Optimizing Call Strategies in RDD: Differential Nonresponse Bias and Costs in REACH 2010

Whitney Murphy, Colm O'Muircheartaigh, Carol-Ann Emmons, Steven Pedlow, and Rachel Harter, NORC

Summary

NORC investigated the relative efficiency of different call strategies for the REACH 2010 project by defining subgroups of cases based on their intermediate call status (*promising* and *unpromising*). Using the relative marginal productivity of additional calls to the two types of cases, we devised a call strategy for the survey. Calls per complete were determined at different stages, as were costs, and cut-offs were established based on the results. Implications for response rates and sample sizes were determined and evaluated.

Next, we compared the interview data generated for promising and unpromising cases to determine whether any differential nonresponse bias would have been introduced into the data by following only promising screener cases past the minimum call cut-off. Using costs and mean square error, we evaluated the call cut-offs we selected in light of the data. The findings suggest that limiting the level of effort to a maximum of seven calls for the unpromising cases would have substantially reduced survey costs without a significant loss in effective sample size.

Background and Problem

In telephone surveys, response rates tend to be positively correlated with the number of call attempts made; i.e., the greater the number of call attempts, the higher the response rate achieved. However, as every survey researcher knows, beyond a certain point, additional call attempts produce diminishing returns in terms of response rates. The research presented in this paper was motivated by lower-than-expected productivity, which resulted in the need to increase efficiency while maintaining high response rates and preserving the integrity of the resulting survey data.

The data come from the first round of the *Racial and Ethnic Approaches to Community Health (REACH) 2010* survey. REACH 2010 is a demonstration project sponsored by the Centers for Disease Control and Prevention (CDC) to eliminate health disparities among minorities. In the first round, NORC conducted sample surveys in 21 REACH communities, 18 of which were surveyed by telephone. In the telephone communities, NORC was contractually obligated to make a minimum of seven call attempts to complete a screening interview. The samples were a mix of pure random-digit dial (RDD) and dual-frame (RDD and listed telephone numbers) designs. At the time of our analyses, we had made a total of approximately 600,000 calls to nearly 90,000 telephone numbers.

Fairly early in data collection, we determined that the number of hours we were devoting to complete each screening interview was higher than we had budgeted. We needed to improve efficiency by determining how best to distribute calling effort across all types of cases. The purpose of this research was threefold: first, to distinguish cases worth pursuing with additional call attempts from those not worth pursuing; second, to identify an optimal maximum number of call attempts that would yield the highest return for the investment; and third, to determine the effect on the resulting interview data of imposing cut-offs on the maximum number of call attempts made.

Design and Methodology

In order to determine an optimal method for improving efficiency, we analyzed the effect of imposing specific cut-off points for the maximum number of calls. We surmised that certain cases would be more productive to follow past our seven call minimum than others. Thus, as a first step, cases were categorized as *promising* or *unpromising* based on the history of the first seven calls.

Several definitions of promising and unpromising were tested in our analyses; however, for brevity, only the definition deemed most appropriate for this study will be presented here. The adopted definition also happened to be the most inclusive, i.e., included more cases than the other definitions. Specifically, the cases categorized as promising were those for which there had been at least one non-negative contact in the first seven calls. A non-negative outcome is one that suggested we had made contact with a household, including soft refusals but excluding hard refusals. Those cases for which the first seven calls all had negative outcomes were deemed *unpromising*. These included cases with unknown household status (e.g., some combination of ring-no-answers and busy signals for the first seven calls), positive household status but no contact, and positive household status with negative contact (e.g., a hard refusal). Along with the promising and unpromising cases, a third type was examined in our analyses, namely, those that were completed on or before the seventh call. We call these the base cases. (Note that none of the cases being analyzed actually had any call cut-offs applied during data collection; the analyses presented in this paper simulate the results had cut-offs been applied,

using the actual data obtained by treating all cases identically.)

Our first set of analyses tested the hypothesis that promising cases were more worthwhile to follow past seven calls than were unpromising cases. We first examined the impact of using the same maximum call cut-off for promising and unpromising cases by computing the results this cut-off had on the screener response rate and cost. Next, we conducted a more refined analysis to test what would have happened if we had followed only promising cases past seven calls and finalized unpromising cases after the seventh call. A comparison of these analyses was used to evaluate our hypotheses about promising cases and to select an appropriate cut-off with respect to efficiency, response rate impact, and cost.

Limiting the number of calls to only the unpromising cases could possibly affect the representativeness of the interview data. Therefore, we examined the interview data to determine whether any nonresponse bias would be introduced by following only promising screener cases past the minimum call cut-off. Three additional sets of analyses were conducted to determine whether there were substantial differences among cases subjected to different maximum call cut-offs. The first set of analyses consisted of significance testing of several screener and interview variables to detect differences among the different types of cases (i.e., base, promising, and unpromising). The second consisted of comparing item response rates for the selected variables by type of case. The third analysis involved calculating the bias that may have been introduced by following only promising cases past the seventh call.

Analyses and Results

To determine the impact of imposed cut-offs on response rates and cost, we determined what would happen if all cases were stopped after a uniform call cut-off level. For this analyses, we chose cut-off levels from 7 to 20 calls and computed the effect that stopping work on cases after that number of calls would have on the response rates, the "marginal cost" (defined here to be the calls per competed screener for each increment up to the cut-off), and the overall cost (the average cost per completed screener). To compute the impact on the *response rate*, we measured the percentage of total (eventual) completes that had been completed by that cut-off level. For "*marginal cost*," we calculated, for each call increment, the ratio of calls to completed screeners for making one additional call to all unfinalized cases. In other words, if an eighth call were made to each of 1,000 cases, and 50 of those calls resulted in completed screeners, then the marginal cost for each of those additional completes was 1,000/50, or 20. This means that we made 20 calls to various households in order to yield one complete. By the same token, *average cost* was computed by taking the ratio of total number of calls to the total number of completed screeners for each cut-off rule.

The results of this analysis are shown in Table 1. By the seventh call (i.e., cut-off equal to seven), close to 80 percent of all the (eventually) completed screeners were completed. At nine calls, that percentage jumped to 86 percent, at 11 calls to 90 percent, and at 14 calls to 95 percent. Thus, if we called each case up to 14 times, we would achieve 95 percent of the completes that we would have achieved by pursuing cases to completion.[1]

Table 1: Results for Uniform Call Cut-offs (All Cases Finalized after k^{th} call)

Cut-off (k^{th} call) for All Cases	Completes as % of all Completes	"Marginal Cost"	Average Cost
7	79	26	18.3
8	83	28	18.8
9	86	31	19.2
10	88	36	19.6
11	90	38	20.1
14	94	49	21.2
17	97	59	22.0
20	99	64	22.6

The next analysis applied variable call cut-offs, based on call status at the seventh call, to illustrate what would have happened if we had implemented different rules for different case types during data collection. The three types of cases, which were described earlier, defined the analysis groups; these include the base cases (those finalized by the seventh call), the promising cases, and the unpromising cases. Just as in the first analysis, the base cases were included in the analysis only inasmuch as they contributed to the number of completed screeners and total calls made at each cut-off level and overall. The promising and unpromising cases were then given different treatments. Promising cases were followed past seven calls, up to 20 calls, while unpromising cases were stopped and finalized after the seventh call. Another comparison of response rates and cost ratios was done, and these results are displayed in

[1] Almost all cases were finalized by the 20^{th} call; less than 5 percent of calls made were after the 20^{th} call, and less than 2 percent of completes were attained after the 20^{th} call.

Table 2. Under this scenario, by extending the call cut-off to approximately 14 calls for promising cases, we would have achieved nearly 90 percent of the total completes while increasing marginal cost by only nine calls (from 16 to 25) and average cost by only a fraction of a call (from 18.3 to 18.5) when compared to finalizing all cases after seven calls.

Table 2: Results for Differential Call Cut-offs (Promising Cases Followed to k^{th} Call, Unpromising Finalized after 7^{th} Call)

Cut-off (k^{th} call) for Promising Cases	Completes as % of all Completes	"Marginal Cost"	Average Cost
7	79	--	18.3
8	82	16	18.3
9	84	17	18.3
10	85	20	18.3
11	87	21	18.3
14	89	25	18.5
17	91	40	18.7
20	91	32	18.8

In reviewing the results of these analyses, we determined that our conjecture had been correct, and that it would be beneficial in the future of this project (and possibly others) to establish a process for classifying cases as promising or unpromising after the first seven calls and then follow the promising cases further. The decision about how much further to extend the cut-off for promising cases was made based on the results of the analyses and on the feasibility and ease of implementation. As a general rule, we wanted to make sure that the response rate was not reduced by more than around 10 percent (in other words, we wanted to choose a cut-off such that at least 90 percent of our total completes would be attained). However, we also wanted to make sure marginal and average costs did not increase so much as to outweigh the benefits of achieving a desired percentage of completes. Any cut-off level between 12 and 14 calls would have been acceptable; we selected a 14-call cut-off because our telephone system was already set up for 14 calls. In implementing our new rules, all calls were run through the initial 7-call cycle, and then promising cases were put through another 7-call cycle for a total of 14 calls. This corresponded well with our stipulations for response rate and cost, as 89 percent of all completes were attained by the 14-call cut-off, and marginal and average costs were reduced noticeably.

Our next step was to determine whether the interview data would be affected by this adjustment to our calling rules. This was achieved through three different analyses: significance testing, an item response rate analysis, and a bias analysis. Several variables were selected from the screener and main interview for analysis purposes. The main criterion for variable selection was the availability of sufficient data. In other words, if a question was intended to be answered by all or most of the respondent population, it was included in the analyses. The screener variables studied were the reported number of adults per household, number of eligible adults per household, ages of household members, race and ethnicity of household members, and household size. The interview variables selected fell into two groups: demographic data (respondent age, income, race, and ethnicity) and health data (physical and mental health status, disease status, smoking habits, eating habits, and exercise habits). Again, base, promising and unpromising cases define the analysis groups used for comparison; for simplicity, they will be denoted as "B," "P," and "U," respectively.

A summary of the significance test results for a subset of the screener variables can be found in Table 3. The significance tests consisted of chi-square tests comparing base, promising, and unpromising cases. Overall, there are quite a few significant differences among the three types of cases. For example, promising cases reported both more adults per household and more eligible[2] adults per household than unpromising. However, base cases reported more eligible adults than either promising or unpromising cases. We also found that unpromising cases were more likely to be one-person households. Comparing these with one-person households among base and promising cases, there were fewer eligible one-person households among the unpromising cases. In fact, we found more eligible adults among the promising cases than base or unpromising cases for all household sizes except the largest, where there were no differences among the three groups. We also found that there were fewer older adults among the unpromising cases than among the base and promising cases. Finally, there were fewer eligible adults found in Hispanic and Asian households among promising and unpromising than among base cases, and fewer eligible adults in African-American households among the unpromising cases.

[2] In this study, eligibility was based on geography (location of the household) and race/ethnicity of the adult.

Table 3: Significance Tests for a Subset of Screener Variables

Variable	B	P	U
Mean # Adults/HH**	2.10	2.00	1.86
Mean # Eligibles/HH**	1.96	1.85	1.72
Mean # Eligible by HH Size			
1-person HHs**	0.53	0.56	0.46
2-person HHs**	0.88	0.95	0.72
3-person HHs**	1.59	1.62	1.37
4+-person HHs*	2.30	1.86	2.41
Mean # HH Members by Age Group			
18-30 years	0.59	0.55	0.57
31-44 years	0.55	0.55	0.57
45-64 years	0.50	0.47	0.42
65 years and older**	0.22	0.23	0.16
Mean # Eligibles by Race			
Hispanic HHs**	1.08	0.97	0.89
African-American HHs**	1.81	1.78	1.59
Asian HHs**	1.70	1.25	1.27

Note: * denotes significant differences among the three case types at the 0.05 level; ** at the 0.01 level.

There are fewer significant differences among the three case types at the interview level than at the screener level. These results are summarized in Table 4. Where differences do occur, the overall findings suggest that the unpromising cases are healthier than the promising cases. For example, the number of days per month that physical health is not good (which ranges from 1 to 30) is significantly lower for unpromising than promising cases. Also, a smaller percentage of unpromising cases report unhealthy smoking behaviors and body mass indexes than promising cases. Likewise, more unpromising cases walk 10 minutes per week than promising cases. Finally, women among the unpromising cases are more likely to have had a mammogram in the past two years than are women among the promising cases.

Results also suggest that the health of the base cases is not as good as either the promising or unpromising cases. For example, general health status, which ranges from 1 (excellent) to 5 (poor) is best among the unpromising and worst among the base cases. Also, income is lower among base cases than among promising and unpromising. Finally, the base cases are less likely to participate in moderate and vigorous exercise activities each week than promising or unpromising cases.

Combining the above results, it appears that the unpromising cases are more likely to be younger adults, live in smaller (single-person) households, and be healthier than the other types of cases. This is consistent with the definition of an unpromising case, given that young, active adults who live alone are less likely to be home to answer the phone. It follows that we risk under-representing this type of adult by not following unpromising cases past seven calls.

Table 4: Significance Tests for Interview Variables

Variable	B	P	U
Age	45.98	45.76	42.89
Income**	2.49	2.60	2.64
General Health Status*	2.90	2.82	2.73
Days/Month Mental Health Not Good	4.32	4.17	4.36
Days/Month Physical Health Not Good*	4.31	4.15	3.24
% Women 50+ had Mammogram in Past 2 Yrs*	90.4	91.5	97.1
% Women had Pap Smear in Past 3 Yrs	93.1	92.5	94.8
% Who Know Signs/Sympt. of Myocardial Infarction	4.1	3.3	4.0
% Told Cholesterol High	33.6	32.3	31.9
% Told have Diabetes	12.2	11.3	9.2
% At Risk: Body Mass Index**	59.8	65.4	59.6
% At Risk: Smoking*	21.5	21.2	16.3
Vegetables Eaten per Day	1.59	1.61	1.60
Fruits Eaten per Day	1.73	1.78	1.80
Fruit and Vegetable Serving Index	2.73	2.69	2.68
% Exercise Moderately 10 min/week*	63.6	66.8	67.7
Days/Week do Moderate Exercise	4.03	4.00	3.92
% Exercise Vigorously 10 min/week	28.6	30.2	32.8
Days/Week do Vigorous Exercise	3.29	3.15	3.03
% Walk 10 min/week*	71.2	70.5	77.0
Days/Week Walk for 10 Minutes	4.73	4.83	4.59

Note: * denotes significant differences among the three case types at the 0.05 level; ** at the 0.01 level.

Across the 21 interview variables analyzed, there was a lower item nonresponse rate for unpromising cases than for promising cases. In fact, counts of "don't know" and "refused" responses across all 21 variables show unpromising cases with the lowest overall item nonresponse rate (6.7%) compared to base (9.1%) and promising (9.3%) cases. Therefore, it appears that although the unpromising cases are more difficult to contact, once

they do cooperate they answer a higher proportion of the questions in the interview than do other types of cases.

Finally, in order to determine whether following only promising cases past seven calls would have an impact on the resulting interview data, we computed the estimated bias for two comparisons:

(1) *B+P vs. B+P+U* (i.e., is it worthwhile to follow the unpromising cases to 14 calls, or is it better to finalize those cases after seven calls based on bias and cost considerations?); and

(2) *B vs. B+P* (i.e., is it worthwhile to follow even the promising cases, or should we just finalize all cases after seven calls?).

The main goal of this set of analyses was to determine whether it is worthwhile to follow promising and unpromising cases beyond seven calls. Following only promising cases reduces the cost by eliminating all work on unpromising cases after seven calls. However, by not following those unpromising cases, we may be biasing the sample estimates. Thus, we need to determine whether the bias introduced is large enough to warrant the extra cost needed to pursue the unpromising cases as aggressively as the promising cases. If no bias is found, it raises the question as to whether it is worthwhile to expend the additional effort on even the promising cases. These analyses are presented in detail below.

The same screener and interview variables were used in these analyses as in the significance and item nonresponse analyses. Results at the interview level will be presented. Screener results follow the same patterns; however, as was true in the significance tests, they show somewhat larger differences.

The results presented in the tables for this analysis are based on comparisons of a treatment group to an "unbiased" reference group. The reference group for each comparison is the more intensively pursued group, and the group which we consider to produce unbiased estimates.[3] The treatment group, then, is the group for which the calling rules are changed, and which we are comparing to the reference group. The measures produced for each variable in this analysis were the bias, the mean square error, and the effective sample size (which takes into account both bias and cost) for the treatment group compared to the reference group.

[3] Note that bias may be introduced by a variety of sources; we refer to this group as "unbiased" in the sense that it is not biased by any changes in data that result from changes in the calling rules.

The numbers presented in the tables below reflect the range, across all interview variables, in the percentage gain or loss in effective sample size that is achieved for the treatment group versus the reference group, given the same cost. A value greater than zero indicates that the treatment group is actually more efficient, based on cost and bias, than the reference group. Less than zero shows that the bias introduced in the treatment group outweighs the cost savings, and thus the treatment group is less efficient than the reference group. Given no bias at all, the treatment group would always be more efficient because fewer cases are followed, thus reducing costs.

Our first comparison was of the treatment group, B+P, to the unbiased reference group, B+P+U. This tells us whether finalizing the unpromising cases after seven calls, but following the promising cases to 14 calls, would be worthwhile, or if the bias introduced by not following the unpromising cases outweighs the cost savings. Table 5, which displays a summary of the results of this comparison, shows that the bias introduced by not following the U cases past seven calls is far outweighed by the cost savings realized. There are only two variables for which the effective sample size for the treatment group (B+P) is smaller than for the reference group (B+P+U) at the same level of cost, and those are only very slightly smaller (1 to 2 percent). This suggests that, considering the bias and cost together, it is much more worthwhile to follow only P cases past seven calls and finalize all U cases at seven.

Table 5. Summary of Effective Sample Size Gain or Loss for B+P vs. B+P+U

Range	Quartile 1	Median	Quartile 3
-2% to +15%	+7%	+11%	+13%

Given that following the U cases past seven calls was found to be inefficient, and that finalizing them after seven calls did not introduce significant bias, we then wanted to determine whether it was worthwhile to pursue even the P cases past seven calls. Table 6 shows the interview results for the treatment group, B, compared to a new reference group, B+P. In this comparison, however, almost all of the variables show a moderate to substantial loss in effective sample size for B cases compared to B+P. This suggests that the reduction in cost that is achieved by not following the promising cases past seven calls is far outweighed by the bias introduced, and that it is not at all worthwhile to stop work on the promising cases after seven calls. Combining this result with the result from the previous comparison suggests that the most detrimental bias is introduced when we do not pursue the promising cases, while not following unpromising cases is actually

beneficial when bias and cost are considered together. In fact, the analysis presented in Table 2 showed that following promising cases to 14 calls was only minimally more expensive than finalizing all work after seven calls; thus, it follows that the reduction in effective sample size here is driven almost entirely by the bias, with cost having almost no effect.

Table 6. Summary of Effective Sample Size Gain or Loss for B vs. B+P

Range	Quartile 1	Median	Quartile 3
-28% to +1%	-7%	-2%	0%

Conclusions

Survey researchers are interested in finding ways to reduce survey costs without compromising data quality. For the REACH 2010 project, NORC lowered survey costs by differentiating between *promising* and *unpromising* cases and pursuing only the *promising* cases beyond the contractual minimum of seven calls. The results presented in this paper suggest that, in doing so, NORC did not sacrifice data quality. On the contrary, for the same cost, we significantly increased effective sample size, even though the response rates were reduced by about one-tenth. Equally, we could have chosen to maintain the effective sample size for a reduction in cost.

Although not reducing the effective sample size, the decision not to pursue the unpromising cases may have affected the composition of the final sample. In this study, the unpromising cases were found to be younger, single-adult households who reported healthier lifestyles and better physical health than the rest of the sample. Therefore, these types of adults are likely to be under-represented in the REACH 2010 sample as a result of not pursuing the unpromising cases beyond seven calls. This effect was mitigated to some degree, however, by the fact that the unpromising cases yielded relatively fewer adults who were eligible for REACH. This effect may be more deleterious to studies with different respondent eligibility criteria.

Finally, the results suggest that imposing a seven-call maximum on all cases would have resulted in highly biased data. Therefore, more calls are necessary, but the same maximum call limit does not need to be applied to all cases.

Suggestions for Future Research

The REACH 2010 sample upon which this study was based is not a nationally representative sample. Future research on nationally representative samples is recommended. For REACH, we also need to look at each of the individual communities to see whether the same general findings emerge for different racial and ethnic groups, different sample designs (e.g., pure RDD versus dual-frame), and different geographic areas. Future research could also refine the definition of a *promising* case and investigate whether it differs in different populations.