Public Opinion Quarterly, Vol. 73, No. 4, Winter 2009, pp. 751–760

A REASSESSMENT OF LIST-ASSISTED RDD METHODOLOGY

MANSOUR FAHIMI
DALE KULP
J. MICHAEL BRICK

Abstract Random digit dial (RDD) sampling methodology was developed over two decades ago when local telephone exchanges used 100-series telephone banks as physical building blocks for telephone number assignment. During the intervening decades the telecommunication industry has undergone a number of fundamental changes that have had drastic effects on the efficiency and coverage of RDD samples. Based on a new study of landline telephone numbers, which was conducted during the second quarter of 2008, this paper reexamines the assumptions that are relied upon for construction of list-assisted RDD samples; quantifies the extent of undercoverage in the corresponding sampling frames; investigates alternative methods of frame construction for RDD samples; and evaluates other sample design options that can offer greater coverage with varying degrees of efficiency.

Introduction

The Mitofsky–Waksberg method of random digit dialing (Waksberg 1978) was a major breakthrough in telephone survey research methodology that improved the efficiency of telephone sampling and allowed researchers to sample from both listed and unlisted telephone numbers. This method introduced the concept of sampling from 100-series banks (telephone numbers with the same first eight digits) as a cluster of telephone numbers. However, operational complexities led researchers to look for further refinements and examine alternative

MANSOUR FAHIMI AND DALE KULP are with Marketing Systems Group, 565 Virginia Drive, Fort Washington, PA 19034, USA. J. MICHAEL BRICK is with Westat, Inc., 1600 Research Blvd., Rockville, MD 20850-3129, USA. Address correspondence to Mansour Fahimi; e-mail: mfahimi@m-s-g.com.

doi:10.1093/poq/nfp066 Advance Access publication November 24, 2009

designs. In particular, Casady and Lepkowski (1993) studied a "truncated" design alternative that only included telephone numbers in 100-series banks with at least one listed residential telephone number (1+ list-assisted design). Since the 1+ list-assisted design excluded residential numbers that were part of the 100-series banks with no listed residential numbers (0− listed banks), the resulting efficiencies were at the expense of reduced coverage. Connor and Herringa (1992) and Brick et al. (1995) estimated that the coverage loss due to residential numbers in 0− listed telephone banks was less than 4 percent for national studies. Moreover, their empirical work revealed little evidence of bias due to exclusion of such households. Consequently, this design and its variants based on larger numbers of listed telephone numbers per bank became the most commonly used method of RDD sampling.

More recently, Tucker, Lepkowski, and Piekarski (2002) reexamined the 1+ list-assisted sampling methodology and concluded that such designs were even more efficient today than RDD samples based on simple random selection. This examination, however, did not reassess some of the assumptions key to the Casady and Lepkowski (1993) study.

This paper provides results from a new study that examines the underlying assumptions for construction of RDD frames and evaluates the coverage and efficiency of the resulting samples. This examination is especially important given the many changes that have been introduced to the telephony environment. For example, while the number of 1+ listed banks increased from 1.7 million to 2.9 million between 1990 and 2008, in the same period the number of 0− listed banks increased from 2.7 million to 6.1 million. Brought on by a digital revolution, this devolution of the telephone system in the US has largely eliminated the utility of 100-series banks for assigning residential telephone numbers, as well as their relevance to the process of RDD frame construction.

As described by Casady and Lepkowski (1993), there are three key parameters for evaluating the coverage and efficiency of stratified list-assisted designs. For a given stratum (i) these parameters are: the proportion of all telephone numbers that are in that stratum (P_i); the proportion of all residential numbers that are in that stratum (Z_i); and the proportion of all numbers in that stratum that are residential, called the hit rate (H_i). While exact values are available for P_i from the sampling frame, the latter two parameters have to be inferred from the sample via point estimates z_i and h_i, respectively. Our study estimates each of these parameters *de novo*.

We begin by presenting an overview of the sample design for a study conducted to examine the current distribution of residential telephone numbers across three different strata (bank types). Next, we provide current estimates for the extent of coverage loss in RDD samples and the percentage of telephone numbers expected to be residential in each stratum. These estimates are then used to evaluate alternative sample designs that can provide greater coverage at the expense of reducing the residential hit rates.

Table 1. Number of 100-Series Telephone Banks, and Sampled Telephone Numbers by Stratum

	100-series telephone banks		Sampled telephone numbers	
Stratum	Number	Percent (100 × P_i)	Number	Percent
0− listed	4,009,944	44.4	9,062	23.8
1+ listed	2,920,039	32.3	20,000	52.6
Remainder	2,099,950	23.3	8,937	23.5
Total	9,029,933	100	37,999	100

Telephone Frame Evaluation Study and Results

Marketing Systems Group (MSG) selected a stratified sample of 37,999 telephone numbers from three strata of telephone numbers that collectively constitute the entire pool of available landline telephone numbers. The first stratum consisted of telephone numbers in 0− listed banks that are part of telephone exchanges with at least one listed number. The second stratum included all telephone numbers in 1+ listed banks, which constitutes the frame used in most of the current RDD designs. The third stratum included telephone numbers in all remaining POTS (plain old telephone service) 100-series and mixed-use banks that are in exchanges with no listed numbers.

Table 1 provides a summary of the sample design used for this study along with values for the parameter P_i as a percentage. In particular, it shows that slightly more than 30 percent of all the telephone numbers are in the 1+ listed stratum, which is the same percentage Tucker, Lepkowski, and Piekarski (2002) reported for 1999. While the sample of 20,000 telephone numbers from this stratum was selected using a disproportionate stratified sample design, those for the 0− listed and the remainder strata were selected in a simple random fashion because no meaningful stratification variables were available. In order to obtain the other two key parameter estimates, z_i and h_i, the sample of 37,999 telephone numbers was processed in three consecutive steps. The three steps were as follows:

1. All numbers were dialed no less than nine times by MSG, resulting in a final status for all but 2,722 numbers;
2. The 2,722 undetermined (no answer or busy) numbers were sent to a vendor that can determine the status of many telephone numbers, as a result of which 825 numbers were assigned a final status and 1,897 still remained undetermined; and
3. The 1,897 undetermined numbers were sent back to the same vendor for a more extensive processing, resulting in a final status for an additional 844 numbers. Ultimately, only 1,053 numbers (less than 3 percent of the sample) ended up with a final undetermined status.

Table 2. Weighted Distribution of the Final Dispositions Across Strata (Coverage Rates)

Disposition	0− listed	1+ listed	Remainder	Total	Sample size
Residential (100 × z_i)	14.5%	80.5%	5.0%	100%	7,868
Business	51.2%	35.7%	13.1%	100%	2,956
Cell	49.9%	15.2%	34.9%	100%	291
Nonworking	49.1%	23.9%	27.0%	100%	23,506
Pager/Fax/Modem	36.5%	28.4%	35.1%	100%	1,620
Undetermined	49.1%	30.5%	20.4%	100%	1,758

Table 3. Weighted Distribution of the Final Dispositions within Strata (Hit Rates)

	Stratum		
Disposition	0− listed	1+ listed	Remainder
Residential (100 × h_i)	4.0%	30.8%	2.7%
Business	9.6%	9.2%	4.7%
Cell	1.0%	0.4%	1.3%
Nonworking	75.8%	50.6%	79.8%
Pager/Fax/Modem	3.7%	4.0%	6.9%
Undetermined	5.8%	5.0%	4.6%
Total	100%	100%	100%
Sample size	9,062	20,000	8,937

Design weights were then applied to each sample telephone number to reflect the stratified design to produce unbiased estimates. Tables 2 and 3 summarize the resulting weighted estimates for each final status across and within bank types. The first row of table 2 provides an estimate of the percentage of residential numbers that are in each stratum (100 × z_i) while the first row of table 3 provides an estimate of the residential hit rate (100 × h_i). The undetermined numbers are included as a legitimate disposition in these tables.

The most striking result shown in these tables is the percentage of the residential numbers that are excluded from the 1+ list-assisted frame (table 2). This percentage has sky-rocketed from less than 4 percent in the early 1990s to almost 20 percent in 2008. Over 70 percent of this coverage loss is attributed to residences whose telephone numbers are now in 0− listed banks within exchanges that have at least one listed number. It should be noted that these coverage losses are from households with landlines, without any consideration for the uncovered households that are reachable only with cell phones.

Another remarkable result is from table 3 showing that the residential hit rate for the 1+ listed stratum is now about 30 percent, which is drastically lower than the 49 percent hit rate reported by Tucker, Lepkowski, and Piekarski (2002). This is partly because the hit rate from that study was not directly estimated; it was extrapolated based on mathematical relationships between the parameters and the screening results from the Survey of Consumer Attitudes of 1999.

The above substantial changes are in part due to the transition in the telephony network. For example, Competitive Local Exchange Carriers (CLECs) are now important providers of telephone service, accounting for 34 percent and 72 percent of all the residential numbers in the 0− listed and the remainder strata, respectively. In contrast, about 95 percent of residential numbers in the 1+ listed stratum are handled by long-time telephony providers such as Regional Bell Operating Companies (RBOC) and Incumbent Local Exchange Carriers (ILEC).

Because of the importance of parameter estimates z_i and h_i in our subsequent calculations, we obtained the reported residential hit rates (h_i) from other large RDD surveys conducted recently as well. Two studies that used the standard 1+ listed design were the 2007 National Household Education Survey, and the 2006 National Immunization Survey. Both reported residential hit rates in the range of 28 to 29 percent. Since both of these surveys were from samples generated by MSG, we also obtained the residential hit rate from the 2007 Motor Vehicle Occupant Safety Survey for which the sample was obtained from Survey Sampling International (courtesy of Abt SRBI). The sample for this survey was selected from 3+ listed banks to improve its efficiency and had a reported residential hit rate of about 31 percent. Since a slightly higher hit rate is expected for 3+ listed as compared to 1+ listed banks, it can be concluded that all three surveys are consistent with an estimated hit rate of about 30 percent for the 1+ listed stratum in 2008.

While these findings show the coverage loss for 1+ listed RDD designs is now close to 20 percent, it does not provide direct estimates of the potential bias because the characteristics of households are only available for the 1+ listed stratum. We are unaware of recent surveys that have included telephone numbers outside of the 1+ listed banks to make such estimates. Moreover, the frame information available for residential numbers in the other two strata is limited to variables such as geography and carrier type, and these provide only anecdotal clues regarding any potential bias. For example, about 63 percent of the residential numbers in the remainder stratum are urban as compared to 47 percent and 32 percent in the 0− listed and 1+ listed strata, respectively. While we discuss the effect of coverage bias later, studies that include samples from the 0− listed and the remaining strata are needed to draw firm inferences about this potential bias.

Alternative Sampling Approaches

In this section we investigate alternative sample designs that can offer greater coverage for the landline households with varying degrees of efficiency. While an alternative based on the Mitofsky–Waksberg scheme seems academically intriguing, for practical considerations our investigation does not include such an alternative. Most notably, our research shows that the first- and second-stage hit rates today are much lower than those discussed by Waksberg in 1978. Moreover, such an alternative would mean having to deal with the operational complexities of this method discussed by Casady and Lepkowski (1993) that lead researchers to look for other options.

A more practical set of design alternatives redefines the strata used by Casady and Lepkowski (1993) to include a greater percentage of residential numbers from the sampling frame. While such alternatives will improve coverage, they are expected to be less efficient because of their reduced residential hit rates. Analogously, our approach relies on the optimal allocation scheme ($n_i \propto z_i\sqrt{h_i(1+(\gamma-1)h_i})$) that depends on the ratio of the total cost of data collection for a completed interview to the cost of sampling and dialing a number that is not residential (γ).

We examined four alternative sample design options, three of which are based on the 100-series banks while the fourth alternative is based on the 1,000-series blocks with any listed numbers. In the latter design alternative, each 1,000-series block consists of ten 100-series banks from the listed exchanges with at least one of these banks being 1+ listed. These design alternatives are as follows:

A. 3-stratum design including the 1+ listed, 0− listed, and the remainder banks;
B. 2-stratum design including the 1+ listed and 0− listed banks;
C. 1-stratum design including only the 1+ listed bank; and
D. 1-stratum design including only the 1+ listed 1,000-series blocks.

With $\bar{h}$ representing the average residential hit rate in the population, the efficiency of a design option can be measured in terms of the ratio of the variance of the stratified sampling estimator to the variance from a simple random sampling (design effect). Using the notation and terminology of Casady and Lepkowski (1993) this estimated design effect is given by

$$R = 1 - \bar{h} \times \frac{\left\{\sum_i \left[z_i\sqrt{\frac{1+(\gamma-1)h_i}{h_i}}\right]\right\}^2}{1+(\gamma-1)\bar{h}}.$$

Table 4 summarizes the input parameters and the resulting projections of efficiency (R) for each design when the cost ratio (γ) is assumed to be 2. For design alternatives based on the 100-series banks, the 3-stratum design option ***A*** has no coverage loss for landline households but is the least efficient design with an estimated variance reduction of about 20 percent. That is, with $R = 0.193$ a variance estimate under this design is expected to be roughly 80 percent of what might result under simple random sampling.

Table 4. Parameter Estimates, Allocation, Projected Efficiency (R) Compared to Simple RDD Sampling, Percentage of Telephone Households not Covered Assuming a Cost Ratio of $\gamma = 2$

Design Option	Stratum	z_i	h_i	Allocation (n_i)	Reduction in variance (R)	Coverage loss
A	1+ listed	0.805	0.308	1.268	0.193	0.0%
	0− listed	0.147	0.040	0.721		
	Remainder	0.048	0.027	0.288		
B	1+ listed	0.850	0.308	1.332	0.294	5.0%
	0− listed	0.154	0.040	0.757		
C	1+ listed	1.000	0.308	1.000	0.532	19.5%
D	1+ listed 1,000-blocks	1.000	0.222	1.000	0.393	13.2%

By comparison, the 1-stratum design option ***C*** excludes about 20 percent of all landline households but is the most efficient option with a variance reduction of more than 50 percent. The compromise 2-stratum design option ***B*** has a coverage loss of only about 5 percent and a variance reduction of about 30 percent. Lastly, the 1-stratum design option ***D*** based on the 1,000-series blocks has variance reduction of about 40 percent and loss in coverage of slightly more than 13 percent. Design option ***D*** is more efficient than the 2-stratum design option ***B*** and less efficient than the 1-stratum design option ***C***.

In order to assess the robustness of these findings with respect to the assumed ratio of the total cost of data collection to the marginal cost of identifying residential numbers at $\gamma = 2$, the variance reduction for different values of this cost ratio is shown in figure 1. This figure suggests that as the relative cost of collecting data increases, the relative benefits of optimal allocation for all design options decrease.

Moreover, one can examine the tradeoff between variance reduction and bias inflation due to increased undercoverage. For a given design alternative, it can be assumed that the population of interest comprises of two strata of sampling units: one that is covered by the corresponding sampling frame; and a second stratum of units not covered by the given frame. Assuming that the covered stratum captures C percent of the total population, any percentage parameter of interest in the entire population, P, can be expressed as $P = CP_c + (1 - C)P_n$ where P_c and P_n are the component parameters from the two strata, respectively. Under these assumptions, the bias of an unadjusted estimator $\hat{p}_c$ obtained from the covered stratum can be expressed by

$$\begin{aligned} Bias(\hat{p}_c) &= P_c - P \\ &= P_c - [CP_c + (1 - C)P_n] \\ &= (1 - C)(P_c - P_n) \end{aligned}$$

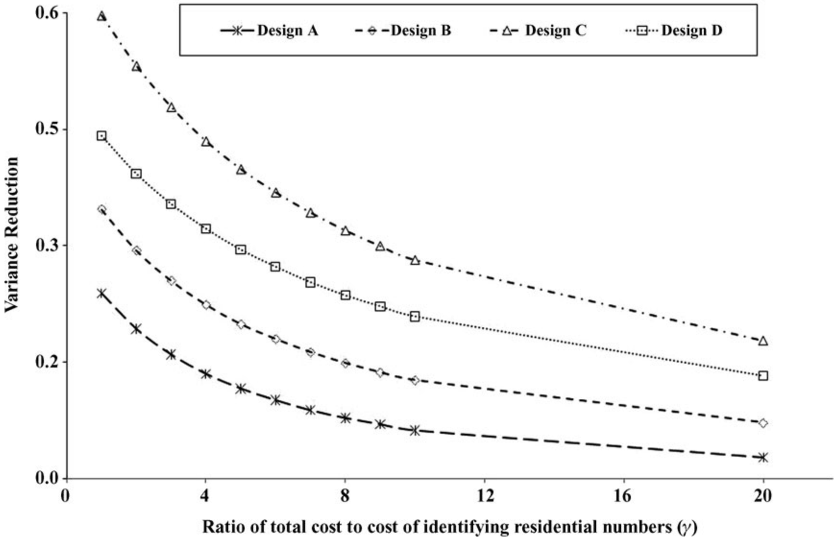

Figure 1. Relative efficiency as compared to simple RDD sampling as a function of cost parameter γ and design alternative.

Should P_c and P_n be available for a particular statistic under each design, a measure of accuracy could be computed to reflect both the resulting bias inflation and variance reduction. The standard approach is to use the root mean squared error (*RMSE*) of the estimate, which with n as the sample size can be approximated by

$$\begin{aligned} RMSE(\hat{p}_c) &= \sqrt{Var(\hat{p}_c) + [Bias(\hat{p}_c)]^2} \\ &\cong \sqrt{(1-R)\frac{P(1-P)}{n} + [(1-C)(P_c - P_n)]^2} \end{aligned}$$

Accordingly, the bias contribution to the *RMSE* can be substantial when n is large or when the difference between P_c and P_n is relatively large. For example, with $P_c - P_n = 0.10$ and $P \cong 0.5$, figure 2 shows that for sample sizes exceeding 1,000 the *RMSE* is dominated by the bias and the unbiased design ***A*** has the lowest *RMSE*. With more modest differences, such as $P_c - P_n = 0.04$, the other designs have somewhat lower *RMSE* until the sample size approaches 2,500. It should be noted that meaningful evaluations of *RMSE* require practical estimates of P_c and P_n for different statistics. However, as discussed earlier, such estimates are currently not available.

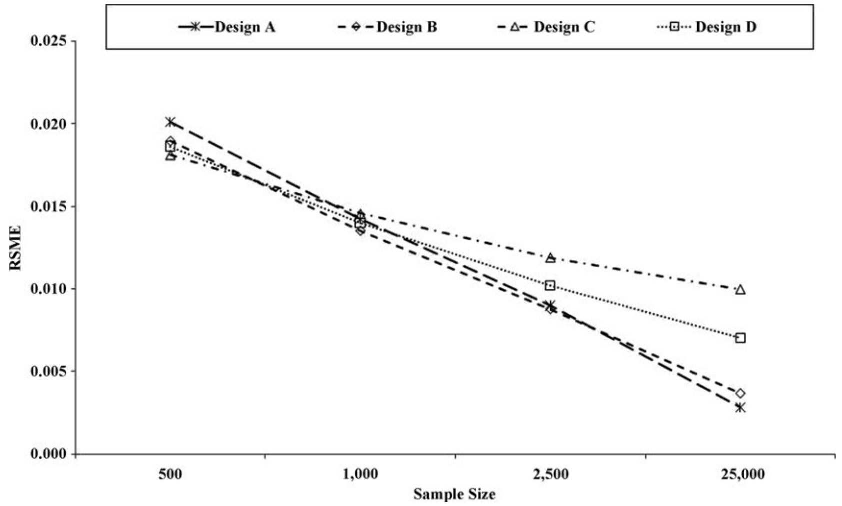

Figure 2. Root mean square error as a function of design alternative and sample size assuming $P_c = 0.5$ and $P_n = 0.45$.

Summary

Digital transition in the US telephony infrastructure has greatly reduced the utility of 100-series banks for constructing RDD sampling frames. As such, samples selected from the traditional frame that includes only 1+ listed 100-series banks entail a much larger coverage loss than suggested by previous studies due to the decrease in residential number assignment density and the increase in alternative providers. An article by Boyle, Bucuvalas, and Piekarski (2009) using somewhat different methodology has estimated the undercoverage in the 1+ listed frame to be lower than what is reported here. We expect continued research on this topic will help clarify these differences.

In order to reduce the loss in coverage associated with sampling from the traditional frame, we considered different stratified design options. Others designs could and should be examined with respect to their efficiency and coverage as well. Furthermore, information about the characteristics of households in the different sampling strata is also needed to better understand the trade-offs between efficiency and coverage. Without such data, direct estimates of coverage bias are not possible.

Lastly, throughout this paper we have discussed coverage of the landline telephone households. Obviously, the effect of cell phones on coverage of the entire population cannot be overlooked. RDD sample designs that ignore cell phones and use the standard 1+ listed design can exclude well over 30 percent of the population with the potential for substantial coverage bias. Furthermore, an increasing percentage of adults live in households where cell phones are predominately used for receiving telephone calls (Blumberg and Luke 2008). Technically, these *cell-mostly* adults are covered in the landline frame, but they may be difficult to reach.

References

Blumberg, Steven J., and Julian V. Luke. 2008. "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January–June 2008." Hyattsville, MD: National Center for Health Statistics.

Boyle, John, Michael Bucuvalas, and Linda Piekarski. "Zero Banks: Coverage Error in List Assisted RDD Samples." *Public Opinion Quarterly* doi:10.1093/poq/nfp068.

Brick, J. Michael, Joseph Waksberg, Dale Kulp, and Amy Starer. 1995. "Bias in List-Assisted Telephone Samples." *Public Opinion Quarterly* 59:218–35.

Casady, Robert J., and James M. Lepkowski. 1993. "Stratified Telephone Survey Designs." *Survey Methodology* 19:103–13.

Connor, Judy, and Steven Herringa. 1992. "Evaluation of Two Cost Effective RDD Designs." Paper presented at the Annual Conference for the American Association for Public Opinion Research. St. Pete Beach, FL, May 18.

Tucker, Clyde, James M. Lepkowski, and Linda Piekarski. 2002. "The Current Efficiency of List-Assisted Telephone Sampling Designs." *Public Opinion Quarterly* 66:321–38.

Waksberg, Joseph. 1978. "Sampling Methods for Random Digit Dialing." *Journal of the American Statistical Association* 73:40–6.