AIR/Office of Scientific Quality and Innovation (OSQI)
Methods Working Paper 2016-01

American Institutes for Research

Survey and Data Sciences Division

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

**Avinash C. Singh**
**Cong Ye**

OCTOBER 2016

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

**October 2016**

**Avinash C. Singh,** Institute Fellow, asingh@air.org
**Cong Ye,** Senior Researcher, cye@air.org

*AIR/ OSQI Methods Working Papers describe work in progress on methodological issues in research, evaluation, policy, practice, and systems change.*

*Challenges in securing district participation in the 2016 C-SAIL (Center for Standards, Alignment, Instruction, and Learning) survey provided the impetus and the first application of the new RSZ design. We would like to thank Mike Garet of AIR for his support and encouragement.*

*An earlier version of this paper was published in the 2016 JSM Proceedings of Survey Research Methods Section.*

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000
**www.air.org**

# Contents

**Page**

Abstract .......... 1

1. Introduction .......... 1
2. Background Review and Motivation .......... 3
3. RSZ: The Proposed Design .......... 6
4. Point and Variance Estimation .......... 7
5. Simulation Results .......... 9
6. Concluding Remarks and a New Application of RSZ .......... 10

References .......... 12

# Tables

**Page**

Table 1. Distribution of Number of Released Cases, Expected Incompletes and Completes .......... 13

Table 2a. Comparison of Estimates for Population Totals ($q$ = .20, Model (a)) .......... 14

Table 2b. Comparison of Estimates for Population Totals ($q$ = .20, Model (b)) .......... 14

Table 2c. Comparison of Estimates for Population Totals ($q$ = .20, Model (c)) .......... 14

Table 3a. Comparison of Estimates for Population Totals ($q$ = .40, Model (a)) .......... 15

Table 3b. Comparison of Estimates for Population Totals ($q$ = .40, Model (b)) .......... 15

Table 3c. Comparison of Estimates for Population Totals ($q$ = .40, Model (c)) .......... 15

Table 4a. Comparison of Estimates for Population Totals ($q$ = .80, Model (a)) .......... 16

Table 4b. Comparison of Estimates for Population Totals ($q$ = .80, Model (b)) .......... 16

Table 4c. Comparison of Estimates for Population Totals ($q$ = .80, Model (c)) .......... 16

# Figure

**Page**

Figure 1. A Simplified Schematic Representation of the Proposed RSZ Design .......... 13

# Abstract

Low response may render a probability sample that behaves like a nonprobability sample. Achieving a 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 takeoff 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; 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 sorting variables deemed as nonresponse predictors are well distributed across 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 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 and zones, if any. The nonresponse adjustment is made through a one-step calibration for nonresponse and poststratification, as the usual two-step approach is not applicable because, in addition to the information needed 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-distributed 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 a suitable 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 concern is growing 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 incentives. In practice, there are three major approaches to the nonresponse problem, as listed below along with their limitations:

A. 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 small follow-up subsample (caused by budgetary constraints). This, in turn, makes the estimator quite unstable.

B. Release of many reserve replicate samples helps to reach the target sample size, but this method burdens the 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.

C. 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 preceding concerns about precision and unbiasedness resulting from ways to reach the target number of completes, there is clearly a need for an alternative to the traditional method of inflating the release sample size to compensate for ineligibility and nonresponse. When faced with a lower number of completes than expected, the initial release is typically followed by the release of reserve replicates. The NRFUS option is also generally not viable due to cost constraints. A natural alternative is to develop ways in which substitution for nonrespondents by similar units can be justified. In this regard, there is clearly a 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, as typically is done in systematic sampling. Both explicit and implicit stratification variables are deemed to be correlated with the outcome as well as with the 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.

In the above rejective sampling approach, we reject nonrespondents in favor of a respondent in the sense that because we do not 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 because of time and budgetary constraints in data collection. Moreover, this method will not be conducive to unbiased variance estimation for general probability proportional-to-size (PPS) sampling designs. Alternatively, with samples of size one the rejective sampling strategy can be implemented relatively easily where replicates correspond to reserve releases. This is where the method of random groups comes in; see Rao, Hartley, and Cochran (RHC) (1962) and Cochran (1977, pp. 266). This method provides exact unbiased variance estimates for PPS designs in addition to a simple approach to implement PPS designs approximately.

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 units) to the case of a respondent subsample (which may come from a single-stage design with no PSUs) to obtain an unbiased estimate under the joint randomization of sampling design and nonresponse model (also known as quasi-randomization) where nonrespondents are replaced at random by respondents.

This problem arose in the context of education surveys in which schools are typically stratified by school level, urbanicity, enrollment size, and percentage of students eligible for free or reduced-priced lunch; and each stratum is further implicitly stratified by sorting variables such as whether or not it is a charter school or Title I school; the percentage of students who are White, non-Hispanic; geographical variables; and other variables that may be relevant to the variables of interest in the survey. Here, in the first phase, schools are PSUs selected using PPS with student enrollment as size measures, and then some schools 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 in general are experiencing high nonresponse.

The proposed method termed randomly split zones (RSZ) for samples of size one partitions each stratum into almost equal-sized zones via implicit stratification, where the number of zones is set equal to half the allocated stratum target 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. The provisions of proportional allocation of number of zones to the allocated stratum sample size, almost equal-sized zones within each stratum, equal-sized groups within each zone, and common sample size of one from each group, ensure the RSZ design to be equal probability selection method (EPSEM) within each stratum. This helps to control the variability of estimates despite formation of many substrata via zones. 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 as well as under the more realistic scenario when response probabilities are unknown and estimated from the sample under a nonresponse model. Empirical results based on a limited simulation study are presented in Section 5, where other EPSEM (simple random sample and systematic random sample) 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 more meaningful 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 allocated to zone $i$. Let $N_i$ be the size or the number of PSUs for zone $i$, $N_{ij}$ be the size or the number of PSUs for the $j$th random group ($j$ = 1 to $n_i$ ), and $m_{ijk}$ 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 (denoted by $k_{ij}$) is selected using PPS from each group *j*. The *i*th zone population total $T_{yi}$ ( $= \sum_{j=1}^{n_i} T_{yij}$) of the study variable *y* is estimated by

$$t_{yi} = \sum_{j=1}^{n_i} t_{yij}, \text{ where } t_{yij} = y_{ijk_{ij}} \pi_{ijk_{ij}}^{-1} \text{ from the selected PSU } k_{ij}. \tag{2.1}$$

and where $\pi_{ijk} = m_{ijk}/m_{ij+}\,,\ m_{ij+} = \sum_{k=1}^{N_{ij}} m_{ijk}$.

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 phase randomization $E_{12}$. Moreover, $V_1 E_2(t_{yi}) = 0$. Now using PPS results for samples of size one, we have

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

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

$$\begin{aligned} E_1 V_2(t_{yi}) &= \sum_{j=1}^{n_i} p_{ij} \sum_{l<l'}^{N_i} (m_{ijl}/m_{ij+})(m_{ijl'}/m_{ij+})\,(y_{ijl}\,(m_{ij+}/m_{ijl}) - y_{ijl'}\,(m_{ij+}/m_{ijl'}))^2 \\ &= \left(\sum_{j=1}^{n_i} p_{ij}\right) \left(\sum_{l<l'}^{N_i} q_{il} q_{il'} \left(\frac{y_{il}}{q_{il}} - \frac{y_{il'}}{q_{il'}}\right)^2\right) \\ &= \left(\left(\sum_{j=1}^{n_i} N_{ij}^2 - N_i\right)/N_i(N_i - 1)\right) \left(\sum_{l=1}^{N_i} q_{il} \left(\frac{y_{il}}{q_{il}} - T_{yi}\right)^2\right) \end{aligned} \tag{2.3}$$

where $q_{il} = m_{il}/m_{i+}$, and the index *l* corresponds to the index *jk* in a given zone *i*. Note that the only difference between $\pi_{ijk}$ and $q_{ijk}$ is due to different denominators. 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 $m_{ij+}$ 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 estimate given by

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

Where $q_{ijk}$ is the same as $q_{il}$ of (2.3), $q_{ijk} = m_{ijk}/m_{i++}$, and $k_{ij}$, as before, is the randomly selected PSU from the group $ij$. The above results for a single stage design can be generalized to multistage or multiphase designs.

Now 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, as in RHC, could be unequal probability (PPS) or equal probability design, 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 search for respondents within a random group as replacements because units within the corresponding zone are similar. This does not imply that nonresponse adjustments would not be needed once we have respondents from each group, because, although units are similar, they still would have differential response probabilities, and the random replacements within a group are drawn until a respondent is obtained subject to a prescribed number of reserve replicates.

For nonresponse adjustment, we will assume a population response model as in Fay (1991) in which, for a given survey, a response indicator $R_k$ is assigned to each unit $k$ in the universe $U$ which takes the value of 1 with probability $\varphi_k$ when the unit is respondent and 0 with probability $1-\varphi_k$ 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. Thus under the joint $\pi\varphi$ −randomization where $\pi$ denotes the random sampling mechanism with selection probabilities $\pi_k$ for inclusion of the $k$th unit in the sample, and $\varphi$ 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 equal-sized 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 on 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 timeframe match the total number of cases released in one stage under traditional designs. Note that once a respondent is obtained from a group, then no further sample replicates are released from that group.

The number $n_i$ of groups is determined using a geometric series formula as shown below. For an unstratified design, let $n_0$ be the target number of completes, $H$ the number of zones, $q$ the anticipated completion rate reflecting unit eligibility and respondent cooperation, 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

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

or

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

In practice, rounding $n_i$'s up or down for some zones would be needed to deal with noninteger values of $n_i$. 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 releases. This is in contrast to the traditional unit-level response rate where the rate level is not under control of the sampler. The feature of higher group response rate helps to make RSZ robust to nonresponse model misspecifications—this observation is supported by the empirical study. With this motivation, the proposed RSZ design is described 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 with natural modifications if the stratum sample size is odd.
>
> 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 $n_i = n_0/H(1-(1-q)^R)$ with natural modifications if $n_i$ is not an integer.
>
> Step IV: Stagewise release of new reserve samples of size one from remaining nonresponding groups from the previous stage until a respondent is drawn or the maximum number of replicate release is reached, whichever comes first. .

As a somewhat realistic but hypothetical example for illustration purposes, consider an unstratified design for public school surveys in the United States, with the total number of public schools in the United States being about 100,000. In RSZ, approximately equal-sized zones (like deep strata) are created by sorting on implicit stratification variables as mentioned in the introduction. Suppose the target sample size of $n_0$ is 1,000, so that the total number $H$ of zones is 500 and the zone size is approximately 200 schools. Next, each zone is randomly split into groups within which replicate samples of size one are selected. If there is 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 at the design stage to account for the completion rate. For our illustration, we assume a somewhat low completion rate $q$ of 50%. 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 releases is constrained by the data collection timeframe and budget.

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 2.065 for our example, which amounts to a total number of 1,033 (i.e., 500 times 2.065) groups. With 1,033 total number of groups, 467 zones can be created with each having two groups, and the remaining 33 zones with three groups each. (The sampler can randomly select 33 zones out of 500 for assigning the additional groups, or assign to the zones that are expected to have low response rates.) Since each zone has approximately 200 schools, the number of schools in each of the two groups in the 467 zones is 100 and is about 67 schools in each of the three groups in the remaining 33 zones. See Figure 1 for a schematic representation of RSZ and Table 1 for stagewise distribution of total released cases, expected number of incompletes and completes. It also shows the distribution of completes and incompletes over the five release stages when the completion rate is reduced to 30%. The total number of groups in this case increases to 1,203, assuming the same number of stages of release; i.e., 5.

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. Stagewise release also allows for interim analysis so that in the intermediate stages, a random subsample of incomplete groups can be selected for release of additional cases in order 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 can be released in a single stage such that the target is achieved in expectation. This may result in excess or shortage of desired number of completes. If completes are less than desired, reserve replicate samples are released; these need to be planned in advance so that they can be integrated with the initial sample release for estimation with appropriate selection probabilities. 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.

The special case of RSZ with $R = 1$ (i.e., no replicates, to be denoted by RSZ(1)) may be of interest in situations where it is desirable to release all cases in one stage in situations where the survey timeframe does not allow for stagewise release. For RSZ(1), the number of zones is, as before, set at half the allocated target sample size in the stratum, but the number of random groups per zone is increased so that the total number of groups equals the total number of released cases suitably inflated at the design stage. Such a design is somewhat similar to the commonly used systematic sampling, as it can take advantage of implicit stratification except that, unlike systematic sampling, it can provide unbiased variance estimation. Moreover, variance of the RSZ(1) estimator, unlike the case of systematic sampling, necessarily decreases with the sample size. In addition, depending on the realized response rate, it provides the option of replicate release among nonresponding groups without having any advance planning.

# 4. Point and Variance Estimation

Before discussing point and variance estimation, it would be useful to make a few observations underlying the RSZ design when faced with the problem of high nonresponse.

A. **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 key 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 poststratification whereby sampling weights are adjusted so that sample estimates for poststratification variables perfectly match the known population totals. This adjustment also has the additional benefit of coverage bias reduction if the sampling frame has either over- or undercoverage imperfections. Sampling weight calibration for poststratification can be achieved by different methods belonging to the class known as generalized raking which includes linear, log linear, and their range-restricted versions (Deville and Särndal, 1992) but they give similar results for large samples.

B. **Low Nonresponse.** In practice, nonresponse is almost always present despite incentives. For this reason, the target sample size is inflated based on 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 the response indicator and generally with the study variable. Under a nonresponse model, sampling weights are adjusted, but the unbiasedness of estimates under the joint sampling design-nonresponse model depends on the correct specification of the model. Although the nonresponse model is difficult to validate, the nonresponse bias in the estimate is not expected to be serious regardless of model misspecification 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).

C. **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 computable. In RSZ, the nonresponse problem is considerably reduced by making several attempts to get a respondent from each group, although some groups are likely to remain nonresponding in any given zone. Moreover, despite responding units within a zone being similar, units from different responding groups are likely to be skewed toward response-prone domains, as different units are likely to have differential response probabilities. Nevertheless, with a high group response rate, a relatively high number of zones would be represented in the respondent subsample; therefore, after suitable nonresponse adjustments, RSZ is expected to be robust to nonresponse model misspecifications.

D. **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; 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) 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. Moreover, the property "that the calibration method adjusts weights so that the estimator with adjusted weights can reproduce perfectly the known population totals for model covariates" also contributes to the robustness of the RSZ estimator to nonresponse model misspecifications.

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

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

where the first term can be unbiasedly 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 follow by summing over all zones. For variance estimation, at least two respondents per zone are assumed; otherwise, suitable collapsing of zones is performed for a conservative variance estimate. Under the more realistic scenario of unknown $\varphi_k$'s, the estimating equations for model parameters $\lambda$ under a commonly used inverse logit model $\varphi_k^{-1}(\lambda) = 1 + e^{-x_k'\lambda}$ are given by

$$\sum_{i=1}^{H_r} \sum_{j=1}^{n_{ir}} (x_{ijk}/\pi_{ijk}) \left(1 + e^{-x_{ijk}'\lambda}\right) = T_x \tag{4.2}$$

where $n_{ir}$ denotes the total number of responding groups within zone $i$, $H_r$ denotes the total number of responding zones, and $\pi_{ijk}$ is defined earlier as in (2.1). The above equations are admissible if the sample weighted totals $t_{xw}$ 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, $t_{xw}$ for some *x* may not be less than $T_x$ due to extreme initial weights, in which case 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 $\lambda$ can be Taylor linearized, and then the variance estimator discussed above for known $\varphi_k$'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 for unstratified designs in relation to simple random sampling (SRS) and systematic random sampling (SYS). Using the Common Core of Data (CCD) from the School District Finance Survey School Year 2012–13, we considered the total federal funding (in millions of dollars) for a school district as the study variable *y* and the total district enrollment (in thousands) as the auxiliary variable *x*. The CCD has 15,471 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 (-0.302, 1.665) and (-0.124, 1.560) and the correlation as .853. Note that the means of log *y* and log *x* could be negative due to concavity of the log function. This completely specifies the bivariate normal distribution and hence the linear regression of log *y* on log *x*. Next, 10,000 values of log *x* were generated and then the corresponding values of log *y* using the regression model and normal errors. With 10,000 pairs of values of (*y*, *x*), the target parameter $T_y$ is obtained as 31,435.29 in million dollars and the control total $T_x$ as 31,606.80 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 corresponding to 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; these 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 three versions: RSZ(1) which does not allow any replications, RSZ(5) which allows for 5 releases, and RSZ(U) with unrestricted number of releases within each group—for RSZ(U), the number of groups per zone was specified as 2. The method RSZ(U) was included as a reference or benchmark for evaluating performance of other RSZ methods if number of replicate releases was not constrained in practice. The nonresponse adjustment was performed under three misspecified models: (a) Simple Hajek-ratio adjustment to ensure sampling weights of respondents sum to $N$, (b) Linear regression model for the adjustment factor which does not necessarily ensure the adjustment factor remains positive, and (c) Log linear model for the adjustment factor which ensures the adjustment factor remains positive. The correct nonresponse model would have been logistic but was not included due to convergence issues because the sample weighted total $t_{xw}$ at times was larger than $T_x$ in simulated samples. This problem could have been overcome easily by trimming the extreme initial weights or smoothing them under a separate model as in Singh, Ganesh, and Lin (2013) but was not pursued, as it was not necessary for the purpose of demonstrating the robustness of RSZ to model misspecifications.

Although none of the nonresponse models was correctly specified, model (c) came closest to the true model except that it was not logit linear and had both intercept and slope parameters unknown. Model (b) came next except that it was linear, and model (a) ranked last in terms of being close to the true nonresponse model because it did not even depend on $x$. For $q$ = .20 (see Tables 2a, b, and c), with 1,000 simulated samples from the same finite population, it was found that RSZ(5) and RSZ (U) estimators were relatively much less sensitive in terms of relative bias (RB) and relative root mean square error (RRMSE) with respect to misspecified models as compared to the other three methods—SRS, SYS, and RSZ(1). For model (c), RSZ(U) exhibits very small RB (less than 5%) for all the models because it had basically no nonresponding groups due to unconstrained replicate release. The RB for all other methods was around 9% for model (c) but increased rapidly for model (b) to around 25% except for RSZ(5) with less than 14%; while for model (a), the RB was extremely high (around 250%) except for RSZ(5) with around 125%.

The above results provide support to the heuristic claim that the more respondents in the sample are from different domains formed by nonresponse predictors, the less is the bias due to misspecified nonresponse models. In other words, despite unit response rate being low, the higher group response rate seems to help reduce bias for RSZ. For variance reduction, it is important to have a large number of respondents regardless of how well they represent different nonresponse predictor domains. It follows that other methods (SRS, SYS, and RSZ(1)) could outperform RSZ(5) in terms of RRMSE when RB is comparable, as is the case with model (c), but their performance drastically deteriorated for models (b) and (a). However, RSZ(U) continued to outperform RSZ(5) in general. The same pattern followed for $q$ = .40 and .80 as seen in Tables 3a, b, and c and 4a, b, and c. The behavior of RSZ(1) was similar to SYS, as expected, but in practice it might be preferable in view of its important desirable properties mentioned earlier. In this simulation study, it was interesting to find out that SYS and SRS behaved quite similarly. It is true that in practice SYS is preferred over SRS due to ease in execution and expected improved precision after implicit stratification; however, it does not follow from theoretical considerations that SYS necessarily outperforms SRS in terms of variance unless the variation within systematic sample clusters is larger than the variation between clusters.

# 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 is different from the purpose originally planned for the RHC methodology. The simulation study, although limited, supported the claim based on theoretical considerations that the new RSZ design is expected to yield robust estimators with regard to moderate misspecifications of the nonresponse model in terms of bias and mean square error (MSE(. The main reason seems to be that it is the high group-response rate that drives the favorable performance of RSZ despite the unit-response rate being low. In practice, the less the number of nonresponding groups, the more we expect RSZ to be robust to model misspecifications. Groups are defined by randomly splitting deep strata (termed zones) within broad design strata such that units within a zone are similar with respect to covariates deemed nonresponse predictors with the goal of obtaining a respondent subsample that represents almost all zones. This goal is somewhat similar to the use of implicit stratification in the usual systematic sampling designs. Based on a generalization of RHC, the feature of samples of size one from each group allows for random replacements for nonresponding units. The unconditional selection probabilities of responding units from each group do not change, but estimation from RSZ requires nonresponse adjustment because response propensity may vary from unit to unit within a group despite units being similar by belonging to the same zone. Having the total number of groups close to the target sample size and more responding groups due to replicate release is expected to render RSZ robust to moderate departures from the true nonresponse model specifications.

The RSZ methodology is applicable to equal or unequal selection probabilities under general stratified multistage cluster designs. For RSZ(1), it provides exactly unbiased variance estimates in the case of full response, but for general RSZ, it provides approximate unbiased variance estimates after nonresponse adjustments when there are at least two responding groups per zone. It may be noted that having many zones or substrata does not increase the unequal weighting effect for RSZ because of its EPSEM design feature. The traditional method of nonresponse adjustment based on response propensity classes is not applicable for RSZ because selection probabilities for rejected units within a group involve unknown response probabilities, but the calibration method for nonresponse adjustment analogous to poststratification is easily applicable. The special case of RSZ(1) provides an interesting alternative to the commonly used systematic sampling designs, as it can provide an exact unbiased variance estimate and ensures the variance of the estimator decreases with the sample size.

Finally, it may be of interest to consider a possible new and important application of RSZ for controlling sample overlap. With cross-sectional multiple surveys and repeated single surveys 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 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 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 the overlap control (see Ernst, Valliant, and Casady, 2000; and Ohlsson, 2000). However, it turns out that with RSZ, the overlap control of schools can be implemented easily 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 units within PSUs, such as teacher selection within selected schools. With repeated surveys over time, where PSU size measures and stratum sample size could change, a Keyfitz-type (1951) rejective sampling can be used with RSZ to retain PSUs over time for partial overlap. Further investigation of the application of RSZ to the problem of sample overlap control is planned.

# References


Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: Wiley

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*(418), 376–382.

Ernst, L., Valliant, R., & Casady, R. J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics, 16*(3), 211–228.

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

Folsom, R. E. Jr., & Singh, A. C. (2000). The generalized exponential model for sampling weight calibration for a unified approach to nonresponse, poststratification, and extreme weight adjustments. In*: JSM Proceedings, Survey Research Methods* Section (pp. 598–603). Alexandria, VA: American Statistical Association.

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

Keyfitz, N. (1951). Sampling with probabilities proportionate to size: Adjustments for changes in probabilities. *Journal of the American Statistical Association, 46*(253), 105–109.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology, 32*(2), 133–142.

Ohlsson, E. (2000). Coordination of PPS samples over time. *Proceedings of the Second International Conference on Establishment Surveys*, June 17–21, 2000, Buffalo, NY (pp. 255–264). Alexandria, VA: American Statistical Association.

Rao, J. N. K., Hartley, H. O., & Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B, 24*(2), 482–491.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology, 33*(2), 99–119.

Singh, A. C., & Folsom, R. E., Jr. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. In*: JSM Proceedings, Survey Research Methods* Section (pp. 610–615). Alexandria, VA: American Statistical Association.

Singh, A. C., Ganesh, N., & Lin, Y. (2013). Improved sampling weight calibration by generalized raking with optimal unbiased modification. In*: JSM Proceedings, Survey Research Methods Section* (pp. 3572–3583). Alexandria, VA: American Statistical Association.

**Figure 1. A Simplified Schematic Representation of the Proposed RSZ Design**

$(N = 100{,}000,\ n_0 = 1{,}000,\ R = 5,\ q = 0.5,\ n_i = 2\ or\ 3)$

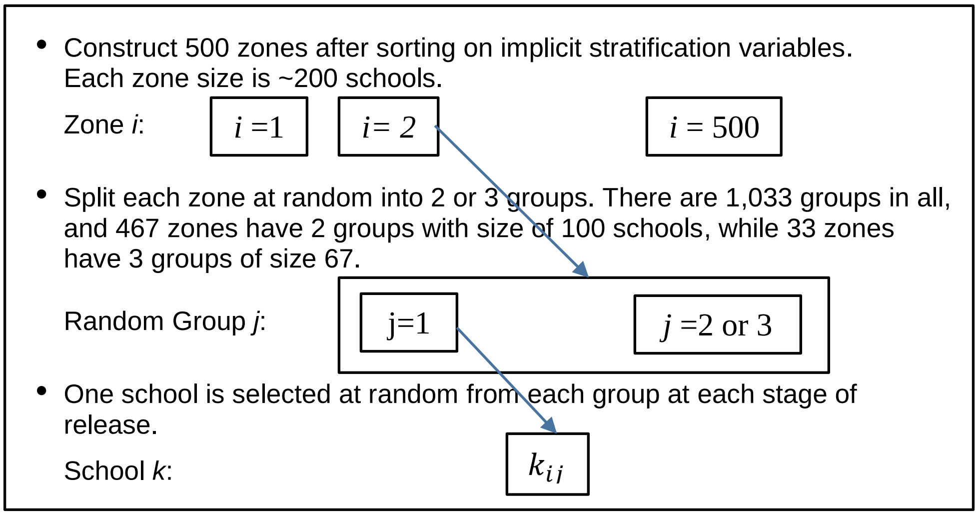


**Table 1. Distribution of Number of Released Cases, Expected Incompletes and Completes**

| Stage | $q = 50\%$ | | | $q = 30\%$ | | |
|---|---|---|---|---|---|---|
| | # Released | # Incompletes | # Completes | # Released | # Incompletes | # Completes |
| **1** | 1,033 | 517 | 516 | 1,203 | 842 | 361 |
| **2** | 517 | 259 | 258 | 842 | 589 | 253 |
| **3** | 259 | 130 | 129 | 589 | 412 | 177 |
| **4** | 130 | 65 | 65 | 412 | 288 | 124 |
| **5** | 65 | 33 | 32 | 288 | 202 | 86 |
| **Total** | **2,004** | **1,004** | **1,000** | **3,334** | **2,333** | **1,001** |

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

| Evaluation Criterion | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| **RB** | 100 | 2.450 | 2.487 | 2.576 | 1.295 | 0.047 |
| | 200 | 2.478 | 2.510 | 2.435 | 1.246 | 0.010 |
| | 400 | 2.425 | 2.443 | 2.447 | 1.335 | 0.036 |
| **RRMSE** | 100 | 2.997 | 2.986 | 3.184 | 2.134 | 0.823 |
| | 200 | 2.777 | 2.741 | 2.681 | 1.837 | 0.457 |
| | 400 | 2.565 | 2.534 | 2.553 | 1.693 | 0.347 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| **RB** | 100 | 0.232 | 0.239 | 0.197 | 0.133 | 0.045 |
| | 200 | 0.259 | 0.280 | 0.265 | 0.138 | 0.044 |
| | 400 | 0.368 | 0.338 | 0.350 | 0.138 | 0.029 |
| **RRMSE** | 100 | 1.087 | 1.024 | 1.080 | 0.637 | 0.501 |
| | 200 | 0.863 | 0.886 | 0.880 | 0.469 | 0.354 |
| | 400 | 0.788 | 0.792 | 0.793 | 0.436 | 0.264 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| **RB** | 100 | 0.091 | 0.074 | 0.092 | 0.087 | 0.046 |
| | 200 | 0.084 | 0.103 | 0.101 | 0.097 | 0.042 |
| | 400 | 0.092 | 0.089 | 0.095 | 0.087 | 0.029 |
| **RRMSE** | 100 | 0.345 | 0.326 | 0.373 | 0.418 | 0.511 |
| | 200 | 0.240 | 0.255 | 0.259 | 0.303 | 0.354 |
| | 400 | 0.185 | 0.173 | 0.181 | 0.217 | 0.265 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 1.060 | 1.081 | 1.060 | 0.263 | 0.022 |
| | 200 | 1.069 | 1.093 | 1.078 | 0.280 | -0.001 |
| | 400 | 1.046 | 1.058 | 1.075 | 0.270 | 0.015 |
| RRMSE | 100 | 1.381 | 1.355 | 1.321 | 0.716 | 0.556 |
| | 200 | 1.241 | 1.224 | 1.222 | 0.551 | 0.358 |
| | 400 | 1.132 | 1.112 | 1.140 | 0.417 | 0.245 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling .

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 0.116 | 0.120 | 0.112 | 0.040 | 0.027 |
| | 200 | 0.115 | 0.137 | 0.114 | 0.050 | 0.019 |
| | 400 | 0.163 | 0.152 | 0.159 | 0.047 | 0.019 |
| RRMSE | 100 | 0.383 | 0.367 | 0.361 | 0.306 | 0.357 |
| | 200 | 0.310 | 0.327 | 0.315 | 0.235 | 0.276 |
| | 400 | 0.304 | 0.304 | 0.308 | 0.172 | 0.199 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 0.072 | 0.073 | 0.080 | 0.039 | 0.026 |
| | 200 | 0.072 | 0.081 | 0.073 | 0.046 | 0.019 |
| | 400 | 0.078 | 0.078 | 0.081 | 0.043 | 0.019 |
| RRMSE | 100 | 0.213 | 0.221 | 0.226 | 0.306 | 0.359 |
| | 200 | 0.160 | 0.165 | 0.162 | 0.232 | 0.278 |
| | 400 | 0.129 | 0.126 | 0.131 | 0.168 | 0.200 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 0.202 | 0.203 | 0.221 | 0.010 | -0.004 |
| | 200 | 0.204 | 0.210 | 0.195 | -0.007 | -0.004 |
| | 400 | 0.191 | 0.195 | 0.187 | 0.007 | -0.002 |
| RRMSE | 100 | 0.500 | 0.454 | 0.530 | 0.360 | 0.331 |
| | 200 | 0.387 | 0.342 | 0.359 | 0.223 | 0.233 |
| | 400 | 0.292 | 0.257 | 0.271 | 0.159 | 0.160 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 0.044 | 0.027 | 0.028 | 0.027 | 0.026 |
| | 200 | 0.021 | 0.041 | 0.031 | 0.007 | 0.012 |
| | 400 | 0.034 | 0.034 | 0.029 | 0.005 | 0.002 |
| RRMSE | 100 | 0.229 | 0.212 | 0.240 | 0.282 | 0.261 |
| | 200 | 0.155 | 0.156 | 0.176 | 0.184 | 0.199 |
| | 400 | 0.119 | 0.105 | 0.118 | 0.132 | 0.130 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

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

| Evaluation | Expected Sample Size Released | Sampling Scheme | | | | |
|---|---|---|---|---|---|---|
| | | SRS | SYS | RSZ(1) | RSZ(5) | RSZ(U) |
| RB | 100 | 0.039 | 0.025 | 0.027 | 0.027 | 0.026 |
| | 200 | 0.019 | 0.039 | 0.029 | 0.006 | 0.012 |
| | 400 | 0.029 | 0.031 | 0.026 | 0.005 | 0.001 |
| RRMSE | 100 | 0.224 | 0.211 | 0.239 | 0.284 | 0.263 |
| | 200 | 0.151 | 0.155 | 0.173 | 0.184 | 0.199 |
| | 400 | 0.113 | 0.102 | 0.115 | 0.132 | 0.130 |

*Notes:* RB, relative bias; RRMSE, relative root mean square error; RSZ, randomly split zones; SRS, simple random sampling; SYS, systematic random sampling.

## ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

**www.air.org**

*Making Research Relevant*

## LOCATIONS

### Domestic

Washington, D.C.
Atlanta, GA
Austin, TX
Baltimore, MD
Cayce, SC
Chapel Hill, NC
Chicago, IL
Columbus, OH
Frederick, MD
Honolulu, HI
Indianapolis, IN
Metairie, LA
Naperville, IL
New York, NY
Rockville, MD
Sacramento, CA
San Mateo, CA
Waltham, MA

### International

Egypt
Honduras
Ivory Coast
Kyrgyzstan
Liberia
Tajikistan
Zambia