Appendix B

Sampling Plan

APPENDIX B. SAMPLING PLAN

The NSV 2000 target population includes veterans living in private households in the US and Puerto Rico. Thus, institutionalized veterans, homeless veterans, and veterans living outside the US are not covered in the survey. Additionally, the survey was also required to provide information on veteran population subgroups of particular interest. The subgroups of primary interest were:

- The seven health care enrollment groups,
- Females,
- African Americans, and
- Hispanic veterans.

In addition, the survey was required to provide information needed for major initiatives that would have a direct effect on veterans, such as benefit eligibility reform and health care benefit reform. The sample design had to accommodate these policy issues.

Sample Design

VA desired to obtain 95 percent confidence intervals of ±5 percent or smaller for estimates of proportion of 0.5 for each of the veteran population subgroups. The resulting design called for 20,000 interviews to be completed by random selection of veterans. We evaluated a number of alternative sample design options and adopted a dual frame design consisting of a random digit dialing sample (RDD Sample) and a List Sample. The selected design resulted in sample allocation of 13,000 RDD completed interviews and 7,000 List Sample interviews. The List Sample design used the VHA Healthcare enrollment file and the VBA Compensation and Pension (C&P) file to construct the sampling frame. VA administrative files alone could not be used for the sample design because the coverage from these files was only about 21 percent.

Veterans living in institutions were included in the survey target population only if they were in the institution for less than 6 months and also had a principal residence elsewhere. Although the list frame contained institutionalized veterans, they were not interviewed as part of the List Sample because these would have to be screened for eligibility. They were, however, included as part of the RDD sample. Veterans living abroad and in the territories were also excluded from the survey target population.

Allocation of Sample across Health Care Enrollment Groups

There were approximately 25 million veterans living in the U.S. in 2000 according to VA projections. VA manages its provision of health care services by assigning veterans to one of seven health care enrollment groups.

The distribution of the total veteran population across the seven enrollment groups is given in Table 1. The law defines two eligibility categories: mandatory and discretionary. Enrollment groups 1 through 6 are termed as mandatory, whereas enrollment group 7 is termed as discretionary.

Table 1. Distribution of total veteran population across health care enrollment groups

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

Three Approaches to Sample Allocation

VA required that the sample design produce estimates of proportions for veterans belonging to each of the seven enrollment groups and for female, Hispanic, and African American veterans. Therefore, different sampling rates had to be applied to the seven enrollment groups to produce estimates with the required levels of reliability.

We considered three approaches to allocate the total sample across the seven enrollment groups: (1) equal allocation, (2) proportional allocation, and (3) compromise allocation.

Approach I – Equal Allocation

Under this approach, the sample is allocated equally to each of the seven enrollment groups. The equal allocation approach achieves roughly the same reliability for the enrollment group estimates of proportions. As a result, the variation between the sampling weights would have been very large and would have resulted in large variances for the national level estimates. We therefore did not choose this allocation because it would not have been very efficient for the national level estimates.

Approach II – Proportional Allocation

For this approach, the sample is allocated to the enrollment groups based on the proportion of the veteran population that each enrollment group represents. Thus, the enrollment groups with larger veteran populations would have received the larger share of the sample. The proportional allocation would be the most efficient allocation for the national level estimates because the probabilities of selection are the same for all veterans irrespective of the enrollment group. We did not choose this allocation because reliable enrollment group estimates would only have been possible for the three largest groups (enrollment groups 3, 5, and 7).

Approach III – Compromise Allocation

As the name implies, the compromise allocation is aimed at striking a balance between producing reliable enrollment group estimates (*Approach I*) and reliable national level estimates (*Approach II*). Because we were interested in both national level estimates and the estimates for each of the enrollment groups, we used the "square root" compromise allocation to allocate the sample across the seven enrollment groups (see Table 2).

Table 2. Allocation of NSV 2000 sample across enrollment groups under "square root" allocation

Enrollment group	1	2	3	4	5	6	7
Percent of sample	7.66	7.25	11.29	4.32	27.61	2.92	38.95

Dual Frame Sample Design

Although it would have been theoretically feasible to select an RDD Sample with "square root" allocation of the sample across enrollment groups, such a sample design would have been prohibitively expensive. The alternative was to adopt a dual frame approach so that all of the categories with insufficient sample size in the RDD Sample could be directly augmented by sampling from the VA list frame. The survey database resulting from this approach would then be constructed by combining the List and the RDD Samples with a set of composite weights.

RDD Sample Design

We used a list-assisted RDD sampling methodology to select a sample of telephone households that we screened to identify veterans. As a result, the RDD sampling frame consisted of all the telephone numbers in the "100-banks" containing at least one *listed* telephone number. (Each 100-bank contains the 100 telephone numbers with the same area code, exchange, and first two of the last four digits of the telephone number.) This type of list-assisted RDD sampling approach has two sources of undercoverage:

- Nontelephone households are not represented in the survey, and
- The loss of telephone households with unlisted telephone numbers in the banks having no listed telephone numbers

Studies show that the undercoverage from these two sources is approximately 4 to 6 percent and an adjustment to correct for the undercoverage was applied for NSV2000.

List Sample Design

The VA constructed the list frame from two VA administrative files, the 2000 VHA Healthcare enrollment file and the 2000 VBA Compensation and Pension (C&P) file. The list frame included information about the enrollment group to which each veteran belonged. Table 3 lists the total veteran population and the percentage of population represented by the list frame for each of the enrollment groups.

Table 3. Percentage of veterans in the VA files by enrollment group

Enrollment 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

The coverage offered by the list frame was advantageous for the dual frame sample design because the sample could be augmented from the list frame for the smaller enrollment groups. The list frame was stratified on the basis of enrollment group and gender, and a systematic sample of veterans was selected independently from each stratum.

Allocation of Sample to List and RDD Frames

Because it was less costly to complete an interview with a case from the List Sample than 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 higher RDD unit cost was due to the additional screening required to identify telephone households with veterans. After analysis, it was determined that 65 percent was the optimum RDD allocation that minimized the cost while achieving square root allocation of the total sample across enrollment groups. (The NSV 2000 cost assumptions were based on the previous RDD studies and the assumption that about one in four households would be a veteran household.)

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 group level and for the population subgroups of particular interest (namely, female, African American, and Hispanic veterans). The 95 percent confidence interval for a proportion equal to 0.5 was required with 5 percent or smaller confidence interval half-width for these population subgroups. The precision requirements meant a sample size of n=768 was needed for each enrollment group, and the total survey would have been 26,000 interviews. This sample size was larger than VA was prepared to select, so it was decided that larger sampling errors for smaller subgroups would be accepted. As

a result, the sample size of 20,000 completed interviews was sufficient to satisfy the new precision requirements.

Alternative Sample Design Options

We evaluated six sample design options with respect to cost and design efficiency for a fixed total sample of 20,000 completed interviews. Two of the designs were based on RDD sampling alone, and the remaining four designs were based on a dual frame methodology using RDD and list sampling. For each of the sample designs considered, we compared the coefficients of variation (cv) of the national estimates and veteran population subgroups, as well as the corresponding design effects. The cv was computed to check the precision requirements for the survey estimates, while the design effects were computed to evaluate the efficiency of each of the alternative sample designs. Cost estimates for the alternative sample designs were also calculated using linear cost models incorporating screener and extended interview unit costs.

Out of the six designs analyzed, the sample design that provided the best solution that satisfied the survey objectives of producing reliable estimates and controlling the overall cost of the survey is summarized below.

The sampling parameters of this selected sample design (sample allocation and sample sizes) are given in Table 4. The table also gives the effective sample size, defined as the total sample size divided by the design effect. The minimum effective sample size must be 384 in order to achieve the required 5 percent half-width for 95 percent confidence interval of the estimate of proportion equal to 0.5. Thus, for this sample design, the only veteran population subgroup for which the precision requirement could not be met was Hispanics.

Table 4. Sample allocation for selected sample design

Characteristic	Sample size			Design effect	Effective sample size
	RDD	List	Total		
All veterans	13,000	7,000	20,000	1.48	13,489
Enrollment group 1	295	1,240	1,535	1.13	1,357
Enrollment group 2	271	1,199	1,470	1.12	1,308
Enrollment group 3	661	1,636	2,296	1.18	1,939
Enrollment group 4	69	931	1,000	2.47	405
Enrollment group 5	3,731	1,231	4,962	1.92	2,589
Enrollment group 6	36	764	800	1.04	773
Enrollment group 7	7,937	0	7,937	1.39	5,712
Male	12,338	6,419	18,757	1.52	12,344
Female	662	581	1,243	2.96	420
African American	1,066	574	1,640	2.52	650
Hispanic	520	280	800	2.57	311

Sample Selection

The samples from the list and RDD frames were selected independently. The RDD Sample consists of a sample of telephone households, and the List Sample consists of veterans sampled from the VA list frame. This section describes sampling procedures for each of the two components.

List Sample Selection

The List Sample is a stratified sample with systematic sampling of veterans from within strata. The strata were defined on the basis of enrollment group and gender. The first level of stratification was by enrollment group and then each enrollment group was further stratified by gender. Thus, the sample had 12 strata (enrollment group by gender).

Under the assumption of an 80 percent response rate to the main extended interview, a List Sample of about 8,750 veterans was anticipated to yield 7,000 complete interviews. We also decided to select an additional 50 percent reserve List Sample to be used in the event that response rates turned out to be lower than expected. With the systematic sampling methodology, we achieved a total sample of 13,129 veterans from the list frame, out of which a sample of 4,377 veterans was kept as a reserve sample.

RDD Sample Selection

National RDD Sample

We selected the RDD Sample of households using the list-assisted RDD sampling method. This method significantly reduces the cost and time involved in such surveys in comparison to dialing numbers completely at random. The general approach we employed was a two-stage sampling procedure in which we initially selected a sample of telephone numbers and successfully screened for households with veterans.

Based on propensity estimates from the 1992 NSV RDD Sample, we estimated that we needed a sample of 135,440 telephone numbers to obtain 13,000 completed extended interviews for the RDD component of the sample. Our assumptions were:

- Residential numbers – 60 percent;
- Response to screening interview – 80 percent;
- Households with veterans – 25 percent; and
- Response to extended interview – 80 percent.

We also decided to select an additional 75 percent reserve RDD Sample to be used in the event that the yield assumptions above did not hold. Thus, a total of 240,000 telephone numbers were selected from the GENESYS RDD sampling frame as of December 2000. From this total, 138,000 telephone numbers served as the main RDD Sample and the remaining 102,000 served as the reserve sample. A supplementary sample of 60,000 telephone numbers were also selected later from the GENESYS RDD sampling frame because of interim RDD sample yields.

Puerto Rico RDD Sample

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 method 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 consisting of 3-digit area code and 3-digit prefix combinations. This resulted in a Puerto Rico RDD sample frame that had 3,250,000 telephone numbers. A systematic sample of 5,500 telephone numbers was drawn from this frame to achieve 176 completed extended interviews.

Sample Management

Successful execution of the NSV 2000 required not only an effective sample design but also careful management of the entire sampling process, from creating the sampling frame to completing data collection. Before each sampling step, project staff identified the goals, designed the process, and prepared detailed specifications for carrying out the procedures. At each stage, quality control procedures were carried out that guaranteed survey data integrity.

To ensure that the sample remained unbiased during the data collection process, we partitioned both the RDD and List Samples into a number of release groups so that each release group was a random sample. The sample was released to data collection staff in waves. Each of these sample waves comprised a number of release groups, which were selected at random. The small size and independence of sample release groups gave precise control over the sample. During data collection, we monitored sample yield and progress toward our targets. When we noticed that a sufficient number of sample cases from the previous waves had been assigned final result codes, we released new waves of the sample.

Sample yield is defined as the ratio of the number of completed extended interviews and the number of sampled cases expressed as a percent. We used chi-square statistics to test for homogeneity of distributions of the sample yield by enrollment group, demographic variables, level of education, and census region across waves and found that none of the chi-square values was significant at 5 percent level of significance. Thus, the time effect introduced by releasing waves of sample at various times during data collection, produced no evidence of bias across the sample waves.