A COMPARISON OF LISTED AND RANDOMLY DIALED TELEPHONE NUMBERS

Bryce D. Landenberger, Robert M. Groves, and James M. Lepkowski
The University of Michigan

Random digit dialing (RDD) sampling methods are used frequently for telephone surveys because they can give all telephone households a chance of selection (i.e., they offer complete coverage of the telephone household population).[1] A cost disadvantage of RDD methods has been the need to dial randomly generated numbers which are nonworking or nonresidential. Further, in many rural areas of the U.S. it is difficult to determine whether a randomly generated number is a working household number, because nonworking numbers are not identified as such by a recording when dialed. Waksberg (1978) has described an RDD method to reduce the number of nonworking numbers generated, but even with that design approximately one in three numbers generated under this procedure is nonresidential (Groves and Kahn, 1979), and each one discovered must be replaced by a newly generated number. This last fact causes lower response rates on surveys with short interviewing periods.

An alternative method for telephone sampling is to select telephone numbers directly from lists of telephone numbers in telephone directories or computerized lists maintained by commercial firms. Not only does this approach reduce the number of nonworking numbers dialed compared to RDD methods, but also information such as the name and address of the subscriber is available for advance mailings to improve response rates compared to the "cold" contact RDD method (Dillman, 1978). In addition, list frames offer the potential of stratified element sampling which may be used to reduce sampling variances of survey estimates.

However, the list frame approach suffers from a major disadvantage : the list frame does not cover the entire population of telephone households. If telephone households which do not have listed numbers differ with respect to the survey measures of interest from those with listed numbers, a survey based on a list frame could be seriously biased. Leuthold and Scheele (1971) in surveys of the state of Missouri found that Blacks, persons in urban areas, and younger persons tended to have unlisted numbers. Rich (1977) in California surveys found that younger and lower income persons tended to have unlisted numbers. Thus, although the list frame approach to telephone sampling has several operational advantages compared to RDD methods, the coverage properties of the list frame limit the utility of this alternate frame.

In this paper we describe an investigation into the coverage and other properties of the list frame. Given these properties, an alternative design is described which combines the list frame and RDD methods in a dual frame sample to provide the complete coverage of the RDD frame and some of the operational advantages of the list frame.

[1] While not denying the important concern of the exclusion of the nontelephone population from household telephone surveys, the focus of this paper is the survey measurement of the telephone household population.

THE INVESTIGATION

In 1982 the Survey Research Center at The University of Michigan conducted approximately 500 telephone interviews with adult U.S. citizens as part of the National Election Survey. A sample of telephone households was selected for the survey using the two-stage RDD procedure described by Waksberg (1978). In the first stage, a systematic sample of 500 central office codes (i.e., the area code plus the first three digits of the seven digit telephone number) was selected from a list of all central office codes in the U.S. The list was sorted by geographic location and by the number of central office codes contained in an exchange. A four digit random number was generated for each selected central office code to form a ten digit number, which was then dialed to determine whether it was a number assigned to a residential household. A total of 111 working household numbers were identified.

In the second stage of selection, the series of 100 consecutive telephone numbers defined by the first eight digits of each working residential number from the first stage was used to define a cluster or bank of telephone numbers for further dialing. The numbers within the cluster were ordered at random and dialed in order until approximately six working households had been identified in each bank of 100 numbers. Interviews were conducted with all cooperating households.

To investigate the coverage properties of a list frame for the same population of telephone households, all numbers lying in the 111 sample clusters generated for the National Election Survey was purchased from the Metromail Corporation, a firm which maintains a computerized file of telephone listings for the entire U.S. The Metromail telephone number file is keypunched directly from current telephone directories from across the country and is updated continuously as new directories are published. Only residential numbers are entered into the file, and data are often available for each listed number such as name of subscriber, address, county, and zip code.

In addition to obtaining listed numbers in the 111 sample clusters, all listed numbers in the banks of 100 consecutive telephone numbers preceding and following the 111 banks in the survey sample were also purchased in order to increase the set of listed Metromail numbers available for studying other properties of the list frame. Thus, a set of listed numbers was selected corresponding to banks of 300 consecutive telephone numbers associated with the 111 clusters used in the National Election Survey. The sample of listed numbers consisted of 11,628 listings from 111 banks of 300 consecutive numbers.

The investigation consisted of two activities. First, the Metromail listed telephone numbers in the 111 clusters were matched to the numbers generated at random and dialed in the second stage of the RDD sample. Estimates of the proportion of working residential numbers that were listed were made. The proportion of <u>listed</u> numbers that were

actually working could be estimated only by weighting the data to account for the unequal probabilities of selection of the bank of 100 numbers. In the second activity, all 11,628 listings were examined for duplicate numbers. Estimates of sampling errors were made using the stratified cluster design in the estimation process.

PROPERTIES OF THE LIST FRAME

Among working household numbers identified in the RDD sample, 44 percent (standard error 3 percent) were matched to listed numbers obtained from the Metromail list for the same set of 111 banks of working numbers. That is, using the results of the RDD sample as a standard, fewer than one-half of the working household numbers were covered by the list frame. In 12 of the 111 banks of 100 numbers there were no listings at all from the Metromail file. Two of these banks were in large urban areas (Queens, New York and Miami, Florida), and, since a higher proportion of telephone households in urban areas have unpublished telephone numbers (Leuthold and Scheele, 1971), finding banks of numbers with a low percentage of working numbers listed in such area is not unexpected. Two other of the 12 banks without listed numbers had low proportions of working household numbers in the RDD sample as well; failure to find any of them listed could be attributed to chance.

It was surprising to find that so few of the working household numbers from the RDD sample had a number that appeared on the Metromail list. Previous research (Rich, 1977; Leuthold and Scheele, 1971) on subnational areas found that over 70 percent of working residential telephone numbers were published. It is possible that the rate nationally is much lower, but the results might also be explained by delays in Metromail's obtaining newly published directories from the exchanges around the country or errors in updating the list once the directories are obtained.

Among the telephone numbers on the Metromail list, 83.6 percent (standard error 1.8 percent) of listed numbers which matched a number from the RDD sample were working household numbers. The remaining 16.4 percent of the matched numbers were divided between nonworking numbers (13.8 percent, standard error 1.7 percent) and nonresidential numbers (2.7 percent, standard error 0.7 percent). As a comparison, approximately 22 percent of the numbers generated in the first stage of the RDD sample were working household numbers, and approximately 65 percent of the numbers generated in the second stage were working household numbers. Thus, the list frame offers a higher percentage of working household numbers than the RDD frame.

Table 1 presents the results of study of duplicate listings in the Metromail list of 11,628 listed numbers in the 111 banks of 300 numbers in the study. (The results have not been weighted to account for unequal probabilities of selection for the banks; these results should not be interpreted as estimates for the entire Metromail file but only for the portion examined in this study.) A total of 302 sets of duplicate numbers were identified, 70 percent of which the last name of the subscriber differed in the two listings but the address was the same. These could, of course, represent unrelated individuals living at the same address paying for a separate listing for each individual or spouses with different last names, and many of these duplicates appeared in geographic locations with universities where students may be sharing residences with unrelated persons. Approximately 20 percent of the duplicates had different last names and different addresses; these may represent numbers that have been reassigned from one subscriber to another but the updating process of the directory or the Metromail list failed to identify them. Another five percent of the duplicates had the same last name but different addresses, perhaps the result of updating errors again. Only 1.3 percent had the same last name and address, the small number perhaps the result of the list assembly process which can easily identify such duplicates on the alphabetically ordered telephone directories.

Table 1

Distribution of Duplicate Pairs of Telephone Numbers Identified on the List Frame

Last Name	Address	Percent of Duplicates	Number of Duplicates
Same	Same	1.3	4
Same	Different	5.0	15
Different	Same	69.9	211
Different	Different	19.5	59
	Missing	4.3	13
Total		100.0	302

COMBINING LIST AND RDD FRAMES

Although rare in household surveys, many survey designs in settings as diverse as farming research (Steinberg, 1965) and studies of professions (Hansen and Tepping, 1978) utilize multiple sampling frames for the same population simultaneously. These frames often have the property that one of them offers complete coverage of the population but expensive access and measurement options for the sample cases, while others are deficient in coverage but offer cheap measurement opportunities. The designs rarely assign equal probabilities of selection to all members of the population and at analysis use selection weights to compensate for different chances of selection. The previous results on coverage properties of the list frame demonstrate that a combined list frame - RDD sample may have many of the properties that make dual frame designs attractive in other settings.

We attempt to answer the question of whether lower standard errors for survey estimates might

Table 2. Optimal Allocations of a Dual Frame List-RDD Sample For Different Estimated Proportions for the Listed and Unlisted Numbers and Cost of List Frame Numbers

Cost Per List Frame Number	Population Parameters		Optimal Allocation	
	Unlisted Numbers	Listed Numbers	Proportion From List	Ratio of Standard Error Optimum:RDD
\$0.10	0.50	0.50	0.66	0.49
\$0.40	0.50	0.50	0.65	0.52
\$5.00	0.50	0.50	0.54	0.83
\$10.00	0.50	0.50	0.49	1.06
\$0.10	0.30	0.50	0.69	0.48
\$1.00	0.30	0.50	0.65	0.57
\$0.10	0.10	0.50	0.78	0.45
\$1.00	0.10	0.50	0.75	0.54
\$0.10	0.01	0.50	0.92	0.39
\$1.00	0.01	0.50	0.91	0.47
\$0.10	0.50	0.30	0.64	0.47
\$1.00	0.50	0.30	0.59	0.55
\$0.10	0.50	0.10	0.50	0.40
\$1.00	0.50	0.10	0.45	0.46
\$0.10	0.50	0.01	0.056	0.29
\$0.40	0.50	0.01	0.033	0.30
\$1.00	0.50	0.01	0.0016	0.32

be obtained at the same total survey cost if a dual frame approach were used instead of simply an RDD survey. The appendix presents in detail the cost model used to answer this question, and describes the form of the estimator of the population mean that could be used with the dual frame approach. Cost estimates used in the model were obtained from experience at the Survey Research Center for the RDD portion and, based on that, likely values for the list frame. The allocation between the RDD and the list frame portions that yields the lowest standard error for a fixed cost depends on two parameters that could vary greatly over applications: the differences in means between listed and unlisted numbers and the cost of the list frame cases. For various differences in proportions between the two frames and costs for the list frame cases, Table 2 presents approximate optimal allocations and the ratio of the standard error for the optimal dual frame design to that of an RDD sample costing the same amount.

The results in the table can be used to make some general observations about when the dual frame approach is most appealing and when it fails. The top panel of the table treats the case when the proportion to be estimated on the total population has the same value, 0.50, for both the numbers on the list and those not on the list. Here, if a dual frame design is used, the optimal allocation lies between 49 percent and 66 percent from the list, but the gains over an RDD design costing the same amount diminish as the cost for each element from the list frame increases from \$0.10 to \$10.00. That highest figure was chosen to identify the point at which the dual frame design becomes inefficient relative to the RDD design. The current costs for the list elements appear to be closer to the \$0.40 per number level than to the \$10.00. The panel therefore demonstrates that given the cost differences in the two frames, when measuring variables that are expected to have similar distributions in the listed and unlisted portion of the population, the dual frame approach is likely to offer substantial reductions in standard errors for the same survey costs.

The remaining panels of the table illustrate the sensitivity of the gains of dual frame sampling to changes in the estimated proportions and the acquisition cost per listed number. The circumstance most advantageous to the dual frame approach is when the variability in the survey variables is very low in the unlisted portion of the population but very high in the listed portion. Here to reduce the standard error of the overall estimate, proportionally greater allocation should be given to the listed portion. This is especially true when the listed elements are very cheap to acquire. Thus, the optimal allocation for the sample where there is a \$0.10 acquisition cost and a proportion of 0.01 for the unlisted and 0.50 for the listed numbers is an allocation of 92 percent to the list frame. The dual frame approach offers a 61 percent reduction in standard errors over the RDD design costing the same amount of money.

Whenever the proportions in the listed and unlisted parts of the population differ, the dual frame design enjoys two advantages, the lower cost of the list frame and the effect of stratifying the estimate, reflecting differences in the two parts of the population. This fact is illustrated in the last panel of Table 2 that presents the

worst case for the dual frame design. When the unlisted part of the population exhibits very large variability on the survey variable and the listed portion is highly homogeneous, the optimal allocation is nearly all to the RDD portion of the design. Despite the fact that the optimally allocated design would have only a few cases from the list frame, the survey estimate using the dual frame approach enjoys a large advantage over the RDD design. Because of the poststratification used in the dual frame estimator, its standard error is less then one third of that from an RDD design that does not reflect the differences in listed and unlisted numbers.

Concluding Remarks

This paper has undertaken an examination of the role of telephone number lists in sample design for telephone surveys. The list examined has been found to cover only a minority of the working telephone numbers (44 percent), but to offer a higher rate of working numbers (84 percent) than is typically achieved in RDD designs. For that reason the costs per number sampled and interviewed using list frames can be lower than those for RDD designs.

The last part of the paper uses developments in other areas of survey research to address whether a combination of RDD and list frame sampling might offer the researcher more precise results than an RDD design. For a fixed cost, and given the SRC experience in administering RDD surveys and the results of this experiment, the dual frame approach could offer large gains over the traditional approach. These gains are so large that they would be expected even if much larger costs per list frame number were encountered than those experienced in the experiment.

In addition to cost advantages leading to higher precision with the dual frame approach, we have observed that the list frame offers freedom from the need to replace nonworking numbers (as is needed in two stage rejection rule sampling now most often used in telephone surveys) and the possibility of advanced letters. Both of these characteristics suggest that response rates in the dual frame approach might be higher than those in the current RDD design.

Appendix 1

Cost and Error Models For List and RDD Frame Samples

In order to determine the optimal mix of two data collection methods that differ in cost and error properties, it is necessary to construct cost and error models that are functions of the numbers of cases sampled in each method. This helps to determine the allocation of the sample between the list frame and the RDD frame which maximizes the precision of survey statistics, given a fixed cost for the survey.

The costs of the sample can be divided into a) costs of sample number acquisition, b) costs of implementing the sample, and c) costs of building the selection weights used in the survey statistics. A simple cost and error model based on experience at the Survey Research Center with the two frames is presented. The reader may replace values used here with those reflecting their own experience in order to apply the models to design decisions facing them.

Costs of Sample Number Acquisition

The costs of the RDD frame must reflect the acquisition of the tape file containing working area codes and prefixes as well as the computer generation of the primary and secondary numbers:

$$[m_r/w_{rm}][P_r + G_r]$$

where

m_r = the number of primary numbers chosen from the tape file,

w_{rm} = the proportion of primary numbers that are working household numbers (we use 0.22), and

P_r = the cost per number of acquiring the tape file (we use \$0.16).

For the list frame the acquisition costs are simpler:

$$[m_l/w_l]P_l$$

where

m_l = the total number of sample numbers which are working household numbers and

w_l = the proportion of list frame sample numbers that are working household numbers (we use 0.84).

P_l = the cost charged per number.

Costs of Sample Implementation

These components include interviewer salary and telephone connect charges required to make initial contact with the sample household. For the RDD sample they are

$$[m_r/w_{rm}][It_{im} + Ct_{cm}] +$$

$$[m_r n_r/w_{rn}][G_r + It_{in} + Ct_{cn}]$$

where

I = the per minute salary costs for interviewers (we use \$0.083),

t_{im} = the average number of minutes that an interviewer spends in determining the working status of a primary number (we use 12); t_{in} is the equivalent for secondary numbers (we use 9.1),

C = the charge per minute of telephone connection (we use \$0.30),

t_{cm} = the average number of telephone connect minutes spent in determining the working status of a number (we use 3.6); t_{cn} is the equivalent for secondary numbers (we use 2.8),

w_{rn} = the proportion of secondary numbers that are working household numbers (we use 0.65),

G_r = the cost of computer generation of sample primary numbers (we use $.05), and

n_r = the number of secondary numbers generated per working cluster (we use 6).

For the list frame:

$$[m_l/w_l][It_{in} + Ct_{cn}].$$

Costs of Weight Construction

For the RDD sample numbers a check of the list frame must be made to determine whether they could have been sampled from that frame as well. The costs of this should be $m_r n_r P_l$. This cost component would not be required for use of the RDD or list frame alone, but is an overhead cost for the dual frame design.

Variance Model

The estimator of the population mean is from Casady, Snowden, and Sirken (1981) that mixes the results from the two frames, for the RDD frame, separating cases that are members of the list frame from those that are not:

$$\bar{y} = p_{11}\bar{y}_{11} + (1 - p_{11})(\theta\bar{y}_{12} + (1-\theta)\bar{y}_2)$$

where

p_{11} = the proportion of RDD cases not on the list frame,

$\bar{y}_{11}$ = the mean of the RDD cases not on the list frame,

θ = an arbitrary mixing parameter,

$\bar{y}_{12}$ = the mean of the RDD cases that are on the list frame, and

$\bar{y}_2$ = the mean of the list frame sample cases.

Other estimators of the mean are also possible (Hartley, 1962). The sampling variance of this estimator is presented in Casady, Snowden, and Sirken (1981).

REFERENCES

Casady, R. J., Snowden C. B., and Sirken, M. G., "A Study of Dual Frame Estimators for the National Health Interview Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, (1981), pp. 444-447.

Dillman, Don A., Mail and Telephone Surveys, John Wiley & Sons, Inc., (1978), New York, New York.

Groves, R. M. and Kahn, R. L., Surveys by Telephone, Academic Press, Inc., (1979), New York, New York.

Hansen, M. H. and Tepping, B. J., "Variance Estimation for Specified Multiple-Frame Survey Design," in H. A. David (ed.), Contributions to Survey Sampling and Applied Statistics, Academic Press, Inc., (1978), New York, New York.

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

Leuthold, David A. and Scheele, Raymond, "Patterns of Bias in Samples Based on Telephone Directories," Public Opinion Quarterly, Vol. 35, Summer, 1971, pp. 249-257.

Rich, Clyde, "Is Random Digit Dialing Really Necessary?" Journal of Marketing Research, Vol XIV, August, 1977, pp. 300-305.

Steinberg, Joseph, "A Multiple Frame Survey for Rare Population Elements," Proceedings of the Social Statistics Section, American Statistical Association, (1965).

Waksberg, J., "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, March, 1978.