A HIERARCHY OF LIST-ASSISTED STRATIFIED TELEPHONE SAMPLE DESIGN OPTIONS

Clyde Tucker, Robert J. Casady and James Lepkowski
Clyde Tucker, BLS, 2 Massachusetts Ave. N.E., Rm 4915, Washington D.C., 20212

1. INTRODUCTION

One of the most difficult tasks in conducting telephone surveys is locating households using a frame of telephone numbers. Only about twenty percent of the telephone numbers in the United States are assigned to residences, and the search for these residential numbers increases the costs of the survey and the length of the required interviewing period. The most popular method for reducing the problem of locating households was first proposed by Mitofsky (1970) and more fully developed by Waksberg (1978). The Mitofsky-Waksberg technique capitalizes on a feature of the distribution of working residential numbers (hereafter referred to as WRNs) in the United States: they tend to be highly clustered within banks of consecutive numbers.

Instead of simply dialing numbers at random, Mitofsky and Waksberg outlined a two-stage design in which banks of 100 consecutive numbers are randomly selected from a frame constructed by appending all 10,000 four-digit suffixes to the list of area code-prefix combinations obtained from BellCore Research (BCR), and a single number from each bank is called. If the number is residential, the rest of the numbers in the hundred bank are sampled until *k-1* additional WRNs are selected; otherwise, the bank is discarded. By restricting calling to within these screened banks, the likelihood of contacting a residence increases threefold to about sixty percent. This procedure produces, in principle, an unbiased sample of telephone households, and one only need know the universe of area code-prefix combinations.

Unfortunately, there are several disadvantages which become apparent when the Mitofsky-Waksberg technique is applied to standard, time-limited cross-sectional surveys. They include the following:

1. The concentration of the sample within certain banks could substantially increase the variance of the estimate if there is a large intra-bank correlation for the characteristics of interest.

2. A total of *k* residential numbers must be contacted in each bank retained for second-stage sampling. This is not usually a serious problem for hundred banks, but it would be for smaller bank sizes. It does mean, however, that only a portion of the numbers in a bank can be used before the bank is discarded

3. Practical limits on the length of the surveying period will prevent finding the requisite number of households in some banks even though they exist.

4. Numbers generated as replacements for non-residential numbers in the original second-stage sample will receive less varied opportunities for calling, especially near the end of the surveying period. A small residual of numbers typically accumulates at the end of the study period for which a final resolution of residential status is impossible within the time constraints.

Several methods have been suggested for streamlining this awkward process. Potthoff (1987) proposed a generalization of the Mitofsky-Waksberg procedure which eliminates the need to contact *k* households in many of the clusters, but this technique has its own complexities. The same is true of a method devised by Burkheimer and Levinsohn (1988) for handling the residual numbers at the end of the survey. Brick and Waksberg (1991) described a modification of the Mitofsky-Waksberg procedure suggested earlier by Waksberg (1984) which eliminates the need to contact the same number of households in every cluster. Instead, a constant number of telephone numbers are contacted in a bank, and weights are assigned to the households found in each of these clusters. The weight for a household is proportional to the reciprocal of the number of households in its cluster.

Although the methodology proposed by Brick and Waksberg does simplify the Mitofsky-Waksberg procedure, it has several problems. Although only a slight bias is introduced, the variances can be affected more substantially. Not only will the variable weights increase the variances (unless they are trimmed), but the cluster sizes (10 or more) necessary to stabilize these weights may limit the number of times the banks can be reused and exacerbate the effects of intra-bank correlations.

Another way to avoid the complexity of the Mitofsky-Waksberg procedure is to select banks of numbers based on external information. Sudman (1973) and Lepkowski and Groves (1986) proposed sampling blocks of numbers using probabilities developed from data on listed residential numbers. This method, however, either restricts the sample to banks with listed numbers or requires that it be supplemented with a sample drawn using the Mitofsky-Waksberg procedure. Furthermore, as Brick and Waksberg observed, the listing rate in the United States is declining to the point that the number of residential listings in a bank may not accurately reflect the total number of households.

Casady and Lepkowski (1991) offered an attractive alternative to the above designs which also uses information on listed residential numbers. They proposed using the counts of listed numbers in banks with one or more listed numbers to stratify the universe of telephone numbers available from BCR into a "high- density" stratum of numbers in banks with at least one listed number and a "low-density" stratum of all other numbers. The estimate of the probability of contacting a residence in the high-density stratum is between 52% and 53% when using ten banks. Only about 2 to 3% of the numbers in the low-density stratum will be assigned to residences. The low-density stratum may be discarded, sampled using an RDD procedure, or further stratified using additional information available from BCR.

This design has several advantages over those previously discussed. Although the information on counts of listed numbers must be purchased, first-stage screening costs are avoided for at least the high-density stratum. For many applications, only a list of the banks with one or more listed numbers is needed. So the declining listing rate becomes less important, and the counts of listed and total residential numbers do not have to be highly correlated. Simple random sampling can be used in the high-density stratum and, possibly, throughout. Thus, variances are not affected by intra-bank correlations, and implementation of the design is relatively straightforward. Finally, stratifying the frame in this way leads to a number of design options.

Casady and Lepkowski discussed some of these options, and Conner and Heeringa (1992) recently tested two designs. However, too little information on the low-density stratum was available to specify all of the alternative designs or fully evaluate the ones that have been considered. A paper by Tucker, Casady and Lepkowski (1992) reported results of a study undertaken to gather the necessary information, but that study was restricted to a convenience sample of six urbanized areas. This paper presents the results of an experiment using a random sample of numbers nationwide. The next section, describing the study design, is followed by the presentation of results. In the concluding section, recommendations concerning alternative designs for use in a variety of situations are offered, and future research needs are discussed.

2. STUDY DESIGN

In order to develop optimal designs using the Casady-Lepkowski methodology, detailed information about the distribution of residential numbers was needed, especially for the low density stratum. The first step was to obtain a file of the counts of listed residential numbers in all of the ten banks on a frame of listed numbers kept by Donnelley Marketing. This information, purchased in April, 1990, was merged with a file containing the universe of ten banks developed from the BCR frame. The ten banks without listed numbers were assigned to sampling substrata using variables previously identified as being related to residential hit rate (Groves, Lepkowski, & Tucker, 1990). These variables, obtained from the BCR file, were (1) whether the area code-exchange of the ten bank was only on the Donnelley frame, only on the BellCore frame, or on both; (2) whether or not the ten banks in area code-exchanges appearing on the Donnelley frame were from thousand banks with listed numbers; (3) whether the wire center in which the area code-exchange was located contained one or more than one exchange (a surrogate for rural-urban).

Table 1 gives the distribution of ten banks across the sampling strata for the entire nation. The initial classification was based on whether or not the ten bank contained listed residential numbers, and the banks without listed numbers (the low-density stratum) were further subdivided as shown. A systematic random sample of 10,500 numbers with stratum identifiers attached were drawn from the complete frame after sorting by geography. An attempt was made to determine the residential status of each number by making up to twelve calls.

3. RESULTS

Based on the combined results from the current study and earlier work at BLS, the strata in Table 1 were collapsed into the six strata shown in Table 2.1. The hit rates h_i were obtained from the current study; the within-bank densities w_i are based on the earlier preliminary work and, thus, are not given as precisely. It should also be noted that:

(1) The estimated value of $\overline{h}$ is based on our sample of 10,500 numbers.

(2) The values of the P_i are known exactly.

(3) The z_i were determined by the equation $z_i = h_i P_i / \overline{h}$.

(4) The t_i were determined by the equation $t_i = 1 - h_i / w_i$.

(5) The "Residual" category contains all telephone numbers from unlisted 10-banks that are found in one, and only one, of the two frames.

The stratum containing listed 10-banks was further stratified using the number of listed numbers in the bank. This was in preparation for sampling proportional to size where the count of listed numbers would be an imperfect, but reasonable measure of size. The results of this operation are shown in Table 2.2. The hit rates for the individual strata were computed by assuming that $h_i = \alpha + \beta(i/10)$ subject to the conditions $h_{10} = .90$ and $\sum_{i=1}^{10} h_i P_i \Big/ \sum_{i=1}^{10} P_i = .5205$. Reasonable guesses were used for the intra-bank densities and, as before, the equation $t_i = 1 - h_i / w_i$ was used to determine the t_i.

Using the information in Table 2.1, the frame was initially partitioned into the four basic sampling strata defined below:

Very High Density Stratum={All telephone numbers in a listed 10-bank}

Moderate Density Stratum={All telephone numbers in an unlisted-bank} $\cap$ {All telephone numbers in a listed 1000 -bank}

Low Density={All telephone numbers in an unlisted 10-bank} $\cap$ {All telephone numbers in an unlisted 1000-bank} $\cap$ {All telephone numbers in a 2+ prefix exchange} $\cup$ {Residual}

Very Low Density= {All telephone numbers in an unlisted 10-bank} $\cap$ {All telephone numbers in an unlisted 1000-bank} $\cap$ {All telephone numbers in a 1 prefix exchange}

As can be seen in Table 3, each of the four strata comprises a significant portion of the frame and can be clearly distinguished from the others on the basis of hit rate. An alternative stratification scheme, pictured in Table 4, was developed by collapsing the moderate and low density strata.

Stratified designs based on the frame stratification given in tables 2-4, as well as the Mitofsky-Waksberg design, were compared to simple RDD sampling of the combined frame using the variance/cost model described by Waksberg (1978) and used by Casady and Lepkowski (1991). Specifically, the sample designs included in the study were

Design 1. Mitofsky-Waksberg sampling applied to combined frame. It is not practical to use 10-bank second stage clusters for this design so 100-bank clusters were assumed. The proportional reduction in variance is from Casady and Lepkowski (1991).

Design 2. Frame stratified by substituting the strata in Table 2.2 for the first strata in Table 3 and referred to as the Full Stratification Design.\ Simple RDD sampling within each of the thirteen strata with stratum sample sizes determined by optimal allocation.

Design 3. Frame stratified as in Table 3.\ Simple RDD sampling within each of the four strata with stratum sample sizes determined by optimal allocation.

Design 4. Frame stratified as in Table 4.\ Simple RDD sampling within each of the three strata with stratum sample sizes determined by optimal allocation.

The proportional reductions in variance or cost , for typical cost ratios, are in Table 5. The cost ratios compare the cost of a productive number (obtaining a completed interview) to an unproductive number, be it residential or not. The Mitofsky-Waksberg design is somewhat more efficient for small cost ratios; but, for those greater than six, all of the designs are similar. There are virtually no differences between the three list-assisted designs.

The sample designs discussed above assume that the sample will be drawn from the entire frame using optimal allocation, but, at the discretion of the researcher, part of the frame can be discarded to further improve efficiency at the risk of some bias.

Designs using only part of the frame are referred to here as "truncated" designs; our attention will be limited to designs that achieve truncation through the elimination of an entire stratum. Several options are available depending on the initial stratification scheme chosen and the amount of potential bias that can be tolerated. The Mitofsky-Waksberg design is not considered in the following because the stratification schemes, and hence the truncation strategies, are based on ten banks. This dictates that the Mitofsky-Waksberg design would of necessity be applied to 10-banks, which is not practical. Had the strata been constructed from hundred banks, truncated Mitofsky-Waksberg designs could have been evaluated.

The four truncated designs compared here discard the "Very Low Density" stratum as defined in tables 3 and 4:

Design 1. Simple RDD sampling applied to the Truncated Frame.

Design 2. Truncated Frame using the Full Stratification Design and Simple RDD sampling within each of the twelve remaining strata ; stratum sample sizes determined by optimal allocation.

Design 3. Truncated Frame stratified as in Basic Stratification Scheme and Simple RDD sampling within each of the three remaining strata; stratum sample sizes determined by optimal allocation.

Design 4. Truncated Frame stratified as in the Alternative Stratification Scheme and Simple RDD sampling within each of the two remaining strata; stratum sample sizes determined by optimal allocation.

Results for these designs are given in Table 6. Just eliminating the "Very Low Density" stratum (about a fourth of all ten banks) from a simple RDD sampling design increases efficiency appreciably, but the gain is even greater when using the other stratified designs. However, this is a gain of no more than 7% over using these sampling strata with the entire frame. In all of these designs, between 1 and 2% of the population is lost. The potential bias is likely to be inconsequential, especially when surveying the general population of telephone households.

4. CONCLUSION

Even if the Mitofsky-Waksberg procedure can be easily administered or the Brick-Waksberg modified design used, potential intra-bank correlation can increase the variance in estimates. This problem is eliminated with the list-assisted designs presented here, making the increase in efficiency for the designs using the whole frame fairly comparable to the second stage of Mitofsky-Waksberg. The list-assisted truncated designs provide additional increases in efficiency if the potential biases can be tolerated. These conclusions hold for most reasonable cost ratios. If the cost ratio is very large, 20 or more, none of the designs, including Mitofsky-Waksberg, are much better than simple RDD sampling.

The cost model used here is relatively simple and does not take into account all costs. More information about these costs is needed and should be incorporated in the cost model. For the Mitofsky-Waksberg design, the additional costs are largely ones accompanying the replacement of nonresidential numbers. For the list-assisted designs, the Donnelley file must be purchased. The processing costs for passing the Donnelley file may be quite large, depending on the available hardware and software. In addition, all of the designs require the BellCore file. Regardless of the design chosen, costs usually can be amortized over several survey administrations.

We soon will have intra-bank densities for the national sample, so the parameters of our designs can be finalized. Other designs could also be considered using this information. For instance, if one were willing to sacrifice some of the variance reduction, samples from each stratum could be drawn strictly proportional to residential hit rate or proportion of the residential population. This would reduce the size of the initial samples of numbers, saving some money and, perhaps, shortening the interviewing period. When the Basic Stratification Scheme is used with the truncated frame and the cost ratio is four or less, the overall residential hit rate can be increased six to eight points if allocation is proportional to hit rate. In these cases, the variance will be no worse than simple RDD, and usually better.

5. REFERENCES

Brick, J.M., and Waksberg, J. (1991), "Avoiding Sequential Sampling with Random Digit Dialing," Survey Methodology, Vol. 17, No. 1, June, pp. 27-42.

Casady, R.J., and Lepkowski, J.M. (1991), "Optimal Allocation for Stratified Telephone Survey Designs," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 444-447.

Conner, J.H., and Herringa, S.G. (1992), "Evaluation of Two New Cost Efficient RDD Designs," paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Groves, R.M., Lepkowski, J.M., and Tucker, C. (1990), "Assessing Telephone Sample Designs That Use Counts of Listed Numbers to Improve Efficiency," paper presented at the annual meeting of the American Association for Public Opinion Research, Lancaster, PA.

Lepkowski, J.M., and Groves, R.M. (1986), "A Two Phase Probability Proportional to Size Design for Telephone Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 73-98.

Mitofsky, W. (1970), "Sampling of Telephone Households," unpublished CBS News memorandum.

Sudman, S. (1973), "The Uses of Telephone Directories for Survey Sampling," Journal of Marketing Research, Vol. 10, No. 2, May, pp. 204-207.

Tucker, C. , Casady, R.J., and Lepkowski, J.M. (1992) "Sample Allocation for Stratified Telephone Sample Designs," Proceedings of the Section on Survey Research Methods, American Statistical Association, forthcoming.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, March, pp. 40-46.

Waksberg, J. (1984), "Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling," unpublished Westat Inc. memorandum.

Table 1. Stratified distribution of ten banks for the United States.

Type\Location of 10-Bank	Number\Percent of 10-Banks
Total	43,770,000
Banks With Listed Numbers	14,835,887
Donnelley Only	00.1%
Donnelley-BellCore	33.9%
Banks Without Listed Numbers	28,934,113
Donnelley Only	
Empty 1000 Bank	0.4%
Non-Empty 1000 Bank	0.3%
BellCore Only	
One Prefix in Exchange	.4%
Two or More Prefixes in Exchange	19.1%
Donnelley and BellCore	
One Prefix in Exchange	
Empty 1000 Bank	24.0%
Non-Empty 1000 Bank	4.5%
Two or More Prefixes in Exchange	
Empty 1000 Bank	9.8%
Non-Empty 1000 Bank	7.6%

Table 2.1. Approximate values of frame parameters when 10-bank characteristics are used to partition the combined BCR/Donnelley frame of telephone numbers. The "Residual" class consists of those telephone numbers found in one, but not both, of the two primary frames. A "Listed Bank" is a bank containing at least one listed number and a "Non -Empty Bank" is a bank containing at least one Working Residential Number.

Location of Telephone Number	Prop. of Frame (P_i)	Prop. of Pop. (z_i)	Hit Rate (h_i)	Prop. of Empty Banks (t_i)	Hit Rate Within Non-Empty Banks (w_i)
Listed 10-Bank	.3390	.8570	.5205	.10	.58
Unlisted 10-Bank, Unlisted 1000-bank, 1 Prefix Exchange	.2397	.0165	.0142	.96	.40
Unlisted 10-Bank, Listed 1000-bank, 1 Prefix Exchange	.0450	.0113	.0516	.90	.55
Unlisted 10-Bank, Unlisted 1000-bank, 2+ Prefix Exchange	.0982	.0225	.0472	.90	.53
Unlisted 10-Bank, Listed 1000-bank, 2+ Prefix Exchange	.0761	.0403	.1090	.78	.50
Residual	.2020	.0524	.0535	.85	.35

$\bar{h} = .2059$ and $\bar{t} = .6248$

Table 2.2. Approximate values of frame parameters for 10-banks with listed numbers.

Number of Listed Telephone Number	Prop. of Frame (P_i)	Prop. of Pop. (z_i)	Hit Rate (h_i)	Prop. of Empty Banks (t_i)	Hit Rate Within Non-Empty Banks (w_i)
One	.0330	.04	.26	.35	.40
Two	.0347	.06	.33	.26	.45
Three	.0433	.08	.40	.18	.49
Four	.0501	.12	.47	.09	.52
Five	.0524	.14	.54	.01	.55
Six	.0484	.14	.62	.01	.62
Seven	.0383	.13	.69	.00	.69
Eight	.0245	.09	.76	.00	.76
Nine	.0113	.06	.83	.00	.83
Ten	.0030	.01	.90	.00	.90

Table 3. Approximate values of the frame parameters for the Basic Reduced Strata Design. Stratum definitions are given below.

Stratum	Prop. of Frame (P_i)	Prop. of Population (z_i)	Hit Rate (h_i)	Prop. of Empty 10-Banks (t_i)	Hit Rate Within Non-Empty Banks (w_i)
Very High Density	.3390	.8570	.5205	.10	.58
Moderate Density	.1211	.0516	.0877	.82	.50
Low Density	.3002	.0749	.0514	.86	.38
Very Low Density	.2397	.0165	.0142	.96	.40

Table 4. Approximate values of the frame parameters for the Alternative Reduced Strata Design; the Moderate Density Stratum and Low Density Stratum have been collapsed into a single Moderate\Low Density Stratum.

Stratum	Prop. of Frame (P_i)	Prop. of Population (z_i)	Hit Rate (h_i)	Prop. of Empty 10-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
Very High Density	.3390	.8570	.5205	.10	58
Moderate\Low Density	.4213	.1265	.0618	.85	.42
Very Low Density	.2397	.0165	.0142	.96	.40

Table 5. Projected proportional reduction in variance\cost (relative to simple RDD sampling of the combined frame) for four alternative sample designs. All four of the alternative designs sample from the entire combined frame and hence cover all of the target population. Cost ratios are typical of research situations.

Sample Design	Proportional Reduction in Variance or Cost			Prop. of Population Not in Scope
	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	
1. Mitofsky-Waksberg	.1832	.1258	.0877	.0000
2. Full Strat. \ OA	.1465	.1021	.0760	.0000
3. Four Stratum \ OA	.1440	.1007	.0751	.0000
4. Three Stratum \ OA	.1419	.0990	.0738	.0000

Table 6. Projected proportional reduction in variance\cost (relative to simple RDD sampling of the entire combined frame) for four alternative sample designs based on sampling from the combined frame less the "Very Low Density" stratum. Cost ratios are typical of research situations.

Sample Design	Proportional Reduction in Variance or Cost			Prop. of Population Not in Scope
	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	
1. Truncated \RDD	.1403	.1118	.0930	.0165
2. Full Strat. Truncated \ OA	.2114	.1590	.1267	.0165
3. Four Strat. Truncated \ OA	.2089	.1576	.1258	.0165
4. Three Strat. Truncated \ OA	.2068	.1560	.1245	.0165