AN EXPERIMENT IN CALL SCHEDULING

Patricia Cunningham, David Martin, J. Michael Brick
Westat, 1650 Research Boulevard, Rockville, MD 20850-3195

Keywords: Random Digit Dial, Telephone Survey, Calling Protocol

1. Introduction

The National Survey of America's Families (NSAF), conducted by the Urban Institute, is part of a multiyear study to assess the New Federalism by tracking ongoing social policy reforms and relating policy changes to the status and well-being of children and adults. The major objective of the study is to assess the effects of the devolution of responsibility for major social programs such as Aid to Families with Dependent Children from the Federal to the state level. The NSAF collects information on the economic, health, and social dimensions of the well-being of children, nonelderly adults, and their families in 12 states and in the balance of the nation. The 12 states, which account for a little more than 50 percent of the country's population, were selected to vary in terms of their size and geographic location, the dominant political party, and key baseline indicators of well-being and fiscal capacity. A sample of the balance of the nation was included so that national estimates could also be produced. Low-income families were oversampled because the policy changes of interest are expected to affect them most. The initial round of the NSAF took place in 1997, with followup rounds in 1999 and 2002. The Urban Institute is being funded by a consortium of foundations, led by the Annie E. Casey Foundation, to carry out the Assessing the New Federalism Project. Westat is responsible for NSAF data collection and related activities.

Summary of Sample Design and Methodology

The survey is a dual frame design with a random digit dialing (RDD) sample to cover the approximately 95 percent of the U.S. population with telephones and an area probability sample to represent households without telephones. A major focus of the survey is to provide reliable estimates for persons and families below 200 percent of the poverty threshold. The area probability sample was included because of concerns about the potential bias in the estimates due to excluding low-income households without telephones.

The RDD portion of the survey used a list-assisted method for sample selection and computer-assisted telephone interviewing (CATI) for screening and interviewing. This component involved screening nearly 200,000 households and conducting detailed 25- to 45-minute interviews with approximately 40,000 to 50,000 persons under the age of 65. The area sample for Round 1 in 1997 and Round 2 in 1999 required listing nearly 40,000 addresses in 138 primary sampling units (PSUs) in order to conduct fewer than 2,000 interviews. For Round 3 in 2002, the state-level area samples were dropped and only a national area sample was conducted. Thus the size of the area sample was reduced by half for Round 3. Cellular telephones were used to connect sample persons in nontelephone households with the telephone center where telephone interviewers administered the questionnaires, reducing any mode effects.

Experiments in Round 3

During Round 3 a series of experiments were conducted to test incentive levels and timing, income screening, and the automatic call-scheduling algorithm. All of the experiments were done in what we called the predictor sample, a random sample of 60,000 telephone numbers that were fielded before all other sampled numbers. The incentive experiment had six conditions that tested whether a $2 prepayment at the screener makes a promise of $10 at the extended more effective than $5 at refusal conversion at the screener. It also tested whether the promise of $10 leads to a higher extended response rate than a prepayment of $5 and a promise of $20 at refusal conversion. The outcome of this experiment is discussed in Canter et al. (2003).

In the income experiment alternative methods of screening for income were compared. The purpose was to assess the most accurate means of asking income for oversampling low-income households without being obtrusive and losing the respondent during the screening interview. Accuracy of reporting was crucial for maintaining effective sample sizes. In one option respondents were asked two or three brief questions to dissagregate their income by thinking about how many people work and about other sources of income. In the second option, respondents were asked if their income was above or below an amount that was well below the low-income cut-off for the study and asked, if necessary, a followup question to work up to 200 percent of poverty.

The final experiment manipulated the automatic call scheduling algorithm. It is the subject of this paper.

2. Background for Experimenting with Call Scheduling

The experiment in call scheduling was focused on the procedures used to contact and determine whether the telephone number was residential or not. The contact procedure is clearly only one of the scheduling procedures used, but it has important implications for the cost of data collection. Other procedures, such as dealing with numbers that were contacted but not completed, were not tested in this experiment. It is also worth noting that the optimal contact procedure may not be the one that yields the greatest participation in the survey. See Brick, Allen, Cunningham, and Maklan (1996) for an example.

In the first round of NSAF in 1997 a call scheduling protocol was used that required the telephone numbers be attempted up to seven times over specific weekday, evening and weekend time periods to establish a first contact. When someone answered the telephone, the residential status for the telephone number was determined. At that point, the noncontact protocol was completed and a different protocol was used, if needed, to obtain the cooperation in the interview. If the telephone number was still not contacted after seven call attempts, it was held for at least a couple of weeks and then re-released for the same seven call sequence. Many of the numbers that were still noncontacts after 14 calls were re-released for seven more call attempts.

Exploratory analysis of the 1997 NSAF calling records found that the percentage of telephone numbers that result in an initial contact (i.e., it was possible to resolve the residency status as residential or nonresidential) decreased with the number of attempts. An interesting result was that the eighth and the 15^{th} (for those noncontact cases that had more than 14 attempts) call attempts had higher than expected percentage contacts, interrupting the essentially monotonic pattern otherwise exhibited. Since the 1997 scheduling protocol was to make seven attempts and then hold the numbers for additional releases, we hypothesized that the delay between the seventh and eighth and the 14^{th} and 15^{th} attempts was the cause of the higher than expected contact rates.

In Round 2 we acted on these findings by revising the calling pattern so that there was a delay between the seventh and eighth calls and a delay between the eight and ninth calls. Note that if the hold period was one week, the way the hold was implemented meant that actual time between the calls was at least one week. The total number of attempts was reduced from the minimum of 14 used in Round 1 to 9 for Round 2. A subsample of telephone numbers was fielded for additional calls to support estimating the residency rate using the survival method.

In Round 3 we revised the scheduling pattern for noncontact numbers to take fuller advantage of these results. We hypothesized that we might achieve the same increase in contacts with even fewer call attempts. The Round 3 approach involved an initial four attempts (over weekdays, weekends, and evenings), a one week hold period, an additional three attempts (completing the requirements for spread of the calls as used in Round 1 and Round 2), a one week hold, and then a release for two additional call attempts. The goal was to more quickly contact households and have fewer numbers in noncontact status. This approach would be used for most of the sample, with a subsample of numbers assigned a greater number of attempts for the survival estimation.

To evaluate the effectiveness of this new approach we conducted an experiment using a more traditional approach of seven initial calls, a one week hold, and two additional call attempts. The experiment is described in more detail in the next section.

3. Scheduler Experiment in Round 3

The hypothesis for this experiment stated that if we reduced the number of calls from seven to four before we held the case for a week we would reduce the number of call attempts required to determine residency. Essentially we expected to find the percentage contacted would increase on the fifth call attempt due to the introduction of the hold period. There were two conditions, named simply 4.3.2 and 7.2, with the dot representing the hold period.

- The 4.3.2 Condition. In this condition the initial four calls to establish contact were automatically scheduled so that one fell on a weekend (either Saturday or Sunday), one during a weekday (between 9am to 6pm), and two on weekday evenings (between 6pm to 7:30 and 7:30 to 9pm) within the respondent's time zone. The calls could be made in any order. If there was no contact, it was held for a period of one week before it was re-released for an additional series of three calls, one on the weekend, one during a weekday, and one on a weekday evening (between 6 and 9pm). If, for example, during the first four calls, an early call (9am-2pm) was made, then we made the other day call (2pm to 6pm) as part of the remaining three calls. If we called Saturday during the first four calls, we called Sunday during the next three calls. Again the calls could be made in any order. If there was no contact after seven calls it was held for another week at which point it was re-released for

an additional two calls, one in the evening (6pm to 9pm) and one on a weekend; and

- The 7.2 Condition. In this condition the initial seven calls to establish contact were automatically scheduled so that there were two on the weekend, two during weekdays (9am to 6pm), and three during weekday evenings (6pm to 7:30pm, 7:30pm to 9pm, and 6pm to 9pm) within the respondent's time zone. These calls could be placed in any order. If there was no contact after seven calls it was held for another week at which point it was re-released for an additional two calls, one in the evening (6pm to 9pm) and one on a weekend.

Notice that after seven attempts the number and placement of calls is exactly the same in both conditions, e.g., two during weekdays, two on weekends, and three on weekday evenings. The only difference is the one week hold period between the fourth and fifth calls in the 4.3.2 condition.

Within the predictor sample of 60,000 telephone numbers we predesignated a subsample of 5,000 telephone numbers to implement the more traditional approach with seven initial calls, a one week hold period, and two additional call attempts (7.2). The remaining 55,000 numbers received the 4.3.2 calling algorithm. Other experiments on income questions and incentives were also carried out using the predictor sample, but a factorial design was used so the sample sizes for the other experiments were balanced for the 4.3.2 and 7.2 conditions. Thus, the other experiments in no way interfered with the scheduler experiment.

4. Findings

Before examining the results of the experiment in terms of the main outcome variable (the number of call attempts to determine residency status), we present some basic results for the two experimental groups. Table 1 shows the sample of 5,000 was randomly assigned to the 7.2 condition and the remaining 55,000 numbers were assigned to the 4.3.2 condition. The percentages in the table show that the two experimental groups were very similar with respect to the number of telephone numbers that were purged (eliminated from dialing because they matched to a business number or were nonworking when they were autodialed). The percentage of numbers that were completed and the percentage of numbers identified as residential were also very similar for the two experimental groups. The differences in the percents for the two groups shown in the last column are all less than the standard error of the difference, so differences this large could be expected by chance. Since the telephone numbers were randomly assigned to the groups, the data in Table 1 essentially verify the random assignment process.

Table 1. Experimental groups outcomes

	Number		Percent		
	7.2 group	4.3.2 group	7.2 group	4.3.2 group	Difference
Sample size	5,000	55,000	100.0	100.0	
Purged	1,892	21,149	37.8	38.5	-0.6
Dialed	3,108	33,851	62.2	61.5	0.6
Contacted	2,643	28,740	52.9	52.3	0.6
Completed	1,292	14,300	25.8	26.0	-0.2
All residential	1,874	20,545	37.5	37.4	0.1

Note: The standard error of the difference of percents is approximately 0.7 so no differences are significant.

Now we examine the distribution of the percent of numbers with resolved residential status by the number of call attempts to determine if the expected relationship holds for the two experimental groups. For this and all further analysis we only include those telephone numbers that were resolved by the ninth call attempt. Figure 1 graphs the cumulative percentage of all sampled telephone numbers that have a resolved residency status for each group. The graph shows that after the first call the percent resolved is about the same for the two groups, but the difference between the 4.3.2 and 7.2 groups is statistically greater than zero at both the second and third call attempts. The difference decreases as the number of call attempts increases. There is no indication of the desired increase in the percent resolved at the fifth call attempt for the 4.3.2 group.

Before exploring the unexpected differences in the percent resolved shown in the second and third call attempts, we more directly examine the anticipated increase in the conditional probability of contact. Figure 2 shows the conditional probability of determining the residential status of a telephone number for the two groups by the number of call attempts. For example, the conditional probability of resolving a telephone number on the second attempt is the number of telephone

numbers that were resolved on the second attempt divided by the number that required at least two calls.

The typical pattern is that the conditional probabilities of resolving the numbers decrease with the number of attempts. Figure 2 shows this general pattern but there are some important features. The conditional probabilities for the 4.3.2 group are substantially greater than the same probabilities of the 7.2 group at the second and third call, resulting in the differences noted in Figure 1. The conditional probabilities for both groups are also not monotonic, and differences from the decreasing rate are clearly seen in the third and eighth call attempts.

Focusing on just the 4.3.2 group, we do not see a spike in the conditional probability of resolving a telephone number at the fifth call, as had been hypothesized. In contrast, consider the eighth attempt that was also delayed for a week, does have a spike that is especially apparent for the 4.3.2 group. If the numbers were not held for at least one week between call attempts seven and eight we would have expected a constant or decreasing conditional probability of resolving at this point. Figure 2 clearly shows that the anticipated relationship in terms of the changes in conditional probabilities for the 4.3.2 group were not obtained, nevertheless the differences in the resolution rates in the early calls are very interesting and these are considered next.

When we examined Figure 1 and saw the differences in the resolution rates at the second and third attempts, we began to investigate the possible source of these differences. One way is to classify the telephone number by the residency status eventually assigned for the number. While we classified the numbers by a number of categories, we only present the one that was most interesting here. Figure 3 is constructed exactly like Figure 1, except it is restricted to telephone numbers that were determined to be residential. Figure 3 shows that residential numbers contributed to the differences in the resolution of the telephone numbers in the second and third call attempts. The resolution rates for the two groups do not converge until the fifth call or later. In fact, most of the differences in Figure 1 are due to the contribution of residential numbers as displayed in Figure 3.

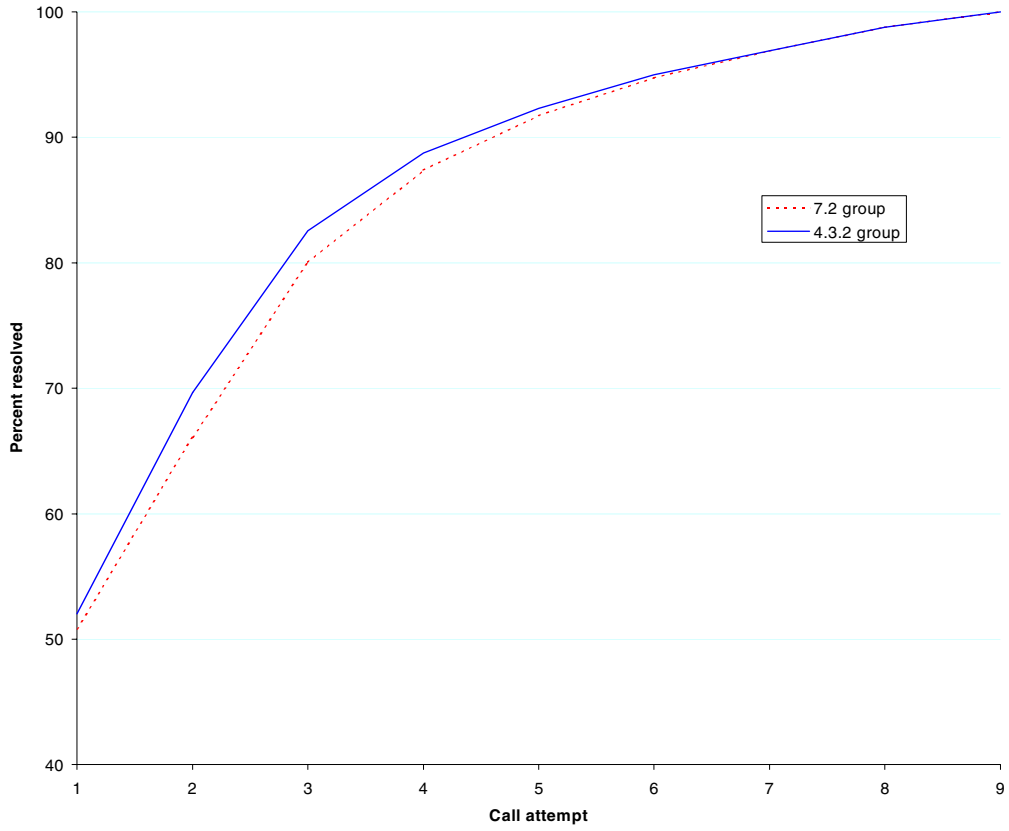

Figure 1. Distribution of all numbers resolved, by call attempt and experimental group

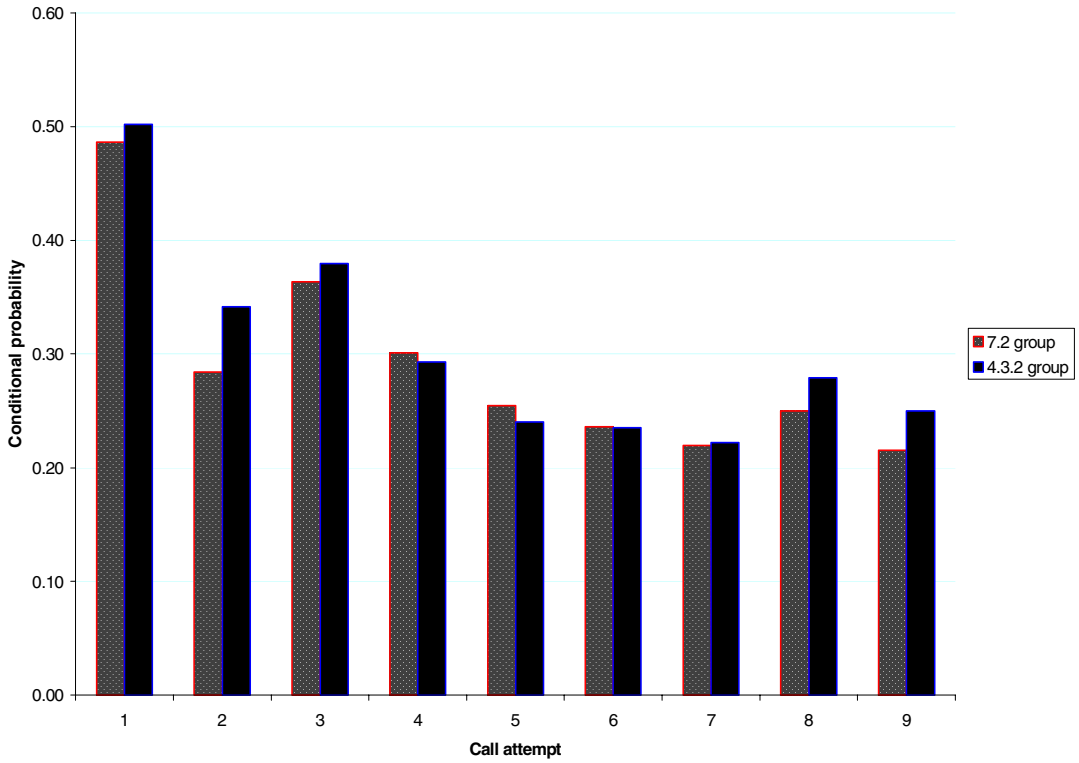

Figure 2. Conditional probability of resolving residency status on a call attempt, by group

Another way of categorizing the telephone numbers is by whether an address could be found for the telephone number that could then be used to send an advance letter to the household. Those numbers associated with an address are called mailable. This classification was considered because mailable numbers have much higher residential rates than nonmailable numbers. When we examined the distributions by mailable and nonmailable status, we found that mailable status did not account for the differences in the experimental groups. The sample size for the 7.2 nonmailable group was too small to do a detailed analysis. However, looking at all 60,000 telephone numbers by mailable status is interesting in its own right. Figure 4 gives the cumulative percentage of all sampled telephone numbers that have a resolved residency status for mailable and nonmailable telephone numbers. The graph shows that mailable numbers are contacted and resolved with fewer attempts than nonmailable numbers. Mailable numbers are much more likely to be residential than nonmailable numbers, so we might expect a slightly lower number of call attempts for the mailable numbers (overall, residential numbers required on average 2.3 call attempts to resolve while nonresidential numbers required 2.4). The mean number of attempts for mailable numbers was 2.2 and for nonmailable numbers the mean was 2.5 attempts. This finding adds some new evidence about the relative efficiency of sampling by mailable status (Brick et al., 2002).

While these results are interesting, they do not explain the difference in the rates of resolving residency status for the two experimental groups at the second and third call attempts. To address this concern, we decided to evaluate the calling patterns for the two groups. We understood that even though the distribution of call attempts by time period (day/evening/weekend) for the 4.3.2 and 7.2 had to be equivalent after seven attempts, they were not forced to follow the same path to this point. In fact, the 4.3.2 group calling pattern was likely to be different from the pattern for the 7.2 group, because numbers assigned to the 4.3.2 group were required to have one day, one weekend and two evening attempts in the first four calls. This restriction was not placed on the 7.2 group. Clearly, this could have an important effect on the rate of resolving the cases.

Table 2 gives the percent of telephone numbers attempted in each experimental group by calling pattern for the first four call attempts. The percentages do not always add to 100 because any pattern that was less than 0.5 percent for both the 4.3.2 and the 7.2 group was excluded from the table. The last column gives the percentage point difference between the estimates for the 4.3.2. group and the 7.2 group. For the first call attempt, the calling patterns are essential the same for the two

groups as should be expected since no restrictions were placed on the first attempts.

On the second attempt, the differences are substantial. The 7.2 group is much more likely to have two attempts on the weekend or during the day, while this pattern was generally not permitted for the 4.3.2 group. The scheduling algorithm did allow some exceptions to the calling rules if it was necessary for workflow reasons. As a result, the percentages in the undesirable patterns are small but greater than zero. For those numbers that required a third attempt, the call distribution for the two experimental groups is still very different. Nearly all of the numbers in the 4.3.2 group have one day attempt and one or two evening attempts. On the other hand, over 83 percent of the numbers in the 7.2 group do not have an evening attempt. Even after the fourth attempt, about 25 percent of the 7.2 group have not had an evening call, while virtually all of the numbers in the 4.3.2 group have two evening calls at the same point. Since the rate of contact for residential numbers is much greater in the evening than during other time periods, the call pattern appears to largely account for the large differences in rates for the two groups overall and for residential numbers.

The findings in Table 2 show that the change in the scheduling algorithm did affect the ability to contact households and determine residential status, but the consequences were not exactly what were expected. The 4.3.2 algorithm required a distribution of call attempts by time period that increased the rate for contacting households in a fewer attempts. As the number of call attempts increased, the probability of having a less desirable pattern for the 7.2 group decreased and by the seventh attempt the two groups had equivalent distributions by time of attempt. This convergence corresponds to the convergence of the cumulative percentage of cases resolved by call attempt for the two experimental groups.

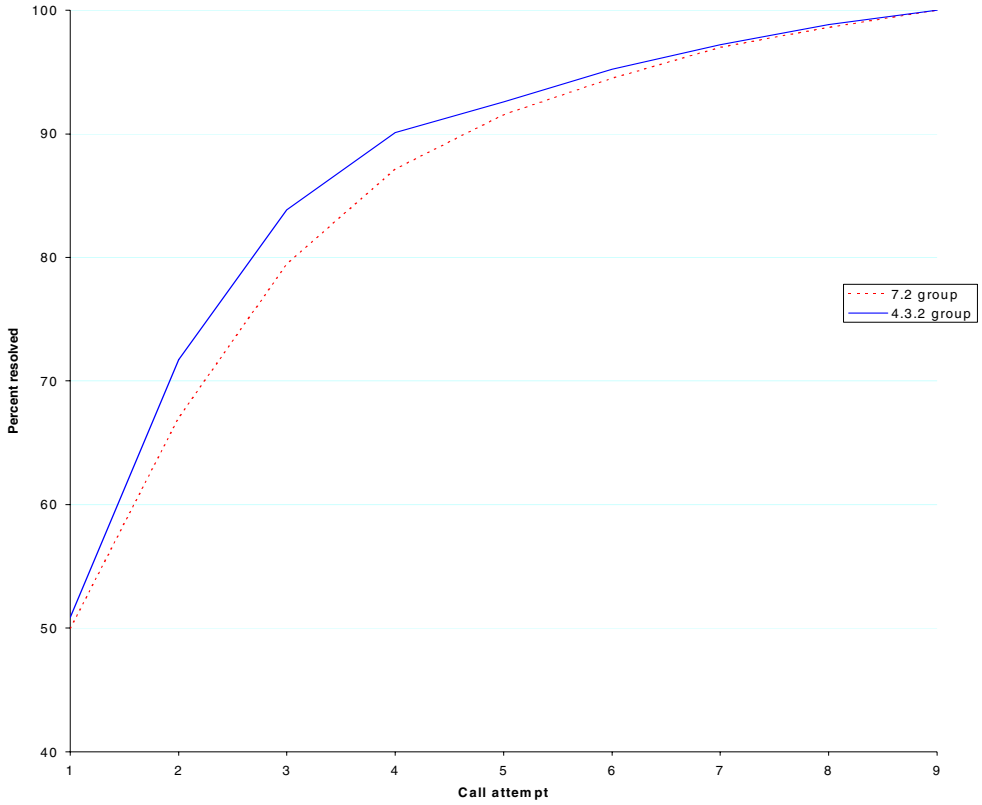

Figure 3. Distribution of residential numbers resolved, by call attempt and experimental group

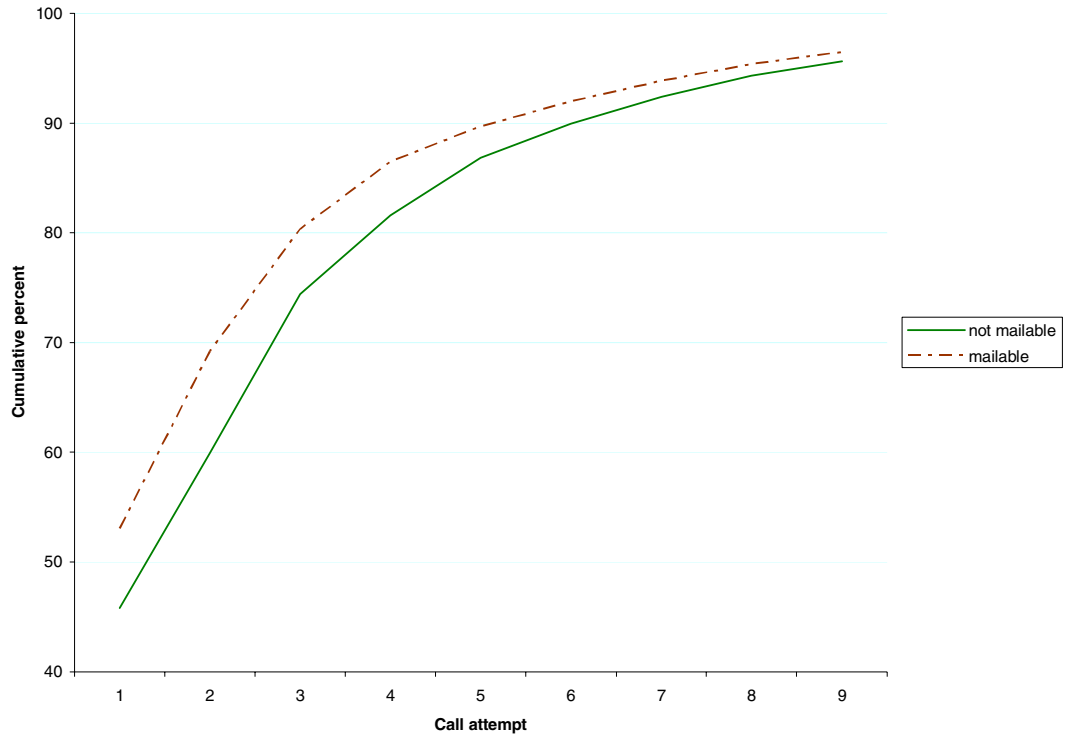

Figure 4. Distribution of percentage of all numbers resolved by call attempt and mailable status

Table 2. Time of attempts by experimental groups

	Time of call			Percent of attempted		
	Day	Evening	Weekend	7.2 group	4.3.2 group	Difference
1st attempt	0	0	1	37.1	37.2	-0.2
	0	1	0	15.9	15.1	0.8
	1	0	0	47.1	47.7	-0.6
2nd attempt	0	0	2	19.1	1.7	17.4
	1	0	1	26.4	48.2	-21.8
	1	1	0	14.2	45.3	-31.1
	2	0	0	37.6	2.5	35.1
3rd attempt	1	0	2	24.8	1.6	23.2
	1	1	1	4.6	70.2	-65.6
	1	2	0	0.4	24.4	-24.0
	2	0	1	23.1	1.9	21.2
	2	1	0	40.2	1.7	38.5
4th attempt	1	1	2	11.8	2.1	9.7
	1	2	1	0.4	92.5	-92.2
	2	0	2	26.5	0.4	26.1
	2	1	1	50.3	2.9	47.4
	2	2	0	6.1	0.9	5.2

Note: Excludes patterns with less than 5% of all attempts in any group.
All differences are statistically significant except those for the 1st call attempts.

5. Discussion

The scheduler experiment compared two different calling algorithms to evaluate a hypothesis that one of the approaches would increase the probability of reaching the household and determining if the number was residential or not. The 4.3.2 procedure required

holding telephone numbers at least one week after the fourth call, expecting that this would increase the conditional probability of resolving the status on the fifth call as compared to the 7.2 procedure that did not have this hold period. The findings showed that using the 4.3.2 procedure did not increase the conditional probability of resolving a number at the fifth call. However, the 4.3.2 procedure was superior to the 7.2 procedure in an unanticipated way.

The 4.3.2 procedure required dialing with calling patterns by time period (combinations of day/evening/weekend calls) that had a higher likelihood of contacting households. In particular, the first four calls in the 4.3.2 procedure had to have one day, one weekend and two evening attempts in the first four calls. The 7.2 procedure did not have this requirement and was much more likely to have few evening calls in the first few dialing attempts. Since the distribution of the first seven calls by time period for the 4.3.2 procedure and the 7.2 were required to be the same, the difference in the percentage of numbers resolved decreased as the number of call attempts increased.

The experiment clearly shows that scheduler algorithm that requires a variety of time periods in early calls reduces the total number of call attempts needed to contact households and to eliminate nonresidential numbers. In a large survey that extends over a relatively long calendar period like the NSAF, this type of scheduler is feasible and recommended. Since the 4.3.2 procedure was used for all the nonexperimental telephone numbers in NSAF in 2002, the benefits were obtained in the survey.

On the other hand, the 4.3.2 scheduler algorithm and others that place greater restrictions on when calls to a number can be made do have implications for survey operations. As more restrictions are placed on when calls can be made, the flow of cases to the interviewers may be affected. In the extreme, a scheduler that places too many restrictions may cause interviewers to have no work to call at certain time periods. Clearly, this would be much less efficient than allowing calls to numbers at less than optimal periods. This type of situation is most likely to happen in surveys that have a short field period or have limited interviewing resources. In such cases, more sophisticated algorithms that prioritize numbers to be dialed so those most likely to benefit from a call attempt at the current period are scheduled first are needed. In fact, the scheduler used for NSAF actually had many of these features.

6. References

Brick, J.M., Allen, B., Cunningham, P., and Maklan, D. (1996). Outcomes of a calling protocol in a telephone survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 142-149.

Brick, J.M., Judkins, D., Montaquila, J., and Morganstein, D. (2002). Two-phase list-assisted RDD sampling. *Journal of Official Statistics*, 18, pp. 203-216.

Cantor, D., Cunningham, P., Triplett, T., and Steinbach, R. (2003). *Comparing Incentives at Initial and Refusal Conversion Stages On a Screening Interview for a Random Digit Dial Survey*. Presentation at AAPOR in Nashville, TN.