ESTIMATING RESIDENCY RATES FOR UNDETERMINED TELEPHONE NUMBERS

J. Michael Brick, Jill Montaquila, Westat and Fritz Scheuren, Urban Institute
Jill Montaquila, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Response rate

1. Introduction

Response rates are important indicators of quality in surveys. In random digit dial (RDD) telephone surveys, a standard definition of response rate has not yet emerged despite many attempts (e.g., Frankel, 1983; Groves and Lyberg, 1988; and AAPOR, 1998). One problem is that the denominator of the response rate must be estimated and there are many ways to do this. The denominator should be the number of residences dialed, but this must be estimated because it is not possible to determine the residential status for all telephone numbers. For example, some telephone numbers ring when dialed (at least that is how it sounds to the person dialing the number) even though the telephone number is not assigned for use. We denote telephone numbers for which residential status is still not resolved at the end of the data collection period "undetermined" telephone numbers. The percentage of undetermined telephone numbers encountered in surveys has been increasing over the last few years as a result of changes in the telephony system. (Piekarski *et al.* 1999)

In the next section, we review some terminology and methods that have been used to estimate the percentage residential for undetermined telephone numbers. The third section presents a new method for estimating the percentage residential. The fourth section extends the method and takes advantage of more information to estimate the percentage. The fifth section applies the methods to two recent RDD surveys conducted by Westat. The final section summarizes the findings and makes recommendations for computing the residency rates needed for estimating response rates in RDD surveys.

2. Methods Currently Used

We assume that each telephone number in an RDD survey is dialed[1], and after these call attempts each number can be categorized as residential (RE), nonresidential (NR), or undetermined (UN). Residential numbers are all those where a person in the household answers irrespective of whether the household agrees to participate in the survey. The nonresidential category includes numbers that are not working and numbers for businesses. Undetermined numbers are those where the only results of call attempts are some combination of ring/no answers, busy signals, or answering machine outcomes.

The report by Frankel (1983) addresses the estimation of response rates. The Council of American Survey Research Organizations (CASRO) published this report and the response rates computed using this methodology are often called CASRO rates. The method distributes the undetermined units in proportion to the distribution of the units that are determined. The percentage of the undetermined numbers that are residential numbers, UNP_{RE}, is estimated as

$$UNP_{RE}(CASRO) = 100 \cdot \frac{RE}{RE + NR}$$

where RE is the number classified as residential and NR is the number classified as nonresidential.

The bounds on this percentage are sometimes used in estimating response rates. If all the undetermined numbers are considered residential, then the response rate is minimized and this method is called the conservative method. The other bound is attained when none of the undetermined numbers are estimated as residential and is called the liberal method

An alternative to the CASRO approach is called the business office method. In this method, a subsample of the undetermined numbers is selected and telephone business offices are contacted to determine whether the numbers are residential. The percentage of undetermined numbers estimated as residential from the business office method is

$$UNP_{RE}(bus.office) = \frac{n_{RESID}}{n_{RSLVD}}.$$

where

n_{RESID} = number of telephone numbers resolved as residential by business office

and

n_{RSLVD} = number of telephone numbers resolved by business office

Shapiro *et al.* (1995) describe an application of the business office method. The business office method appears to overestimate the percentage of the undetermined numbers that are residential. The results of the business office method are not portable from one survey to the next unless the same procedures are used in the studies (Keeter and Miller, 1998).

3. New Method for Estimating Residency Rate

In this section, we describe a new approach for estimating the residency rate for undetermined numbers that overcomes many of the shortcomings of the CASRO and business office approaches. We begin by

[1] Methods of classifying a number as nonresidential other than having an interviewer dial the number are included in this process but are not counted as a call attempt.

explicitly defining the residency rate for all telephone numbers and then discuss the new approach to estimating this rate.

For ease of description, we refer to the t-th call attempt as "trial t." The trials do not refer to fixed points in time, since one case might be receiving its 18th call attempt while another case is receiving its 2nd. Let r_t denote the number of cases (telephone numbers) resolved as residential at trial t, and let n denote the total number of telephone numbers. The residency rate at trial t, R_t, is estimated by

$$\hat{R}_t = \frac{\sum_{k=1}^{t} r_k}{n}, \qquad (1)$$

where k denotes a trial (i.e., call attempt number) at which there are non-censored cases (i.e., cases newly resolved as either residential or nonresidential).

In RDD surveys, the residential status of a large proportion of cases is usually resolved within the first few call attempts. Note that $\{\hat{R}_t\}_{t=1,2,3,\ldots}$ is a nondecreasing sequence that converges to the asymptote R_∞, the overall residency rate. If the residential status of all cases was resolved by some trial T, then $\hat{R}_T$ could be used as an estimate of the overall residency rate. However, in practice, it is neither feasible nor cost-effective to resolve the residential status of all cases. Even after a large number of calls, some cases will remain undetermined, with a status of "no answer" or "no answer, answering machine." Some of these cases are nonworking numbers (numbers that have not been assigned); others include telephone numbers connected to home computers, etc.

The estimated residency rate among cases with undetermined numbers is the difference between the estimated number of residential numbers (an unbiased estimate of R_∞ multiplied by the number of telephone numbers) and the resolved number of residential telephone numbers, divided by the number of undetermined numbers.

All the variables needed to apply this approach are known, except the estimate of R_∞. A scheme for estimating R_∞ is to consider cases with undetermined numbers at the end of data collection as right-censored data, with a varying number of call attempts. When cast in this light, techniques from survival analysis can be used. In particular, we first estimate the overall survival function from the data, where the survival function is the probability that a telephone number is not resolved as either residential or nonresidential by a specific trial. We then partition the survival function into a separate function that describes the probability of a number being classified as residential. This function, evaluated at an infinite number of call attempts, is an estimate of R_∞. These two steps are described below.

The Kaplan-Meier estimator (also known as product-limit estimator) is a nonparametric procedure to estimate the survival function, $S(t) = \Pr\{T \geq t\}$, where T is a nonnegative random variable that denotes the "lifetime" of the case. In our application the lifetime is the number of call attempts until the number is classified as residential or nonresidential. The Kaplan-Meier estimate (Lawless, 1982)[2] is

$$\hat{S}(t) = \prod_{i:t_i<t} \frac{n_i - d_i}{n_i} \qquad (2)$$

where

- i indexes the trial or call attempt at which there are non-censored "deaths". (In this context, "deaths" are cases resolved to be residential or non-residential.);
- n_i is the number of cases "at risk" just prior to trial t_i. (In this context, being "at risk" just prior to trial t_i corresponds to still being called at the t_i^{th} call attempt.); and
- d_i is the number of "deaths" or resolved cases at trial t_i.

The determination that a telephone number is residential and the determination that a telephone number is nonresidential may be thought of as the two "causes of death[3]." The survival function given in (2) does not estimate R_∞ because it is the survival function for the resolution of cases due to any reason. The survival functions for the two causes of death are estimated (Lawless, 1982) by

$$\hat{S}_{RES}(t) = \sum_{i:t_i \geq t} \frac{d_{RES,i}}{n_i} \hat{S}(t_i) \qquad (3)$$

and

$$\hat{S}_{NONRES}(t) = \sum_{i:t_i \geq t} \frac{d_{NONRES,i}}{n_i} \hat{S}(t_i) \qquad (4)$$

where $d_{RES,i}$ is the number of cases determined to be residential at trial t_i; $d_{NONRES,i}$ is the number of cases determined to be nonresidential at trial t_i, and the summations are defined only at those trials t_i where $n_i > 0$.

---

[2] $\hat{S}(t)$ is defined only for those t for which $n_i > 0$. Also, note that the terms in the product correspond only to those trials at which there are non-censored observations (i.e., trials for which all observations are censored are not included in the product).

[3] In personal correspondence, Jerry Lawless noted that the problem is more accurately a mixture model than a competing risks model. However, he pointed out the approach described was appropriate because Larson and Dinse (1985) showed that the two models give the same estimates in this case.

The overall residency rate is then estimated as

$$\hat{R}_{\infty} = \frac{\hat{S}_{RES}(0)}{\hat{S}_{RES}(0) + \hat{S}_{NONRES}(0)}. \quad (5)$$

With this estimate of R_{∞}, we can apply the approach described above to estimate the residency rate for cases with undetermined telephone numbers. The estimate is

$$\hat{R}_{UN} = \frac{\left(\hat{R}_{\infty} \cdot n_{TOT} - n_{RES}\right)}{n_{UN}}, \quad (6)$$

where n_{TOT} is the number of total number of cases, n_{RES} is the number resolved as residential, and n_{UN} is the number undetermined.

The survival method, like the CASRO and other approaches, relies on the observed data to predict the residency rate for the undetermined numbers. In the special case in which all undetermined numbers are censored at the last trial at which there are non-censored observations and none are censored sooner, the survival method is equivalent to the CASRO method. Thus, to obtain any benefit from the survival approach in this situation, a sample of the undetermined numbers should be dialed additional times to estimate the residency rate. The survival method typically estimates a lower residency rate than the CASRO method because the percentage of telephone numbers that are resolved as residential tends to decrease with the number of call attempts.

4. Conditioning Using Auxiliary Data

The survival method procedures described above use only the number of call attempts to estimate the residency rates for undetermined numbers. In many surveys, auxiliary data associated with residential status are also available and could be used in the estimation. In this section, we extend the procedure to take advantage of these data.

For standard RDD surveys, two auxiliary items are considered although other variables might also be used. One of these items is whether the telephone is listed or not. A second auxiliary item available in most Westat surveys for the "no answer, answering machine" cases is the interviewer's coding of the type of answering machine for each answering machine call result. Each answering machine call result was coded by the interviewer as either likely to be residential, likely to be nonresidential, or undeterminable. For the purpose of this analysis, we derived a variable that summarizes the call attempt-level codes. For each telephone with at least one answering machine outcome a variable was created with the values: "residential" if at least as many call attempts were coded "likely to be residential" as either of the other two codings; "nonresidential" if more call attempts were coded "likely to be nonresidential" than either of the other two codings; and "unclassified" otherwise.

With items that are likely to be highly associated with residential status such as these two items, we expect to improve the estimate of R_{∞} by using them in addition to the number of call attempts in the analysis. To do so, we fit separate survival curves and derive separate residency rate estimates for each of the groups defined by combinations of these auxiliary variables.

The estimate of the percentage undetermined numbers that are residential is subject to sampling error. The sampling error may be larger than expected because the survival function at trial t is computed from the cases that do not have a resolved residential status before the t-th call attempt and the number of cases is smaller once censoring begins. In addition, the estimator of the residency rate for undetermined numbers is a complex function of the survival function. To properly compute the sampling error of this statistic a jackknife replication method and the WesVar software (Westat, 1998) was used. This approach takes into account both the sample design of the survey and the complexity of the estimator.

5. Application

The survival method procedures described in Sections 3 and 4 were applied to data from two large scale RDD surveys conducted by Westat in 1999. One of the surveys is the 1999 National Household Education Survey (NHES:1999) and the other is Cycle 2 of the National Survey of America's Families (NSAF:1999). A brief description of the two surveys follows and then the survival method is applied.

Conducted by Westat between January and April 1999, the NHES:1999 was one cycle of a periodic household survey sponsored by the National Center for Education Statistics (NCES) that addresses a variety of education issues. The NHES:1999 was a list-assisted RDD survey that covered the 50 states and the District of Columbia. A total of 167,347 telephone numbers were sampled. Telephone numbers in high-minority exchanges (those in which at least 20% of persons are black or at least 20% of persons are Hispanic) were sampled at twice the rate of telephone numbers in low-minority exchanges. Additionally, during the field period, some cases with no answer after eight call attempts and without a mailable address (i.e., either no mailing address could be obtained for the telephone number or mailings to the address were returned by the postmaster) were subsampled. Only half of such cases were refielded for additional call attempts. The estimates given here have been weighted to account for the oversampling of telephone numbers in high-minority exchanges and to account for the subsampling of nonmailable "no answer" cases. Additional details on the NHES:1999 are in Nolin *et al.* (2000).

The other survey is the NSAF:1999, a project of the Urban Institute in partnership with Child Trends. The survey is part of an effort to assess the effects of the devolution of social programs to the states. The

first survey was conducted in 1997 and data for the second survey, discussed here, were collected from February 15 until October 3 of 1999. The importance of states in the assessment resulted in a design with large sample sizes for 13 specific states and a large national sample. The sample design for the NSAF:1999 was further complicated because some of the telephone numbers sampled in the 1997 survey were retained for the Cycle 2 survey at different rates depending on the result of the 1997 survey. A new sample of telephone numbers was also selected. Weights were applied to adjust for the differential sampling rates. In total, 380,037 telephone numbers were sampled for the NSAF:1999. Additional details on the NSAF are in Brick *et al.* (1999).

The general approach described in Section 3 was applied to the NHES:1999 data. Using equation (6), the residency rate is estimated to be 35.3 percent (with a standard error of 3.1%) for the "no answer" and "no answer, answering machine" cases. The overall residency rate estimate is 46.5 percent. The 35.3 percent estimate for the undetermined numbers is lower than the rate computed using either the CASRO (47.5%) or the business office (40.5%) methods.

The conditional approach described in Section 4 was next applied to the NHES:1999 data. In this application, the listed status of the telephone number and the interviewers' primary coding of answering machine call results were used in combination to create eight different categories of telephone numbers. Separate residency rate estimates were obtained for each of the eight categories.

The residency rate asymptote, R_∞, was estimated for telephone numbers in the NHES:1999 sample for each of the eight categories. These estimates are given in Table 1. Based on the estimated residency rates and the distributions of undetermined numbers in Table 1, the overall estimated residency rates for "no answer" cases and for "no answer, answering machine cases" are 25.1 percent and 20.6 percent, respectively. The overall estimated residency rate for undetermined numbers (i.e., "no answer" and "no answer, answering machine" cases combined) is 24.2 percent, lower than the 35.3 percent for the unconditional survival method. The overall residency rate using the conditional survival function method is 45.7 percent.

The unconditional survival function approach was also applied to the NSAF:1999 data and the residency rate for the undetermined numbers is estimated to be 5.8 percent (with a standard error of 1.2%). The rate for the undetermined numbers using the CASRO method is 47.0 percent and using the business office method is 40.5 percent. The overall residency rate estimate using the unconditional survival function estimate is 43.6 percent, while for the CASRO method it is 47.0 percent and for the business office method it is 46.5 percent. The differences between the estimates in the NHES:1999 and the NSAF:1999 are discussed in the next section.

Because of the sample design of the NSAF:1999, variables other than listed status and primary answering machine outcome were considered as auxiliary variables. One auxiliary variable examined was the site or location, defined to be the states[4] with large samples and the balance of the U.S. Another auxiliary variable was the sampling stratum that was based on the outcome of the number from the 1997 sampling.

The results of doing the survival function approach conditioning only on the site are shown in Table 2. The residency rates for undetermined telephone numbers vary substantially by site, ranging from 3 percent to 14 percent, which is consistent with an observation of Piekarski *et al.* (1999). The estimated standard errors for some states are large indicating some instability in these estimates. The overall national residency rate for the undetermined telephone numbers is 7.2 percent when the survival function is estimated by site. The table also shows the estimated overall residency rate by site, once again demonstrating the differences by geography in the telephone system.

When the conditional survival function method was applied to the NSAF:1999 using the sampling strata as an auxiliary variable in addition to listed status and primary answering machine outcome, the estimates became very unstable. Deleting one or two observations resulted in dramatic changes in the estimates of the residency rates. This finding indicates that the survival function method is sensitive and may perform poorly with small samples.

6. Discussion

The estimation of the percentage of the undetermined telephone numbers that are residential in RDD studies is an important issue because it is a component in the denominator of the response rate. The survival function method provides a new approach to this problem by using more information in the estimation. Applying the survival function method to two RDD surveys conducted in 1999, we found that the residency rate for the undetermined numbers is lower than estimated using either the CASRO or business office methods. For some surveys, the difference may make a substantial difference in the estimated response rate. For the NHES:1999, we estimated the difference would be about 3 percent and in the NSAF:1999 the difference is nearly 5 percent.

The survival function methods do have some limitations and technical difficulties. One assumption of the survival method that is not fully satisfied is the requirement that after an infinite number of calls all of the telephone numbers will be classified as residential or nonresidential.

[4] There were 13 states with large samples and in Wisconsin a large sample was taken within Milwaukee County. Counting the balance of the U.S., the number of sites is 15.

Table 1. Estimated residency rates and distribution of undetermined numbers in the NHES:1999 sample by listed status and primary coding of answering machine call results

Listed status	Primary coding of answering machine call results	Estimated residency rate asymptote (R_∞), percent	Estimated residency rate for undetermined numbers $100 \cdot \hat{R}_{UN}^{(g)}$	Standard error of estimated residency rate of undetermined numbers	Distribution of undetermined numbers $n_{UN}^{(g)}$ [1]	
					"No answer" cases[2]	"No answer, answering machine" cases
Listed	None	75.4	32.5%	25.1%	1,689	
Listed	Residential	90.4	21.1	4.0	2	1,633
Listed	Nonresidential	38.1	2.2	5.3		89
Listed	Unclassified	89.3	62.1	23.0		108
Unlisted	None	30.0	24.2	2.1	13,864	
Unlisted	Residential	86.4	22.8	7.3	6	1,540
Unlisted	Nonresidential	7.7	0.8	0.6		339
Unlisted	Unclassified	61.4	19.9	8.3		337
Overall		45.7	24.2	2.7[3]	15,561	4,046

[1] Counts given here are weighted to reflect the differential sampling of telephone numbers by minority stratum and the subsampling of nonmailable "no answer" cases for follow-up.

[2] The weighted total of eight cases that finalized as "no answer" but had an answering machine call result in their call histories are cases that were refielded and the answering machine counter used to finalize a case as "No answer, answering machine" was reset.

[3] The standard error of the overall residency rate estimate for undetermined numbers was computed assuming the eight strata (listed status by answering machine classification) were independent.

Source: U.S. Department of Education, National Center for Education Statistics, National Household Education Survey, 1999.

Table 2. Estimated residency rate for undetermined telephone numbers and overall in the NSAF:1999 by site

Site	Residency rate for undetermined numbers		Overall residency rate	
	Estimate	Standard error	Estimate	Standard error
Total	7.2%	0.9%	43.6%	0.2%
Alabama	14.2	3.2	49.5	0.5
California	4.0	1.0	42.9	0.4
Colorado	2.8	1.1	41.8	0.3
Florida	3.0	0.6	43.2	0.3
Massachusetts	4.6	1.3	46.9	0.3
Michigan	9.9	3.0	42.9	0.4
Minnesota	2.8	2.0	42.8	0.3
Mississippi	11.9	3.5	49.6	0.7
New Jersey	4.5	1.1	43.6	0.3
New York	7.6	2.4	45.7	0.4
Texas	6.5	2.5	40.9	0.4
Washington	8.3	2.6	42.7	0.4
Milwaukee County	3.8	8.3	39.6	0.7
Balance of Wisconsin	8.9	3.8	45.7	0.4
Balance of U.S.	9.1	1.9	43.8	0.3

Source: Urban Institute, National Survey of America's Families, 1999.

Another issue that could distort the estimates of the residency rate using the survival method is related to the nature of the call attempts. Suppose a telephone number was attempted only one day, but up to 40 calls were made during a three-hour period in that day to reach the household. In this situation, the number of call attempts may not be a very good measure of the exposure to "risk" or classification. The choice of a good measure of exposure is a common problem in life-testing.

A third issue in the estimation of residency rates using the survival method is whether to use weighting adjustments to account for censoring. In the approach we used in this paper, no weighting adjustments were

made to account for censoring; censored cases retained their weights and were used in the estimation of the residency rates. Although there are other alternatives, we recommend this approach.

An advantage of the survival method is that it can be applied to many studies, using data collected from the particular study. However, the specific estimates are not very portable. The results from one study cannot be directly extrapolated to other studies because the distribution of call attempts (across days of the week, times of the day, and over calendar time) may affect the estimates. Instead, all the data needed for the analysis and estimation should be captured in the specific study.

A related topic is how the number of telephone call attempts affects the estimation of the residency rate for the undetermined numbers. Having more than a few call attempts at each number and distributing those attempts over a wide range of times is essential for estimating R_∞ accurately and for making sure that the number of call attempts is a reasonable measure of exposure. In addition, we have done some investigations to show that it is important to have a sufficiently large sample of telephone numbers that are dialed more times than the standard censoring point in order to estimate the residency rate well using the survival function method.

If the number of telephone numbers that are attempted many times is small, then the estimate of the percentage of undetermined numbers that are residential may be unstable and sensitive to the outcomes from just a few telephone numbers. If the sample of numbers dialed many times is small, then one possibility is to form fewer subgroups and ignore some of the potential auxiliary variables, as was done in the NSAF:1999. Another possibility is to use a parametric survival function instead of the Kaplan-Meier method.

The survival method provides a formal and statistically defensible method of using the data collected for the undetermined telephone numbers in estimating the residency rate for these numbers in RDD surveys. Existing approaches, such as the CASRO approach, ignore this information and produce less reliable estimates. In our examination, we found that the listed status of the telephone number, whether or not an answering machine was ever detected, and the interviewer's classification of the answering machine messages were important conditioning variables. Further investigations, or other RDD studies, might find other important conditioning variables.

The results of the survival method also provide an opportunity to further reduce the problem of estimating the residential status for undetermined numbers. By examining the residency rates for the never answered and answering machine cases in the format of Table 1, it is possible to identify specific subgroups that contribute heavily to the estimated residency rate. Additional call attempts can then be targeted to those subgroups.

As a final note, we urge that other indicators such as the number and distribution of call attempts for the undetermined numbers be included in methodological reports if the survival function method is adopted in practice. This information can be evaluated by data users if they are concerned about the validity of the response rate calculations.

7. References

The American Association for Public Opinion Research. 1998. *Standard definitions: Final disposition of case codes and outcome rates for RDD telephone surveys and in-person household surveys*, AAPOR, Ann Arbor.

Brick, P. D., Kenney, G., McCullough-Harlin, R., Rajan, S., Scheuren, F., Wang, K., Brick, J. M., and Cunningham, P. 1999. *1997 NSAF survey methods and data reliability, Report No. 1*, Urban Institute, Washington, D.C.

Frankel, L. 1983. The report of the CASRO task force on response rates, in *Improving Data Quality in a Sample Survey*, edited by F. Wiseman, Cambridge, MA: Marketing Science Institute.

Groves, R. M., and Lyberg, L. 1988. An overview of nonresponse issues in telephone surveys, in *Telephone Survey Methodology*, edited by Groves *et al*, John Wiley & Sons, New York.

Keeter, S., and Miller, C. 1998. Consequences of reducing telephone survey nonresponse bias – or what can you do in eight weeks that you can't do in five days. Paper given at the AAPOR meetings in St. Louis, MO.

Larson, M. G., and Dinse, G. E. 1985. A mixture model for the regression analysis of competing risks data. *Applied Statistics*, 34:201-211.

Lawless, J. F. 1982. *Statistical models and methods for lifetime data*, John Wiley & Sons, New York.

Nolin, M. J., Montaquila, J., Nicchitta, P., Kim, K., Kleiner, B., Lennon, J., Chapman, C., Creighton, S., and Bielick, S. 2000. *NHES:1999 Methodology Report.* U.S. Department of Education, Office of Educational Research and Improvement.

Piekarksi, L., Kaplan, G., Prestegaard, J. 1999. Telephony and telephone sampling: The dynamics of change. Paper given at the AAPOR meetings in St. Petersburg, FL.

Shapiro, G., Battaglia, M., Camburn, D., Massey, J., and Tompkins, L. 1995. Calling local telephone company business offices to determine the residential status of a wide class of unresovled telephone numbers in a random-digit-dialing sample. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 975-980.

Westat 1998. *WesVar Complex Samples 3.0 User's Guide.* SPSS, Chicago, IL.