistic scenario when response probabilities are unknown and estimated from the sample under a nonresponse model. Empirical results based on a simulation study are presented in Section 5 where equal probability selection methods (simple random sample, systematic random sample, and RSZ) are compared under a single stage unstratified design. Finally, Section 6 contains concluding remarks and a new application of RSZ for controlling the sample overlap among multiple cross-sectional surveys.

2. Background Review and Motivation

We first review briefly the RHC method for a simplified PPS selection of PSUs. For our purpose, instead of splitting strata into random groups, it is better to split zones into random groups where zones (or substrata) partition the strata via implicit stratification. Also for illustrating RHC, it is sufficient to consider a single zone *i* (out of a total of *H* zones or substrata) which is randomly split into approximately equal-sized groups of PSUs; the number of groups being n_i, the size of the sample. Let N_i be the size or the number of PSUs for zone *i* and let N_{ij} be the size or the number of PSUs for the *j*th random group (*j*=1 to n_i), and x_{ijk} be the PPS size measure for the *k*th PSU in the *j*th random group of the *i*th zone. Now to draw a PPS sample of n_i PSUs from the *i*th zone, one PSU (denote by k_{ij}) is selected using PPS from each group. The *i*th zone population total T_{yi} (= $\sum_{j=1}^{n_i} T_{yij}$) of the study variable *y* for the *i*th zone is estimated by

$$t_{yi} = \sum_{j=1}^{n_i} t_{yij}, \text{ where } t_{yij} = y_{ijk_{ij}} \left(x_{ij+}/x_{ijk_{ij}}\right) \text{ from the selected PSU } k_{ij}. \qquad (2.1)$$

Conditional on a given random split (denote the expectation operator under the first phase randomization by E_1), t_{yij} is unbiased for T_{yij} under the second phase randomization of PPS selection (denote the expectation operator here by E_2), and, therefore, t_{yi} is unbiased for T_{yi} under the two stage randomization E_{12}. Moreover, $V_1E_2\left(t_{yi}\right) = 0$. Now using PPS results for samples of size one, we have

$$\begin{aligned} V_2\left(t_{yij}\right) &= \sum_{k=1}^{N_{ij}} \left(x_{ijk}/x_{ij+}\right) \left(y_{ijk}\left(x_{ij+}/x_{ijk}\right) - T_{yij}\right)^2 \\ &= \sum_{k<k'}^{N_{ij}} \left(x_{ijk}/x_{ij+}\right)\left(x_{ijk'}/x_{ij+}\right) \left(y_{ijk}\left(x_{ij+}/x_{ijk}\right) - y_{ijk'}\left(x_{ij+}/x_{ijk'}\right)\right)^2 \end{aligned} \qquad (2.2)$$

Since probability of any two units (*k, k'*) belonging to the same random group *j* in the *i*th zone is $\left(N_{ij}/N_i\right)\left(N_{ij}-1/N_i-1\right)$, and denoting it by p_{ij}, we have the unconditional variance

$$\begin{aligned} E_1V_2\left(t_{yi}\right) &= \sum_{j=1}^{n_i} p_{ij} \sum_{l<l'}^{N_i} \left(x_{ijl}/x_{ij+}\right)\left(x_{ijl'}/x_{ij+}\right) \left(y_{ijl}\left(x_{ij+}/x_{ijl}\right) - y_{ijl'}\left(x_{ij+}/x_{ijl'}\right)\right)^2 \\ &= \left(\sum_{j=1}^{n_i} p_j\right) \left(\sum_{l<l'}^{N_i} q_{il}q_{il'} \left(\frac{y_{il}}{q_{il}} - \frac{y_{il'}}{q_{il'}}\right)^2\right) \\ &= \left(\left(\sum_{j=1}^{n_i} N_{ij}^2 - N_i\right)/N_i(N_i-1)\right)\left(\sum_{l=1}^{N_i} q_{il}\left(\frac{y_{il}}{q_{il}} - T_{yi}\right)^2\right) \end{aligned} \qquad (2.3)$$

where $q_{il} = x_{il}/x_{i+}$. The minimum value is obtained when all the N_{ij}'s are equal to a common value N_{i0}. Then the V(t_{yi}) is given by the familiar PPS with replacement formula $(1/n_i)\sum_{l=1}^{N_i} q_{il}\left(\frac{y_{il}}{q_{il}} - T_{yi}\right)^2$ except for the reduction factor $(1-(n_i-1)/(N_i-1))$. The RHC yields approximate PPS selection probabilities if the total group size measures x_{ij+} for different groups are approximately equal within a zone *i*. This slight

relaxation in the PPS requirements allows for considerable simplicity. In particular, an important property of the RHC method is that V(t_{yi}) admits an exact unbiased variance estimate given by

$$v(t_{yi}) = \left((\sum_{j=1}^{n_i} N_{ij}^2 - N_i)/(N_i^2 - \sum_{j=1}^{n_i} N_{ij}^2)\right)\left(\sum_{j=1}^{n_i}(\sum_{k'=1}^{N_{ij}} q_{ijk'})\left(\frac{y_{ijk_{ij}}}{q_{ijk_{ij}}} - t_{yi}\right)^2\right) \qquad (2.4)$$

where $q_{ijk} = x_{ijk}/x_{i++}$ and is identical to q_{il} if jk corresponds to the index variable l, where k_{ij} is the randomly selected PSU from the group ij. The above results for a single stage design can be generalized to multi-stage or multi-phase designs.

We need to generalize RHC to the problem of finding random replacements for nonrespondents within zones (or deep strata) where units are similar with respect to explicit and implicit stratification variables—these are deemed to be good predictors of nonresponse. Here the underlying design could be unequal probability (PPS) or equal probability design as provided by RHC but there is the additional goal of being able to draw alternate units from the random group with known selection probabilities to serve as replacements. It is natural to look for respondents within a random group as replacements because units within a zone are similar. This does not imply that nonresponse adjustments would not be needed because although units are similar, they still would have differential response probabilities. To this end, we will assume a population response model as in Fay (1991) in which a response indicator R_k is assigned to each unit *k* in the universe *U* which takes the value of 1 with probability φ_k when the unit is respondent and 0 when nonrespondent. It is also assumed that given known auxiliary variables (deemed good predictors for response), the $R_k's$ are independent of the study variables y_k's and that units respond independently. Thus under the joint $\pi\varphi$ −randomization where π denotes the random sampling mechanism with selection probabilities π_k for inclusion of the *k*th unit in the sample, and φ denoting the random response mechanism, we have the standard estimator $\sum_{k\in U} y_k R_k I_k/\varphi_k\pi_k$ based on the respondent subsample as an unbiased estimator of the population total T_y.

Now for the proposed generalization of RHC to RSZ, we need to specify the number of equal-sized zones partitioning each stratum and the number of groups per zone. The number of zones is set equal to half the allocated sample size to the stratum so that there are at least two random groups per zone needed for variance estimation. The number of random groups per zone depends on the inflated sample size based the anticipated response rate at the sample design stage so that the total number of sample cases released in stages (the initial stage and through replicate release) within the data collection time frame match the total number of cases in one stage in traditional designs. This is determined using a geometric series formula as shown below. For an unstratified design, let n_0 the target number of completes, *q* the anticipated completion rate reflecting unit eligibility and interview response, and *R* the number of replicate release per group including the initial release, then the constant number n_i of groups per zone is given by

$$n_i H(1-(1-q)^R/q = n_0/q$$

or

$$n_i = n_0/H(1-(1-q)^R) \qquad (2.5)$$

In practice, some rounding up would be needed to obtain an integer value for the number of groups. It may be remarked that the feature of random replacements for nonrespondents within each group under RSZ leads to the concept of group response rate which can be made higher depending on the number of replicate release. This is in contrast to the traditional unit level response rate whose level is not under control of the sampler. With this motivation, the proposed RSZ design is illustrated in detail in the next section.

3. RSZ: The Proposed Design

RSZ(*R*) can be described in the following steps where *R* is the number of stages of release.

Step I: Partition the universe *U* into strata and allocate sample to each stratum.

Step II: Partition each stratum further into equal-sized zones after sorting. The number of zones is set equal to half the allocated stratum sample size.

Step III: Specify the total number R of release (e.g., R=5) and define equal number n_i of groups per zone within a stratum using the relation $n_i = n_0/H(1-(1-q)^R)$.

Step IV: Stage-wise release of new reserve samples of size one from remaining nonresponding groups from each stage.

As an illustrative realistic but hypothetical example, consider an unstratified design for household surveys in the United States with the total number of housing units (HUs) in US the population being about 131 million. In RSZ, approximately equal sized zones (like deep strata) are created by sorting on implicit stratification variables such as census block level information about household income, number of children, marital status, age group and housing tenure which can be obtained from a vendor such as CLARITAS Inc. Suppose the target sample size of n_0 is 4500 so that the total number H of zones is 2250 and the zone size is approximately 58,222 HUs. Next, each zone is randomly split into groups within which replicate samples of size one are selected. If there was no nonresponse, then only two random groups are needed per zone to meet the target sample size and for unbiased variance estimation. However, the number n_i of groups per zone is inflated to account for the completion rate which is typically a product of the cooperation rate (such as 50%) for the main questionnaire, screener eligibility (such as 33%), and screener response in conjunction with the validity of the HU being residential (such as 50%). Thus, for our illustration, the completion rate q is 1/12 or 8.3%. In RSZ, the inflated sample can be released in stages as replicate samples of size one from each incomplete random group after interim review of remaining target completes. In practice, the number of such release is constrained by the data collection timeframe and cost. Suppose the total number R of replicates including the original release feasible in the timeframe is 5. Then the number n_i of groups per zone can be easily obtained using the formula (2.5) as 5.7 for our example. After rounding up to 6 groups per zone, the number of HUs per group is about 9704 and the total number of released cases or HUs in Stage I is 13,500; see Figure 1 for a schematic representation of RSZ and Table 1 for stage-wise distribution of total released cases, expected incompletes and completes. It also shows the distribution of completes and incompletes over the five release stages when the completion rate is reduced by half to 4.2% which might happen if the screener response and HU validity rate is reduced to 25%. The value of n_i in this case increases to 10.53 or 11 assuming the same number of stages of release.

It might be of interest to note that in RSZ, it is advantageous to release random replicate sample cases in stages to the extent possible within the timeframe for data collection in order to obtain completes essentially from each and every zone and, as a result, making the final sample representative of the population like the initially designed sample. Stage-wise release also allows for interim analysis so that in the intermediate stages, a random subsample of cases from incomplete groups can be released to reduce excess completes. However, assuming the anticipated response rate does not change considerably over the collection period, there is no such advantage under the traditional approach because there the sample inflation is not governed by zone representation, Therefore, the inflated sample of cases is designed to be released in a single stage such that the target is achieved in expectation. This may result in excess or shortage of desired completes. If completes are less than desired, reserve replicate samples are released which need to be planned in advance so that they can be integrated with the initial sample release for estimation purposes. Under RSZ, however, there is no such need for advance planning of reserve sample release in view of readily available replicate samples of size one from each random group.

4. Point and Variance Estimation

It would be useful to summarize first the key points underlying the RSZ design when dealing with high nonresponse.

1. (No Nonresponse) For unbiased estimation, random sampling is needed to obtain a representative sample of the population. For efficient estimation, often stratification (explicit and implicit) is employed using auxiliary variables deemed correlated with several outcome or study variables. In the absence of nonresponse; i.e., in the full sample case, and in the absence of noncoverage, there is no bias in the usual estimators. However, their efficiency can be further improved by calibration for post-stratification whereby sampling weights are adjusted so that sample estimates for post-stratification variables perfectly match the

known population totals. This adjustment also has the additional benefit of coverage bias reduction if the sampling frame had either over- or under-coverage imperfections. Sampling weight calibration for post-stratification can be achieved by different methods under the class known as generalized raking such as linear, log linear, and their range-restricted analogues (Deville and Särndal, 1992) but they give similar results for large samples.

2. (Low Nonresponse) In practice, nonresponse is almost always present despite incentives. For this reason, the target sample size is inflated in light of the anticipated response rate. The realized subsample of respondents is likely to be skewed toward response-prone domains defined by auxiliary variables deemed correlated with response indicator as well as the study variable. Under a nonresponse model, sampling weights are adjusted but the unbiased of estimates under the joint sampling design-nonresponse model depends on the correct specification of the model. Although the nonresponse is difficult to validate, the nonresponse bias in the estimate is not expected to be serious if the nonresponse rate is low and the model has good response predictors. However, if nonresponse is high, the bias could be serious unless the model can be correctly specified (Groves, 2006).

3. (High Nonresponse) RSZ provides a new way of replacing nonrespondents at random by selecting a responding unit from each group after several draws if necessary. The unconditional selection probabilities for the responding unit in a random group regardless of units rejected before is the same as the selection probability at the first draw which is easily justifiable and computable. Although in RSZ, the nonresponse problem is considerably reduced by making several attempts to get a respondent from each group, some groups are likely to remain nonresponding while units from all responding groups are likely to be skewed toward response-prone units because units may have differential response probabilities although they are from the same zone. Nevertheless, a relatively higher number of zones would be represented in the respondent subsample and therefore after a suitable nonresponse adjustment, RSZ is expected to be robust to nonresponse model misspecifications.

4. (One Step Sampling Weight Adjustment for Nonresponse) With RSZ, traditional methods for nonresponse adjustment are not applicable because selection of additional units within a group depends on whether the previously drawn unit responds or not and hence their selection probabilities are unknown due to unknown response probabilities. However, the calibration method for nonresponse adjustment (Folsom and Singh, 2000; see also Kott, 2006 and Särndal, 2007, and Haziza and Lesage, 2016) works with only the respondent subsample and population control totals (or their reliable estimates) for the auxiliary variables in the model. In this case, since only responding units from each group contribute in the estimating equations for model parameters, it is sufficient to work with unconditional selection probabilities for responding units from different groups. Thus, if the group response rate for RSZ is not too low, there is less dependence on the model for bias adjustment and the fact that the calibration method adjusts weights so that the estimator with adjusted weights can reproduce perfectly the known population totals for model covariates, the RSZ estimator after nonresponse adjustment is expected to be robust to model misspecifications.

We now can derive expressions for point and variance estimates under RSZ. If the response probabilities φ_k's were known, then denoting $y_k R_k/\varphi_k$ by z_k, the RSZ estimator after the nonresponse adjustment for the total T_{zi} for zone i is given by $t_{zi} = \sum_{j=1}^{n_i} t_{zij}$ and its variance $V_{\pi|\varphi}$ is analogous to the expression in (2.3) when y is replaced by z. The unconditional variance $V_{\pi\varphi}(t_{zi})$ is given by

$$V_{\pi\varphi}(t_{zi}) = E_\varphi V_{\pi\varphi}(t_{zi}) + V_\varphi(T_{zi}) \qquad (4.1)$$

where the first term can be unbiased estimated analogous to (2.4) and the second term is of much smaller order if the total number $n_i$of groups in the zone i is much smaller than the population size N_i and hence negligible.

The point estimator for the total T_z and its variance readily follows by summing over all zones. Under the more realistic scenario of unknown φ_k's, the estimating equations for model parameters γ under a commonly used inverse logit model $\varphi_k(\gamma) = 1 + e^{-x_k'\gamma}$ are given by

$$\sum_{i=1}^{H_r} \sum_{j=1}^{n_{ir}} (x_k/\pi_k)\left(1 + e^{-x_k'\gamma}\right) = T_x \qquad (4.2)$$

where n_{ir} denotes the total number of responding groups within zone *i* and H_r denotes the total number of responding zones. The above equations are admissible if the sample weighted totals t_x of x's are less than the population totals T_x. This is needed for the adjustment factors to be greater than 1. In practice, t_x for some x may not be less than T_x due to extreme initial weights and initial smoothing of weights (Singh, Ganesh and Lin, 2013) can be used to overcome this problem as an alternative to weight trimming. For variance estimation, the RSZ estimator with estimated γ can be Taylor linearized and then the variance estimator discussed above for known φ_k's can be used. Alternatively, an improved estimator using a sandwich formula (Singh and Folsom, 2000) can be obtained.

5. Simulation Results

A limited simulation study was conducted to test performance of RSZ with a single stratum in relation to simple random sampling (SRS) and systematic random sampling (SYS). Using the Common Core of Data (CCD) School District Finance survey School Year 2012-13, we considered the total federal funding (in millions) for a school district as the study variable *y* and the total district enrollment ((in thousands) as the auxiliary variable *x*. The CCD has 15471 school districts with positive values of *y* and *x*. Due to skewed nature of distributions of *y* and *x*, we consider the log transformation and assume that the joint distribution of log *y* and *log x* is bivariate normal for generation of the finite population. The mean and standard deviation of log *y* and *log x* were obtained respectively from CCD as (-.302, 1.665) and (-.124, 1.560) and the correlation as .853. This completely specifies the bivariate normal distribution and hence the linear regression of log *y* on *log x*. First 10000 values of log *x* were generated and then the corresponding values of log *y* using the regression model and normal errors. With 10000 pairs of values of (*y,*x), the target parameter T_y is 31435.29 in million dollars and the control total T_x is 31606.80 in thousand students. The nonresponse was induced via Poisson sampling with response probabilities given by a logistic model using *x* as a covariate. The slope parameter was set to 1 while the intercept was set empirically to obtain mean response rates *q* of .2, .4, .8 respectively for three scenarios of low, medium and high response rates. For each of the three sampling designs, SRS, SYS, and RSZ, three sample sizes *n*=100, 200, 400 were considered which correspond to the total number of released cases. Thus, with *q* =.20, the target number of completes is 20, 40, and 80 respectively for *n*=100, 200, 400. For RSZ, we considered two versions: RSZ(5) which allows for 5 releases and RSZ(U) with unrestricted number of releases within each group. The nonresponse adjustment was performed under three misspecified models: (i) Simple Hajek-ratio adjustment to ensure sampling weights of respondents sum to *N*, (ii) Linear regression model for the adjustment factor which does not ensure the adjustment factor is positive, and (iii) Log linear model for the adjustment factor which ensures the adjustment factor remains positice.

None of the nonresponse model is correctly specified but model (iii) comes closest except that it is not logit linear and has both intercept and slope parameters unknown. Model (ii) comes next except that it is linear and model (i) ranks last in terms of being close to the true nonresponse model because it does not even depend on *x*. With 1000 simulated samples from the same finite population, it was found that RSZ estimator was quite robust in terms of bias and MSE with respect to misspecified models but the other two methods SRS and SYS were quite sensitive being worst for model (i) but quite well for model (iii); see Tables 2(a,b,c), 3(a,b,c) and 4(a,b,c).

6. Concluding Remarks and a New Application of RSZ

In this paper, a generalization of the RHC method for random replacement of nonrespondents in the presence of high nonresponse was developed which was different from the original purpose of RHC. Nonresponse adjustments to RSZ for nonresponding groups and nonresponding zones were suggested via calibration. It was found based on a limited simulation study that RSZ was quire robust with regard to model misspecification in comparison to SRS and SYS in terms of bias and MSE. It is remarked that RSZ can also be used for sampling on successive occasions by using Keyfitz (1951) for updating random groups.

It may be of interest to consider a possible new and important application of RSZ for controlling sample overlap. With multiple cross-sectional surveys that are also repeated over time, a natural question to consider is how to select PSUs (such as schools in education surveys) at the first phase in a coordinated manner across different surveys such that the overlap of PSUs can be controlled cross-sectionally and also over time. Having such a control would help in distribution of workload in an equitable manner across PSUs (schools) and in reducing response burden on any given school in that a school can be given the option of time out after having participated in a number of surveys. Also with any repeated survey over time, having a partially overlapping design is especially useful in an efficient estimation of trend. The problem of overlap control in sampling, also known as collocated sampling, is difficult in general even for simple random samples because suitable random selection for each survey needs to be maintained for unbiased point and variance estimation after collocation; see Ernst, Valliant, and Casady (2000) and Ohlsson (2000). However, it turns out that with RSZ, overlap control of schools can be easily implemented by considering the analogy between nonresponding units and units already in use by other surveys. The basic idea can also be extended to the second phase of second stage of units within PSUs such as teacher selection within selected schools. It is planned to investigate further the application of RSZ to the problem of sample overlap control.

References

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Ed., New York: John Wiley

Deville, J.-C. and Särndal, C.-E.(1992). Calibration estimators in survey sampling. *JASA*, 87, 376-382

Ernst, L., Valliant, R., and Casady, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths, Journal of Official Statistics, Vol. 16, No. 3, 211-228

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual* Research *Conference,* pp. 429-440

Folsom, R.E. Jr., and Singh, A.C. (2000). A Generalized Exponential Model for Sampling Weight Calibration for a Unified Approach to Nonresponse, Poststratification, and Extreme Weight Adjustments. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 598-603.

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

Haziza, D. and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Keyfitz, N. (1951). Sampling with probability proportionate to size: adjustments for changes in probabilities, JASA, 46, 105-109.

Kott, P. (2006). Using calibration weighting to adjust fir nonresponse and undercoverage. *Survey Methodology,* 32, 133-142

Ohlsson, E. (2000). Coordination of PPS samples over time. In Business survey Methods, New York: Wiley, 255-264

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, Series B, 24, 482-491.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119

Singh, A.C. and Folsom, R.E. Jr. (2000). Bias Corrected Estimating Functions Approach for Variance Estimation Adjusted for Poststratification. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 610-615.

Singh, A.C., Ganesh, N., and Lin, Y. (2013). Improved sampling weight calibration by generalized raking with optimal unbiased modification. *Proceedings of the American Statistical Association, Survey Research Methods Section, 3572-3583*

Figure 1: A Simplified Schematic Representation of the Proposed RSZ Design

($N = 131 \times 10^6$, n_0=4500, $R = 5$, $q = 1/12$, $n_i = 6$)

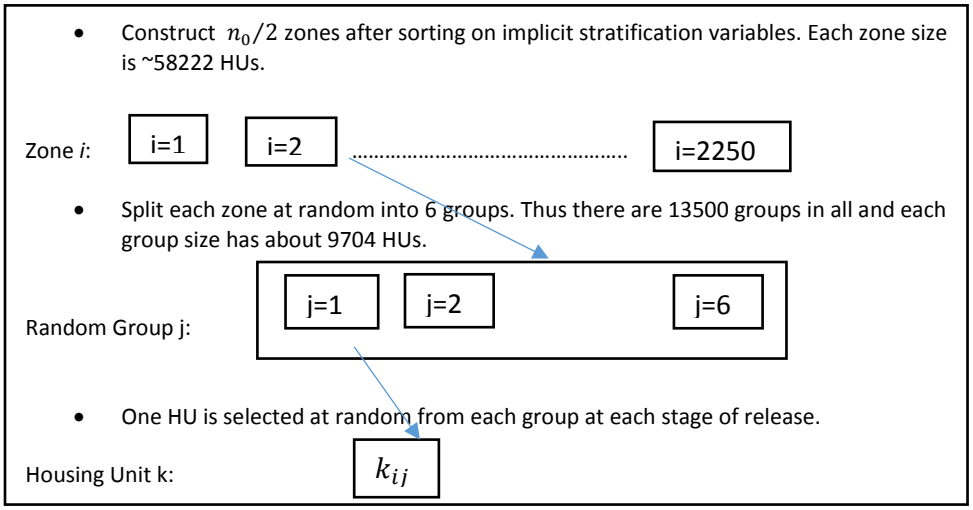

Table 1: Distribution of Number of Released Cases, Expected Incompletes and Completes

Stage	$q = 1/12$			$q = 1/24$		
	# Released	# Incompletes	# Completes	# Released	# Incompletes	# Completes
1	13500	12375	1125	24750	23719	1031
2	12375	11344	1031	23719	22731	988
3	11344	10399	945	22731	21784	947
4	10399	9532	867	21784	20876	908
5	6750	6188	562	15750	15094	656
Total	54368	49838	4530	108734	104204	4530

Table 2(a): Comparison of Estimates for Population Totals (q=.20, Model (i))

Evaluation Criterion	Expected Samp. Size (released)	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	2.431	2.472		-0.095	0.052
	200	2.411	2.478		-0.096	0.014
	400	2.483	2.464		-0.099	0.019
RRMSE	100	2.975	2.971		0.610	0.829
	200	2.672	2.693		0.534	0.516
	400	2.632	2.571		0.344	0.370

Table 2(b): Comparison of Estimates for Population Totals (q=.20, Model (ii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.230	0.239		0.128	0.047
	200	0.341	0.272		0.112	0.039
	400	0.308	0.353		0.120	0.028
RRMSE	100	1.060	1.017		0.625	0.501
	200	0.890	0.931		0.449	0.387
	400	0.770	0.765		0.429	0.269

Table 2(c): Comparison of Estimates for Population Totals (q=.20, Model (iii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.090	0.075		0.083	0.048
	200	0.102	0.104		0.081	0.037
	400	0.098	0.100		0.082	0.027
RRMSE	100	0.342	0.325		0.413	0.511
	200	0.254	0.256		0.279	0.389
	400	0.183	0.186		0.211	0.270

Table 3(a): Comparison of Estimates for Population Totals (q=.40, Model (i))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	1.057	1.076		-0.026	0.024
	200	1.039	1.073		-0.029	0.006
	400	1.084	1.076		-0.010	0.002
RRMSE	100	1.381	1.349		0.483	0.557
	200	1.200	1.190		0.324	0.346
	400	1.176	1.139		0.238	0.234

Table 3(b): Comparison of Estimates for Population Totals (q=.40, Model (ii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.071	0.074		0.040	0.030
	200	0.152	0.135		0.043	0.025
	400	0.145	0.160		0.041	0.014
RRMSE	100	0.211	0.221		0.307	0.359
	200	0.329	0.345		0.233	0.278
	400	0.300	0.299		0.164	0.194

Table 3(c): Comparison of Estimates for Population Totals (q=.40, Model (iii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.068	0.067		0.039	0.028
	200	0.084	0.082		0.040	0.025
	400	0.081	0.082		0.038	0.014
RRMSE	100	0.207	0.215		0.306	0.361
	200	0.174	0.168		0.231	0.280
	400	0.130	0.131		0.160	0.195

Table 4(a): Comparison of Estimates for Population Totals (q=.80, Model (i))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.202	0.203		0.007	-0.007
	200	0.188	0.205		-0.006	-0.003
	400	0.209	0.207		0.000	0.002
RRMSE	100	0.503	0.454		0.359	0.326
	200	0.364	0.331		0.225	0.222
	400	0.311	0.277		0.158	0.156

Table 4(b): Comparison of Estimates for Population Totals (q=.80, Model (ii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.038	0.025		0.026	0.024
	200	0.040	0.046		0.015	0.021
	400	0.037	0.038		0.004	0.007
RRMSE	100	0.224	0.211		0.282	0.260
	200	0.163	0.161		0.194	0.194
	400	0.119	0.112		0.136	0.133

Table 4(c): Comparison of Estimates for Population Totals (q=.80, Model (iii))

Evaluation	Expected Sample size	Sampling Scheme				
		SRS	SYS		RSZ5	RSZU
Relative Bias	100	0.037	0.027		0.026	0.024
	200	0.035	0.043		0.015	0.021
	400	0.034	0.035		0.003	0.007
RRMSE	100	0.220	0.214		0.284	0.262
	200	0.159	0.158		0.195	0.194
	400	0.113	0.108		0.136	0.133