Single Phase Simplified Variance Estimation Approach to Two Phase-Stage Hybrid Designs

Avinash C. Singh
Statistical Research and Innovation Division, Statistics Canada, Ottawa K1A 0T6 avi.singh@statcan.gc.ca

ABSTRACT

Two phase designs are useful for increasing estimation efficiency (i.e., reduced mean squared error) for reasons including deep stratification, improved nonresponse adjustment and calibration controls, and reduced coverage bias. They are also useful for increasing sample efficiency (i.e., reduced operational cost and implementation difficulties) for reasons including reduced screening cost, current contact information, and integrating surveys by using the first phase as a master sample. Despite these benefits, such designs are not in wide use due to difficulty in variance estimation unlike the two stage single phase designs under the with-replacement-primary-sampling-unit (*wrpsu*) assumption. A nesting modification of two phase (cluster followed by element sampling) designs is proposed in the sense that the second phase sampling is nested within each first phase cluster and is performed pps (probability proportional to size) with size measures determined by sample allocation rates to domains defined by second phase stratification variables. It also allows for implicit stratification within each cluster with respect to primary stratification and other auxiliary variables. However, due to clusters being treated as design strata, actual sample sizes in the original second phase design strata are random but in expectation nearly match the target sample allocations. With this modification, which is a two phase-stage hybrid, the usual single phase simplified variance estimation under the *wrpsu*-assumption remains applicable even if the second phase variance is not estimable. Comparison with alternative methods is discussed.

Key Words: Conditionally Unbiased and Conditionally Independent Cluster Total Estimates; Two Phase Variance Estimation; With Replacement PSU Assumption.

1. Introduction and Summary

In a fundamental paper on survey sampling, Neyman (1938) introduced double (or two phase) sampling for stratification for more sample efficiency (i.e., reducing operating cost and implementation difficulties) as well as for more estimation efficiency (i.e., reducing MSE-mean squared error). For a study variable with high variability in the population, it may be expensive to obtain a reasonable parameter estimate because of the need of a large sample size. This is even more so if information on the variable is expensive to collect. The success of two phase sampling rests on the fact that it may be inexpensive to collect data on auxiliary variables well-correlated with the study variable in a large first phase sample, and then this is used for stratification (for over- or under- sampling) in the second phase to collect a much smaller sample of data on the study variable. Although the two phase sampling may seem to be more expensive to implement because of two phases of interviewing and field operational cost, the resulting gain in efficiency may offset the extra cost. The large first phase sample can produce a frame with rich information about correlated auxiliary variables that could not only be used for deep stratification at the second phase, but also for calibration control totals (random but based on a large sample) for more efficient estimation. It is also known that a rich frame (i.e., having good auxiliary variables) is highly advantageous for reducing nonresponse bias; see Scheuren (2006) on the use of para data and Groves (2006) for an interesting discussion on the importance of good auxiliary variables versus efforts to reduce nonresponse rates.

Two phase designs have the potential of offering new solutions to some challenging and pressing problems. As an example, a cost effective alternative to dual frame designs can be suggested to deal with the well known problem of bias due to noncoverage of cell-only households in telephone surveys, see e.g., Wolter and Chowdhury (2008). Suppose in the first phase, clusters defined as geographic neighborhoods are selected by pps (probability proportional to size) and a rich frame containing basic information on age, gender, race along with contact information and interview preference (mail, telephone-landline or cell, or internet) for each household is constructed. For sample efficiency, the first phase data can be collected by neighbor-administered personal interview where using a telephone list a person in the neighborhood can be recruited for a moderate compensation somewhat similar to part-time student employment. Note that respondents' trust and cooperation are expected for the interview conducted by a

neighbor provided there are no sensitive questions raising privacy concerns. Next the primary auxiliary variables collected in the first phase are used for stratification in the second phase and then households in the second phase strata are selected pps perhaps with implicit stratification on secondary auxiliary variables. Finally data on study variables are collected by trained interviewers using respondent's preferred mode. This two phase design with possibly different interview modes in the two phases may be useful in reducing high screening costs associated with rare or specific target populations (such as households with special health care needs) by collecting appropriate basic information in the first phase. An example of an alternative to two phase is the single phase approach considered by Srinath (2004) using dual frame designs for telephone surveys by random digit dialing in the context of National Immunization Survey (re: Smith et al., 2001). Two phase designs may also offer a cost effective alternative for updating contact information for the second phase interview if the basic information can be collected inexpensively in the first phase as in the case of neighbor-administered personal interview mentioned earlier. Recently, two phase designs are also being used for secondary surveys in the interest of sample and estimation efficiencies, where the sample from the main survey (or surveys) serves as the first phase, i.e., as a master sample for integrating secondary surveys; for example, the American Community Survey conducted by the US Census Bureau. However, issues such as respondent fatigue may be of concern with the master sample approach because data collected at the first phase consist of more than just the basic information about respondents.

Despite a long list of possible advantages of two phase designs, such designs are not commonly used unlike single phase two (or higher) stage designs, the main reason being that the golden rule of *wrpsu*-based simplified variance estimation formula for single phase designs is not applicable in general to two phase designs. Consequently, various popular replication methods for variance estimation, which are particularly convenient for taking account of variability due to nonreponse and poststratification adjustments, are rendered invalid in general. With more design information, two phase variance estimation is of course possible. However, the lack of computer software suitable for two phase designs compounds the problem further for practitioners. As a result, survey data analysis using models becomes even more difficult. Another important reason for common use of single phase two stage designs is operational efficiency as it allows for fixed interviewer load per cluster or psu. However, this may be offset by using two phase designs in view of the potential of reducing biases due to nonresponse and noncoverage as well as the associated cost to overcome these problems.

The problem of variance estimation for two phase designs has generated considerable interest among researchers in recent years; see e.g., Binder et al. (2000), and Hidiroglou, Rao and Haziza (2008) among others. Important contributions on the use of replication methods related to the problem considered here for designs with cluster sampling at the first phase and element at the second phase have also been made, the main references being Kott and Stukel (1997) and Kim, Navarro, and Fuller (2006). However, their results are restricted to simple random sampling in the second phase strata which makes it possible to neglect the covariance between cluster-level estimates under mild conditions. Although often in practice, the second phase sample is indeed simple random within strata, it would be useful to have a method that could allow for implicit stratification (with respect to auxiliary variables not used for second phase stratification) and possibly unequal sampling rates in the second phase such as pps of elementary units (within strata) where each unit's size measure is collected at the first phase. It may be noted that for the basic two phase design where there is no second phase stratification with the sampling design not depending on the first phase sample and the second phase sampling being nested within first phase clusters, the single phase simplified variance estimator remains directly applicable. This is for the reason that the basic two phase design satisfies the conditions of invariance and independence (Sarndal, Swensson, and Wretman, 1992, Ch. 4, pp. 134); see Section 3 for more details. Such designs are also in common use as, for example, in the case of unequal probability selection of individuals within selected households. In this paper, for two phase designs (with cluster sampling at the first phase and element at the second phase), we generalize the above result for basic two phase designs to more sophisticated ones where pps sampling of elementary units is nested within first phase clusters, and the size measures correspond to sample allocation rates determined for domains defined by primary auxiliary or stratification variables; all auxiliary variables (primary and secondary) can also be used for implicit stratification within each cluster. As a byproduct, for simplified variance estimation, no extra assumptions are needed to neglect covariance contributions from cluster level estimates. However, due to clusters being treated as design strata, actual sample sizes in the original second phase design strata are random but in expectation nearly match the target sample allocations.

The original idea of the proposed design was developed in the context of analyzing disclosure treated data due to subsampling (for suppression) of records by Singh, Yu, and Dunteman (2003). Thus the need for a simpler analysis in the back end motivated suitable modifications of the second phase design in the front end. With the proposed

design, cluster level total estimates satisfy the sufficient conditions of conditional unbiasedness (*CU*) of the cluster totals and conditional independence (*CI*) between clusters given the first phase sample for the simplified variance estimation under the *wrpsu* assumption. It turns out that the *CUCI* conditions are sufficient, and the stronger conditions of invariance and independence in addition to *CU* are not needed. The proposed design is two phase but the second phase design is nested within the first phase cluster as in two stage designs. Thus it is a hybrid design and is termed as a two phase-stage design. In Section 2, we present a review of variance estimation for two phase designs and a motivation for the proposed design. Section 3 describes the proposed two phase-stage design while a comparison with other methods is given in Section 4. The final Section 5 contains concluding remarks.

2. Variance Estimation for Two Phase Designs: Review and Motivation

Consider a two phase design with cluster level sampling at the first phase followed by element level sampling at the second phase. Let U_1 denote the first phase finite population of clusters with total number of clusters $|U_1|(=N_1)$, s_1 denote the first phase sample of size $|s_1|(=n_1)$, and $\{\pi_{1i}\}_{1\le i\le N_1}$ denote the sample inclusion probabilities at the first phase. The sample s_1 with (random) total number of elementary units $N_2\,(=\sum_{s_1} N_{2i}$, N_{2i} being the size of the *i*th cluster selected) is stratified into H strata at phase two based on the auxiliary variable information from s_1 where it is assumed that the stratum definition does not depend on s_1. Now conditional on s_1, from each stratum h $(1\le h\le H)$, a second phase sample s_{2h} of size $|s_{2h}|(=n_{2h})$ is selected with sample inclusion probabilities $\{\pi_{2hk}\}_{1\le k\le n_h}$. Thus the total sample size of the resulting two phase sample is $n_2(=\sum_h n_{2h})$. The parameter of interest is the population total T_y of the study variable y and is defined as $\sum_{i=1}^{N_1} T_{y(i)}$ where $T_{y(i)}$ is $\sum_{k=1}^{N_{2(i)}} y_{k(i)}$ which also equals $\sum_{h=1}^{H}\sum_{k=1}^{N_{2h(i)}} y_{hk(i)}$, $y_{hk(i)}$ being the *y*-observation on the *k*th unit in the *h*th stratum subgroup of the *i*th cluster.

The standard double expansion (not Horvitz-Thompson but similar) estimator of T_y is given by

$$t_y = \sum_h \sum_{s_{2h}} \left(y_{hk} / \pi_{1(hk)} \pi_{2hk} \right) = \sum_h t_{y(h)}\,, \tag{2.1}$$

which can alternatively be expressed in terms of estimated cluster totals as in two stage designs by

$$t_y = \sum_{s_1} t_{y(i)} / \pi_{1i}, \quad t_{y(i)} = \sum_h \sum_{s_{2h(i)}} y_{hk(i)} / \pi_{2hk(i)}\,. \tag{2.2}$$

Note that the sample size $|s_{2(i)}|(=n_{2i})$ for the *i*th cluster is random but the total sample size in the second phase $n_2(=\sum_{s_1} n_{2i})$ is fixed under the design.

2.1 Standard Variance Estimator for Two Phase Designs

First we observe that t_y is unbiased for T_y because the domain estimate $t_{y(i)}$ at the *i*th cluster level is conditionally unbiased (*CU*) for $T_{y(i)}$ given s_1. Now the variance of t_y under two phase designs is given by

$$\begin{aligned} V(t_y) &= V_1\left(\sum_{s_1} T_{y(i)} / \pi_{1i}\right) + E_1 V_2\left(\sum_h t_{y(h)}\right) \\ &\triangleq\ V_1 + V_2 \end{aligned}\,, \tag{2.3}$$

where V_1 is the phase one component and V_2 is the phase two component of the total variance. An unbiased variance estimate can be suitably obtained as

$$\hat{V}(t_y) = \hat{V}_1 + \sum_h \hat{V}_2(t_{y(h)})\,. \tag{2.4}$$

2.2 Nonstandard Variance Estimator for Two Phase Designs

Here with the alternative expression (2.2) for t_y and again using the fact that $t_{y(i)}$ is *CU* for $T_{y(i)}$, we obtain

$$V(t_y) = V_1 + V_2^*; \quad V_2^* = V_2\left(\sum_{s_1} t_{y(i)}/\pi_{1i}\right). \tag{2.5}$$

It follows that $V(t_y)$ can be alternatively expressed as

$$V(t_y) = V_1 + E_1\left(\sum_{s_1} V_2\left(t_{y(i)}/\pi_{1i}\right)\right) + E_1\left(\sum_{i\neq j\in s_1} C_2\left(t_{y(i)}/\pi_{1i}, t_{y(j)}/\pi_{1j}\right)\right), \tag{2.6}$$

where $C_2(\cdot,\cdot)$ is the conditional covariance given s_1. Clearly, for all clusters *i*, we need the realized sample size n_{2i} in the *i*th cluster of the second phase sample at least 2 for unbiased estimation of variances $V_2\left(t_{y(i)}/\pi_{1i}\right)$ and covariances $C_2\left(t_{y(i)}/\pi_{1i}, t_{y(j)}/\pi_{1j}\right)$ in the phase two variance component.

2.3 Single Phase Type Simplified Variance Estimator: Issues

Suppose we blindly use the single phase two stage simplified variance estimator under the commonly made wrpsu assumption. We have

$$\tilde{V}(t_y) = n_1(n_1-1)^{-1}\sum_{s_1}\left(t_{y(i)}/\pi_{1i} - n_1^{-1}\sum_{s_1}\left(t_{y(i)}/\pi_{1i}\right)\right)^2. \tag{2.7}$$

The above estimator is not unbiased due to nonzero conditional covariance between cluster level total estimates. More specifically,

$$E(\tilde{V}) = E_1\left(n_1(n_1-1)^{-1}\sum_{s_1}\left(T_{y(i)}/\pi_{1i} - n_1^{-1}\sum_{s_1}\left(T_{y(i)}/\pi_{1i}\right)\right)^2\right) + E_1\left(\sum_{s_1} V_2\left(t_{y(i)}/\pi_{1i}\right)\right) - (n_1-1)^{-1}E_1\left(\sum_{i\neq j\in s_1} C_2\left(t_{y(i)}/\pi_{1i}, t_{y(j)}/\pi_{1j}\right)\right), \tag{2.8}$$

which simplifies to

$$E(\tilde{V}) = V(t_y) - n_1(n_1-1)^{-1}E_1\left(\sum_{i\neq j\in s_1} C_2\left(t_{y(i)}/\pi_{1i}, t_{y(j)}/\pi_{1j}\right)\right). \tag{2.9}$$

For simple designs at the second phase such as simple random sampling without replacement, the $C_2(\cdot,\cdot)$ terms can be neglected under mild conditions, and so $\tilde{V}$ becomes approximately unbiased; see Section 4 for further discussion. It follows from (2.6) that if sampling designs $\{p(s_{2i})\}_{i\in s_1}$ at the second phase are chosen such that the cluster level estimates $\{t_{y(i)}\}_{1\le i\le n_1}$ are *CI*, then the $C_2(\cdot,\cdot)$ terms disappear and the two phase variance reduces to single phase two stage type variance. Moreover, it follows from (2.9) that under the *wrpsu*-assumption, the usual simplified variance estimator can be used without any bias. We observe that for simplified unbiased variance estimation, we only need cluster level estimates to be *CUCI*--conditionally unbiased and conditionally independent. As mentioned in the introduction, the two conditions of *CU* and *CI* are satisfied by basic two phase designs. This motivates the hybrid design proposed in the next section which allows for unequal sampling rates at the second phase for units within the same second phase stratum, but the nesting modification to the traditional two phase design consists of treating clusters (and not the original second phase strata) as design strata for the second phase.

3. Proposed Hybrid Design: Two Phase-Stage

3.1 Description of the Proposed Two Phase-Stage Design

It consists of the following steps.

Step I: Draw the phase one sample s_1 of n_1 clusters with sampling rate π_{1i} for the *i*th cluster.

Step II: Given s_1, stratify the total number N_2 of elementary units as in second phase stratification and obtain sampling rates π_{2hk} for the unit *k* in stratum *h*. Note that π_{2hk}'s need not be equal within stratum *h*.

Step III: For second phase sampling, define clusters as design strata as in two stage designs and not the original second phase strata used for determining π_{2hk}'s. Allocate sample sizes n_{2i}^* to each cluster such that

$$n_{2i}^* =_{approx} \sum_{h=1}^{H}\sum_{k=1}^{N_{2h(i)}} \pi_{hk(i)} = E_2(n_{2i}), \tag{3.1}$$

where n_{2i}^* are subject to controlled rounding to satisfy $\sum_{i=1}^{n_1} n_{2i}^* = n_2$.

Step IV: Perform implicit stratification of all units within each selected cluster via serpentine sort with highest priority assigned to phase two stratification variables.

Step V: Draw a pps (with size measures π_{2hk}) sample of size n_{2i}^{*} nested within each cluster and independently across clusters given s_1 so that *CUCI* estimates of cluster-level totals can be obtained.

Observe that for the hybrid two phase-stage design, first phase clusters are treated as explicit design strata at the second phase, and the original phase two stratification is essentially preserved in expectation through size measures for pps selection and through implicit stratification within each cluster via serpentine sort. Thus it allows for deep stratification (implicit) within each cluster with fixed sample sizes, but the realized sample sizes in the original second phase strata become random, but in expectation nearly match the target sample allocations.

3.2 Simplified Variance Estimation for Two Phase-Stage Designs

For two phase-stage designs, it follows by construction that the cluster level estimates $\{t_{y(i)}\}_{1\le i\le n_1}$ satisfy the *CUCI property*. If we also make the *wrpsu*-assumption, then the simplified variance estimator $\tilde{V}(t_y)$ is unbiased. We note that as in single phase two stage designs, we can also use for two phase-stage designs, the easily implementable *pps*-systematic sampling with implicit stratification within each cluster without requiring unbiased estimation of $V_2\left(t_{y(i)}/\pi_{1i}\right)$ at the cluster level and still can justify use of the simplified variance estimator under the *wrpsu*-assumption. We also remark that for single phase designs, the usual conditions of invariance and independence along with *CU* of cluster level estimates are invoked to render $t_{y(i)}/p_{1i}$ (where $p_{1i}=n_1^{-1}\pi_{1i}$) as *iid* under the *wrpsu*-assumption. This is used to justify an SRS-type simple variance estimator $\tilde{V}(t_y)$ of (2.7). Note that the term invariance signifies that the design at the second stage is same regardless of the outcome of the first stage sample, while independence signifies that the second stage sampling is done independently from *psu* to *psu*, i.e., *psu*'s are treated as design strata; see Sarndal et. al (1992, Ch. 4, p. 134). However, it follows directly from (2.8) that the weaker condition of *CI* along with *CU* cluster estimates is sufficient to justify unbiasedness of $\tilde{V}(t_y)$ under *wrpsu.* In this case, $t_{y(i)}/p_{1i}$'s continue to have common mean but no longer have common variance, and thus are no longer *iid* because $V_2\left(t_{y(i)}/p_{1i}\right)$ depends on s_1 and not just s_{1i} part that belongs to the cluster *i*.

4. Comparison with Other Methods

In this section, we present a brief review as well as comparison with some alternative methods. For simpler two phase designs (with general first phase cluster sampling but simple random sampling at the second phase), Kott and Stukel (1997) considered the important problem of justifying the use of a replication method (such as jackknife) for variance estimation of post-stratified or reweighted expansion estimator (REE). This estimator is obtained from t_y (double expansion estimator-DEE) by adjusting the sampling weight $w_{1hk}w_{2hk}$ by the stratum-specific multiplicative adjustment factor a_h which is given by

$$a_h=\sum_{k=1}^{N_{2h}} w_{1hk}\Big/\sum_{k=1}^{n_{2h}} w_{1hk}\,w_{2hk};\ w_{1hk}=\pi_{1hk}^{-1}, w_{2hk}=\pi_{2hk}^{-1}, \tag{4.1}$$

where π_{1hk} is π_{1i} if unit *hk* belongs to cluster *i*, and π_{2hk} is constant for each stratum and equals N_{2h}/n_{2h} under simple random sampling. It follows that w_{2hk} cancels out in the REE. Kott and Stukel raised concerns about the applicability of jackknife replication method for DEE. Kim, Navarro, and Fuller (2006) addressed this concern and showed that replication methods are applicable to DEE as well. They considered the following alternative expressions for REE and DEE:

$$\begin{aligned} t_{y,REE}&=\sum_{h=1}^{H}\Big(\sum_{k=1}^{N_{2h}} w_{1hk}\Big)\Big(\sum_{k=1}^{n_{2h}} w_{1hk}y_{hk}\Big/\sum_{k=1}^{n_{2h}} w_{1hk}\Big),\\ t_{y,DEE}&=\sum_{h=1}^{H}\Big(\sum_{k=1}^{N_{2h}} w_{1hk}x_{hk}\Big)\Big(\sum_{k=1}^{n_{2h}} w_{1hk}x_{hk}\tilde{y}_{hk}\Big/\sum_{k=1}^{n_{2h}} w_{1hk}x_{hk}\Big);\ x_{hk}=w_{1hk}^{-1},\tilde{y}_{hk}=w_{1hk}y_{hk}\end{aligned}. \tag{4.2}$$

The above expressions show easily for both REE and DEE how the jackknife replicate will perturb the estimates when a cluster is dropped because it basically adjusts the weight w_{1hk}. It uses a clever way of re-expressing DEE as a poststratified estimator which makes it transparent how a replication strategy would work. However, in the

traditional expression of DEE, w_{2hk} appears which makes it nontrivial for cluster-based replication as pointed out by Kott and Stukel.

For justification of replication methods, besides the *wrpsu*-assumption, we need to express the estimator at least approximately as a linear statistic t_z (where z_{hk} is the linearized variable corresponding to y_{hk} in t_y) such that cluster level total estimates $t_{z(i)}$ satisfy (approximately) *CUCI* conditions, the same conditions required for a simplified variance estimator. The DEE estimator $t_{y,DEE}$ is clearly a linear statistic but $t_{y,REE}$ is not. However, for n_{2h} sufficiently large, REE can be approximately expressed as a linear estimator $t_{z,REE}$ using standard results on linearization of regression calibration estimators. Now, following the arguments of Kott and Stukel (1997) and Kim et al. (2006), if the second phase design were Poisson with sampling rates π_{2hk}, then clearly the *CUCI* conditions would be satisfied by both DEE and REE. However, for simple random sampling designs at the second phase, the above result continues to hold approximately using the asymptotic equivalence of stratum mean estimators (such as $\sum_{k=1}^{n_{2h}} w_{1hk} y_{hk} \Big/ \sum_{k=1}^{n_{2h}} w_{1hk}$) with those under the Poisson design. Thus both $t_{y(i),REE}$ and $t_{z(i),REE}$ satisfy *CUCI* conditions approximately. Unlike the above case of simpler two phase designs, the proposed two phase-stage design allows for implicit stratification and unequal sampling rates for units in the domains defined by stratification variables, but this is achieved by defining clusters as design strata and targeted sample allocations to the original design strata are nearly satisfied in expectation. The cluster level estimates, of course, satisfy *CUCI* by construction. So, replication methods remain applicable for variance estimation of t_y under two phase-stage designs.

It may be of interest to note that the key idea of satisfying *CUCI* conditions under the proposed two phase-stage design has a direct link with methods for finding imputation adjusted variance estimator; in particular, with the significant technique of Rao and Shao (1992) on an adjusted jackknife method for variance estimation in the presence of hot deck imputation. This analogy was also noted by Kott and Stukel (1997). The imputed estimator can be viewed as an estimator under a two phase design by regarding the occurrence of a responding unit as a second phase unit selection under a random nonresponse mechanism—the analogy being that the information on the study variable is obtained only at the second phase except that the nonresponse mechanism is unknown here unlike the case of known sampling design $\{p(s_{2i})\}_{i \in s_1}$ for standard two phase designs. Now assuming that the nonresponse mechanism is ignorable for the imputation model, and that it does not depend on the realized sample s_1 (this is like the invariance assumption and is also assumed under Fay's population response model), the imputation model parameters can be consistently estimated using data from respondents only while accounting for the sampling design. For example, use of mean imputation over imputation classes requires sampling design-weighted class means, and is analogous to post-stratification of t_y for which replication methods can be used for design-based variance estimation assuming negligible sampling fraction and the usual *wrpsu*. Here, clusters correspond to usual *psu*'s in the first phase, and the unknown nonresponse mechanism is assumed for convenience to be uniform, i.e., selection of respondents for the second phase design follows Bernoulli sampling. Thus the *CI* condition is satisfied at the cluster level estimates, and the *CU*-condition is satisfied under the imputation model. If imputation is stochastic as in hot deck, then the value of a donor selected from the imputation class is used instead of the mean. So to ensure *CU* of imputed values, the weighted hot deck procedure was suggested by Rao and Shao. It may be of interest to note that the need of weighted hot deck for satisfying *CU* motivated the work of Singh, Folsom, and Vaish (2004) to propose centering of the predictive mean neighborhood method— a compromise between the predictive mean matching method and the hot deck method. Centering of the random mechanism used for stochastic imputation implies that the second phase random selection of the imputed value for a first phase unit is unbiased given the first phase sample. It follows that one can also use other stochastic imputation methods as alternatives to weighted hot deck suggested in the Rao-Shao procedure as long as the mean of the random distribution of imputed values matches with the mean under the imputation model.

5. Concluding Remarks

The proposed two phase-stage hybrid design was developed to take best of both single phase two stage designs and two phase designs. It addresses the back-end need of a simple variance estimator for two phase designs by a simple modification of the design itself at the front-end. The proposed design assumes a general probability cluster sampling

at the first phase, and a general element level sampling at the second phase. However, the modification consists of pps selection of units in the second phase nested within first phase clusters where the size measures correspond to sample allocation rates of domains defined by stratification variables. Thus the target sample allocations for these domains are nearly achieved in expectation. Nevertheless, in comparison to existing methods which assume simple random sampling within strata at the second phase, and rely on approximations for validity of simplified variance estimator under the *wrpsu*-assumption, the two phase-stage design allows for implicit stratification with respect to all auxiliary variables collected at the first phase, and makes the validity of the simplified variance estimator exact. The proposed two phase-stage design does not, however, apply to the case when general element level sampling is used at both phases. In this case, it is not clear how a simplified variance estimator can be developed. This problem should be investigated in future. Another interesting problem for future study is the application of the proposed two phase-stage design to analyse the precision of estimates obtained after adjusting for coverage bias when neighbor administered personal interview is used to create a complete and rich frame containing auxiliary variables for the second phase design. This method can be compared with the alternative based on single phase dual frame methodology in terms of point, variance, and interval estimates.

Acknowledgment

The author would like to thank Harold Mantel of Statistics Canada for useful discussions.

References

Binder, D. A., Babyak, C., Brodeur, M., Hidiroglou, M.A., and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *Canadian Journal of Statistics*, 28, 751-764.

Groves, R.M. (2006). Nonreponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* (Special Issue), 70, 646-675.

Hidiroglou, M.A., Rao, J.N.K., and Haziza, D. (2008). Variance estimation in two-phase sampling. *Australian Journal of Statistics,*

Kim, J.K., Navarro, A., and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association,* 101, 312-320.

Kott, P.A., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey methodology*, 23, 81-89.

Neyman, J. (1938). Contributions to the theory of sampling human populations. *Journal of the American Statistical Association,* 33, 101-116.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey under hot deck imputation. *Biometrika*, 79, 811-822.

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

Scheuren, F. (2006). The use of paradata in statistical inference. *2005 Statistics Canada Methodology Conference Proceedings*, Ottawa.

Singh, A.C., Yu, F., and Dunteman, G.H. (2003): MASSC: A new data mask for limiting statistical information loss and disclosure. *Monograph of Official Statistics, Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, pp 373- 394 (www.unece.org)

Singh, A.C., Grau, E.A.,and Folsom, R.E., Jr. (2004). Imputation and unbiased estimation: use of centered predictive mean neighborhood method. *JSM Proceedings of Section on Survey Research Methods*.

Smith, P.J., Battaglia, M. P., Huggins, V.J., Hoaglin, D.C., Roden, A.-S., and Khare, M., Ezzati-Rice, T.M., and Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20, 17-24.

Srinath, K.P., Battaglia, M.P., and Khare, M. (2004). A dual frame sampling design for an RDD survey that screens for a rare population. *JSM Proceedings of Section on Survey Research Methods.*

Wolter, K.M., and Chowdhury, S. (2008). Design, conduct, and anlysis of large RDD surveys. *In D. Pfeffermann and C.R. Rao (eds.), New Handbook of Statistics: Sample Surveys: Theory, Methods, and Inference*, Amsterdam: Elsevier.