ESTIMATING COVERAGE BIAS IN RDD SAMPLES WITH CURRENT POPULATION SURVEY DATA

Lee H. Giesbrecht, U.S. Bureau of the Census
Dale W. Kulp, Amy W. Starer, GENESYS Sampling Systems

Key Words: Coverage, List-assisted, RDD

Credible random digit dialing (RDD) methods have been in use for nearly thirty years, but the list of methodologies with tractable probabilities and definable sample frames is relatively short. Specification of telephone interviewing for both social science and commercial survey research purposes has increased apace with telephone penetration and the costs of face-to-face interviewing methods. At the same time, the development of sophisticated computer assisted telephone interviewing systems and a long-term decline in telephone long distance rates, makes the argument for telephone interviewing ever more compelling.

Early efforts at defining telephone sampling frames and procedures utilized two complimentary approaches to the problem of designing an efficient equal probability of selection (*epsem*) RDD sample of telephone households. Although both employed two-stage designs, the first employed an equal probability sample of Area Code-Exchange combinations -- NPA-NXXs -- as first stage sampling units. For each NPA-NXX selected in the first stage, the specific hundred series bank (i.e.. the first two-digits of the four-digit suffix) that were designated as residential by the telephone company were then determined. Finally, an *epsem* sample of telephone numbers was selected within the NPA-NXX and hundred series banks deemed residential [Chilton, 1972]. The downside to this approach was essentially economic: since only about 20% of all possible numbers were assigned to residences, significant efforts were required to identify the hundred-series banks in which residential numbers had been assigned. Moreover, required investment was ongoing, since both new exchanges (and new working banks within existing exchanges) needed to be incorporated into the sampling frame in order to maintain representation and coverage over time.

The second approach has become known as the Mitofsky-Waksberg Method of RDD sampling [Waksberg, 1978]. This procedure utilizes an unrestricted sample of hundred series banks and the generation of a single two-digit suffix within each selected bank to provide a set of ten-digit numbers which represent the Primary Number sample. Each of these numbers is then screened to determine whether or not it is a working residential number. Those first stage banks corresponding to primary numbers which are not residential are discarded. The remaining banks represent the PSUs for the second stage of sampling. The result is a probability proportionate to size (PPS) sample of hundred-series banks (probabilities being proportionate to the number of assigned residential numbers in each bank). But, since the objective is an overall *epsem* sample of telephone numbers, all first stage PSUs will nominally require equal sample sizes in the second stage. These requirements resulted in complex and time consuming data collection procedures, involving multiple PSUs but did eliminate the prohibitive ongoing financial investment in sample frame definition and maintenance required by the former method.

By the late 1970's, many versions of what now is termed "list-assisted" RDD sampling were adopted for use by commercial survey research firms. This new method of defining the RDD sample frame was dependent upon the consumer direct marketing industry's nationwide White Page Telephone Directory databases which contained the names and addresses of those with published telephone numbers. The database of residential telephone listings is used to determine the number of listed telephone numbers in each 100-bank [e.g., (301) 457-3800, 3801,..., 3899]. Hundred-banks with no listed numbers (zero-listed 100-banks) are typically left out of list-assisted sampling frames. The remaining 100-banks (1+ listed, 2+ listed, etc., depending on the truncation threshold chosen) make up the list-assisted sampling frame.

Until the late 1980's, these list-assisted methods were virtually constrained to market research applications, where issues relating to sample frame coverage and potential bias, lack of tractability, indeterminate probabilities of selection, etc., were less important than increasing the efficiency of data collection efforts. Meanwhile, the Mitofsky-Waksberg procedure remained the clear methodology of choice for statistical sampling applications.

Although the use of list-assisted RDD methods has increased, the issue of potential non-coverage bias has

remained, stimulating research efforts to quantify the extent of measurement biases. The purpose of this research effort is to provide additional insights into non-coverage bias issues that result from the sample frame truncation inherent in list-assisted frame construction.

Past Research on List-assisted Frame Non-Coverage

Although the sample sizes available for previous research have been large enough to make reasonable estimates of the size of the non-coverage bias, the characteristics of this population are less well established.

Connor and Heeringa (1992) utilized thirty(30) months of data from the Survey of Consumer Attitudes to assess the impact of sample frame truncation. They found that 3.4% of the working residential numbers (WRNs) in Primary Number Sample (Mitofsky-Waksberg design) were not covered in the 1+ listed frame, and that these WRNs not contained in the 1+ listed frame contained 2.7% of eligible households. They were able to make some comparisons of respondent characteristics, the most notable being that younger households, especially in the 18-24 year old range, tended to be over represented in the zero-bank frame.

Research conducted in 1994 by Biemer and Akin, although limited in geographic scope, estimated non-coverage proportions of 2.0% in California and 4.8% in Texas.

The 1995 study by Brick, et al, combined both reclassification of existing first stage selections from two studies using Waksberg designs with an independent study based on direct screening of an *epsem* sample of 10,000 RDD numbers selected within the zero-bank stratum itself. The two estimates of non-coverage based on reclassifying the 5,708 and 10,000 first stage selections yielded estimates of 3.5% and 6.5%, respectively. The third approach, utilizing a single-stage *epsem* RDD sample within the zero-listed stratum, yielded an estimate of 3.7% non-coverage.

It should be noted that Tucker, Casady, and Lepkowski (1993) also investigated zero-bank non-coverage, but their research was limited to ten-series banks and consequently is not comparable to the current work.

Previous research would indicate national zero-bank non-coverage rates of between 3% to 4% of all telephone households. With the exception of the direct screening effort in the Brick collaboration however, all estimates were derived from reclassification of previously screened 1st stage primary numbers from Mitofsky-Waksberg designs. Unknowns vary, but it is unclear in all cases what BELLCORE exchange-types were included in the original frame construction, and what impact and bias unsuccessful/inclusive first-stage screening results may have had on the final estimates.

The current study takes a different approach, utilizing data from personal, face-to-face interviews, during which telephone numbers are requested for subsequent follow-up interviews by telephone whenever possible. Although any biases related to the original telephone screening effort are eliminated, there are other sources of bias relating to the impact of the original response rates, and what impact if any this may have on the results. The major limitation of previous efforts relates to the minimal sample sizes available for zero-listed respondents and the tentative assessment of potential bias that resulted -- this should be eliminated in the current study.

Finally, research efforts into transient telephone/non-telephone households (Keeter, 1995) raises interesting questions regarding a significant component of non-coverage. Concentrating on the population with intermittent telephone service, as defined in successive CPS Annual March Demographic Files, the research found that during a one-year period, nearly 60% of non-telephone households at either point in time, had phone service at the other time. These results argue quite strongly that a significant proportion of non-telephone households are not permanently out-of-scope from a telephone sampling perspective.

Detailed Findings

The data presented here are based on CPS data appended with information by GENESYS Sampling Systems. Each month's CPS study is a nationally representative address-based sample of households. The appended information categorizes the CPS household's telephone number into zero-listed and 1+ listed hundred-series banks, as well as whether the specific telephone number is listed or unlisted in the Donnelley DQI[2] Database.

We used 25 months of CPS data (January 1994 - January 1996) for our analysis. Each month had an average of 49,000 interviewed households. For most of our analyses, all 25 months of data were included, for an aggregate sample size of about 1.2 million household records. However, the CPS is a longitudinal survey. Since interviews are attempted 8 times in 16 months with each household, our aggregate sample

includes multiple interviews with the same household. We derived variances for our estimates based on generalized variance functions which, in addition to accounting for the CPS's clustered sample design, took the overlap of households in monthly samples into account.

Our estimates are averages of the 25 monthly weighted estimates. Comparisons between proportions were made using *t*-tests at the 10% significance level (90% confidence interval). Unless otherwise noted, all differences between proportions are significant at a 10% confidence level.

Each household was categorized as to telephone status based on the answer to the question, "Is there a telephone in this house?" If the answer was no, the household was classified as not having a telephone (for the month in question.)

The CPS does not ask whether there are additional telephone numbers in the household, or what they are. Consequently, some households were coded as not in the 1+ listed frame, when in actuality they may have another phone number that is in the frame. The result of this would be a slight overestimate of the proportion of households excluded from the 1+ list-assisted telephone sampling frame.

Total Household and Telephone Frame Status

Approximately 3.9% of the telephone household records could not be coded as to the type of hundred series bank or listed status primarily because of missing and/or invalid telephone numbers.

Table 1 details the results of this initial classification process. Based on the aggregate 25 months of CPS data, an estimated 6.1 percent of households do not have a telephone. Of the remaining 93.9 percent of households that indicated having a phone, 3.9 percent could not be coded. Based on telephone households that were coded, 2.2% were in zero-listed banks.

The estimated proportion of zero-listed non-coverage is somewhat lower than estimates in previous research, but a number of factors may have influenced those results. Specifically, if we examine the data by calendar quarter over the 25 months, estimates of zero-listed non-coverage range from 1.6 to 2.8 percent. Zero-bank non-coverage bias appears to be larger for list-assisted frames constructed in the Fall and Winter, than other times of the year; similar seasonal variation was also found by Connor and Heeringa in their 1992 study. This variation may be an artifact of the seasonal variations in White Page Directory publication schedules, compounded by the 60-90 day lag following publication required for compilation, verification and updating of the DQI[2] database.

Table 1
Components of Sample Frames

	All Households	*All Telephone Households*	*Coded Telephone Households*
	100.0%		
Non-Telephone	6.1%		
Telephone	93.9%	100.0%	
Not Coded		3.9%	
Coded		96.1%	100.0%
Zero-Listed			2.2%
1+ Listed			97.8%

Potential Influence of Uncoded CPS Records

There were three reasons a household's telephone number could not be classified by hundred-series bank type. The most prevalent (65.8%) was that the telephone number was missing (the respondent refused to give his or her telephone number to the interviewer) or the number given was incomplete or invalid. The second largest group (31.2%) were numbers which were designated by the respondent as their day-time business number. The final group of numbers (3.0%) were assigned to an area outside the sampled household's PSU. This apparently occurred when the interviewer obtained a telephone number for a second home or vacation home for the follow-up interview.

All the differences between the coded and not-coded telephone households were within a few percentage points. Although most of the differences were statistically significant, we do not believe that these records are systematically different from the coded records. Therefore, they were excluded from the remainder of the analyses.

Non-Coverage Estimates for Geographic and Demographic Characteristics

Table 2 provides estimates of non-coverage for a variety of geographic and demographic characteristics. For each characteristic the table provides an estimate of the proportion of telephone households in zero-listed hundred series banks. For comparative purposes we have also included the proportion of non-telephone households for each variable.

Regionally, an RDD 1+ list-assisted frame has estimated non-coverage ranging from a low of 1.4% in the Northeast to a high of 2.9% in the West. This variation by region could be a result of differences in listed status within the underlying database used to define the 1+ list-assisted frame. As will be detailed later, the directory listed rate of CPS telephone households was among the highest in the Northeast, and lowest in the West. The logical explanation would be that as listed rates decrease, marginally populated NPA-NXXs and hundred-series banks have lower likelihood's of identification since the number of assigned residential numbers decreases.

Table 2
Households in 0-Listed 100-Banks as a Proportion of Total Telephone

Region	*0-Listed*	**No Phone*
Northeast	1.4%	5.5%
Midwest	2.1%	5.3%
South	2.4%	8.0%
West	2.9%	5.5%
Household Income (March 1995)		
Less than $5,000	5.0%	24.1%
5,000 to 7,499	3.7%	18.3%
7,500 to 9,999	4.2%	13.9%
10,000 to 14,999	4.6%	11.0%
15,000 to 24,999	4.2%	7.1%
25,000 to 34,999	4.2%	4.7%
35,000 to 49,999	3.5%	3.2%
50,000 to 74,999	2.9%	1.5%
75,000 or more	2.9%	1.4%
Length of Residence (November 1994)		
< 1 Month	11.6%	21.0%
1-6 Months	11.6%	13.8%
7-11 Months	6.6%	10.2%
1-2 Years	3.7%	8.6%
3-4 Years	2.3%	5.8%
5 Years +	1.8%	3.3%
Refused/DK	3.7%	11.7%
Employment Status of Reference		
Employed	2.5%	5.1%
Unemployed	3.5%	16.0%
Not in Labor Force	1.5%	8.0%
Refused/DK	3.5%	2.1%

Household Composition	*0-Listed*	**No Phone*
Husband & Wife	1.7%	3.4%
Single Male w/Family	2.9%	12.0%
Single Female w/Family	2.6%	11.8%
Male Individual	3.6%	12.5%
Female Individual	2.2%	5.5%
Group Quarters	40.9%	23.9%
Children 3 or Under in the		
No Kids 3 or Under	2.1%	5.8%
One 3 or Under	3.2%	9.7%
Two 3 or Under	4.5%	15.9%
Three or More	5.8%	27.9%
Age of Reference Person		
Age 15-17	7.7%	28.1%
Age 18-24	7.7%	15.8%
Age 25-34	3.8%	9.2%
Age 35-54	1.8%	5.7%
Age 55-64	1.3%	4.3%
Age 65 and Up	0.8%	3.6%
Race of Reference Person		
White	2.2%	5.0%
Black	2.3%	14.6%
Am. Indian, Aleut, Eskimo	3.6%	16.8%
Asian/Pacific Islander	3.3%	4.7%
Other	3.6%	14.5%
Hispanic Origin of Reference Person		
Hispanic	2.8%	14.6%
Non-Hispanic	2.2%	5.6%

Differences between categories within characteristics are significant at alpha = .10.
**Proportion without phones is calculated as a percent of total households.*

Sampling Frame Coverage

In general, non-coverage rates appear to be correlated with both income and length of residence [Note: length of residence data estimates are restricted to the November 1994 CPS Supplement]. Of those households with incomes under $5,000, 5.0% of the telephones were in zero-listed banks, compared to just 2.9% of those with incomes over $75,000. Recent movers, have the highest probability of having a telephone number in a zero-listed bank; those in

residence less than six months 11.6%. Of less mobile households, those residing in their current residence for more than 5 years, just 1.8% are estimated to be excluded from a 1+ listed frame. These results are not surprising since telephone number assignment to households who move are much more likely to be within more recently established NPA-NXXs.

Obviously, length of residence and income are not independent. Households with lower incomes tend to be more mobile than those with higher incomes and, as discussed by Keeter (1994), are more likely to be "transient telephone households" -- households with interruptions in telephone service over time. Employment status, as it can be related to income, indicates that non-coverage among Unemployed households is higher than among Employed (3.5% vs. 2.5%). Non-coverage of households where the reference person is Not in the Labor Force is lowest (1.5%), but it should be noted that these households are more likely to be older, and the reference person retired.

Husband-Wife headed units are estimated to have just a 1.7% level of non-coverage, while households headed by single males and male individuals have zero-bank non-coverage levels of 2.9% and 3.6%, respectively. In comparison, over forty percent of families residing in Group Quarters are estimated to not be covered within the 1+ list assisted frame.

Presence of young children, under three years of age, is again highly correlated with younger, lower income, more mobile households. Consequently, it is not unexpected to see higher levels of non-coverage for such households. An estimated 3.2% of households with one child under three, increasing to 5.8% of those with three or more children under three years of age can be assumed to be excluded from a list-assisted frame.

Like Conner and Heeringa (1992), we also found that households with younger heads were more likely to be in the zero-listed 100-banks. Finally, non-coverage estimates by race show little difference for White and Black households: 2.2% vs. 2.3%. However, the rate of Hispanic non-coverage is about 25% higher than non-Hispanic households (2.8% vs. 2.2%).

Table 3
Comparison of 100% Telephone Frame and 1+ Listed Frame

	1+ Listed Hhlds.	*All Phone Hhlds.*	*% Diff.*
Employment Status of Reference Person			
Employed	65.88%	66.05%	-0.3%
Unemployed	2.74%	2.78%	-1.3%
Not in the Labor Force	30.62%	30.40%	0.7%
Number of Children Aged 3 or Under in the Hhld			
None Aged 3 or Under	90.09%	89.97%	0.1%
One Aged 3 or Under	8.76%	8.85%	-1.0%
Two Aged 3 or Under	1.09%	1.12%	-2.4%
Household Composition			
Husband & Wife	56.07%	55.79%	0.5%
Single Male w/Family	3.23%	3.25%	-0.6%
Single Female w/Family	11.56%	11.60%	-0.4%
Male Individual	12.55%	12.72%	-1.4%
Household Income (March 1995)			
Less than $5,000	3.26%	3.30%	-1.4%
10,000 to 14,999	8.54%	8.61%	-0.9%
15,000 to 24,999	16.41%	16.51%	-0.6%
25,000 to 34,999	14.36%	14.44%	-0.5%
35,000 to 49,999	16.98%	16.95%	0.2%
50,000 to 74,999	17.65%	17.49%	0.9%
75,000 or more	14.47%	14.34%	0.9%
Age of Reference Person			
Age 18-24	4.64%	4.91%	-5.6%
Age 25-34	19.08%	19.40%	-1.6%
Age 35-54	41.31%	41.15%	0.4%

Differences are statistically significant at alpha = .10.

Estimates of Total Sample Bias

Table 3 details estimates of household characteristics one would expect from the Total Telephone frame, and from a 1+ list-assisted sampling frame. Most of the relative differences were found to be less than 1.5%. However, due to the large sample sizes involved, all of these differences are significant at the 90% level. For illustrative purposes, we have highlighted five demographics with the most significant differences between the two frames. We would expect 2.78% to be unemployed where the sample frame is all telephone households, compared to 2.74% using a 1+ list-assisted frame. With a list assisted frame one would expect 1.4% fewer respondents with incomes under $5,000;

2.4% fewer interviews among households with two(2) children under age three; etc.

Summary of Results and Conclusions

Differences between households with and without telephones are related to several sociodemographic characteristics. This has been demonstrated by Thornberry and Massey, Brick et. al. and Keeter. Our data coincide with past research.

Although there are many differences between households with and without telephones, the percentage of households in zero-listed 100-banks across most sociodemographic characteristics appears to be fairly stable. The only sizable differences were found for young households (18-24) and recent movers. Overall, the modest level of 2.2% non-coverage in a 1+ list-assisted frame found in this study is encouraging.

However, it should be noted that the non-coverage bias inherent in a list-assisted telephone frame appears to compound the bias introduced by eliminating non-telephone households. For the most part higher levels of list-assisted non-coverage occur among the same demographic groups that exhibit high proportions of non-telephone households.

Income and mobility effects seem to be pervasive influences in determining the level of non-coverage. However, for even large scale surveys, the differences in expected sample distributions is small, and highly amenable to correction with minimal weighting.

To the extent recent movers who change their telephone numbers are assigned to new exchanges and hundred series banks, there is an inherent, systematic mechanism at work that will guarantee some minimum level of non-coverage. Short of a more comprehensive, more current source, households in new exchanges will continue to be systematically underrepresented in list-assisted frames.

Future Research

Our primary findings so far lead us to believe that the bias introduced by truncating zero-listed hundred-series banks is small. Given that, however, what is the magnitude of this bias for different estimates? If we can measure the bias, can we improve telephone survey estimates by adjusting for it based on our data? Since the non-telephone population is a much larger source of bias in a telephone survey, can we quantify this bias and adjust telephone survey data to further improve estimates?

A critical follow-up to this detailed demographic analysis will involve exploitation of the longitudinal aspects of the underlying database. How stable is the sample frame? Are the characteristics of the "out-of-scope" population highly variable over time, or are the changes reflective of overall telephone household characteristics?

Again, the work to date has served to confirm and extend previous research efforts limited by sample size. The preliminary analyses however will form a firm basis for examining the implications of alternative frame construction parameters as they relate to potential non-coverage bias, while providing reliable estimates for the development of adjustment factors to compensate for that non-coverage.

References

Biemer, Paul and D. Akin. 1994. "The Efficiency of List-Assisted Random Digit Dialing Schemes for Single and Dual Frame Surveys", *1994 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1:1-10.

Brick, J. Michael, J. Waksberg, D. W. Kulp, and A. Starer. 1995. "Bias in List-Assisted Telephone Samples", *Public Opinion Quarterly* 59(10):218-35.

Casady, Robert J. and J. M. Lepkowski. 1993. "Stratified Telephone Survey Designs", *Survey Methodology* 19(1): 103-13.

Chilton Research, Inc. (1972). "National Telephone Household Probability Sampling Methodology", Unpublished Manuscript.

Connor, Judy, and S. Heeringa. 1992. "An Evaluation of Two Cost Efficient RDD Designs", Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, Fl.

Keeter, Scott. 1995. "Estimating Noncoverage Bias from a Phone Survey", *Public Opinion Quarterly* 59(10):218-35.

Thornberry, O.T. Jr. and Massey, J.T. 1988. "Trends in US Telephone Coverage Across Time and Subgroups." In Groves, R.M., et al, (eds). 1988 Telephone Survey Methodology. New York: John Wiley & Sons. pp 25-49.

Tucker, Clyde, R. M. Casady, and J. M. Lepkowski. 1993. "A Hierarchy of List-Assisted Stratified Telephone Sample Design Options", Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Charles, IL.

Waksberg, Joseph. 1978. "Sampling Methods for Random Digit Dialing", *Journal of the American Statistical Association*, 73(361):40-46.