Survey Research Methods (2013)
Vol.7, No.1, pp. 45-55
ISSN 1864-3361
http://www.surveymethods.org

Adaptive Contact Strategies in Telephone and Face-to-Face Surveys

James Wagner
University of Michigan

In attempting to contact households for a survey, it is necessary to determine the timing of each call. Often, average "best" times to call are used in order to determine when to place the first call(s). The timing of subsequent calls is then governed by very general rules. This paper tests an experimental method that uses multi-level models to predict the times that have the highest probability of contact for each household and uses the predictions from these models to prioritize cases for calling. The predictions are updated each day in real-time as additional data are gathered. The method is evaluated through a series of experiments on a telephone survey that used automated call scheduling and an experiment on a face-to-face survey, where a recommended calling time was delivered to interviewers. The results of these experiments are used to suggest directions for future research.
Keywords: calling strategies; responsive design; household nonresponse

1 Introduction

Since the 1970s, the computerization of telephone interviewing has offered the prospect of improving the efficiency of conducting surveys. It was hoped that computerized call-scheduling algorithms would lead to higher contact and interview rates. There is a body of literature that focuses on methods for improving contact rates in telephone surveys. Much of this literature focuses on determining the average best times to call; or the sequence of calls that have, on average, the highest contact rates. Unfortunately, it would appear that little progress has been made in this area of research.

An alternative approach to this problem would be to develop household-specific estimates of the best time to attempt contact. These estimates would be built upon to the extent that they are available – the call-level data from the household. Such estimates might provide the basis for a more efficient contact strategy. In addition, this approach would allow us to "tailor" the contact strategy to the household. If successful, a tailored strategy would allow data collection organizations to increase control of the composition of the set of respondents.

In this paper, I will outline such a strategy and present results from a series of randomized experiments on both telephone and face-to-face calling strategies. The strategy proposed here is described as "adaptive" since it learns from sequentially gathered data while also directing how those data will be gathered from future calls. It develops an estimate of the best calling time for the next call using the current call history data. The prescribed strategy is then attempted, the result is added to the data, and a new estimate is derived from these supplemented data. Unexpected results from the initial experiment illustrate the importance of considering both operational constraints as well as the possibility of interactions between treatments across phases of the survey process. Some combination of these issues may have led to these unanticipated results. A series of modifications to the experimental treatments were implemented in an attempt to counteract these effects. Finally, the results from these experiments will be used to suggest future directions for research.

Contact information: James Wagner, University of Michigan, Survey Research Center, e-mail: jameswag@isr.umich.edu

2 Background

Efficient call-scheduling algorithms have long been a subject of research for survey methodologists. Much of the research in this area has focused on the average best times to call. This research may consider the placement of a single call or short sequences of calls (for example, the first three or five calls). In an early article in this area, Weeks at al. (1980) looked at the best times to place a call using data from an in-person survey. This research was extended by Weeks et al. (1987) to a telephone survey and the timing of the first three calls was considered. Other research on telephone surveys has looked at the efficiency of various calling patterns for the first few calls (Kulka et al. 1988; Massey et al. 1996; Cunningham et al. 2003). This research has to led to general recommendations about how to most efficiently establish contact. The European Social Survey (ESS) provides an example of such a contact protocol (Stoop et al. 2010). The guidelines for the ESS suggest that a minimum of four calls be placed to each household and that these calls be spread over different times of day and days of the week, with at least one call in the evening and one on the weekend. The guidelines also suggest that calls be placed in at least two different weeks to aid in contacting households that may be away temporarily.

Another strand of research has attempted to recommend timing for the next call based on characteristics from the sampling frame and the history of previous calls. In the case of Random-Digit Dial (RDD) surveys in the US, the information on the sampling frame is generally limited to character-

istics of the geographic area with which the telephone number is associated. Greenberg and Stokes (1990) employed a Markov Decision Process model that used the history of previous calls as well as the frame data to determine the best time to place the next call. The transition probabilities were estimated using logistic regression. Their model suggested that 30% of the calls should be placed on the first evening of the survey. This recommendation is beyond the capability of most telephone research facilities and has never been experimentally tested. Brick et al. (1996) considered a similar approach that used logistic regression models to identify the best time of day, day of week and lag time between calls. Predictors in the model included contextual data as well as information about the results of previous attempts. Groves and Couper (1998) recommend calling unlisted telephone numbers first in order to allow for more attempts since these cases are expected to be more difficult to complete. They also suggest that cases with answering machines might be given a special protocol involving more calls during the evening. In general, this second strand of research produced protocols that differ from the average best time protocols in that they require that the call scheduling adapt to incoming information (the outcomes of previous calls).

In contrast to research on telephone surveys, most of the research on establishing contact in face-to-face surveys focuses on variation in contact rates across interviewers (Campanelli et al. 1997; Purdon et al. 1999; Pickery and Loosveldt 2002; Durrant and Steele 2009). For instance, Campanelli et al. (1997) observe that more experienced interviewers would prefer to call during the daytime – even though these calls are less efficient. Eventually, these experienced interviewers will switch to calling at other times. In face-to-face surveys, it is common for interviewers to be given a general training in the times that are best for establishing contact. They are encouraged to call at different times of day and days of week, but they are left to work the sample as they see fit. As a result, the emphasis of the research into contact strategies for face-to-face surveys has been focused on how well different interviewers do at this task. There is almost no research into how to improve contact rates in these settings. For face-to-face surveys, Groves and Couper suggest contacting managers of locked buildings as early as possible and possibly switching to telephone for these kinds of cases. Although they do not offer a specific strategy, they do suggest that call record data and observations can help "managers guide interviewers in their calling strategies" (Groves and Couper, p. 117).

Durrant et al. (2011) find that household characteristics are useful in predicting the best times for contact. They also note that in the absence of these characteristics, interviewer observations about sampled housing units and characteristics of the area (such as Census data) can be predictive of contact rates. The results of their analysis suggest that strategies such as calling at different times and leaving a note may help improve contact rates. They also suggest that cases that are more likely to be contacted during the day can be identified and prioritized for effort during that time.

In panel surveys, some research indicates that higher contact rates for the second wave and following may be achieved by calling households at the time of day and day of week in which they were initially interviewed (Laurie et al. 1999; Lipps 2012).

Previous research has also considered the relationship between contact and cooperation. Since the ultimate goal is to attain cooperation, one might adopt a strategy that has lower contact rates if it would lead to higher cooperation rates. For example, we might expect high contact rates for calls placed at 3am. We would, however, expect quite low cooperation rates for such calls. In general, prior research has identified only weak correlations between the time of contact and cooperation. Weeks et. al. (1987), for example, found that the effect of times of day and day of week was similar for both contact and cooperation, with the exception that Sundays tended to have higher cooperation rates than other time of day and day of week combinations. Brick et al. (1996) report a similar finding. Lipps (2012) reports that cooperation appears to be unrelated to time of day and day of week in the panel survey setting, but calling a household back at the time of day and day of week at which they were first interviewed does increase cooperation rates.

The concept I will employ for exploring the best calling algorithm is a learning model. Over time, as we accrue more data on any particular household, we are learning about that household's patterns for being at-home and willingness to answer the telephone or a knock at the door. Successfully contacting a household at one time of day or day of the week (i.e. in a particular calling "window") increases our estimate of the chance of success in that window for that household.

Unsuccessful attempts to contact a household decrease our estimate of the probability of achieving contact in that window.

For many households, contact and interview result relatively early in the process – within the first two or three calls. This means that for many households we have little or no data with which to estimate the probability of being home and willing to accept a telephone call within any window. For these households, we will need to "borrow strength" from the data generated by telephone calls to other households. This can be done using multi-level models (Gelman and Hill 2007). In those households where we have no data, we are essentially using conditional means where we condition on the sampling frame data that are available for all households (see below for a description of sampling frame data).

Research in other fields has addressed similar problems in different contexts. In the area of marketing research, Rossi, McCulloch, and Allenby (1996) consider a similar problem. Their goal was to customize or tailor the face value of a coupon to a specific household. They attempted to estimate household-level parameters using demographic and purchase history data. They used multi-level models to do so. Bollapragada and Nair (2010) considered the problem of improving "right party contact" rates at credit card collection calling centers. Their goal was to estimate contact probabilities for each household the call center is attempting to reach. Their algorithm assigns the overall average contact rate to each household and adjusts these starting values for each household upward when a call attempt is successful and

downward when the attempt fails. In both these papers, the authors have attempted to estimate characteristics of households using historical data.

In my application, I will attempt to estimate household characteristics – the probability of being contactable (i.e. at home and willing to answer a telephone call). I will use multi-level models where the household is a grouping factor. The fixed effects are frame variables available for all cases. These variables are time invariant for each household (i.e. they do not vary over calls). The household-level estimates will be used to decide which cases have their highest probability of contact (not necessarily the highest of all cases) in the current window. Those cases will receive the highest priority for calling. Cases for which the current window has the second highest probability of contact will be prioritized after those cases, and so on until all active cases have been prioritized. The models are re-estimated daily and the entire, active sample is re-prioritized at the beginning of each call window. The prescribed time for the next call may change for any given household after the success or failure of previous calls.

Although the ultimate goal is to complete interviews, I focused on establishing contact for three reasons. First, in telephone surveys, establishing contact is difficult and noncontacts generally compose a large proportion of the nonresponse (Curtin et al. 2005). Second, prior research (cited above) indicates that there is not a strong connection between time of contact and cooperation. Third, in the telephone survey used here, we noted that the number of contacts required to complete the survey each month was fairly constant, while the number of calls tended to vary a great deal. This seemed to indicate that greater efficiency was possible.

This paper presents the results of several experiments. Each experiment generated new hypotheses that led to further refinements of the experimental treatment. These are the first steps toward the eventual goal of producing a strategy for establishing contact and completing interviews that is more efficient for any given household.

3 Data and Methods

The data come from two surveys. The first is an RDD telephone survey that is conducted on a monthly basis – the Survey of Consumer Attitudes (SCA). The survey collects approximately 300 RDD interviews per month. The survey has a fixed field period (about 4 weeks) and, as a result, quite frequently generates multiple calls per day during the latter part of the field period to each active case in an effort to meet production targets. The sample is prepared by a vendor that attaches contextual data to the sample file. The ZIP code of each telephone number is estimated using listed numbers from the same 100-bank (sets of 10-digit telephone numbers with the first 8 digits in common). Census data for the associated ZIP Code Tabulation Area (ZCTA – the Census Bureau's attempt to match Census geography to ZIP codes) are then attached to each telephone number. Table 1 lists several of the key context variables that are available. Of course, given the estimated geography of the case, any data that are reported for particular geographies can be attached to the sample in a similar manner (see Johnson et al. 2006 for a more detailed description).

Table 1: SCA Contact Propensity Predictor Variables

Context Variables
Listed/Letter Sent
% Exchange Listed
% of Telephone Numbers in Exchange that are Listed
Total Households
Household Density (households per 1000 Sq ft.)
Median Years Education
Median Income
Log (Median Income)
Census Region
% 18-24, 25-34, 35-44, 45-54, 55-65, 65+
% White
% Black
% Hispanic
% Owner Occupied

Prior research indicates that household-level covariates (Durrant et al. 2011) or, in the case of panel studies (Lipps 2012), information from prior waves (including call history data) will be more predictive than these context variables. However, in the absence of household-level data, these context variables have been shown to be predictive of contact (Durrant et al. 2011). Other research has found that the urbanicity and median income of the estimated geographic area (Dennis et al. 1999; Brick et al. 1996) are predictive of contact rates in telephone surveys. As part of the model fitting exercise, different transformations on some of these variables were tried. The natural logarithm of the median income sometimes produced a better fit. Brick et al. (1996) reported using a similar strategy. Other research has reported that the proportion of the population that is Black, the proportion Hispanic, and the median years of education of the estimated geography of the telephone number are predictive of contact rates as well (Brick et al. 1996).

The second set of data comes from a large area probability sample with face-to-face interviewing. The National Survey of Family Growth 2006–2010 (NSFG) was a continuous survey that released new samples of about 5,000 housing units each quarter. The sample was worked to completion in 12 weeks and then a new sample was released. Since the survey is an area probability sample, data from Census 2000 were available at the Census Block level. Additional variables are available at higher geographic levels, such as Census Block Group, Census Tract, and the Census ZIP Code Tabulation Areas (ZCTA). These are the context variables described in Table 2. Some of these variables (percent working 16 years of age and older, percent working in the evenings, and percent that commute 30 or more minutes) are available from the long form of the Census (since replaced by the American Community Survey) at the Census Tract or ZCTA level. In addition, field interviewers visit the neighborhood before attempting contact with any households. During those

initial visits, interviewers make observations about the neighborhood and housing units. Those observations are listed in the "Interviewer Observations" column of Table 2.

The data from the telephone and face-to-face surveys include the records of every call. Each call is a record in the dataset used in the analysis. These calls and, hence, the number of data records increase each day as new calls are made. In the case of the telephone survey, these call records are automatically generated and time-stamped by the sample management system. The interviewer assigns a result code and leaves a note about the call. In the case of the face-to-face survey, the call records are generated manually. This leaves the possibility that some calls may not have a record, or the time assigned to the call might be inaccurate (Biemer et al. 2011). The fact that interviewers determine when to place the call in the face-to-face survey also opens the possibility that estimates of contact rates may be biased. The time recorded for each call was recoded into clusters of time known as "call windows". These call windows are defined by the day of the week and time of day (see Table 3) Call windows were defined by reviewing contact rates for prior quarters/months of data collection. The number of windows was set to four since more windows would leave fewer calls in each window. Using more than four windows might have increased the homogeneity within any window, but there would also be less data in each window.

Some calls were excluded from the dataset. Any calls that were set as appointments were deleted since the purpose is to predict the probability of contact for a "cold" call, not an appointment. The call number did not enter the models as a predictor. Estimating the average probability of being at home after eight calls, for example, was not the goal. The goal was to provide household specific estimates. For example, if we were to call a household 8 times and have contact on all 8 calls, we would expect to have contact on a 9^{th} call for that household. The contact rate for all 9th calls is not particularly informative for this purpose.

In addition, in the first experiment, for operational reasons related to the sample management software in the telephone facility, refusal conversion and Spanish language calls were not included in the experimental algorithm. Cases in these groups were only identified during calling and were then treated separately from the calls to the majority of cases. It was thought to be simpler to ignore these groups at first. The separate treatment of these cases proved to be important when the results of the experiment became available and was the basis for further modifications.

At the beginning of the field period, there are no call histories for the current sample. Therefore, data from prior months or quarters were used. Specifically, in the case of the telephone survey, the call records from the same month in the prior year (in order to capture any seasonal effects in the data) and the month prior to the current month were used. In the case of the NSFG, data from the prior quarter were used. Data from the current month or quarter are analyzed daily. The models are re-estimated daily, and the results are updated with all of the call records from the first day through the prior day included. Estimates of coefficients from these models were monitored on a daily basis to determine if estimates might change over time.

The models are multi-level logistic regression models predicting contact (R_{ijl} with the household being a grouping factor. The models provide household-specific estimates of the probability of contact for each of the call windows. The predictor variables in this model are the variables described in Tables 1 and 2. Let X_{ij} denote a $k_j \times 1$ vector of predictor variables for the i^{th} household and j^{th} call. The data records are calls. There may be zero, one, or multiple calls to a household in each window. The outcome variable is an indicator for whether contact was achieved on the call. This contact indicator is denoted R_{ijl} for the i_{th} household on the j^{th} call to the l^{th} window. Then for each of the four call windows denoted l, a separate model is fit where each household is assumed to have its own intercept which is from a $N(0, \sigma^2)$ distribution. The model is estimated:

$$Pr(R_{il} = 1) = log\ it^{-1}(\beta_{0l} + \beta_{0il} + \sum_{j=1}^{p} \beta_{jl} X_{ijl})$$

In these models, the coefficients and their standard errors are nuisance parameters – that is, parameters that need to be estimated in order to estimate the quantity of interest, but in which we are not directly interested. Nevertheless, in order to give a sense of these models, we briefly describe some of the estimated coefficients from one of the estimated models. For the SCA at the end of September 2009, an increase in the square root of the proportion of the households in the estimated ZIP code of the telephone number will increase the estimated probability of contact. This was true in all of the windows. On the other hand, in Window 1, neighborhoods with a higher proportion of persons 35–44 years of age have lower estimated rates of contact. This effect was not observed in Windows 2, 3, or 4 when the models were fit.

Figure 1a shows the predicted contact rate for households in window 1 (y-axis) by the empirically observed contact rates for those households (x-axis) for September 2009. The solid 45-degree line is the set of points at which the observed and predicted contact rates would be equal. It is worth noting that there are small sample sizes (i.e. the number of calls) for many cases – those cases for which contact and interviewing is achieved early. In those cases, we do not have a very reliable estimate of the true contact rate since the sample sizes are quite small. The marginal logistic regression model (the light gray squares and dotted line) is averaging contact rates over all the households conditional on the fixed characteristics (i.e. the sampling frame and context variables) in the model. It appears that these characteristics provide very little information about the true contact rates as the predictions vary only a small amount around the mean. A leastsquares regression line fit to the points for the marginal logistic regression is nearly flat (indicating that the covariates do not help us differentiate households with high contact rates from those with low contact rates). It is possible to add information to the marginal model (number of calls, number of contacts, etc.) to improve the fit.

Table 2: NSFG Contact Propensity Predictor Variables

Context Variables	Interviewer Observations
% Urban	Multi-Unit Structure
% Black	Physical Impediments to entry
% Hispanic	Residential or Residential/Commercial Area
% with Children	Evidence of Spanish Speaking
% Work (16+)	Access Problems
% Work Evenings	Safety Concerns
% Commute 30+ mins	
% Owner occupied	
Census Region	

Table 3: Call Window Definitions

Window	SCA Definition	NSFG Definition
1	Sat-Sun-Mon 4pm-9pm	Fri-Sat-Sun 4pm-9pm
2	Tues-Fri 5pm-9pm	Mon-Thurs 4pm-9pm
3	Sat-Sun 9am-4pm	Sat-Sun-Mon 9am-4pm
4	Mon 9am-4pm, Tues-Fri 9am-5pm	Tues-Fri 9am-4pm

The multi-level model (the dark circles and dashed line), on the other hand, does provide useful information in differentiating households. Presumably, the call records specific to the household provide most of the information for these estimates. Since most of the cases with observed contact rates of 1.0 were resolved in only one call, the model fit may be better than implied by Figure 1a. In Figure 1b, we see that the predictions for window 1 are quite good using the multilevel model for households that have 5 or more calls. When we have sample sizes as small as 5, the predictions are very good.

Since, in this case, the covariates add very little to the model, it may be that an even simpler approach would work as well. For instance, Bollapragada and Nair (2010) assign average contact rates to each household and then modify those up or down by a specified amount depending upon the outcome of each call. A simple Bayesian update procedure for a binomial probability may also work. A prior could be chosen that would be equivalent to 2 or 3 calls. This prior would then be updated with the incoming data to create new posterior estimates.

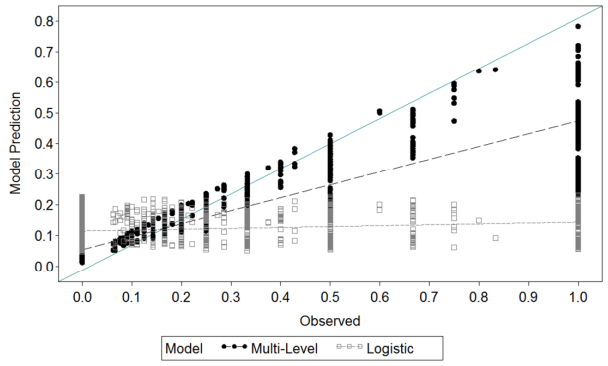
In the NSFG model estimated at the end of quarter 16, cases that were observed by interviewers to be in multi-unit structures had lower estimated probabilities of contact in all four windows. In Windows 1 and 2, sampled units in Census Regions 2, 3, and 4 had higher estimated probabilities of contact than those in Census Region 1. For Window 4 (Tuesday–Friday during the day), households in neighborhoods with higher proportions of the population that commute 30+ minutes had lower contact rates.

The benefit of the multi-level modeling procedure for this situation is that as more data are added for any household-window combination, these covariates will play a less dominant role in the estimates and the actual contact rate for each household-window combination will play a more dominant role. Household-window combinations with fewer calls will borrow strength from other households with similar characteristics (Gelman and Hill 2007). Stronger predictors on the sampling frame would allow more extensive differentiation of households in terms of their contact rates.

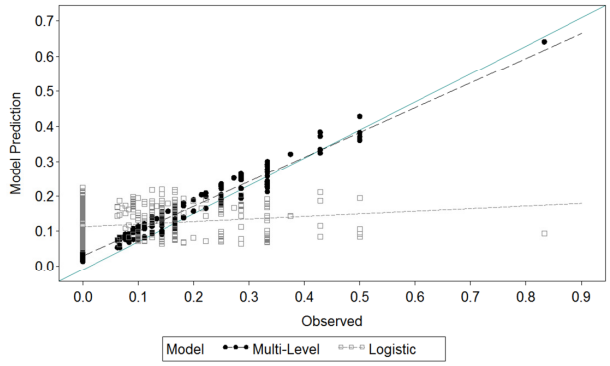
The next step is to compare the estimated contact probabilities within a household and find the window with the highest estimated probability of contact for that household. For the telephone survey, during that window, the case – along with all other cases that meet this criterion – will be sorted to the top of the list by the call scheduling algorithm used in the telephone call center. Each case had a window with the second highest estimated probability of contact. During that window, the case would be sorted on the list after the cases for which that window had the highest probability and so on. In this way, all active cases were available for calling in every window.

Table 4 presents an example with four sampled units. The estimated probabilities are relatively large for cases 2 and 4, and quite small for case 1. For case 1, window 1 (highlighted) has the highest estimated probability of contact. For cases 2 and 4, the highest probability of contact occurs in window 2. At the beginning of window 2, the sample would be sorted such that cases 2 and 4 would be called first. Between cases 2 and 4, one case would be randomly selected for the first call. After cases 2 and 4 had been called, case 3 would be in the second group since it has its second highest probability of contact in window 2. Finally, case 1 would be the last case to be called since it has its lowest probability of contact in window 2. Of course, the number of cases called in any window would be determined by the number of interviewers working and the length of each call they make.

Under this sorting approach, a case with a low estimated probability of contact could be sorted to the top of the list in any given call window – as long as the estimated probability of contact for that window was the highest probability for that case. Cases within the group that were prioritized in the current window were sorted randomly. Future research could address what method of sorting within this group may work



(a) for All Households



(b) for Households with at Least 5 Calls

Figure 1. Window 1 Model Predictions versus Empirical Contact Rates: Random Effects vs Marginal Logistic Regression Model

Table 4: Call Window Definitions

	Contact Probability			
Case	Window 1	Window 2	Window 3	Window 4
1	0.05	0.01	0.03	0.02
2	0.25	0.35	0.20	0.15
3	0.05	0.10	0.15	0.08
4	0.40	0.50	0.30	0.20

best. For example, an interesting research question might be whether sorting lowest to highest estimated probabilities improves the performance of the algorithm.

The call scheduling algorithm was implemented experimentally. On the telephone study, a random half of the sample was assigned to the experimental treatment. The other half of the sample was the control group. The experimental design required that the experimental and control groups be sorted in an interleaving fashion. The past practice had been to sort only at the beginning of the day. The control group sort was based on an algorithm that assigned weights to various factors and then sorted based on the sum of these weights. There were two weighting schemes. The first was used for the first part of the month. It prioritized cases by time zone, those that had fewer than 5 calls, and those that were not called already on that day. Later in the month, the weighting scheme included the following factors: time zones, number of calls, whether the case had already been called that day, whether contact had previously been made with the number, and whether a household listing had been taken.

The experimental design required frequent sorting of the list as the call windows were specific to the time zone. For example, on a Tuesday, the list was sorted before calling began in the morning, at 5pm EST, 6pm EST, 7pm EST, and at 8pm EST as the various time zones included in the study crossed the call window boundaries.

For the face-to-face survey, the window with the current highest estimated probability of contact for each household was delivered to the interviewer as a recommended call time. Interviewers were told about the experiment in a briefing before it began. They were told that the recommendations would be delivered on a sample of their cases. It was explained that they recommendations were designed to aid them in the search for better times to establish contact and were based on the available call record data. They were urged to make use of the recommendations, but also to plan efficient trips.

Interviewers are also provided with a general training on establishing contact before beginning work on the study. They are told which times are generally better for establishing contact (evenings and weekends – "peak" calling hours). They are also given general strategies for households that are more difficult to contact – for example, trying different times of day and days of week. Supervisors monitor interviewers (including proportion of calls made during "peak" times) and intervene when interviewers have unusually low contact rates.

Every case had a recommendation developed. Since this was an experiment, within each sampled second-stage area unit (neighborhood), a randomly selected half of the housing units were assigned the treatment. For these cases, the recommendation was revealed. This experimental design allowed us to evaluate whether interviewer's followed the recommendation at a rate greater than would occur by chance (i.e. the rate at which the interviewer and the model select the same window).

As with the telephone survey, these recommendations were updated daily by an automated procedure. The recommendations were delivered to interviewers over the internet. The recommended window appears in the interviewer's view of the sampled unit in the computerized sample management system (see Figure 2). Interviewers are urged to download this information from the central database and upload their updated call records daily. Since daily updates are important for several features of the responsive design for this survey, compliance is quite high.

4 Results

Experiment 1. The first experiment was applied to English-language calls prior to any refusal. This approach

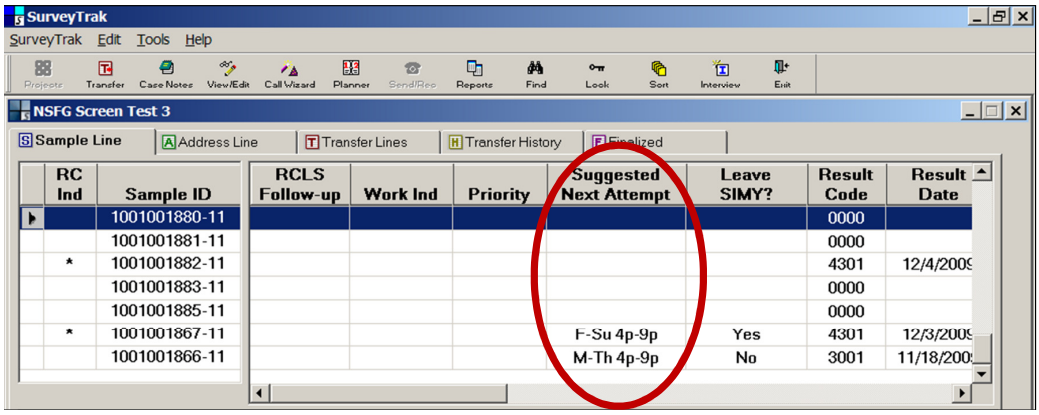


Figure 2. Sample Management System Delivery of Recommended Call Times

was taken largely for technical reasons. The sample management system grouped cases based on key characteristics, including sample for which there had never been a refusal, refusal conversion sample, and Spanish speaking cases. In order to simplify the implementation, the focus of the initial experiment was on the calls prior to any refusal. The results were an improved efficiency of establishing contact for calls governed by the algorithm. This experiment was run for 2 months. The results are presented in Table 5. Contact rates are significantly higher for the experimental group compared to the control group [p=0.008]. Unfortunately, the algorithm had the undesirable effect of reducing contact rates for calls made after any refusal was taken.

Not all of the calls were placed in the window estimated to have the highest probability of contact. If the telephone facility had staff available to make more calls, they would call cases that had the second highest prioritization, third highest, and so on as long as staff were available. In this survey, with a fixed field period, there is pressure to call the entire sample as quickly as possible. In this circumstance, it is possible that all of the active sample may be called on any given day. In general, this occurred more frequently in the second half of the field period. If we only look at calls that were made during the window with the highest estimated probability of contact (33.8% of all calls to the experimental sample), the contact rate is slightly higher – 12.5%.

The results of the experiment seemed to indicate that improvement was possible, but that the approach needed to be extended to those calls not governed by the algorithm in Experiment 1.

Experiment 2. The desired technical changes were made that allowed the inclusion of refusal conversion calls in the prioritization scheme. The same models were used to estimate contact probabilities over all the calls. Essentially, from the modeling standpoint, the distinction between refusal conversion and other types of calls was ignored. This seemed to be a plausible approach under the assumption that the process of contact had very little to do with the process of obtaining cooperation. In other words, whether a call was a refusal conversion or not should not have anything to do with the best time to call in order to establish contact. The sample management system does, however, segregate the two types of calls.

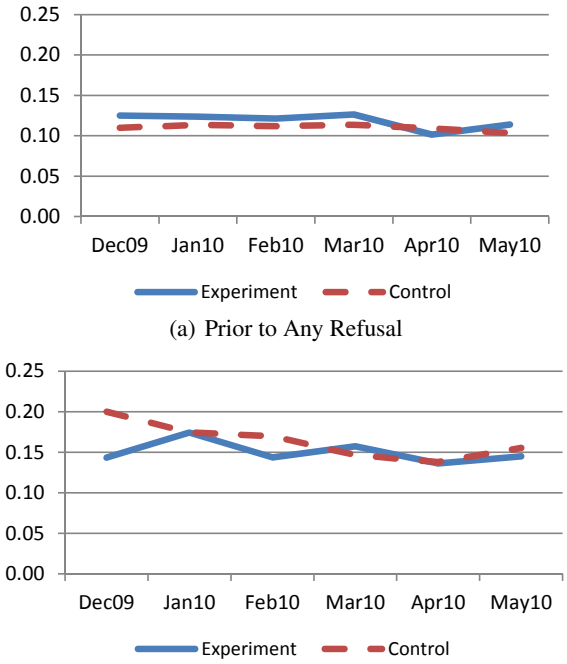
Experiment 2 was allowed to run for 6 months. The results were similar to those of Experiment 1 (see Figures 3a and 3b). Despite the change, the calls placed after any refusal were still less efficient for the experimental treatment group. As a result, the control group averaged 43.6 calls per complete while the experimental group was slightly higher with 46.1 calls per complete. The contact rates for the two groups were a bit closer than in Experiment 1. However, for the experimental group, when cases were called in the window with the highest probability of contact (26.6% of all calls to the experimental group), the contact rate was higher – 13.4%. The opposite was true for the refusals. Cases that were called in the two windows of lowest probability (48.4% of all refusal conversion calls to experimental cases) had the highest rate of contact – 16.3%. This suggests that the model prediction for refusal conversion calls was not very good.

The results of this second experiment made it appear as if establishing contact after a refusal conversion might be different than prior to any refusal. A hypothesized explanation of this result was that caller-ID, which enables the person being called to identify who is calling before answering, was being used to screen our calls by those households that had already refused. The experimental method could increase the chance of calling back at the time of the first refusal (since a refusal is a contact, and successful contacts increase the estimated probability of contact). If caller-ID was used to screen calls, then perhaps the person who gave the first refusal would be likely to be at home and to screen the call. Calling at a different time, on the other hand, might reach another person in the household and result in contact. This hypothesis led to Experiment 3.

Experiment 3. The third experiment changed the algorithm for the refusal conversion calls. The refusal conversion cases had the window in which the first refusal occurred recoded as the window least likely to lead to contact, regardless

Table 5: Results from Experiment 1 (Telephone survey)

	Prior to First Refusal			Post First Refusal			Total		
Group	Contact Rate	Interviews	Calls Per Interview	Contact Rate	Interviews	Calls Per Interview	Contact Rate	Interviews	Calls Per Interview
CON	0.099	180	54.9	0.160	116	30.9	0.122	296	45.5
EXP	0.120	195	47.5	0.156	106	38.6	0.126	301	44.4



(a) Prior to Any Refusal

(b) After Any Refusal

Figure 3. Contact Rates

of the estimated probabilities. All other windows would be prioritized over this window.

Experiment 3 was allowed to run for 4 months. The results were interesting in that the advantage held by the experimental condition for calls prior to any refusal was lost. On the other hand, calls after any refusal were slightly more efficient for the experimental treatment. As a result, the two methods were about equal in overall efficiency. If we only look at calls to the experimental group that were made during the window of highest priority for the calls prior to a first refusal (27.8% of all calls priot to any refusal to experimental cases), then the contact rate for this group is 13.0%. This is higher than the rate for the control group. On the other hand, the calls made during the window of highest priority for the refusal conversion calls (25.2%) had a much lower contact rate of 9.7%. Again, this suggests much poorer prediction for the "modified" model prediction.

There are two hypothesized explanations for these results. First, the design of the calling algorithm does not account for the fact that cases will be called at less than optimal times. Perhaps another algorithm or approach would provide an "optimal" approach to this more highly constrained situation. Such an approach would either account for the current practices of scheduling interviewers or change those practices.

A second hypothesis is that there is an interaction between the treatments offered before any refusal and those that came after the first refusal. The experimental group did well, but through some unidentified mechanism, was creating more difficulty in the later phase. This was addressed in the third experiment by avoiding the time at which the initial refusal was taken. However, another mechanism could be behind this interaction. In other words, it may be the case that another modeling approach that accounts for this interaction may be more productive.

Experiment 4. The protocol was further modified to model the refusal conversion calls separately. The models for the refusal conversions calls were fit separately from the models for calls prior to any refusal and the predictions were based on data from the refusal conversion phase only. An additional predictor available for the refusal conversion calls was whether the first refusal had occurred in the current window. This predictor was quite frequently significant and included in the model. In all other respects, the protocol was the same.

Experiment 4 was allowed to run for 3 months. The results were similar to the results from Experiments 1, 2, and 3. Overall, the contact rate for refusal conversion calls in the experimental group was 14.0% and 16.0% for the control group. In addition, the cases in the experimental group that were called during windows other than the two highest priority windows for each case had the highest contact rate. Again, the model prediction for the refusal conversion phase is quite poor.

At this point, the experiments were concluded. It appeared as if there is the possibility of interactions between the actions taken across the two phases. If so, another modeling strategy is needed that aims at maximizing contact and response across the separate phases. It is also possible that the experiments did not adequately account for the operational constraints. A successful strategy may need to account for the number of interviewers, the distribution of their hours across the days of the field period, and other important features of the manner in which the sample is called.

Experiment 5. The final experiment is based on a survey that does face-to-face interviewing – the NSFG. Interviewers on this survey determine which cases to call and when to call them. They are given a general training about which times

Table 6: Results from Experiment 3

	Prior to First Refusal			Post First Refusal			Total		
Group	Contact Rate	Interviews	Calls Per Interview	Contact Rate	Interviews	Calls Per Interview	Contact Rate	Interviews	Calls Per Interview
CON	0.110	452	48.3	0.141	141	56.6	0.119	593	50.3
EXP	0.106	464	48.6	0.136	140	53.8	0.114	604	49.8

are, on average,better for establishing contact (evenings and weekends). They are also encouraged to call as many sampled units as they can on any visit to the second-stage area unit (neighborhood) in order to be efficient. Their work is reviewed by supervisors. If an interviewer is having trouble contacting a household, a supervisor might recommend a strategy (trying other call windows, leaving a note for the household, etc.). Interviewers report any calls made and the result of those calls into laptop computers that are synchronized with a central database as frequently as every day.

As with the SCA experiments, multi-level logistic models predicting the probability of contact were run daily. The predictions were used to determine which call window had the highest probability of contact for each case. The window with the highest estimated probability of contact was recommended to the interviewer as a good time to call the case. This was a randomized experiment. The sampling was done within sample segments. As a result, approximately half the lines in any segment received the experimental treatment and half received the control. A recommended call window was estimated and stored for each case each day. However, for the experimental group, these recommendations were shown to the interviewer (see Figure 2). For the control group of sampled housing units, the recommendation was not shown to the interviewer.

In order for this experiment to be effective, two things need to be true: 1) the interviewers need to follow the recommendation; and 2) the recommended strategy needs to be more efficient than others. If the first condition is not met, then the second condition cannot be tested. We would conclude, however, that the method is not effective since it is not implemented by the interviewers. This is similar to an intent-to-treat experiment. If the experimental group does not follow the prescribed treatment regimen, then the experimental method is judged to be ineffective.

The results of this experiment were that interviewers did not follow the recommendations. For the control group, for which the interviewers did not see the recommendation, the proportion of calls that coincided with the recommended call time was 23.0%. For the experimental group, the timing of the call occurred at the recommended time 23.6% of the time. Since these rates are approximately the same [p=0.29], we can conclude that viewing the recommendation did not increase the interviewer's chances of following the recommendation. In other words, the interviewers did not follow the recommendations.

In debriefings following the experiment, the interviewers repeatedly mentioned that for area probability samples calls are scheduled in groups. This reflects the clustered nature of the sample. As a result, they felt that recommendations for individual housing units were not useful. Of course, they could have used these recommendations to help determine which set of cases should be called on any given trip, or when to schedule trips. For instance, if there were a group of cases for which the recommendation was to call on a weekday, they could have called those cases on a daytime trip. If time allowed, they could then have called additional cases. This would have allowed them to follow the recommendation while remaining efficient. The evidence shows that they did not do this.

5 Discussion

Improving contact strategies continues to be a difficult problem. The results of the experiments reported here indicate that perhaps more complex solutions are required. Future research should explore all the factors that may impact contact rates in both the telephone and face-to-face setting. The approach taken here for improving contact rates in telephone surveys did not specifically account for the scheduling of interviewers, nor did it explore other parameters of the call scheduling algorithm – such as when to call back busy signals, how long to wait after the first refusal, etc. In the case of face-to-face surveys, the constraints imposed by travel to sample clusters and the low marginal cost of any single call relative to that travel surely need to be considered. For face-to-face surveys, the problem of call scheduling is really a trip planning problem.

It may be that methods from operations research and machine learning are better able to incorporate the constraints faced by either telephone or face-to-face surveys. Applying these methods to the problem of establishing contact could lead to important breakthroughs. For example, the Markov Decision Process model proposed by Greenberg and Stokes allows for constraints to be added to the problem specification. This approach allows for the development of solutions that accommodate the real-world requirements of each situation. Further, it may be possible to account for the variability in these process inputs (scheduling and hours constraints) in defining an optimal solution. For example, it might be possible to design an algorithm for maximizing contact rates under a "worst-case" scenario for staffing.

Further, the method for defining call windows could be improved. One key assumption of the methods proposed here is that the call windows are homogenous. In other words, the assumption is that the estimated contact rate applies equally well to all times within the window. If the assumption is wrong, this would likely lead to inefficiencies in the ap-

proach. Perhaps more sophisticated clustering algorithms could be used to define both the number and boundaries of these call windows.

In the case of telephone surveys, an alternative hypothesis also needs to be investigated. It is possible that the methods used for establishing contact in telephone surveys – embodied in the staffing procedures and call scheduling algorithms currently in use – have evolved over time (in a largely undocumented fashion) to the point where they are already quite efficient. One could imagine a process where managers "tweak" the system at irregular intervals. If contact rates improve after the "tweak", then the change is kept. Otherwise something else is changed. If this is an accurate description of an ongoing evolution in these algorithms, then it may be the case that only marginal gains in efficiency are available. A comparative study of the sample management algorithms and staffing policies of various call centers would help to illuminate this question.

Another area for future research is the potential for interactions between treatments offered at different phases of the survey process. It appears that what is done in one phase may be reducing the effectiveness of treatments in another phase. If this is the case, then the goal of increasing the overall efficiency cannot be met by choosing the method that works best in the first phase and then, separately, choosing the method that works best in the second phase. The treatment strategy needs to account for this interaction and select a strategy – encompassing both phases – that is overall most efficient. An algorithm that improves contact strategies should account for these potential interactions.

We can extend this logic even further. Our ultimate goal is not just to establish contact. Our ultimate goal might be to increase response rates, or even to minimize the risk of nonresponse bias. In either case, maximizing contact rates might be counterproductive in relation to our ultimate goal. If so, we might choose a less efficient contact strategy in order to obtain a higher response rate or a higher value on some statistic that we expect to be related to the risk of nonresponse bias. "Tuning" our models and methods to outcomes such as a minimized risk of nonresponse bias might be more beneficial.

In the world of adaptive treatment regimes (Murphy 2003; Murphy 2005; Lavori and Dawson 2008; Collins et al. 2004), interactions between treatments offered at different times is a well-studied phenomenon. As an example of the adaptive treatment regimes approach, Thall, Millikan, and Sung (2000) consider competing, multi-course treatments for prostate cancer. There are four single-course treatments. If the first treatment fails, then a second treatment from the remaining three is tried. This means there are twelve possible two-course treatments. They note that some treatments can create a 'cross-resistance' with other treatments. Cross-resistance occurs when the treatment used for the first treatment causes the probability of success with the second treatment to be lower than if another treatment had been used as the first treatment. In other words, there is an interaction between the two treatments that reduces the effectiveness of the second treatment. If the best single course treatment creates cross-resistances with the best second treatment, then it may be better to start with a less effective first treatment that has a lower probability of success and retain the best second treatment in case of failure. Wagner (2008) has suggested that these methods may be useful for surveys. In the case of these call scheduling experiments for the telephone survey, there appears to be a cross-resistance created from the first treatment to the second treatment. Future research will be devoted to addressing this problem.

In the face-to-face experiment, the results were that interviewers did not follow recommended call times. In this case, call scheduling is much more complicated due the clustering of the sample. Interviewers could have used the recommended call times as inputs to their process for planning trips. They did not do so. It may be useful to understand how interviewers currently make those decisions. Wagner and Olson (2011) provide a preliminary analysis of interviewer decisions about trip planning and travel. Additional research is needed to understand how these decisions are currently made and what centrally-provided inputs to these decisions could be helpful for increasing contact rates, improving response, or minimizing the risk of nonresponse bias.

As briefly mentioned earlier, it may be the case that errors in the call records in the face-to-face survey are biasing estimates of contact rates. Biemer et al. (2011) provide evidence that interviewers may underreport calling for a variety of reasons. For example, interviewers may drive or walk by households without ringing the door because of visual clues indicating that no one is home (car is gone, lights off, etc.) If these sorts of "driving by" events are not included in our call records, it can bias estimated contact rates. Biemer et al. (2011) show that these biases can be substantial. Further research is needed to determine how extensive these issues may be on other face-to-face surveys and how to correct and repair them.

There are open research questions in the area of call scheduling. For telephone surveys, a formalization and documentation of current practices and knowledge would allow for an assessment of their optimality. In the development of new scheduling algorithms, it may be that considering staffing practices and other operational aspects are an essential component of any optimization. In face-to-face surveys, our understanding of how interviewers currently plan trips to sample clusters needs to be improved in order to allow us to develop information that will be useful for them in making those plans while also improving the survey's ability to meet its objectives.

Acknowledgements

Thanks to Richard Curtin for allowing the experiments on SCA to be conducted. Thanks also to Joe Matuzak, Dave Dybicki, and Rebecca McBee for working on implementing this experimental design. Thanks to James Lepkowski and William Mosher for permission to run the experiment on NSFG and to Nicole Kirgis, Shonda Kruger-Ndiaye, and Brad Goodwin for work on the implementation of the experiment.

References

Biemer, P., Chen, P., & Wong, K. (2011). *Errors in the Recorded Number of Call Attempts and Their Effect on Nonresponse Adjustments Using Callback Models.* Paper presented at the International Statistical Institute Congress, Dublin. Downloaded May 15, 2012 at http://isi2011.congressplanner.eu/pdfs/450164.pdf.

Bollapragada, S., & Nair, S. K. (2010). Improving Right Party Contact Rates at Outbound Call Centers. *Production and Operations Management*, *19*(6), 769-779.

Brick, J. M., Allen, B., Cunningham, P., & Maklan, D. (1996). *Outcomes of a Calling Protocol in a Telephone Survey.* Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria: 142-149.

Campanelli, P., Sturgis, P., & Purdon, S. (1997). *Can you hear me knocking?: investigation into the impact of interviewers on survey response rates.* National Centre for Social Research.

Collins, L. M., Murphy, S. A., & Bierman, K. L. (2004). A conceptual framework for adaptive preventive interventions. *Prevention Science*, *5*(3), 185-196.

Cunningham, P., Martin, D., & Brick, J. M. (2003). *An Experiment in Call Scheduling.* Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria: 59-66.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly*, *69*(1), 87-98.

Dennis, J. M., Saulsberry, C., Battaglia, M. P., Roden, A., Hoaglin, D. C., Frankel, M., et al. (1999). *Analysis of Call Patterns in a Large Random-Digit-Dialing Survey: The National Immunization Survey.* Conference website of the International Conference on Survey Nonresponse 1999: 1-23.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(4), 1029-1049.

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(2), 361-381.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, New York: Cambridge University Press.

Greenberg, B. S., & Stokes, S. L. (1990). Developing an Optimal Call Scheduling Strategy for a Telephone Survey. *Journal of Official Statistics*, *6*(4), 421-435.

Groves, R. M., & Couper, M. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Johnson, T. P., Cho, Y. I. K., Campbell, R. T., & Holbrook, A. L. (2006). Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey. *Public Opinion Quarterly*, *70*(5), 704-719.

Kulka, R. A., & Weeks, M. F. (1988). Toward the development of optimal calling protocols for telephone surveys: a conditional probabilities approach. *Journal of Official Statistics*, *4*(4), 319-332.

Laurie, H., Smith, R., & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, *15*, 269-282.

Lavori, P. W., & Dawson, R. (2008). Adaptive Treatment Strategies in Chronic Disease. *Annual Review of Medicine*, *59*(1).

Lipps, O. (2012). A Note on Improving Contact Times in Panel Surveys. *Field Methods*, *24*(1), 95-111.

Massey, J. T., Wolter, C., Wan, S. C., & Liu, K. (1996). *Optimum calling patterns for random digit dialed telephone surveys.* Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria: 485-490.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 331-355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, *24*(10), 1455-1481.

Pickery, J., & Loosveldt, G. (2002). A Multilevel Multinomial Analysis of Interviewer Effects on Various Components of Unit Nonresponse. *Quality and Quantity*, *36*(4), 427-437.

Purdon, S., Campanelli, P., & Sturgis, P. (1999). Interviewers Calling Strategies on Face-to-Face Interview Surveys. *Journal of Official Statistics*, *15*(2), 199-216.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The Value of Purchase History Data in Target Marketing. *Marketing Science*, *15*(4), 321-340.

Stoop, I. A. L., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: lessons learned from the European Social Survey*. Chichester, West Sussex, U.K.; Hoboken, N.J.: Wiley.

Thall, P. F., Millikan, R. E., & Sung, H. G. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, *19*(8), 1011-1028.

Wagner, J., & Olson, K. (2011). *Where Do Interviewers Go When They Do What They Do? An Analysis of Interviewer Travel in Two Field Surveys.* Survey Research Methods Section, Joint Statistical Meetings, Miami.

Wagner, J. R. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias.* Program in Survey Methodology. Ann Arbor, University of Michigan. Doctoral Dissertation.

Weeks, M. F., Jones, B. L., Folsom (Jr.), R. E., & Benrud, C. H. (1980). Optimal Times to Contact Sample Households. *Public Opinion Quarterly*, *44*(1), 101-114.

Weeks, M. F., Kulka, R. A., & Pierson, S. A. (1987). Optimal Call Scheduling for a Telephone Survey. *Public Opinion Quarterly*, *51*(4), 540-549.