DUAL FRAME SAMPLE DESIGN FOR A NATIONAL SAMPLE OF VETERANS

G. Hussain Choudhry, Inho Park, and Tiandong Li, Westat
Inho Park, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: RDD Sampling, List Sampling, Dual Frame Sample Design, Sample Allocation, Composite Weights, Replication Variance

The 2000 National Survey of Veterans (NSV) was conducted to obtain current information relevant to planning and budgeting of the Department of Veterans Affairs (VA) programs and services for veterans in general, as well as for certain subgroups of veteran population. The sample design employed was a dual frame design with list-assisted Random Digit Dialing (RDD) sample of telephone households and a sample of veterans from the VA administrative files; health care enrollment file, and pension and compensation file. We determine the sample allocation between the RDD and list sampling frames for the dual frame sample design. We also discuss the weighting strategy for the dual frame sample design.

1. Introduction

The 2000 National Survey of Veterans (NSV) was intended to provide estimates for the entire U.S. population of veterans, as well as for veteran population subgroups of special interest to the Department of Veterans Affairs (VA). The subgroups of primary interest were the seven health care enrollment priority groups. The VA was also particularly interested in data for female, African American, and Hispanic veterans. In Section 2, we describe the dual frame sample design that was implemented for the 2000 national survey of veterans. The sample size and sample allocation are discussed in Section 3. We discuss the weighting and estimation methodology in Section 4. Finally, in Section 5, we compare the efficiency of the dual frame sample design with that of an RDD sample design for a fixed cost.

2. Dual Frame Sample Design

The VA desired to obtain 95 percent confidence intervals of ±5.0 percent or smaller for estimates of proportion of 0.5 for each of the veteran population subgroups. The target population for the NSV was the noninstitutionalized veteran population living in the continental United States and Puerto Rico. According to year 2000 projections of the veteran population provided by the VA, approximately 25 million veterans were living across the country. Based on criteria determined by the VA, veterans are classified as belonging to one of seven health care enrollment priority groups. The distribution of the total veteran population across the seven priority groups is given in Table 1. Priority groups 1 through 6 are termed as mandatory, whereas priority group 7 is termed as discretionary.

Table 1. Distribution of total veteran population across priority groups

Priority group		Percent of total
Mandatory	1	2.31
	2	2.06
	3	5.01
	4	0.73
	5	29.96
	6	0.34
Discretionary	7	59.59

The VA required that the sample design produce estimates for veterans belonging to each of the seven priority groups and for female, African American, and Hispanic veterans. Due to very small population sizes of priority groups 4 and 6 veterans, these priority groups had to be sampled at relatively higher sampling rates to produce estimates with the required levels of reliability. Since estimates were required both at the priority group level and the national level, we used “square root” allocation to allocate the sample across priority groups. Under the “square root” allocation the sample would be re-allocated from the larger priority groups to the smaller priority groups as compared to under the proportional allocation.

Although it would have been theoretically feasible to select an RDD sample with “square root” allocation of the sample across priority groups, such a sample design would have been prohibitively expensive. The RDD sample design is an Equal Probability Selection Method (epsem) design, meaning that all telephone numbers are selected with equal probability. Thus, a very large RDD sample would have to be selected in order to yield the required number of veterans in priority group 6, the priority group with the smallest proportion of veterans. The alternative was to adopt a dual frame approach so that all of the categories with insufficient sample sizes in the RDD sample could be directly augmented by sampling from the VA list frame. The corresponding survey database would be constructed by combining the list and the RDD samples with a set of composite weights. This approach allowed us to use both samples to achieve the desired level of precision for subgroups of interest to the VA. Next, we describe the RDD and List sample designs.

RDD Sample Design

We used a list-assisted RDD sampling methodology to select a sample of telephone households that we screened to identify veterans (Casady and Lepkowski, 1991, 1993, and Potter et al., 1991). In list-assisted sampling, the set of all telephone numbers belonging to operating telephone exchanges is considered composed of 100-banks. We restricted the sampling frame to the "one-plus listed telephone banks" only and selected a systematic sample of telephone numbers. The "one-plus listed telephone banks" are the banks with at least one residential telephone number that is listed in a published telephone directory. Thus, the nonlisted telephone numbers belonging to "zero-listed telephone banks" were not represented in the sample. In addition, the nontelephone households are not represented in the survey. It should be emphasized that nonlisted telephone numbers belonging to "one-plus listed telephone banks" are included in the list-assisted RDD sampling frame.

The list-assisted RDD sampling methodology significantly reduces the cost and time involved in such surveys in comparison to dialing numbers completely at random. The list-assisted RDD sampling methodology is implemented in the GENESYS sampling system, which employs a single-stage equal probability sampling methodology to select a sample of the telephone numbers. The "one-plus listed telephone banks" are initially sorted by geographic variables, such as state, metropolitan, and nonmetropolitan, and also by area codes and five digit prefixes. These sorts construct the sampling frame. The frame is then divided into implicit strata (almost) equal in size while preserving the sort ordering. The total number of such implicit strata is the same as the desired sample size. Then a single telephone number is selected independently from within each implicit stratum.

No listed household information was available for Puerto Rico. As a result, we used a naïve RDD sampling approach called "RDD element sampling" (Lepkowski, 1988) instead of the list-assisted RDD procedure that we used for the national RDD sample. With this methodology, all possible 10-digit telephone numbers were generated by appending four-digit suffixes (from 0000 to 9999) to known 6-digit exchanges (i.e., 3 digit area code and 3 digit prefix combinations), and a systematic sample of telephone numbers was drawn. Before sampling, the frame file was sorted by 6-digit exchange and place name (or service name). This implicit stratification permitted a better representation of the population of households.

List Sample Design

We constructed a list sampling frame from two VA administrative files; the 2000 Veterans Health Administration (VHA) Healthcare Enrollment file and the 2000 Veterans Benefit Administration (VBA) Compensation and Pension file. The files were crossed against each other, and a single composite record was created for each veteran by matching the social security numbers. Each veteran was then identified to belong to one of the seven priority groups. Table 2 lists the total veteran population and the percentage of population represented by the list frame for each of the priority groups.

Table 2. Percentage of veterans in the VA files by priority group

Priority group	Veteran population (thousands)	Percentage of veterans in the list frame
1	577.5	100.0
2	516.4	100.0
3	1,254.1	100.0
4	183.6	94.7
5	7,501.4	25.5
6	83.8	100.0
7	14,920.3	5.9
All veterans	25,037.1	21.6

As can be observed from Table 2, the two largest priority groups (priority groups 5 and 7) have very low coverage of the veteran population in the list frame, whereas four out of the remaining five priority groups (priority groups 1, 2, 3, and 6) have 100 percent coverage. The list frame provides almost 95 percent coverage for priority group 4 (the second smallest priority group). This feature of the list frame was advantageous for the dual frame sample design because the sample could be augmented from the list frame for the smaller priority groups. The VA lists covered 21.6 percent of the overall veteran population including the priority group 7 veterans. Because of the very large proportion of priority group 7 population, no List Sample was required to augment this group of veterans. After excluding priority group 7 veterans, the list frame contained a total of over 4.5 million veterans, accounting for 44.7 percent of the mandatory veteran population, namely those belonging to the priority groups 1 through 6.

The list frame was stratified on the basis of priority group (priority groups 1 through 6) and gender. Thus, the veterans on the list frame were assigned to one of 12 design strata and a systematic sample of veterans was selected independently from each stratum. The allocation of the list sample to the strata is discussed in Section 3.

3. Sample Size and Sample Allocation

Sample Size Determination

The decision on the sample size of completed extended interviews was guided by the precision requirements for the estimates at the health care enrollment priority group level and for the population subgroups of particular interest (i.e., female, African American, and Hispanic veterans). The sample size required for the 95 percent confidence interval with desired half-width (w) for a proportion of p=0.5 can be determined by solving the following equation for the sample size n_c

$$1.96\sqrt{\left(\frac{0.25}{n_c}\right)\times(deff)} = w,$$

where *deff* is the design effect for the corresponding survey estimate. For example, the sample size would be 768 for 95 percent confidence interval with 5.0 percent margin of error for a sample design with design effect (*deff*) equal to 2.0. In order to assign a sample of 768 completed interviews to priority group 6 (the priority group with smallest proportion of veterans), while maintaining "square root" allocation across priority groups, we would have to complete approximately 26,000 interviews. This sample size was larger than the budget permitted and it was decided to reallocate the sample across priority groups by departing slightly from the proposed "square root" allocation and accepting larger sampling errors for some veteran population subgroups. As a result, the sample size of 20,000 completed interviews was sufficient to satisfy the new precision requirements.

Sample Allocation Between List and RDD Frames

Because it was less costly to complete an interview with a case from the List Sample than from the RDD Sample, the goal was to determine the combination of list and RDD sample cases that would achieve the highest precision at the lowest cost. The largest proportion of veterans belongs to priority group 7, which accounts for 59.6 percent of the total veteran population. The proposed "square root" sample allocation scheme meant that we would allocate 38.9 percent of the total sample to priority group 7 veterans. Because no sample augmentation from the list was required for priority group 7 we needed to allocate 65.3 percent (38.9 divided by 59.6) of the total sample to the RDD frame. Any smaller proportion allocated to the RDD frame would have had an adverse impact on the reliability of the estimates, and a larger RDD proportion would have increased the cost. Thus, 65.3 percent was the optimum allocation, which was rounded to 65 percent for allocating the sample to the two frames.

List Sample

It turned out that among the seven priority groups, the highest design effects were for priority groups 4 and 5. This was because coverage of these priority groups by the VA lists was less than 100 percent. In spite of the high design effect for priority group 5, the precision requirement was satisfied because of the larger sample size. First, we allocated the List Sample to the priority groups 1 through 6 to achieve "square root" allocation of the total sample across priority groups. We then re-allocated the List Sample from priority group 5 to priority groups 4 and 6, the two smallest priority groups. The design effects for female, African American, and Hispanic veterans were also larger than 2. The female veterans account for 5.1 percent of the total veteran population, and Hispanic and Black Americans are respectively 4.0 percent and 8.2 percent of the total veteran population. The precision requirements for the estimates for female veterans were achieved through over-sampling the list frame but the precision requirements for the estimates for Hispanic veterans could not be satisfied. The precision requirement for the estimate for African Americans was met because of larger sample size. A sample of 13,129 veterans was selected from the list frame to yield approximately 7,000 completed interviews. Female veterans were sampled at twice the rate as compared with male veterans while keeping the List Sample size fixed at 13,129.

RDD Sample

We initially selected an RDD sample of 240,000 telephone numbers from the December 2000 GENESYS RDD sampling frame. Based on the result of the interim RDD sample yields, we also selected a supplementary sample of 60,000 telephone numbers from the GENESYS RDD sampling frame as of June 2001 to yield 13,000 completed interviews from the two RDD samples. Another RDD sample of 5,500 telephone numbers was also selected from the Puerto Rico RDD frame.

4. Sample Weighting

After the data collection and editing, we constructed the sampling weights for the data collected from the sampled veterans so that the responses could be properly expanded to represent the entire veteran population. The weights were the result of calculations involving several factors, including original selection probabilities, adjustment for nonresponse, households with multiple residential telephones, and benchmarking to veteran population counts from external sources. A separate set of weights was produced for the List and the RDD Samples, which were then combined to produce the composite weights for use with the combined list and RDD samples.

A set of replicate weights was also constructed for each respondent veteran for use in variance estimation. The calculation of the composite weights and replicate weights is described below.

4.1 List Sample Weights

The steps involved in constructing the List sample weights are the calculation of a base weight, poststratification adjustment to known list frame population counts, an adjustment to compensate for veterans with unknown eligibility and an adjustment for nonresponse. Eligibility status of each and every sampled veteran could not be determined, e.g., it could not be ascertained whether a sampled veteran was alive or deceased. Thus, nonrespondents were classified into two categories: (1) eligible nonrespondents, and (2) nonrespondents with unknown eligibility, and the nonresponse adjustments were applied in two steps. The adjustments were applied within homogeneous adjustment classes, which were determined using *CHAID* software.

4.2 RDD Sample Weights

The steps in RDD sample weighting included computing the base weight and various adjustments at the screener interview level and the extended interview level. These steps can be summarized as follows.

- Computation of base weight as the inverse of the probability of selection of the telephone number associated with the household;
- Adjustment to account for household level nonresponse during screening;
- The reciprocal of the number of "regular residential" telephone numbers used by the household (excluding telephone numbers used only for business purposes, fax machines, cellular phones, pagers, or mobile phones); and
- Adjustment to account for the nonresponse to the extended interview.

The nonresponse adjustment factors at the screener and extended interview levels were calculated separately for homogeneous classes defined with *CHAID* analysis. The steps followed for calculating the adjustment factors for the Puerto Rico RDD sample were similar to those for the national RDD sample.

After applying these adjustments, the national (list-assisted) RDD and the Puerto Rico RDD samples were combined into one RDD sample. The weights were further adjusted in a two-dimensional raking procedure. The raking ratio estimation procedure is based on iterative proportional fitting procedure developed by Deming and Stephan (1940), and involves simultaneous ratio-adjustments to two or more marginal distributions of the population counts. The purpose of the raking procedure was to improve the reliability of the survey estimates, and to correct for the bias due to missed households, i.e., households without telephones and households with unlisted telephone numbers belonging to "zero-listed telephone banks." We formed the two raking dimensions from the cross-classification of veterans according to the demographic/education/region characteristics of the veterans. The first dimension was formed from the cross-classification of three age categories (under 50, 50-64, over 64) with four education levels (no high school diploma, high school diploma, some college, bachelor's degree or higher) and four race categories (Hispanic, African American, Other, and White), resulting in 48 cells. The second dimension was formed from the cross-classification of gender (male, female) and the four Census regions (Northeast, Midwest, South, and West), resulting in 8 cells. Three cells for the first raking dimension with smaller sample sizes were collapsed to achieve sufficient cell sample size. Thus, the number of cells for the first raking dimension reduced to 46 after collapsing the three cells with deficient sample sizes.

We used the Census 2000 Supplementary Sample (C2SS) data from the Bureau of Census to define the control totals for the raking procedure. The Puerto Rico RDD sample was also included in the raking procedure. Since the C2SS did not include Puerto Rico in the survey target population we estimated the Puerto Rico veteran population counts for the year 2000 from the Census 1990 population counts based on a model.

4.3 Composite Weights

In order to compute the composite weights we need to know which RDD sample veterans belong to the list frame. The social security numbers (SSNs) of all the veterans on the list frame were known. In order to identify the RDD sample veterans on the list frame, we needed to obtain their SSNs during data collection so that the RDD overlap sample would be identified by matching the SSNs of the veterans in the RDD sample with the list frame. However, out of the 12,956 completed extended RDD interviews we were able to obtain the SSN from only 6,237 veterans, which is 48.1 percent of the RDD completed extended interviews. The veterans sampled as part of the RDD sample could thus only be categorized as belonging to the "overlap" RDD sample or "nonoverlap" RDD sample if the SSN was reported. For others (those who did not report their social security numbers) we used the following prediction model to impute the "overlap" status of the veterans who did not report their social security numbers.

$$\begin{aligned} prob(Overlap) &= prob(SSN)\times prob(Overlap \mid SSN) \\ &\quad + prob\left(\overline{SSN}\right)\times prob\left(Overlap \mid \overline{SSN}\right). \end{aligned}$$

In the above formula $prob(Overlap)$ is the probability that a veteran in the completed RDD sample belongs to the overlap domain, $prob(SSN)$ is the probability that a veteran in the completed RDD sample reported the SSN, $prob(Overlap \mid SSN)$ is the conditional probability that a veteran in the completed RDD sample with reported SSN belongs to the overlap domain, $prob\left(\overline{SSN}\right)$ is the probability that a veteran in the RDD completed sample did not report the SSN, which is equal to $1-prob(SSN)$, $prob\left(Overlap \mid \overline{SSN}\right)$ is the conditional probability that a veteran in the RDD completed sample with unreported SSN belongs to the overlap domain.

We need to determine the probability of overlap conditional on not reporting the SSN, i.e., $prob\left(Overlap \mid \overline{SSN}\right)$, and this can be computed from the above expression because all other probabilities are known. We used *CHAID* analysis to determine homogeneous classes for imputing the overlap status for those not reporting the SSN. We used demographic and socio-economic variables, such as age, gender, race, education, income, etc., and priority level as predictor variables in the *CHAID* model. The probability of overlap conditional on not reporting the SSN, i.e., $prob\left(Overlap \mid \overline{SSN}\right)$, was determined independently for each cell and the "overlap" status was imputed by taking a random sample of the veterans out of those who did not report the SSN. In other words, the “overlap” status of the veterans with unreported SSN within a class was imputed as belonging to the “overlap” domain such that the proportion belonging to the “overlap” domain was as close to the desired probability as possible.

A composite weight was created for the identified “overlap” RDD Sample (both observed and imputed) and List Sample cases using the Hartley (1962) approach as this approach could be adopted to take into account the design effects of the RDD and List Sample designs when combining the two samples. The List and RDD samples were combined into one file consisting of 12,956 completed extended interviews from the RDD sample and 7,092 completed extended interviews from the List Sample resulting in a combined sample of 20,048 completed extended interviews. The parameter λ for constructing the composite weights was chosen to minimize the variance.

The composite weight for each veteran in the RDD sample and List sample was calculated as

$$w_{comp} = \begin{cases} \lambda \times w_1 & \text{if veteran is in the list sample} \\ (1-\lambda) \times w_2 & \text{if veteran is in the RDD overlap sample} \\ w_2 & \text{if veteran is in the RDD nonoverlap sample} \end{cases}$$

where

w_1 = original list sample weight; and

w_2 = original RDD sample weight.

The parameter λ ($0 < \lambda < 1$) defines the composite weight that is used to produce the composite estimate as a linear combination of the List sample estimate and the RDD overlap domain estimate. The optimum value of the parameter λ for estimating a proportion is given by

$$\lambda = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2},$$

where

σ_1^2 = variance of a proportion from the List sample; and

σ_2^2 = variance of a proportion from the RDD overlap sample.

The composite weighting gives increased weight to the estimates with smaller variance, i.e., smaller value of σ^2. In practice, the survey estimates of proportions are produced for several characteristics. As a result, there was an optimum value of the parameter λ corresponding to each of the characteristics. It would not be practical to have separate set of weights for these characteristics. Therefore, the λ values corresponding to these estimates were averaged according to the formula

$$\lambda = \frac{\sum_i \lambda_i \left(n^{(RDD)} \Big/ deff_i^{(RDD)} + n^{(List)} \Big/ deff_i^{(List)} \right)}{\sum_i \left(n^{(RDD)} \Big/ deff_i^{(RDD)} + n^{(List)} \Big/ deff_i^{(List)} \right)},$$

where

λ_i = λ for the *i*-th estimated proportion;

$deff_i$ = design effect for the *i*-th estimated proportion;

n = number of responding veterans;

RDD = overlap RDD sample; and

$List$ = list sample.

In the above formula, the sample size when divided by the design effect represents the effective sample size as compared with simple random sampling due to such design features as clustering and unequal probabilities of selection. Thus, the value of parameter λ is obtained by taking the weighted-average of the individual λ values where the weights are proportional to the corresponding effective sample sizes. The average λ value turned out to be 0.7272 and was used to construct the composite weights for the combined sample.

Raked Composite Weights

The composite weights obtained by combining the List and RDD samples were also raked using the same two-dimensional raking procedure that was used for the RDD sample raking. The only difference was that we did not need to collapse the cells in the first raking dimension, which was defined by cross-classification of age, education and race/ethnicity. The combined RDD and List sample sizes were more than 30 for all 48 cells used for the first raking dimension and hence we did not collapse any cells.

4.4 Replicate Weights

A set of 51 delete-one jackknife (JK1) replicate weights was also created for the List sample as well as RDD sample for use in variance estimation. To create the replicate composite weights, each replicate weight from the List sample was multiplied by the same value of parameter λ (=0.7272) that was used for creating the full sample composite weight. For the RDD overlap sample cases, each replicate weight was multiplied by a factor of $(1-\lambda)$. The non-overlap RDD sample cases were assigned replicate composite weights equal to their original RDD sample replicate weights. Finally, the replicate composite weights were raked to the veteran population counts in a two-dimensional raking procedure as was done for the full sample composite weights.

5. Cost-Variance Efficiency of Dual Frame Design

We obtained the cost-variance efficiency of the dual frame design relative to RDD sample design. The cost-variance efficiency was defined as the ratio of the products of survey costs and the corresponding variances for the two designs. Let $C^{(RDD)}$ and $C^{(Dual)}$ be the costs of the RDD and dual frame sample designs, and $Var^{(RDD)}$ and $Var^{(Dual)}$ be the corresponding variances. Then the cost-variance efficiency of the dual frame design relative to RDD design is given by

$$Eff\,(Dual \text{ vs. } RDD) = \frac{C^{(RDD)} \times Var^{(RDD)}}{C^{(Dual)} \times Var^{(Dual)}}.$$

The cost-variance efficiency can also be interpreted as the ratio of the variances corresponding to the two sample designs for a fixed total cost.

The variances of the survey estimates, e.g., estimates of totals, ratios (or means) and difference of ratios, can be obtained using the JK1 replication method. The corresponding variance is given in Wolter (1985). We obtained the costs of the two designs using a linear cost model. The survey cost of the dual frame design with 20,048 completed extended interviews was 143 percent of the RDD sample design with only 12,956 completed extended interviews. Table 3 gives the cost-variance efficiency of the estimate of proportion of a certain veteran characteristic for a number of domains defined by demographic subgroups and mandatory/discretionary healthcare enrollment priority groups.

Table 3. Cost-variance efficiency of estimates of proportions of veterans with certain characteristic

Domain	Cost-variance efficiency (%)
Hispanic	288.8
African American	194.1
White	157.3
Others	127.8
Mandatory Priority	149.3
Discretionary Priority	83.2
Male	189.6
Female	130.3
All Veterans	181.4

The cost-variance efficiency values of more than 100 percent indicate that the dual frame design was more efficient than the RDD sample design when both the survey cost and the precision of the estimate are taken into consideration. We note that the cost-variance efficiency of the estimate of the proportion is more than 100 percent for all the domains except for the domain discretionary priority group. The reason for less than 100 percent efficiency for the priority group 7 estimate is that no list sample was allocated to priority group 7 veterans.

6. References

Casady, R.J., and Lepkowski, J.M. (1991). Optimal Allocation for Stratified Telephone Survey Designs. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 111-116.

Casady, R.J., and Lepkowski, J.M. (1993). Stratified Telephone Survey Designs. *Survey Methodology*, Vol. 19, pp. 103-113.

Deming, W.E., and Stephan, F.F. (1940). On a Least Square Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, Vol. 11, pp. 427-444.

Hartley, H.O. (1962). Multiple Frame Surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, pp. 203-206.

Lepkowski, J.M. (1988). Telephone Sampling Methods in the United States. *Telephone Survey Methodology* (Eds. Groves, R.M., Biemer, P.P., Lyberg, L.E., Massey, J.T., Nicholls II, WL., and Waksberg, J.). New York: Wiley, pp. 73-98.

Potter, F.J., McNeill, J.J., Williams, S.R. and Waitman, M.A. (1991). List-Assisted RDD surveys. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 117-122.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag: New York.