ACHIEVING AN OPTIMUM NUMBER OF CALLBACK ATTEMPTS: COST-SAVINGS VS. NON-RESPONSE ERROR DUE TO NON-CONTACTS IN RDD SURVEYS

Brian E. Harpuder and Jeffery A. Stec, The Ohio State University
Brian E. Harpuder, Center for Survey Research, The Ohio State University, 3045 Derby Hall, 154 N. Oval Mall, Columbus, OH 43210-1330

Key Words: Unit non-response, RDD Survey, Buckeye State Poll

I. Introduction

There is no doubt that conducting surveys costs money. Money often is the one parameter that drives, or perhaps better stated, limits survey research because it sets constraints on effort and breadth. Because money is a limiting factor, we must often cut corners or be willing to tolerate a level of inaccuracy because the benefit of extra precision is outweighed by the dollar costs of having to obtain that precision. One of the most costly expenses in survey research is making numerous callback attempts to sampled respondents who we are not able to reach on the first call attempt. Many of these call attempts will be futile and not result in a completed interview. However, we also know that making callbacks is one of the best ways to increase response rates and hopefully reduce the amount of unit non-response bias that may exist in our data. In conducting survey research, we have these two competing goals. The first goal is to minimize survey costs and the second goal is to minimize the potential non-sampling errors in our sample estimates of population statistics.

Taking survey costs into account may suggest a more appropriate mode of conducting surveys or may allow resources to be better directed so as to minimize, as much as possible, *all* of the sources of error or at least the sources that produce the most. The Total Survey Error (TSE) approach, which involves thinking about sampling errors and different non-sampling errors simultaneously, is well suited to a cost-benefit analysis because one can think about where extra money spent will have the greatest effect in limiting biases. TSE is the driving force behind this paper's findings.

In this paper, we are concerned with unit non-response that results from non-contacts. Simply put, "it is the failure to obtain measurements on sampled units (Groves and Lyberg, 1998)." However, one cannot reiterate enough that the existence of non-respondents to a survey is not synonymous with unit non-response error. One *may* have unit non-response bias only to the extent that those interviewed are systematically different from those who are sampled but are not interviewed.

Non-response poses several major problems to survey research.

> First, to the extent that the non-respondents are different from the respondents on the survey measures, statistics based on respondent data alone will be biased estimates of the full telephone population parameters of interest. Second, non-response reduces the size of the sample, which forms the basis of the estimates. Third, survey costs are increased by efforts to reduce non-response (Groves and Lyberg, 1998).

It is these different effects of unit non-response error that survey researchers must always keep in mind when considering how they will implement their particular survey.

If one of the best ways of reducing the potential for unit non-response error in our data is by making numerous callback attempts, the question is then how many callbacks should one make. Even with numerous callbacks, there will be those who refuse to participate; however, what about those who are just more difficult to reach because of the hours they work or where they live. Previous research has shown there are differences in respondents' answers depending on the number of call attempts that were necessary to reach the respondent, though some of this research (Merkle et. al. 1993) suggests that these differences may be small. Unfortunately, we have not really had an understanding or a way of developing an understanding of how different numbers of callback attempts affect survey data. Moreover, we have also not had a way to connect the benefits of reducing the potential for unit non-response error to the costs that are associated with doing that.

We propose a method that leads us in the direction of being able to quantify the error due to unit non-response; and, moreover, draws a direct connection between unit non-response and costs. We will show where, in our experience, the costs of making additional calls are not outweighed by the benefits of reducing the potential of unit non-response error. We recognize that the optimal cut-off point may not be the best one for every survey research organization. However, we feel that a replication of our methodology that takes into account the experiences and costs of other organizations will allow those organizations to make a determination based on statistical calculations and not *ad hoc* decisions as to when to terminate additional callback attempts. Previous research has shown the need for cut-off rules for dialing in surveys that use Mitofsky-Waksberg sampling methods, however, this

same research has not been available for more basic RDD-based surveys (Alexander 1998).

Our methodology is useful to the extent that a researcher has experience in working with the target population. This methodology may not be appropriate for an RDD survey of a very small geographic area or one that seeks a rare population, but it is appropriate for the broad population surveys that most organizations conduct.

We should also note that our research does not take into account unit non-response error due to refusals. Research has shown that there are real differences (Kojetin 1993; Stec, Lavrakas, & Stasny, 1999, Groves and Couper 1998) between refusal and hard-to-reach respondents.

II. The Costs of Survey Research

As Groves (1988) emphasizes so well in his book, *Survey Errors and Survey Costs*, it is important to think about survey research from a cost standpoint because of its importance as a limiting factor. It is also important that when one thinks about the costs of survey research, it is from the TSE perspective since spending more money to fix one problem will mean that there is less money to help resolve another problem. Groves states that "From one perspective survey costs and errors are reflections of each other; increasing one reduces the other" (Groves, 1988). It is important for survey researchers to accept that reductions in bias can easily be outweighed by the costs of the reduction. Sebold (1998) found that it is very questionable if the reduction in bias associated with lengthening the field period of survey was justified by the costs. As researchers, we should not consume ourselves with concerns about the amount of potential bias in our surveys, but we should understand how it manifests itself and what ability we have to reduce it within a reasonable cost framework.

The first issue to address is what are the costs of doing survey research? Second, which of those costs are fixed no matter what survey is conducted? Finally, which costs are variable depending on the conditions imposed on the study? One would think that the fixed costs could be ignored for our research since there is not much that one can do about them, however, these fixed costs are important because they take away from the total budget that can be used to complete the survey. Implicit in this strategy of determining callbacks is that there is a cost that an individual, corporation, organization, etc. is willing to pay for their research to be conducted. This cost includes all charges to be made by the survey organization. Most likely, there is a range of costs that would be tolerable and a level of tolerable bias. A fundamental goal of our methodological research is to be able to associate levels of unit non-response bias with given costs.

Which costs are fixed? Since most survey organizations now use some form of a CATI system, the costs of programming that system are largely fixed. The costs of developing the survey such as questionnaire development are also fixed. The cost of purchasing the sample from a company that sells samples is also fixed for RDD surveys. There are additional fixed costs such as the day-to-day management of the survey as well as the costs of producing all of the data and materials at the end of the survey period. Overhead is not a fixed cost because overhead really works as a fixed percentage of total expenses so that it will be a larger burden the larger the amounts of administrative work that are needed in order to administer the survey. For our research, we roughly estimated fixed costs at $9,000 per survey due to the many different things described above.

Variable costs for a survey include the amount of person-hours needed to complete the survey in terms of administrative management, supervision, and, of course, interviewing time. Simply put, the more interviews, the higher the costs. The size of the target population and the level of tolerable sampling error most often determine the number of interviews needed.

Of course, the largest variable costs are telephone time and interviewers' and supervisors' time. The number of interviews required will have a major effect on these costs as will, of course, the number of callbacks allowed. What inflate these costs are calls that result in no information being gained, such as a good time to reach the respondent. In many cases, we will not know if the number that we are dialing is even a working number and in those situations we are wasting time with numbers which should never have been called. Doing this a few times is prudent, but if one were to allow up to fifteen or twenty callback attempts the costs would increase rather quickly. It is for this reason that we believe our findings are so important. By being able to quantify the dollar costs of additional callbacks with respect to the benefits, we hope to convince researchers that at some point other survey errors could be more effectively reduced with the additional expenditure that is being wasted on high numbers of callbacks. We have estimated that these variable costs can range from a little over $9,000 to more than $11,000 for the example of 500 interviews in the state of Ohio that we are using. The variation in these figures can vary dramatically across surveys and certainly across survey research firms. Each survey organization will need to compute their own costs for the purposes of using our methodology.

III. Methodology

In order to develop our model of when callback attempts should be terminated in a RDD study, we are using data that was gathered by the Ohio State

University's Center for Survey Research (OSU-CSR). The OSU-CSR conducts a poll every month called the Buckeye State Poll (BSP). The OSU-CSR interviews approximately 500 respondents within a given month on the state of the economy and the state of each household's own personal credit card finances. The sample design, target population, and construction of the sampling pool are consistent across all months. The target population is English-speaking, Ohio residents over the age of 18.

Since April 1998, we have had our CATI system automatically track the number of calls that it takes to achieve a completion. We also calculate how many calls it takes to make contact with an individual. As a result, we are using data gathered by the OSU-CSR for the BSP between April and December of 1998. A total of 4,974 interviews were completed during the time period.

In order to complete the minimum of 500 interviews on a monthly basis, we will generally release on the order of 2,500 phone numbers. This is necessary in order to filter out non-working numbers, businesses, privacy manager protected numbers, and refusals. If we were to restrict the number of call attempts we make, an increase in the amount of phone numbers released would be necessary and would thereby reduce response rates. It is critical to remember, however, that response rates are *irrelevant* if there is minimal non-response error. At the OSU-CSR we are concerned with non-response error *not* non-response rates.

In our research, we focus on non-response error that is a result of non-contacts. In order to do this we are separating out individuals based on the call attempt number at which they were contacted. The logic is simple. Had we stopped calling at, for instance, five callbacks, then we would know that all of those individuals reached at calls above five would have been non-contacts. Hence, we can tell whom the people are that we would have missed. We are using past evidence of what non-response error would have been from non-contacts to predict what effect we expect non-response due to non-contacts to be in the future. We believe that this assumption is supported as long as there was no significant change in the sampling methodology used and as long as the data are left unweighted.

One of the problems of previous research is that it has focused on bias in individual variables. While it is important to understand how certain variables are affected by bias, when one makes a cost-benefit calculation, one is generally not interested in just a single variable. We are interested in limiting the effect of non-response error due to non-contacts across a series of questions. For our study we have looked at how non-response affects the variables that are of central interest to us in the analyses that we do on a monthly basis. These variables are questions that pertain to the state of the economy, credit card usage, as well as some central demographic variables.

Our assessment of bias was done by incrementally looking at the bias with each successive number of allowed call attempts. We compared completions on one call attempt to all those completed after two or more attempts across a number of survey variables. We then compared completions made on the first or second call to all those completed after more calls, and so on and so forth. We computed z-scores for each of these variables by subtracting the mean (or proportion) for a variable for all callback attempts from the mean (or proportion) for a variable for the restricted number of call attempts. We then deflated each of these differences by the appropriate standard deviation of the sample estimate and summed the squares of these z-scores ($\sum(z-score)^2$).

The z-score here is a measurement of bias, standardized by the sample estimates standard deviation. For the demographic variables where we had Census projection data, it is a measurement of deviation from the true population mean. For the other variables it is a measurement of deviation from the overall mean of all data collected. The other variables are simply those that we frequently utilize in our analyses or are those that are the subject of our analyses.[1] For continuous data, means were utilized, whereas categorical variables are computations based on proportions. Implicit in our research is the assumption that the overall sample mean (proportion) approximates the true population value. If one does not accept this assumption then one can still see if the standardized bias for particular variables changes with more callbacks. This would be important to look at because if the bias does not change much, then why continue making callbacks?

It is important to use a z-score because it is a measure that is standardized and allows one to compare across different variables and add together z-scores from different variables. For example, the z-score for household income would be the mean of household income from a given number of callbacks subtracting the overall mean for all call attempts. Dividing that difference by the standard error of the mean estimate obtained for the mean at the given number of callbacks gives the z-score for household income[2]. This z-score can then be squared which allows us to add these values across variables. This statistic is useful for looking at

[1] Readers interested in the particulars of the variables used here should contact the lead author.

[2] For this study, we are using population values for income because they are readily available. We would suggest that if population values are available for the variables one is interested in than those values should be used for comparison rather than the overall values from the survey.

the relative magnitudes of bias, but, unfortunately, does not allow for any further statistical testing. The sum of these squared z-scores would be distributed chi-squared with n degrees of freedom (n being the number of variables considered) if each set of variables were pair-wise independent. However, variables such as income and education are not independent of each other.

We strongly believe the costs that are incorporated into our analyses are appropriate and are proportionately reflective of the costs that many other survey centers would incur. We computed costs by determining 1) the total number of calls that would need to be made for a given sample size using historical data about how many calls of each disposition we obtain after each callback attempt and 2) the length of time for a call of the given disposition. By multiplying these factors together with the costs of each call, we were able to compute total costs for each number of callbacks allowed.

After having computed the sum of the squared z-scores, we plot this magnitude against the number of callback attempts and costs in order to see the difference that can arise in costs and bias. While no formal statistical test is used, the visual display is striking and suggests the importance of specifying the number of callbacks that are allowed an incorporating this information into the budgeting process of a study.

IV. Findings

Most completions occur after a relatively small number of call attempts. The mean number of call attempts necessary to achieve a completion is 4.8 calls. However, this mean number of calls is pulled higher by several outliers where we, on a one-time basis, allowed for a very large number of callback attempts and, in some cases, completed interviews after more than 60 call attempts. Are those vast majority of cases where completions result after just a few tries the same as the others which take far longer to complete?

The results of our research suggest that high numbers of callbacks are not justified under any cost situation given the type of survey that we are conducting because the reduction in bias does not compensate the survey researcher for the additional cost of more callback attempts. The results of our research show how dramatic an effect the number of callbacks can have on the total costs of a survey even considering the fixed costs that are inherent in any survey regardless of callback attempts. These costs are incurred in large part, from the compensation of interviewers, supervisors, and phone costs.

Figure 1 shows how bias and costs inversely move together. Recall that the sum of the squared z-scores can be used as a means to gauge the magnitude of the bias involved in the data collection across numerous variables. We are most often interested in numerous variables simultaneously and this measurement allows us to gauge the bias in all of these simultaneously. This methodology would lend itself to a more statistically rigorous test if it were not for the violation of the independence of variables assumption. Nonetheless, it remains useful for gauging the magnitude of bias.[3]

As one can see from Figure 1, it appears that the appropriate number of callbacks that one should make is between four and six. This is determined by simply seeing that survey costs go monotonically higher while bias remains stabile and is minimized after four call attempts. Until we can compute more precise confidence intervals, prudence would suggest that a callback restriction of six to seven is appropriate even though bias appears to reach a global minimum after 4 calls. We would suggest to fellow researchers with regard to bias that it is better to be on the conservative side and to make more phone calls rather than less if there is any doubt.

After six or seven calls there does not seem to be sufficient justification in terms of reduction of bias to increase the costs of the survey. It should be noted, however, that the benefit gained in terms of reduction of non-contact non-response bias may be offset by increases in bias from other sources and should, as always, be considered from the TSE framework.

If we look at some of the specific variables where we have Census data and the biases associated with them, we find that our sample statistics, unweighted, do not match Census population values. Table 1 below shows the comparisons of age obtained from the BSP to the population values obtained from 1997 Census projections. Specifically, the values shown are a measure of how much the sample mean for that category deviates from its population mean normalized by the standard error of the callback mean. In other words, the values in each cell of the table are z-scores, so the statistical significance of the difference between the callback mean and the overall mean can be gauged by referring to a z-score table. Both age and race (not shown) proportions from the sample do not reflect the proportions in the population. It suggests that the sample is not representative of the population with respect to age or race even after allowing for twenty callbacks. Post-survey weighting procedures can correct for non-response bias present in sample estimates of the age and race variables.

If we look at a substantive survey variable at different callback attempts, we see that there is little variation after five or six calls (Table 2). Since we are using the overall sample proportion for these variables

[3] We should note that we could bootstrap in order to determine what the distribution of these summed squared z-scores is. Once that was done we could then use statistical procedures to test formally for differences. The next version of this paper will do just that.

as a proxy for the unknown population proportion, we know that the difference between the callback proportion and overall sample proportion will go to zero as the number of callbacks increases. What we do find here, though, is that additional callbacks past five or six attempts are not going to yield statistical significant differences between the callback proportion and the overall proportion. It appears that after these five or six calls there is nothing more that can be gained from continued callbacks.

V. Conclusions

Our research does not help us to understand how to correct for non-response once it occurs but it does suggest a methodology to help one limit the extent to which there is potential for non-response bias. At the OSU-CSR, this research will be used as part of an evaluation of how many callback attempts we should make. A replication of our methodology and the use of the resultant findings in the overall TSE framework will help survey researchers to minimize costs while understanding how unit non-response that is a result of non-contacts affects the overall data quality of the research.

We find that six or seven calls would be the most appropriate cutoff point. We are erring on the conservative side since our research seems to suggest that as few as four callbacks may be appropriate. It is important to note that these findings will vary from organization to organization and even project to project. While there may be variations, the methodology we applied should be universally applicable.

It is our hope that other researchers will replicate our methodology and determine for their own survey's purposes, the optimum number of callbacks that should be attempted. There seems to be little justification for attempting high numbers of callbacks when time and money could be saved with fewer callbacks.

VI. References

Alexander, Charles H. 1998. "Cutoff Rules for Secondary Calling in a Random Digit Dialing Survey." In Mick P. Couper et al., eds. Computed Assisted Survey Information Collection. New York: John Wiley and Sons, Inc.

Groves, Robert M. 1988. *Survey Errors and Survey Costs*. New York: John Wiley and Sons.

Groves, Robert M. and Lars E. Lyberg 1998. "An Overview of Nonresponse Issues in Telephone Surveys." In Mick P. Couper et al., eds. Computed Assisted Survey Information Collection. New York: John Wiley and Sons, Inc.

Kojetin, Brian A. 1993. "Characteristics of Non-Respondents to the Current Population Survey (CPS) and Consumer Expenditure Interview Survey (CEIS)", Proceedings of the Section on Survey Research Methods, 1993 Annual Meeting of the American Statistical Association, San Francisco, California

Merkle, Daniel M., Bauman, S.L. and Lavrakas, P.J. 1993. "The Impact of Callbacks on Survey Estimates in an Annual RDD Survey." In Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Sebold, Janice. 1998. "Survey Period Length, Unanswered Numbers, and Nonresponse in Telephone Surveys. In Mick P. Couper et al., eds. Computed Assisted Survey Information Collection. New York: John Wiley and Sons, Inc.

Stec, Jeffery A., Paul J. Lavrakas, and Elizabeth A. Stasny. 1999. "Investigating Unit Non-Response in RDD Surveys". Paper presented at the May 1999 Conference of the American Association for Public Opinion Research, St. Pete's Beach, Florida.

Figure 1: Costs vs. Bias for Given Callbacks: All Data

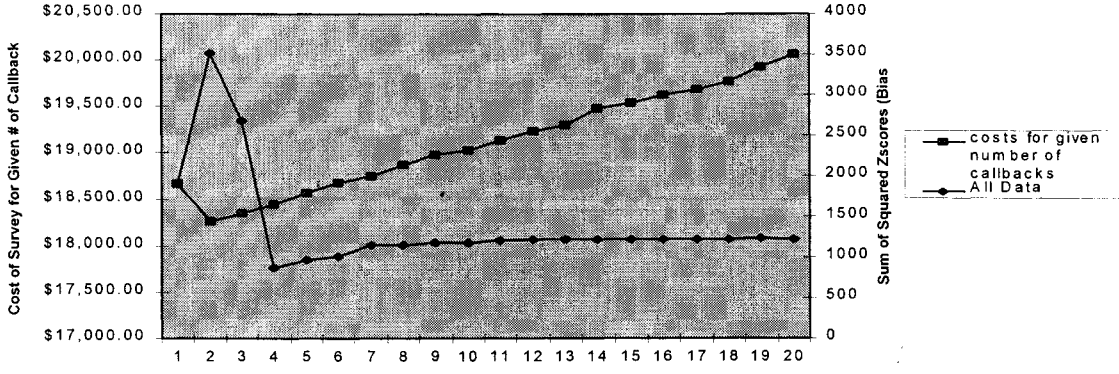

Table 1: *AGE*

	18-24	25-34	35-44	45-54	55-64	>65
Call1 vs. all	-8.58192*	-10.0335*	-0.581890	2.804421*	1.708354	10.66523*
Call2 vs. all	-12.0982*	-10.9924*	1.844424	5.528620*	2.260827*	9.112307*
Call3 vs. all	-3.47851*	-3.91766*	0.982755	3.017135*	0.708187	3.057539*
Call4 vs. all	-3.70333*	-3.87152*	1.395124	3.462548*	0.577798	2.429373*
Call5 vs. all	-3.76448*	-3.74243*	1.604637	3.564672*	0.606515	2.060671*
Call6 vs. all	-3.88286*	-3.86012*	1.655097	3.965980*	0.625587	1.839006
Call7 vs. all	-8.63608*	-3.76173*	1.849985	4.075988*	0.712614	1.524084
Call8 vs. all	-4.25659*	-3.68557*	2.185405*	4.144823*	0.724648	1.253807
Call9 vs. all	-4.24501*	-3.59170*	2.213862*	4.275583*	0.662311	0.971751
Call10 vs. all	-4.22116*	-3.55707*	2.393039*	4.319811*	0.596730	0.831628
Call11 vs. all	-4.25514*	-3.36809*	2.491225*	4.354585*	0.528603	0.687310
Call12 vs. all	-4.21358*	-3.38967*	2.586721*	4.461298*	0.458676	0.615865
Call13 vs. all	-4.30400*	-3.40677*	2.599766*	4.483797*	0.460989	0.542832
Call14 vs. all	-4.24699*	-3.41655*	2.607231*	4.496671*	0.462313	0.468127
Call15 vs. all	-4.26136*	-3.35410*	2.61605*	4.511882*	0.463877	0.469711
Call16 vs. all	-4.27299*	-3.28896*	2.623194*	4.524203*	0.390881	0.394356
Call17 vs. all	-4.21015*	-3.29446*	2.627581*	4.451715*	0.391535	0.395016
Call18 vs. all	-4.21454*	-3.29789*	2.630319*	4.536491*	0.391943	0.395427
Call19 vs. all	-4.22023*	-3.30235*	2.633874*	4.542622*	0.392473	0.319107
Call20 vs. all	-4.22636*	-3.30714*	2.718904*	4.549216*	0.393043	0.319570

*An asterisk cell indicates that the sample proportion of respondents in that category after a given number of calls is significantly different from population proportions at .05 level or better. The values in each cell are z-scores.

Table 2: MICH1A: *We are interested in how people are getting along financially these days. Would you say that you and your family living there are better off or worse off financially than you were a year ago?

	Worse off	Same	Better off
call1 vs. all	0.046914	2.448831*	-2.69451*
call2 vs. all	0.961969	3.690637*	-4.64651*
call3 vs. all	0.771632	1.375896	-2.49106*
call4 vs. all	0.640284	0.915045	-1.81449*
call5 vs. all	0.469953	0.961206	-1.65111
call6 vs. all	0.415242	0.760982	-1.35002
call7 vs. all	0.284068	0.703133	-1.11302
call8 vs. all	0.288885	0.396099	-0.75772
call9 vs. all	0.292702	0.320832	-0.67246
call10 vs. all	0.221729	0.324247	-0.67962
call11 vs. all	0.148899	0.326812	-0.49030
call12 vs. all	0.149833	0.246469	-0.49338
call13 vs. all	0.150586	0.165019	-0.29814
call14 vs. all	0.075463	0.165491	-0.19954
call15 vs. all	0.075718	0.166049	-0.20021
call16 vs. all	0.075924	0.166501	-0.20075
call17 vs. all	0.076058	0.166796	-0.20111
call18 vs. all	0.076145	0.166986	-0.20134
call19 vs. all	0.076248	0.083545	0
call20 vs. all	0.076350	0.083657	-0.20188

*An asterisk cell indicates that the sample proportion of respondents in that category after a given number of calls is significantly different from population proportions at .05 level or better. The values in each cell are z-scores.