LIST-ASSISTED SAMPLING: THE EFFECT OF TELEPHONE SYSTEM CHANGES ON DESIGN[1]

Clyde Tucker, Bureau of Labor Statistics
James Lepkowski, University of Michigan
Linda Piekarski, Survey Sampling, Inc.

Abstract

List-assisted RDD designs became popular in the late 1980s and early 1990s. Work done by the Bureau of Labor Statistics and the University of Michigan resulted in the development of the underlying theory for these designs as well as the evaluation of various alternative sampling plans to optimize the method. Robert Casady and James Lepkowski document this work in an article in the June 1993 issue of *Survey Methodology*. Recent research to re-evaluate these designs in light of the significant changes in the telephone system over the last decade is presented in this paper. The paper provides background on the development of list-assisted designs, and recent changes in the U.S. telephone system are reviewed. Using 1999 data from Survey Sampling, Inc., an analysis of the current state of the telephone system is presented, and a re-optimization of the earlier designs is undertaken. Results from the earlier work are compared to findings from the 1999 data.

1. Introduction

The Mitofsky-Waksberg random digit dialing (RDD) method (Mitofsky 1970 and Waksberg 1978) was a major innovation in the design of telephone sample surveys. A two stage sampling procedure, the method was widely used because of the simplicity of implementation and reduced cost through more efficient screening of telephone numbers.

The method selects clusters of numbers (100-banks defined by area code, prefix, and first two digits of the suffix) with probabilities proportional to the number of working residential numbers in the cluster, despite the fact that this number is not known at the time of selection. Numbers are selected from clusters identified in the first stage as having at least one working residential number. The method is thus a two-stage probability proportional to size (PPS) equal probability selection of working residential numbers. Further, for ongoing survey operations, a set of 100-banks with at least one working residential number could be used to generate subsamples across several successive studies.

Although widely used for a number of years, the Mitofsky-Waksberg method had several disadvantages that led to a search for other methods. While simple conceptually, there were features of the design cumbersome to administer. Further, because it is a two-stage sample design, variances were larger than those from a simple random or stratified random sample of the same size. To overcome these problems, and retain equal probability sampling in the design, list-assisted methods began to be used in the late 1980s and early 1990s. These methods utilized a frame of listed telephone numbers constructed from telephone directories used by commercial mailing firms. The listed telephone number frame itself was not suitable for direct sampling of telephone numbers because a substantial share of telephone households do not appear in the frame. However, by sampling numbers from 100-banks that contained listed telephone numbers, efficiencies obtained in the second stage of the Mitofsky-Waksberg could nearly be achieved. Sample selection could be simple random, or stratified random, selection of telephone numbers from across 100-banks containing one or more listed telephone numbers. The loss in precision due to cluster sampling was eliminated, and samples generated from list-assisted methods proved less cumbersome to implement than the two-stage cluster method of Mitofsky-Waksberg.

List-assisted methods were examined by Casady and Lepkowski (1993), who laid out the statistical theory underlying the design and presented empirical analysis on the properties of stratified list-assisted design options. Brick and colleagues (1995) showed that the potential bias resulting from the loss of residential telephones in 100-banks without listed numbers was small. Government agencies, academic survey organizations, and private survey firms subsequently adopted list-assisted designs.

The empirical examination of stratified designs in Casady and Lepkowski's work used estimates of parameters from the underlying structure of the 1990 telephone system in the US. Potential gains in efficiency under list-assisted designs depended on the distribution of residential numbers across different types of 100-banks in the system.

[1] Presented at the Annual Conference of the American Association for Public Opinion, Montreal, Canada, May 18, 2001. The findings and opinions expressed are those of the authors and do not necessarily reflect those of the Bureau of Labor Statistics or Survey Sampling, Inc.

The telephone system, however, has changed in dramatic ways since Casady and Lepkowski completed their work almost ten years ago. For example, the number of area codes, and thus the total number of telephone numbers in the system, has almost doubled in the last ten years. There are today 90% more available telephone numbers. On the other hand, the number of households has increased by only a little over 10%. As a result, the proportion of all telephone numbers assigned to a residential unit has dropped from over 0.20 to no more than 0.15, and perhaps lower. The number of "active" prefixes, those with one or more listed telephone numbers, has increased since 1990, but these prefixes now are a smaller percentage of all prefixes. The proportion of unlisted numbers is now approaching 30%, and much larger in some urban areas.

There is also evidence that telephone companies now appear to be less systematic in the assignment of residential numbers across 100-banks. While the number of residences has grown by only 10%, the number of 100-banks with residential numbers has increased by over 50%. The increase in the unlisted rate and the unusually large number of residential banks has resulted in a decline in the proportion of listed residences in 100-banks, from 1990 percentages in the low and middle 50's to percentages in the upper 30's today.

In addition, there has been substantial growth in the number of households with multiple lines. Second lines dedicated to computers, fax machines, and home businesses have made it more difficult to distinguish non-contacts from non-working numbers. Finally, there has been an increase in the assignment of whole prefixes to a single firm. The identification of business numbers, and the separation of those numbers from residential ones, has become more problematic.

Given all of these changes, it is time to reconsider the Casady and Lepkowski designs that were optimized for a telephone system with different underlying parameters than the one we have today. This paper first reviews the basic features of the Casady-Lepkowski approach. It then compares features of the current telephone system and the one existing at the time Casady and Lepkowski did their original work. The set of designs Casady and Lepkowski optimized using 1990 data are then optimized for the current telephone system. Finally, the current efficiencies of these designs will be contrasted to their efficiencies in the past.

2. The List-Assisted Method

The list-assisted method assumes that the entire frame of telephone numbers is available, and stratified on the basis of several auxiliary variables to improve the efficiency of the samples selected. Chief among these auxiliary variables is whether the particular telephone number is in a 100-bank with at least one listed residential number. Two strata are created: one with telephone numbers in 100-banks with one or more listed numbers and a second or all remaining numbers. Further stratification of the "remaining number" stratum could be achieved by knowing characteristics of the prefixes and sets of 10 100-banks comprising a 1000-bank (a set of 1000 consecutive telephone numbers with the same area code, prefix, and first digit of the four digit suffix).

Several alternative stratified designs can then be examined, each optimally allocated for efficiency with respect to a given set of user needs. The optimal allocation minimizing the sampling variance of an estimate for a fixed expected cost (*C**):

$$m_i = \frac{z_i \sigma_i}{\sqrt{h_i}} \left(\frac{1+(1-h_i)\lambda_i}{1+(\gamma-1)h_i} \right)^{1/2}$$

The allocation depends on the proportion of the population in each stratum (z_i), the within-stratum variances (σ_i^2), the within-stratum hit rates (h_i), the proportion of the variance in the estimate of a characteristic accounted for by between stratum differences (λ_i), and the ratio of the total cost of data collection to the cost of just identifying the residential numbers (γ). To assess the relative efficiency of these optimally allocated designs, the sampling variance of the mean under optimal allocation was compared to that which would have been obtained under a simple random selection of all telephone numbers. This proportional reduction in the variance relative to simple random sampling is approximately

$$\left(1-\bar{h}\right) \frac{\left(\sum_{i=1}^{H} \frac{z_i \sigma_i}{\sqrt{h_i}} \left[\left(1+(1-h_i)\lambda_i\right)\left(1+(\gamma-1)h_i\right) \right]^{1/2} \right)^2}{\sigma^2 \left(1+(\gamma-1)\bar{h}\right)}$$

Casady and Lepkowski examined several two- and three-stratum list-assisted designs. For the two-stratum designs, banks of numbers were assigned to a high or low density stratum according to whether or not the bank contained at least one listed residential number. In three-stratum designs, the low density stratum was divided into those with moderate to low residential hit rates (Low density) and those expected to have very few residential numbers (Very low density).

3. Study Design

Casady and Lepkowski developed their designs using counts in 100-banks purchased in 1990 from Donnelly Marketing, Inc. These data were merged with auxiliary information from the BellCore Research telephone frame of all telephone numbers. The current research uses data on 100-banks from 1999 data supplied by Survey Sampling, Inc. containing all auxiliary information. Both data sets were stratified using the auxiliary variables already discussed, facilitating the comparison of designs in the telephone system over the past decade. The relative efficiencies of 1990 and 1999 based designs are then compared.

4. Results

4.1. Changes in the Telephone System

Table 1 illustrates recent changes in the telephone system. The percentage of 100-banks that contained one or more listed numbers declined from 38% in 1990 to 30% in 1999. There has been a corresponding decline in the "density" of listed numbers within these listed 100-banks as well. Figure 1 shows the distribution of the number listed numbers in listed 100-banks over several years since 1986. The distributions for later years through 2000 have shifted to the left, indicating a decrease in the proportion of telephone numbers in listed 100-banks that are listed numbers.

Table 2 contrasts the distributions of telephone numbers between 1990 and 1999 across types of numbers. As will be noted subsequently, there is a substantial shift over time to prefixes with no listed numbers. This trend reflects changes in the phone system introducing more numbers for number portability and nonresidential purposes.

Table 1. Distribution of 100-banks by number of unique listed numbers, 1990 and 1999

Number of	1990		1999	
	All banks	Listed banks	All banks	Listed banks
Total banks	4,350,164	1,656,627	7,715,800	2,316,446
0	62.0	--	70.0	--
1	1.0	2.6	1.1	3.8
2	0.4	1.0	0.6	1.9
3-9	2.3	6.2	2.7	9.0
10-19	2.7	7.1	4.3	14.5
20-29	4.0	10.5	5.9	19.5
30-39	5.7	15.3	6.5	21.8
40-49	7.0	18.3	5.3	17.5
50-59	6.7	17.6	2.8	9.2
60-69	5.2	13.7	0.7	2.5
70-79	2.6	6.8	0.0	0.3
80-89	0.4	1.1	0.0	0.0
90-100	0.0	0.0	0.0	0.0

Figure 1. Incidence of listed phones in working blocks Survey Sampling, Inc

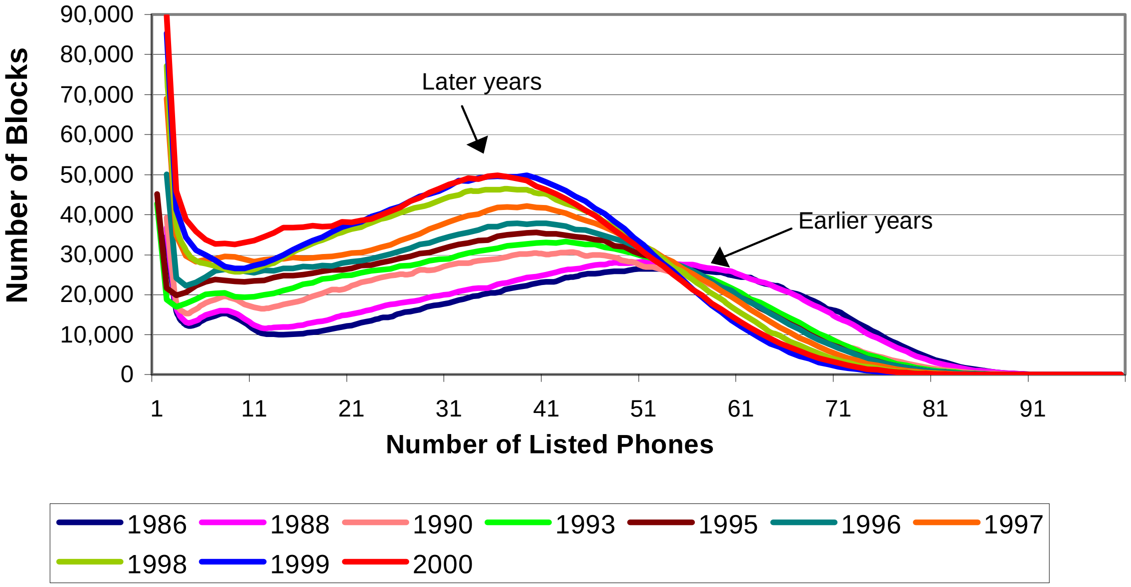

Table 2. Distribution of 100-banks by telephone sampling strata, DMIQ, 1990, and Survey Sampling Database, 1999

Description (Stratum)	1990		1999	
Total	4,350,164	100.0	7,715,800	100.0
100-banks with 1+ listed numbers (1)[a]	1,656,627	38.1	2,316,446	30.0
Urban	- - -		468,429	6.1
Rural	- - -		1,090,729	14.1
Suburban	- - -		757,283	9.8
100-banks with no listed numbers	2,693,337	61.9	5,399,354	70.0
Area code-Prefix with no listed numbers (3)	855,200	19.7	2,949,100	38.2
Exchange class “No listings”	- - -		1,986,300	25.7
Exactly one prefix in the exchange	- - -		42,300	0.5
Two or more prefixes in the exchange	- - -		1,944,000	25.2
All other Exchange classes	- -		962,800	12.5
Exactly one prefix in the exchange	15,900	0.4	26,600	0.3
Two or more prefixes in the exchange	839,300	19.3	936,200	12.1
Area code-Prefix with 1+ listed numbers	1,838,337	42.3	2,450,254	31.8
Exactly one prefix in the exchange	1,196,666	27.5	923,940	12.00
1000-bank with no listed numbers (3)	1,048,960	24.1	765,220	9.9
1000-bank with 1+ listed numbers (2)	147,706	3.4	158,720	2.1
Two or more prefixes in the exchange	641,671	14.8	1,526,314	19.8
1000-bank with no listed numbers (2)	429,660	9.9	955,450	12.4
1000-bank with 1+ listed numbers (2)	212,011	4.9	570,864	7.4

[a] Denotes the stratum to which the 100-banks are assigned: (1) listed or high density 100-banks, (2) unlisted and low density 100-banks, and (3) unlisted and very low density 100-banks.

4.2. The Initial Stratification Scheme

The stratification scheme used in the current analysis matches that used by Casady and Lepkowski. Initially, the frame was divided into the three strata pictured in Table 3. The High-density stratum, containing 30% of all the 100-banks, includes banks with one or more listed numbers. This stratum has almost 97% of the residential numbers, and a residential hit rate of approximately 49%, as estimated from recent screening results from the University of Michigan Survey of Consumer Attitudes.

The second consists of unlisted banks that are in area code and prefix combinations with one or more listed numbers and are in either 1000-banks with a listing or in exchanges with two or more prefixes (urban areas). This stratum makes up about 22% of the frame and contains 2.5% of the residential numbers. Based on information originally presented in Tucker, Casady, and Lepkowski (1992), the low density stratum has an estimated hit rate of 1.7%.

The third stratum contains the remaining 48% of the 100-banks. These banks are either in area code and prefix combinations with no listed numbers or in exchanges with only one prefix (rural area) and a 1000-bank with no listed numbers. This stratum has about 1% of the listed residential numbers and a hit rate of 0.3%.

Table 4 provides a comparison of the parameters in the current three-stratum design to those used by Casady and Lepkowski. There is a decline in the proportion of banks in the high-density stratum over the decade, but a large gain in the proportion in the very low density stratum. On the other hand, the proportion of all residential numbers is a little higher in the high-density stratum compared to ten years ago. Given the decline in the densities within listed 100-banks, it is not surprising that the hit rate in the high-density stratum is somewhat lower now. In fact, the hit rates have dropped across all three strata.

Table 3. Three stratum design

High Density	**Low Density**	**Very Low Density**
Listed 100-banks	*Unlisted 100-banks*	*Unlisted 100-banks*
30% tel. nos.	1+ listed in AC/Prefix	1+ listed in AC/Prefix
49% hit rate	*1 prefix, listed 1000-bank*	*1 prefix, unlisted 1000-bank*
96.5% of pop.	2.1% tel. nos.	9.9% tel. nos.
	1.3% hit rate	0.4% hit rate
	0.2% of pop.	0.2% of pop.
	2+ prefix, unlisted 1000-bank	0 listed in AC/Prefix
	12.4% tel.nos.	*1 prefix in exchange*
	1.2% hit rate	0.9% tel. nos.
	1.0% of pop.	0.2% hit rate
	2+ prefix, listed 1000-bank	0.01% of pop.
	7.4% tel nos.	*2+ prefix in exchange*
	2.7% hit rate	37.3% tel nos.
	1.3% of pop.	0.3% hit rate
		0.7% of pop.

Table 4. Three stratum design: 1999 v. 1990

Stratum/ density	Prop. Frame	Prop. Popn.	Hit rate	Prop. Empty 100-banks	Hit rate in non-empty banks
1: High	0.300	0.965	0.490	0.058	0.520
	0.380	0.940	0.521	0.030	0.537
2: Low	0.219	0.025	0.017	0.955	0.383
	0.200	0.040	0.042	0.914	0.490
3: Very	0.481	0.010	0.003	0.992	0.383
low	0.420	0.020	0.010	0.980	0.490

4.3. Analysis

Five designs, four list-assisted and the Mitofsky-Waksberg sample designs, are examined (as in the work of Casady and Lepkowski). There is a design based on all three strata. A two-stratum design uses the high-density stratum and all remaining numbers (a collapsing of the low- and very low-density strata). The two-stratum design acknowledges the fact that little efficiency is gained by separating the second and third strata, and that implementation would be simplified if only two strata were used.

Two “truncated designs” were formed by eliminating the very from the three-stratum design, and by eliminating the low- and very low-density collapsed stratum from the two-stratum design. These truncated designs do not attempt to cover the small number of households missed by eliminating the least productive stratum from the two and three stratum designs.

Table 5 shows the reduction in variance (compared to simple random sampling) for the five designs for a fixed total cost across the designs. The results are presented when the ratio between total data collection costs and the costs of identifying or screening to find residences is two, 10, and 20, respectively. The proportion of the population covered under each design also is shown.

Table 5. Comparison of five designs: 1999 v. 1990

Design	Prop. Reduction in Variance			Prop. not covered
	$\gamma = 2$	$\gamma = 10$	$\gamma = 20$	
2 stratum	0.427	0.159	0.078	0.000
	0.283	0.077	0.032	0.000
2 stratum	0.598	0.291	0.177	0.035
(truncated)	0.492	0.206	0.119	0.050
Mitofsky-	0.384	0.117	0.044	0.000
Waksberg	0.281	0.060	0.014	0.000
3 stratum	0.445	0.173	0.089	0.000
	0.300	0.087	0.039	0.000
3 stratum	0.531	0.241	0.141	0.010
(truncated)	0.410	0.157	0.088	0.020

For 1990 and 1999, the reduction in variance is quite large when the cost ratio is two, reflecting the larger relative importance of telephone household screening costs in short interview surveys. The reduction in variance is greater for all of the list-assisted designs compared to the Mitofsky-Waksberg procedure.

Further, in all cases, the reduction in variance is greater now than a decade ago. This finding is due to the decline in the residential hit rates from over 20% to under 15% for the base simple random sampling over the time period. The truncated two-stratum design is the most efficient, followed closely by the truncated three-stratum design. In the latter case, only 1% of the population is not covered.

The cost ratio is an important factor to consider. The largest gains in precision are, as noted, for the lowest cost ratio. By the time the ratio becomes as large as 20 (a much longer interview period), none of the designs does substantially better than simple random sampling.

Table 6 provides the relative efficiencies of each list-assisted design compared to Mitofsky-Waksberg for 1990 and 1999. For both time periods, all of the list-assisted designs are more efficient than the Mitofsky-Waksberg design. In some cases the relative efficiencies have increased over time, and in other cases they have decreased. Again, the truncated designs perform better than those that cover the whole population. Thus, for users willing to disregard a small loss in coverage, the truncated designs are quite attractive, especially when the cost ratio is low.

Table 6. Efficiency compared to Mitofsky-Waksberg (1999 v. 1990)

Design	Relative Efficiency (%)		
	$\gamma = 2$	$\gamma = 10$	$\gamma = 20$
2 stratum	9.9	26.2	43.2
	0.7	22.1	56.3
2 stratum	35.8	59.8	74.9
(truncated)	42.9	70.9	88.2
3 stratum	13.6	32.2	49.9
	6.3	31.0	64.1
3 stratum	27.6	51.5	68.5
(truncated)	31.5	61.8	84.1

5. Discussion

It is unclear what other changes will occur in the telephone system in the coming years, and how telephone sampling might be affected. While the total size of the system should not grow as rapidly in the coming years as it has in the last decade, the allocation of numbers to different types of 100-banks shown in Table 2 may continue to change. For instance, a small but growing number of residences have only cellular service, and these numbers have rarely been included in current telephone survey designs. The designs discussed in this paper could incorporate them, but should residences with both regular and cellular service be included and then considered to have multiple lines? Furthermore, as long as the billing algorithm remains the same, cellular users will be reluctant to pay to do survey interviews. Assuming change in the telephone system slows down, the listing rates (and, presumably, the unlisted numbers, too) within 100-banks should increase. With the increasing densities would come an increase in efficiency for all designs. Of course, this assumes that numbers will be assigned as they have been in the past.

Other factors, however, will continue to affect the efficiencies of RDD designs. The continued increase in computer usage could make the identification of residential numbers even more difficult. Noncontact rates have climbed for even personal visit surveys, and this problem will be more severe for telephone surveys. This situation is compounded by the rising use of technologies such as Caller ID to screen calls. The public's increasing reluctance to participate in surveys could result in higher refusal rates, even if they answer their phones. Thus, the continued feasibility of conducting telephone surveys may depend less and less on the ease of locating a residential number and more and more on the respondent's willingness to cooperate.

References

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

Casady, R.J. and Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.

Mitofsky, W. (1970). Sampling of telephone households. Unpublished CBS News memorandum.

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*, 291-296.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.