DYNAMICS OF "LIST-ASSISTED" RANDOM DIGIT DIALING(RDD) FRAME COVERAGE

Dale W. Kulp, GENESYS Sampling Systems
565 Virginia Drive, Fort Washington, Pennsylvania, 19034

KEYWORDS: RDD, telephone sampling, non-coverage bias.

Interest in sampling applications using list-assisted RDD frames has increased over the past few years. Their use by academic and social science researchers has also heightened interest in research on list-assisted frame limitations and variations (Casady, et al, 1993). And, recent efforts have provided reliable estimates of non-coverage resulting from the truncated sampling frame integral to the list-assisted definition process (Connor and Herringa, 1991; Brick, et al, 1994).

Less understood are the dynamics of the list-assisted sampling frame over time. This paper focuses on the incorporation of "new" NPA-NXX (i.e., Area Code/Exchange) combinations into the list-assisted frame, through the use of a longitudinal database designed to track listed telephone household assignments. The investigation pays particular attention to the time-lags inherent in the underlying listed household databases, as well as the impact of the listed household cut-off criterion on the accurate incorporation of new NPA-NXX's into the list-assisted frame.

Background. Since the late 1970's, the preferred method for RDD telephone sampling among social science and academic researchers has been the Mitofsky-Waksberg method for random digit dialing (Waksberg, 1978). Over the past few years, there has been increasing interest in a class of methods known generically as "list-assisted" RDD methodologies. Although a wide range of both proprietary and commercial "list-assisted" variants have been used for over twenty-years, these methods have been almost exclusively limited to commercial survey and market research applications.

To a large extent these two classes of RDD methods developed and matured in virtual isolation. Historically, "list-assisted" RDD sampling was restricted to a handful of large survey research organizations and a few commercial RDD suppliers. With some exceptions, the primary concerns related to data collection efficiency while maintaining some statistical face-validity. Moreover, documentation of procedures, standards, sources, and sampling methodologies themselves were minimal and often non-existent. Few commercial researchers had a full understanding of RDD sampling or the limitations and potential biases of the methods they advocated and employed. List-assisted RDD sampling was little understood, little researched, and virtually ignored by social scientists.

The popularization of the Mitofsky-Waksberg(M-W) method accomplished two significant things. First, it provided a reliable and cost-efficient means for organizations of all sizes to produce and implement RDD samples. But most importantly, it eliminated problems with commercially available list-assisted frames resulting from lack of standards and the resulting indeterminacy of frame coverage. In comparison to list-assisted RDD methods, significant research efforts have been devoted to M-W methods, with a primary focus being the exploration of various procedural changes to minimize difficulties inherent in the sample administration and data collection phase. These difficulties specifically relate to the two-stage design and the sequential nature of the second stage sampling process, which often result in cost and timing difficulties.

In contrast, list-assisted sample frames are based on a subset, or truncated version of the M-W sample frame. Although the exact procedural details and parameters used to accomplish this truncation vary, the data sources used to determine the portions of the frame to be retained are fairly standard, and are derived from household databases compiled from local telephone directories. Hence, the term list-assisted is derived from the method's reliance on lists, or databases of directory-listed telephone households to define the sample frame. It is important to note however that the term list-assisted does not infer that the resultant sample contains only listed telephone households. Rather, the term list-assisted is descriptive of the data sources used to define, truncate, and restrict the sample frame itself.

A distinct advantage of list-assisted frames is the ability to utilize single stage, unclustered sample designs.
Most research on list-assisted sampling has focussed on the characteristics of the truncated portion of the sampling frame. In other words, the bias one can potentially introduce by using a truncated frame. This work has fairly well defined the proportion of telephone households contained in the truncated portion of the frame, and provided some directional information regarding the total potential bias involved.

Both the included and excluded portions of the full frame are far from static. New residential exchanges are established; residential assignments are made in these exchanges; some of these are listed households which are then included in the local telephone directory; these new listings are then incorporated by the database compiler; and ultimately, they identify the portions of those new exchanges which are then included in the list-assisted frame. As one might surmise, there are significant time lags in this process.

The process by which telephone exchanges migrate between the non-covered and covered portions of the total frame has escaped research. How quickly do portions of the truncated, excluded frame migrate to the list-assisted frame. How do the definitional parameters used to construct the frame influence that rate of change? And, what does this imply for the frequency of updating the list-assisted sample frame itself?

Constructing the List-Assisted RDD Frame. To understand the dynamics of the two frames and the ultimate impact on coverage issues, it is necessary to review the critical steps involved in defining a list-assisted RDD frame. [For reference purposes this overview is based on the procedures employed by GENESYS Sampling Systems in their twice-annual update process - see GENESYS Sampling Methodology.]

The basic sources required for constructing the list-assisted frame are the BELLCORE V&H Coordinate Tape and the Donnelly DQI2 Database. The BELLCORE File is available from Bell Labs on either a monthly, quarterly or annual basis. This database contains every authorized NPA-NXX (Area Code and telephone exchange) combination in North America. In addition, it provides specific information about each NPA-NXX, including the "type(s) of service" the exchange provides, its geographic location, etc. Currently there are 21 types of service, of which only five(5) are utilized in constructing the GENESYS RDD frame. The Service-Type Codes, their description and relative frequencies are shown in Table 1.

The five(5) specific service types relate to NPA-NXX's which nominally can provide residential telephone service - POTS or "plain old telephone service." This set of 45,566 unique combinations typically represents the untruncated sample frame from which the 1st stage sampling units (PSU's) would be constructed and selected for a Mitofsky-Waksberg design. However, in constructing a list-assisted frame the next step is to identify the portion to be truncated, or excluded.

A typical Mitofsky-Waksberg sample will further subdivide the 45,566 potentially residential exchanges into "hundred-series" blocks, or banks. These blocks represent the first eight digits of the normal ten-digit telephone number. Each exchange contains 100 such blocks (i.e., NPA-NXX-00xx, NPA-NXX-01xx, ... NPA-NXX-99xx), with each hundred-series containing 100 two-digit extensions, from 00 to 99. Consequently, the full frame contains 4,556,600 unique hundred-series blocks, and 455,660,000 unique residential numbers.

TABLE 1. Frequency distribution, NPA-NNX codes from March 1994 Bellcore file.

BELLCORE Code	Description	Frequency
00	Regular(POTS)* Service	44,349
50	Shared among 3 or more Service Types	132
51	Shared between POTS and Mobile	81
52	Shared between POTS and Paging	821
54	Shared between POTS and Cellular	183
-	Non-residential Service Types	5,456
	Total	51,022

*(Plain Old Telephone Service)

Most list-assisted frames also employ the hundred-series construct, but restrict the Mitofsky-Waksberg frame by classifying the individual blocks as nominally "working" or "nonworking." A block is determined to be "working" based on whether or not it contains residential telephones which are known to be assigned and in service. Determination of working status is accomplished via a list, or database of residential telephone number assignments. Hence the name list-assisted is derived from the use of a list of residential assignments which identifies the nominal "working portion" of the comprehensive frame.

A "list-assisted" sampling frame is thus a subset, or truncation of the Mitofsky-Waksberg frame. The subset being all hundred series blocks with some predetermined level or threshold of residential service. Usually this threshold is expressed in terms of the number of listings which occur within a hundred-series bank and could be set at one, two, three, or more residential assignments. Commercial RDD sample suppliers usually have a standard threshold, or "cut-off," set at two or three listings, while others offer databases constructed to the users specifications, ranging from one listed household upwards.

The process of identifying the "list-assisted" frame is fairly straightforward once one is familiar with the anomalies present in the listed household source databases. The Donnelly DQI2 Database is our preferred source for defining the frame, and it contains approximately 65,000,000 individual listed (published) telephone household records, along with geographic codes such as census tract/block group and ZIP+4 for each residence.

These databases are far from ideal for the purpose of constructing a list-assisted frame. There are regulatory restrictions which of course limit the information contained to directory listed telephone households, and even among these, there are millions of records for which usable geographic location data is either missing, imputed, or assigned. Even more important, the compilation and verification processes affect the timeliness of incorporating new and/or changed residential listings into the database. The sheer size of the listed household database provides numerous opportunities for partial or inaccurate incorporation of changes in area codes, exchanges and underlying geographic codes.

Many of these limitations are relevant to the research below and will be discussed in more detail, but it is important to realize that these databases will always contain some variance relative to the BELLCORE file and significant time-lags resulting from the telephone directory publication and lengthy database compilation processes. Detailed and time-consuming screening/correction procedures can eliminate the classification problems, but the time-lags inherent in the underlying source database, results in a list-assisted frame which is anywhere from about twelve to eighteen months out of date when constructed.

The time-lag in the underlying data source compounds the difficulty of accurately identifying the exchanges and associated working banks which actually provide residential service. Consequently, the composition of the truncated portion of the sampling frame contains hundred series blocks of three types:

I hundred-series banks with no current or future plans for residential assignments or service;

II banks with residential service, but no "listed households;"

III banks with listed telephone households, but those households have not as yet been incorporated into the listed household database.

Unfortunately, it is impossible to distinguish between these types of hundred-series banks prior to construction of a list-assisted frame. But if it were possible, the potential non-coverage bias attributable to using a truncated, list-assisted RDD frame could be significantly reduced by simply including Types II and III, or assigning them to a separate stratum.

An important aspect of list-assisted frame construction is the time-consuming and computer resource intensive operations required. To that extent the operational focus is primarily on the current, point-in-time database sources required for the periodic frame definition efforts. Over the years, we have devoted some intermittent resources and efforts at analyzing changes in the list-assisted frame over time; the time-frames involved have been very short, typically one year at most. And, the primary focus has been to demonstrate the relative stability of the frame over the short-term (e.g., between semi-annual updates).

One thing that is perfectly clear is that little is known regarding the underlying dynamic process that begins with authorization of a new NPA-NXX, and its partial or complete incorporation into the list-assisted frame some time in the future. The objectives then, of this investigation are really three-fold:

1) To provide an initial set of longitudinal data that enables one to track the residential assignment process over time;

2) To provide a preliminary look at the dynamics of this process, the time frames involved and the implications for improving list assisted frame construction, update frequencies, etc.; and,

3) To examine the extent to which retrospective data might provide useful information for classifying NXX's as to their likelihood of providing residential service.

It should be clarified that this paper is based on results from the initial development phase of a comprehensive longitudinal database, tracking the listed household assignment histories of all NPA-NXX combinations and their associated hundred-series banks. This investigation was limited to a subset of all such NPA-NXX combinations, the reasons for which will become clear as the process of building this database is reviewed below.

A Longitudinal Database of New NPA-NXX Combinations. The objective was to create a longitudinal database of all "new" NPA-NXX combinations and their "listed" residential assignment histories, as derived from periodic Donnelly DQI2 Files. To accomplish this, two sets of critical databases were required.

The first is an ongoing series of BELLCORE V&H Coordinate Files, covering an arbitrary baseline period forward to the present. The characteristics of the BELLCORE file precludes the reliable identification of "new" NPA-NXX combinations without direct comparison to their presence on the prior file. Although the BELLCORE tape provides "active dates" and "change codes" these have proved unreliable since the date field is re-used for recording the "effective date" of various changes. Consequently, the only reliable method of identification is the actual first appearance of an NPA-NXX code with an associated "Effective Date."

There are further complications introduced by the nearly 15 new NPA's established since the late 1980's. In both parts of split NPA's, the number of new NXX's created is greater than in non-split NPA's and consequently the dynamics of household assignments are potentially quite different. The decision was made to limit the investigation to NPA's which have not undergone a split either prior to, during, or at anytime after the baseline period. This limited the database and investigation to new NXX's in 94 out of the current total of 125 NPA's.

Although this restriction alleviated significant efforts involved in tracking NXX's through these NPA splits, it raises issues as to the generalizability of these results. NPA splits have primarily occurred in high density urban and/or high growth areas throughout the US (e.g., New York, Northern New Jersey, Los Angeles, San Francisco, Atlanta, Chicago, etc.) However, our experience has been that where NPA's are split, the rate of establishing new NXX's increases, resulting in a decrease in the density of listed residential assignments across the board. Thus, although the number of new residential exchanges increases, their incorporation into the list-assisted frame may actually span a longer time frame than in non-split NPA's. Moreover, the recency of many of these NPA-splits precludes the establishment of sufficient history to support any longitudinal analysis.

The baseline period was arbitrarily established as a two-year time-frame, from 1 January 1990 through 31 December 1991, which was subsequently split into four six-month time periods. The relevant BELLCORE files were sequentially screened and a total of 1574 "new" NXX's were identified with Effective Dates falling in the two-year period. Table 2 provides a frequency distribution of the new NPA-NXX combinations along with their distribution across the four six-month baseline groups.

TABLE 2. Frequency Distribution of NPA-NXX Baseline Groups

Effective Date	Newly Established NPA-NXX Combinations
1990-1 (1/1 to 6/30/90)	324
1990-2 (7/1 to 12/31/90)	477
1991-1 (1/1 to 6/30/91)	364
1991-2 (7/1 to 12/31/91)	409
Total	1574

Data was then extracted from five successive Donnelly DQI2 databases spaced approximately six months apart. The first DQI2 database was April 1992, followed by October 1992, and at subsequent six month intervals to April 1994. The number of residential listings were recorded by NPA-NXX combination and associated hundred series banks. This provides the ability to construct a number of important summary variables at both the hundred series and NPA-NXX levels at multiple points in time.

Summary variables were constructed for each of the four baseline groups at each of the five points in time. The time-series variables for each baseline group were then related to each other based on the approximate number of months between each DQI2 file and the center of the group's baseline period. This created a series of datapoints ranging from six to forty-eight months, each containing a varying number of baseline groups, from one to four.

Results of Analyses. Table 3 summarizes the cumulative proportion of NPA-NXX combinations with one or more (1+) directory listed telephone households by elapsed time between the effective date baseline periods and the date of the DQI2 database extract.

The results clearly indicate that incorporation of a "new" NPA-NXX combination into the list-assisted frame will span a significant time period. The incorporation process is primarily dependent upon the publication schedule for the telephone directory(s) associated with a particular NPA-NXX, and the subsequent time required for the database compilation process. The data leads to both some expected and unexpected results.

It is no surprise that the incorporation process is relatively slow for the first six to twelve months. Table 4 shows that, on average, about 24 listings per NXX have appeared on the listed household(LHH) database after eighteen months. As mentioned previously, there is an expected time-lag between actual assignment and appearance on the DQI2 file:

- Telephone directory publication frequencies are typically once per year. Assuming a constant publication schedule throughout the year, one would expect that the average time-lag between the assignment of a residential number in a new NPA-NXX combinations, and its appearance in the next published directory will be six months.

- Time between publication date and receipt by Donnelly averages one month, about 20% are not received for two months or more, and a few, up to a year.

- Processing and verification of each directory typically requires 45 to 60 days, but can be delayed much longer if there are associated geo-coding problems in the directory's coverage area.

On average, new assignments to the underlying LHH database will not be reflected for a minimum of nine to twelve months. This lag, especially early on, has the greatest impact on non-coverage, as both actual residential NPA-NXX's and more specifically, their associated hundred-series residential banks will remain unidentified and outside the frame.

Somewhat surprising however, are the small increases in the number of new NXX's entering the frame beyond an approximate 30 month mark. This is again evident in Table 4 where the proportions of NXX's with selected total numbers of listed households by time period are

TABLE 3. Percentage of NPA-NXX Combinations With 1+ DirectoryListed Households by Effective Date and Elapsed Time from Baseline Period

Baseline Period for NXX Groups	Elapsed Time (Months) From Baseline Period							
	6	12	18	24	30	36	42	48
1990-1 (n=324)	-	-	-	30.6	41.7	41.7	43.8	46.0
1990-2 (n=477)	-	-	14.9	39.4	41.9	47.2	50.1	-
1991-1 (n=364)	-	3.0	34.3	36.0	42.0	44.2	-	-
1991-2 (n=409)	2.0	17.8	23.2	42.1	48.9	-	-	-
TOTAL	2	10.9	23.3	37.5	43.7	44.7	47.6	46

TABLE 4. Cumulative Percent of NPA-NNX Combinations By Number of Directory Listed Telephone Households(LHH)

Number of Listed Households	Elapsed Time (Months) From Baseline Period							
	6	12	18	24	30	36	42	48
> 1	2	10.9	23.3	37.5	43.7	44	47.6	46
> 50	0.5	7.1	19.2	32.7	39.1	40.5	42.8	39.5
> 100	*	6.0	15.9	29.3	36.3	38.8	40.9	38.0
> 500	*	1.2	3.7	9.8	15.1	19.2	23.2	24.4

*<0.05%

shown. An obvious conclusion here is that complement of NXX's, determined by the various telephone companies to provide residential service, have been reliably identified within about two and a half years of their Effective Date on the BELLCORE File.

This is further strengthened by evidence that most NXX's established as residential continue to grow beyond the 30 month period in terms of listed household assignments (table 5). At that point, about one-third of identified residential NXX's have 500 or more listings, this increases to around 50% at the end of four years. In fact less than 5% of the increase in listed households after 30 months is attributable to NXX's subsequently identified as containing a residential listing.

A critical variable related to the timely incorporation of "new" hundred series residential banks into a list-assisted frame, is the cut-off criterion employed in classifying a bank as being residential and working. In Table 6, the mean number of residential hundred-series banks per NXX is shown by elapsed time and the cut-off criterion employed. After 12 months, a one listed household criterion has identified 2.2% of all banks as residential. This compares to just 1.8% for a two LHH cut-off and 1.5 for a three-listed household cut-off. In other words, 82% of one-LHH banks can be expected to be included using a two-LHH criterion, 68% using three, and dropping to 55% for a five-LHH criterion. As the time from Effective Date lengthens, the criterion gap and resulting non-coverage

TABLE 5. Cumulative Mean Business and Listed Residential Assignments per NPA-NNX Combination

	Elapsed Months							
	6	12	18	24	30	36	42	48
Residential Listings	2.3	23.8	63.9	144.6	197.4	239.9	279.8	288.4
Business Listings	0.3	1.5	3.9	9.6	14.9	21.0	27.4	33.3
TOTAL	2.6	25.3	67.8	154.2	212.3	260.8	307.2	321.7

differential narrows. At 30 months, a three LHH cut-off will miss only 11% and a five LHH criterion, just 15%.

The substantial early-on impact of the listed household cut-off criterion, and the associated increase in non-coverage is primarily a result of non-published residential listings. Assuming 60% of new listings in a given NXX are requested by telephone subscribers to be non-published, and a one listed household criterion, the expected number of new residential listings needed to identify any given bank as residential, is 1.7. With a two listed household criterion, this increases to 3.3, and to 5.0 with a three LHH criterion.

Implications for Further Research. The dynamics of the list-assisted frame are little understood, especially issues related to the migration of newly established NPA-NXX combinations from the "non-residential", truncated portion of the frame to the residential portion. A primary reason for this lack of research is the relative scarcity of the requisite data needed to establish a longitudinal perspective over extended periods of time.

These first steps need to be expanded to the full telephone frame by incorporating the more historically volatile NPA's, which have undergone "splits" during the past few years. Ideally, this extension needs to be both backward and forward in time, providing both increased generalizability and confirmation of the patterns identified above:

- Specifically, the ability to reliably classify NPA-NXX combinations as "non-residential" given historical, or retrospective evidence is compelling. This ability has design implications as well as the potential for improving operational efficiency in Mitofsky-Waksberg frames.

- Such a longitudinal capability should provide similar insights into the non-residential "hundred-series" banks of typical list-assisted RDD sampling frames, and possibly suggest stratification strategies to decrease non-coverage with minimal cost impacts.

- Incorporating variables related to geographic location, urbanicity, demographics, etc., may result in a reliable means to identify and predict where non-coverage predominates in the "non-residential" portion of the list-assisted frames.

The dynamics of list-assisted frames are a relatively new area of inquiry. Establishment of this initial longitudinal tool is but the first step towards the creation of a comprehensive database which will support research in this area in future years.

References.

Brick, J.M., Kulp, D.W., Starer, A, and Waksberg, J., "Bias in List-Assisted Telephone Samples", paper presented at Annual Conference of American Association of Public Opinion Research (AAPOR), 1994.

TABLE 6. Percentage of Hundred-Series Banks Classified Residential as a Function of Listed Household Cut-off Criterion

	Elapsed Time (Months)							
	6	12	18	24	30	36	42	48
1 LHH	0.2	2.2	4.7	8.6	10.6	11.6	13	13
2 LHH	0.1	1.8	4	7.8	9.7	10.9	12.1	12
3 LHH	0.1	1.5	3.7	7.4	9.4	10.6	11.8	11.8
4 LHH	0.1	1.4	3.4	7.1	9.2	10.4	11.6	11.6
5 LHH	0.1	1.2	3.2	6.8	9	10.2	11.4	11.4

Casady, R., Lepkowski, J., Tucker, C., "A Hierarchy of List-Assisted Stratified Telephone Design Options", paper presented at Annual Conference of American Association of Public Opinion Research (AAPOR), 1993.

Connor, J., and Herringa, S., "Evaluation of Two Cost Efficient RDD Designs", paper presented at Annual Conference of American Association of Public Opinion Research (AAPOR), 1993.

GENESYS Sampling Systems, "Methodology", internal GENESYS document, 1994.

Waksberg, J., "Sampling Methods for Random digit Dialing", Journal of the American Statistical Association, Vol. 73, No. 361, 1978.