Adaptive Survey Design to Reduce Nonresponse Bias

by

James R. Wagner

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in the University of Michigan
2008

Doctoral Committee:

Professor Trivellore E. Raghunathan, Chair
Professor Robert M. Groves
Professor Susan A. Murphy
Research Professor Mick P. Couper

To Angela and Caden

Acknowledgements

I would like to thank my committee for their time and effort. In addition, I would like to thank the following persons for their permission to use data: Richard T. Curtin (SCA), Robert Groves and William Axinn (NSFG), and David Weir and Mary Beth Ofstedal (HRS).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1. Introduction

Nonresponse bias of the unadjusted respondent mean is a function of two elements – the response rate and the differences between responders and nonresponders. However, the response rate continues to be the most widely used measure of quality (Biemer and Lyberg, 2003). The response rate is just one component of this bias. On its own, it cannot really tell us much about the potential nonresponse bias. At best, it can provide a boundary for the maximum nonresponse bias. Considering the standard expression for estimated nonresponse bias in the estimated population mean $\frac{m}{n}(\bar{y}_R - \bar{y}_{NR})$, where *m/n* is the nonresponse rate, it is possible that we may increase the response rate and at the same time increase the difference between the respondents and nonrespondents thereby worsening the nonresponse bias. This would occur when the distribution of converted nonrespondents on the *y* variable changes over different response rates. The risk inherent in this situation is that surveys are working to produce higher response rates while, in fact, not addressing nonresponse bias at all.

Thus, the field needs a better indicator for nonresponse bias which will allow us to change current practices. Unfortunately, we rarely have a direct measure of the bias.

However, there are indirect measures available. This dissertation proposes an alternative measure – the fraction of missing information, developed for assessing the impact of missing data – and assesses its potential impact on current practices. This measure is a model-based assessment of the nonresponse bias and our uncertainty about that assessment. I then propose methods for identifying protocols tailored to the characteristics of the sampled case. There are three goals for this dissertation:

1. Establish the need for alternative measures of nonresponse bias and motivate the fraction of missing information as a useful indicator.
2. Identify the need for tailoring survey methods in order to reduce nonresponse bias and suggest statistical approaches for identifying efficient tailored methods.
3. Suggest methods for integrating measures of nonresponse bias with tailored protocols in order to develop stopping rules for surveys.

The "rules of the game" for surveys have created an environment which effectively stifles research aimed at reducing nonresponse bias through data collection. The response rate is a key indicator for the risk of nonresponse bias. The use of the indicator has created a climate that overemphasizes the importance of the response indicator in analyses of nonresponse bias. Another indicator which involves more of the available data might lead to a climate in which concerns over who is responding could also be emphasized. In such a climate, survey designs that are tailored to the characteristics of the respondent would be favored.

2. Chapter 2: Fraction of Missing Information

2.1 Need for Alternative Measures of Nonresponse Bias

Despite the fact that the response rate is not a direct measure of nonresponse bias, it has become the standard measure of quality (Biemer and Lyberg, 2003). Babbie, in a widely used textbook on social research, suggests that "a review of the published social research literature suggests that a response rate of at least 50 percent is considered adequate for analysis and reporting. A response rate of 60 percent is good; a response rate of 70 percent is very good" (2007, p. 262). In addition, the Office of Management and Budget (OMB) sets standards for statistical agencies and data collections sponsored by the Federal government. They set as a standard that organizations "plan for a nonresponse bias analysis if the expected unit response rate is below 80 percent" (OMB, 2006, p.8). Other statistical agencies set their own policy in this regard. For instance, The National Center for Educational Statistics (NCES) specifies in its statistical standards that "any survey stage of data collection with a unit or item response rate less than 85 percent must be evaluated for the potential magnitude of nonresponse bias before the data or any analysis using the data may be released" (NCES, 2002, p. 88). Beyond Federal government-sponsored surveys, it is common practice for groups that contract the services of survey organizations to specify target response rates in addition to target sample sizes.

However, nonresponse bias is a more important outcome than merely nonresponse rates. In a reformulation of the expression above, this bias is a function of the correlations between the propensity to respond and the survey variable. The "stochastic" formulation of nonresponse (Bethlehem, 2002) bias shows this most clearly:

$$B(\bar{y}_R) \approx \frac{\sigma_{y\rho}}{\bar{\rho}},$$

where $\bar{y}_R$ is the unadjusted estimate of a mean from the responding units, $\sigma_{\rho y}$ is the correlation between the response propensities and the survey variable, and $\bar{\rho}$ is the average response propensity. It should be clear from this formulation that a higher response rate does not imply a reduction in bias as the correlation term, $\sigma_{\rho y}$, can change in a nonlinear fashion as response rates change. If, for example, a higher response rate also had higher correlations between the response propensities and the *y* variable, then the bias will be higher. Empirically, there have been special studies of nonresponse bias where a reduction in response rate does not have much impact on the bias (Curtin and Singer, 2000; Keeter et al., 2000; Merkle and Edelman, 2002; Groves, 2006; Keeter, et al., 2006). There have also been examples where an increased response rate leads to greater bias (Merkle and Edelman, 1998).

2.2 Fraction of Missing Information

The response rate, as an indicator of nonresponse bias, implies a particular model of the nonresponse bias. This simple model suggests that as the response rate increases, the nonresponse bias decreases. If this model is wrong and this measure is not producing the result for which we had hoped, is there an alternative measure that allows us to do so? The answer is no, unless we utilize a model of the relation between known characteristics of all sampled units and the survey variables that allows us to "fill-in" the nonresponding values. If we assume that the model was correctly specified, then we would have a more direct assessment of the quantity of interest.

There are many methods for dealing with missing data (see Little and Rubin, 2002). Imputation is one such method. Any method for dealing with missing data relies on the assumption that conditional on the model the missingness does not depend on the survey variable. In other words, the method assumes that the missing data are Missing at Random (MAR). Under the specified model, the estimate from the fully-imputed dataset is the estimate of the population mean while the complete case estimate is the estimate of the respondent mean. The difference is the estimate of the nonresponse bias, given the imputation model. In addition to the variability of the survey variable, there is an additional uncertainty introduced into estimates developed using imputation due to the variance associated with the model used to impute the values. The fraction of missing information captures the ratio of the latter variance to the total variance. Conditional on the model, it is a measure of the precision with which we are able to "fill-in" the missing data.

2.3 Chapter 2 Overview

In chapter 2, I will discuss problems with the response rate as an indicator for nonresponse bias. I will define the fraction of missing information and describe why it is a useful alternative to the response rate. I will then provide two applications where I have implemented this indicator with real surveys.

3. Chapter 3: Adaptive Survey Design

3.1 Survey Methods Largely Not Tailored

The reliance on the response rate as a key metric has created a climate to which survey organizations have adapted themselves. Response rate targets have become a key

measure of success. Therefore, maximizing the response rate is the goal of these organizations. This is not necessarily the same thing as minimizing the nonresponse bias. Since nonresponse bias is also a function of the differences between responders and nonresponders, the question of which units respond is also important. However, in the pursuit of the highest response rate, survey organizations have largely ignored this question and are instead looking for the next easiest case to interview. These organizations have become expert at finding the portions of the sample that have the highest probability of responding. From the logic of the situation, this makes sense. But is this the outcome we really want?

I will suggest that this imbalance has led survey organizations to focus on methods that have the largest impact on overall response rates. These methods may increase response rates for the majority – or even just many of the sampled cases. However, there may still be significant subgroups for which these methods are less than optimal. Many experiments have been done that compare two methods – for example, a five dollar incentive compared to a ten dollar incentive. The outcome of the experiment is a difference in response rates between the two conditions. There is very little understanding of who is responding to each level of the incentive. Of course, there are exceptions. Groves et al. (2004), for example, consider for whom an incentive is more important and for whom the saliency of the topic is more important. Roose et al. (2007) find that a follow-up procedure is more effective for those who are less interested in the survey topic. However, in general, much of current methodological research is designed as if it does not matter who responds as long as the largest proportion possible responds.

This raises the question, do current methods simply bring in more of the same kind of respondents? Curtin, Presser, and Singer (2005) show that late and early respondents to the Survey of Consumer Attitudes are not significantly different on the key measures collected by this survey. Heerwegh, Abts, and Loosveldt (2007) consider this question using data from the Flemish Housing study. This survey had the unusual condition where the survey variable (type of housing) was also on the frame, so that nonresponse biases were known. They show that after about 30% of the final response had been achieved, the estimates of a key survey statistic had achieved the same level of bias that was achieved when 100% of the interviews had been completed. In their words, "the nonresponse error and its components remain relatively stable throughout the entire fieldwork period. This indicates that simply collecting more data (increasing the sample size) does not decrease the nonresponse error" (p. 9).

Of course, these are only case studies. It is always possible that there are only "more of the same" types of respondents to collect. In which case, low response rates and high response rates should produce similar nonresponse biases. On the other hand, if there are differences between those who have responded (or are likely to respond) and those who are less likely to respond, and our methods bring in only those who are likely to respond, then increasing the response rate by interviewing mainly those who are more likely to respond will only perpetuate or – if the response rate differences between the two groups become greater -- even exacerbate the bias. Merkle and Edelman (1998) provide an example from an incentive study done for exit polling where an incentive increased response rates among Democrats but not Republicans. The groups had relatively equal response rates before the incentive. The result of the incentive was to

create differential response rates that led to increased bias. In this type of situation, researchers may want to be aware of who is being brought in and what methods are bringing them into the survey, and conversely who is not being brought in and what method (obviously not yet used) would lead them to respond.

3.2 Need for Tailored Methods

When achieving the highest possible response rate is the goal, going after the easiest to interview cases is a natural approach. Would we go after the easiest cases if we were instead monitoring the estimated bias and our uncertainty about this estimate under an assumed model? We could begin to approach it as an optimal allocation problem. Or, we could incorporate our uncertainty about the values we would impute for the missing cases, and attempt to interview the cases with the most variation in their imputed values. Certainly, data collection would be opened up to the possibility of attempting to interview cases other than the "lowest hanging fruit."

In the Flemish Housing survey example cited above, if refusals were cheaper to be interviewed than those who were very difficult to contact, current strategies would go after the refusal conversions. This approach would have the largest impact on the response rate, but it would not reduce the bias of the unweighted estimates since the bias was associated with the noncontacts. Another approach might allow us to pursue the noncontacts at the expense of doing refusal conversions.

The question is, what methods will we need if we allow the possibility of not going after the easiest case at each step? In these examples, our original design is no longer working. The next step should be something different. What is that next step? More of the same, or a design tailored to the case?

Under an approach governed by a model-based metric, the goal would be to reduce our uncertainty about the nonresponse bias. The survey would be monitoring both the model estimate of the bias and our uncertainty about that estimate. These measures of data quality might lead to the development of many more methods. The new methods would be linked not to the ability to increase the response rate, but to the ability to bring in specific respondents. For example, in general, incentives increase response rates. However, they may do so by increasing response probabilities for many persons while decreasing them for some. Under a response rate metric, the use of incentives is generally recommended. However, if we are interested in bringing in specific persons then we would want to identify the methods most successful at bringing those persons into the survey.

In some cases, this might mean always interviewing the next easiest case; for instance, when the model predicts that current responders and nonresponders are very similar. In other cases, this might mean something other than interviewing the easiest case; for instance, when the model suggests that an important subset of nonresponders have values different from estimates of key survey statistics based on current responders. If you want specific cases, how do you interview them? The current strategy might be seen as an iterative approach where we try a protocol on all cases, and then switch to something else if the previous method does not succeed. In some cases, this might be the best approach. However, it seems likely that a better approach would match the protocol to the person. It might be that initial strategy, which failed, reduces the probability that someone will respond relative to another sequence of protocols. At the very least, the strategy that learns what is effective most quickly would be preferred.

3.3 Chapter 3 Overview

In Chapter 3, I will provide justification for tailored methods of data collection. I will point to methods developed in the field of clinical trials – specifically, research into dynamic treatment regimes – as useful tools for the development of new survey methods. I will then demonstrate two applications of this approach with data from real surveys. I will also suggest how "learning" algorithms might be used to increase the efficiency of data collection. Finally, I will recommend a "computerized decision support system" approach for the implementation of adaptive protocols to field surveys.

4. Chapter 4: Stopping Rules for Surveys

4.1 Need for Stopping Rules

If the response rate is no longer the target, then the question is when should we stop collecting data? Traditional sample size calculations start from the sample size needed to estimate a quantity with a specified precision. These sample sizes are then adjusted to account for nonresponse. This approach usually does not account for the uncertainty created when we use weighting or other adjustments to account for known differences between responders and nonresponders. Nor does this approach address the risk of nonresponse bias.

Another approach would be to use the imputation uncertainty as a metric for stopping rules. When a prespecified level of uncertainty about values imputed for nonresponders has been met that is sufficient to meet the analytical purpose of the survey, then data collection can stop. Going beyond that point will be an inefficient use of resources.

4.2 Chapter 4 Overview

In Chapter 4, I will propose that new methods for dynamically determining when to stop data collection are needed. I will present a derivation for a stopping rule that attempts to account for uncertainty due to nonresponse. I will demonstrate this rule using simulations. I will then apply the method to real survey data.

5. Conclusion

Survey methods have become adapted to the "rules of the game." Those rules have been molded by an environment where the response rate is a key indicator for the risk of nonresponse bias. The advantage of this indicator is that it can be calculated without assumptions about missing data. The limitation of this indicator is that it is only a component of the nonresponse bias. The empirical demonstration of cases where it is in fact a poor indicator has further called this measure into question.

New indicators are needed. A useful indicator will allow us to assess and reduce the risk of nonresponse bias. The fraction of missing information may fill this gap. It has the disadvantage of involving a model. In order for this measure to be accepted, model-based indicators will have to be accepted. It has the advantage of using all of the available data. These data will likely be used in adjustment strategies. By examining the relationship among the complete data on the frame and the incomplete survey data during data collection, we may alter the priorities of that data collection.

In order to implement these new priorities, new methods will be required. If we can evaluate the information that each case might contribute, then we could prioritize particular cases. An effective way to implement these priorities will be to tailor our survey design to the characteristics of this case. "Tailoring" can also include the decision

about when to stop collecting data. Current “stopping rules” are based on the response rate. A more effective stopping rule would also consider the characteristics of nonresponders and their potential for changing current estimates.

Chapter 2

The Fraction of Missing Information as a Monitoring Tool for Survey Data Quality

1. Introduction

Quality survey data collected on each member of a representative sample from a well-defined population are essential for empirical social science research. However, response rates have been steadily declining (de Leeuw and de Heer, 2002; Atrostic et al., 2001; Petroni et al., 2004; Curtin, Presser, and Singer, 2005), calling into question the validity of inferences drawn from these data. Survey methodologists have responded by seeking new ways to reduce nonresponse. There have been many experiments, for example, on the impact of incentives on response rates (see Singer et al., 1999, for a review). The outcome of the experiment is expressed as a difference in response rates between the two conditions. Many other design features of surveys – including prenotification letters, call scheduling, and wording of introductions (Link and Mokdad, 2005; de Leeuw et al., 2005; Greenberg and Stokes, 1990; Weeks et al., 1980; Weeks et al., 1987; Kulka and Weeks, 1988; Houtkoup-Steenstra and van den Bergh, 2000) – have been tested experimentally as means to improve response rates. These studies rarely investigate the impact of the changes in the survey protocols on the quality of survey estimates.

In this sense, the focus of survey methodology has been on reducing nonresponse rates – as opposed to minimizing nonresponse bias. Theoretically, it has always been known that nonresponse bias is the product not only of nonresponse rates, but also differences between responders and nonresponders. It is understandable why surveys have focused on response rates as an indicator of data quality. First, the response rate is a feature of a survey while nonresponse bias is specific to a statistic. Second, it is nearly always the case that the survey values for nonresponders are unknown.

This focus, however, has led to a set of decision rules that may be harmful to the data that are collected and, hence, the estimates that are produced. When the response rate is the key metric by which a survey is judged, then the data collection strategy aims to collect data on the easiest cases to interview among the current set of nonresponders. This rule has the potential to drive survey organizations to collect more data from persons like those they have already interviewed. A more rational strategy would attempt to track the risk of nonresponse bias for the data being collected. If such a metric existed, it would lead survey organizations to search for data collection strategies that produce data with the lowest risk of nonresponse bias – as opposed to the highest response rate.

Unfortunately, since we do not know how nonresponders would respond if interviewed, we must employ statistical models in order to say anything about the quality of the data that we do have compared to the quality of the data that we could have (with "quality" referring to the nonresponse bias property). Under an assumed statistical model relating the data on our frame to the survey variables, we could track the quality of the data that we are collecting. In addition, we could identify nonrespondents who would add the most information to the current set of respondents.

In this chapter, I argue that alternatives to the response rate as a measure of data quality are needed. I will review some alternative measures. I will then propose a new measure drawn from research on methods for dealing with missing data — the fraction of missing information — and demonstrate applications of this measure in real survey settings.

1.1 Response Rate as a Measure of Data Quality

Response rates have become a near universal measure of data quality (Biemer and Lyberg, 2003). Their ubiquitous nature can be seen in the fact that contracts for conducting surveys often specify target response rates. In addition, while leading journals no longer specify response rate requirements for publishing papers that report survey data, they do often have unstated response rate targets below which they are unlikely to consider articles for publication (Johnson and Owens, 2003).

However, there is not necessarily a direct link between nonresponse rates and nonresponse bias. This should be readily apparent since the nonresponse rate is the property of a survey, and yet we know that nonresponse bias is the property of a statistic. Assuming that the population consists of two strata – responders and nonresponders — there are two components of the bias of the mean: the nonresponse rate and the difference in the population means between responders and nonresponders:

$$B(\bar{Y}_r) = (\frac{M}{N})(\bar{Y}_r - \bar{Y}_m),$$

where $\bar{Y}_r$ is the mean of the respondent population, $\bar{Y}_m$ is the mean of the nonrespondents population, M is the number of nonrespondents in the population, N is the population size, and $B(\bar{Y}_R)$ is the bias of the respondent mean.

Even in this simple formulation, it should be clear that a lower nonresponse rate does not necessarily lead to lower nonresponse bias. If the second component (the difference in means) is greater when the first component (the response rate) is larger, then nonresponse bias could be higher with a higher response rate.

Several recent empirical studies have shown instances where lower response rates do not lead to increased nonresponse bias (Keeter et al., 2000; Curtin et al., 2000; Merkle and Edelman, 2002; Keeter et al., 2006). In these cases, for the range of response rates considered, there does not seem to be a correlation between the overall response rate and nonresponse bias. Groves (2006) review of studies of nonresponse bias shows that there is very little correlation between the response rate and the nonresponse bias. Of course, this may not be true for every survey and every survey statistic, but when this is true the response rate is a poor indicator of data quality. In addition, if subgroup response rates are not correlated with the survey variables, then nonresponse adjustments based on these subgroups may not lead to much bias reduction, but they will inflate variance estimates (Little and Vartivarian, 2005).

In the worst case, it is possible that lower nonresponse rates might lead to higher nonresponse bias. A well-known example is provided by Merkle, Edelman, Dykeman, and Brogan (1998) from the exit poll setting. Without an incentive, Democrats and Republicans responded at similar rates. They provided a pen as an incentive in one experimental treatment group. The incentive, however, worked more effectively among Democratic Party voters than among Republicans. The impact of the incentive, in this case, was to increase the bias. Again, in this case, the response rate is a poor or even misleading indicator of data quality.

The obvious problem is that while the nonresponse rate is known, the difference between responders and nonresponders on a statistic of interest is not usually known. However, to the extent that nonresponse rates are not a good indicator for the nonresponse bias, actions or post-survey adjustments based on the response rate will be either inefficient, biasing, or both. Something is needed to fill this gap between nonresponse rates – which are known – and nonresponse biases, which are unknown but are the thing about which we care.

1.2 Alternatives to the Response Rate

The growing recognition that response rates are an incomplete diagnostic when it comes to nonresponse bias is spurring research into alternatives to the response rate. Groves et al. (2008) consider a range of alternatives. They consider two types of indicators: those that are calculated at the survey level and those that are estimated at the statistic level. Those estimated at the survey level include variance functions of nonresponse weights, poststratification weights, response rates of subgroups, goodness of fit statistics on propensity models, and R-Indexes. These indicators, since they are calculated at the survey level, depend on the characteristics on the frame or from control totals (in the case of poststratification weights) and not on the actual survey data. As such, there is an implicit model relating these variables to the survey data. For example, the variables used to define the nonresponse weighting cells are assumed to be predictive of the survey variables.

One of these alternatives, R-Indexes, has been proposed by van der Grijn, Schouten, and Cobben (2006, also Cobben and Schouten, 2007; and Schouten and Cobben, 2007). These R-Indexes are meant to measure "the similarity between the

response to a survey and the sample or the population under investigation" (van der Grijn, et al., p.1). They propose four indices. These indices measure the variability of subgroup response propensities. A survey with less variability in these response propensities has a better match between the characteristics of the respondents and the population they are meant to represent along the dimensions of the variables used in the model to estimate the propensities. The authors also suggest that these indices can be monitored during data collection in order to direct effort to cases with lower response propensities, thereby reducing the variability among subgroup response rates.

The strength of the approach is that these indices can be calculated in the presence of partially missing survey data. These response propensities are calculated with complete information available on the frame. In order for the R-Indices to be comparable across surveys, they need to be estimated with the same variables. This would suggest that all surveys should have a common set of frame data. In some European countries that sample from registries of the population, this approach may be feasible. However, in the US, where much less is known about sampled units (particularly in RDD surveys), setting such a standard may be difficult. In addition, if these frame variables became involved in the sample design – for example, through oversampling an important subgroup – how would this impact the estimation of R-Indices? Nevertheless, the R-Indices would encourage survey organizations to build better (and more uniform) sets of frame data.

This particular strength of the approach is also its weakness. The indices do not involve any information from the partially missing survey data. Although these data are incomplete, it is still a loss of information to ignore them. Since the R-Indices measure the "representativeness" of the sample with respect to variables on the frame, the user is

left to presume that these variables have some, unspecified relationship to the survey variables. If they did not, then it would not make sense to use them as indicators of potential nonresponse bias. They may predict nonresponse over many surveys, but if they do not predict the survey variables, then these frame variables will not be useful for adjustment strategies. In fact, they may only inflate variance estimates if used in statistical adjustment (Little and Vartivarian, 2005). Models relating frame information to the survey variables are more direct in giving us what we want and make explicit the underlying assumptions.

The authors also note that since these indices are based on the variances of the response propensities, they are affected by the size of the sample. Since these variances are related to the sample size, the indices will also be a function of sample sizes.

Rancourt (2002) discusses the use of variance components for monitoring the quality of the data. He analytically decomposes the variance into components due to sampling error, nonresponse, edits and imputation. He suggests that the data collection funds ought to be directed at reducing the largest component of the variance. For example, if the sampling variance is the largest component of the variance, then additional cases should be added. If the editing and imputation variance is the largest component, then the procedures used to develop these edits and imputations should be modified.

This approach assumes that the variance is the only component of the error. While it certainly is useful to strategically reduce the variance, ignoring potential bias is a shortcoming of this approach. The recommendation for a problem with edit and imputation variance is to improve the data collection procedures or reduce the amount of

editing. It may be that this uncertainty reflects a potential nonresponse bias. If the nonresponse were ignorable, but a full range of data had not been collected, then a more satisfying approach might be to collect data from a subset of nonresponders to improve the range in the data.

A second set of indicators suggested by Groves et al. (2008) includes those estimated at the level of individual survey estimates. The correlation between survey nonresponse weights and the survey variable is one such measure. This correlation estimates the bias prior to adjustment. If there is a high correlation, then the weights are more likely to change the estimate relative to one estimated without the nonresponse weighting factor. There is an implicit model underlying the weighting adjustment. Another closely related measure they suggest is the variation in survey means across deciles of the survey weights.

In the preceding paragraphs, I have reviewed other potential measures of the quality of survey data. Each of these measures has strengths and weaknesses. None of the indicators at the level of the survey involve the survey data in their estimates. Even though we may have data on a large proportion of the cases, these data are ignored in favor of complete data on the frame and the response indicator. Involving the survey data in the estimate of a quality indicator is a difficult task because in order to evaluate missing data, we need to assume some sort of model. These sorts of models are used in nearly every survey for statistical adjustment after the data has been collected. Methods involving the correlation of the weights and the survey variable have an implicit model. I propose to bring these models into the data collection process in order to improve the quality of the data that feeds into these adjustment models.

1.3 The Fraction of Missing Information

One measure of the uncertainty about unknown or missing values has been developed in the framework of missing data analysis and multiple imputation (Rubin, 1987; Dempster, Laird, and Rubin, 1977). Creating multiple imputations under an assumed model allows us to assess our certainty about those imputed values. A higher quality dataset would have less uncertainty about the imputed values. In other words, the quality of the data we do have improves to the extent that it allows us to fill in the data we do not have. In addition, the use of imputations allows us to assess the potential nonresponse bias of an estimate as the difference between the estimate based on the fully-imputed dataset and the complete cases only. Thus, the approach is also informative of the potential impact of nonresponse on survey estimates.

1.4 Chapter Preview

In the rest of this chapter, I will define the fraction of missing information and discuss how it could be used as a measure of data quality. I will then present two applications of this measure to actual survey data.

2. The Fraction of Missing Information as a Measure of Data Quality

The concept of the information in data dates back to Fisher (Fisher, 1925). In Fisher's definition, the information can be written in the following manner:

$$I(\theta|X) = E\left\{\left[\frac{\delta}{\delta\theta} l(\theta|X)\right]^2 |\theta\right\},$$

where X is a data value assumed to be generated by a function $f(X|\theta)$, where θ is a vector of parameters that index the function. The likelihood ($L(\theta|X)$) is proportional to

this function and $1(\theta|X)$ is the natural logarithm of this likelihood. The inverse of $I(\theta|X)$ is a measure of precision of the estimates. In maximum likelihood estimation, the inverse of the observed information is often used as a variance estimate. The problem is that this variance is understated in the presence of missing observations of *X*.

In the early 1970s, Orchard and Woodbury (1971) formalized the problem with a "missing information principle" that argued for a method of incorporating uncertainty due to missing values in variance estimates. Dempster, Laird and Rubin (1977) then developed the fraction of missing information as a tool for predicting the speed with which the Expectation-Maximization (EM) algorithm will converge. It was a natural step from the EM algorithm to imputation. Rubin (1987) suggested the fraction of missing information could be used to judge the efficiency of multiple imputation. He also noted that the fraction of missing information is "equal to the expected fraction of observations missing in the simple case of scalar Y_i with no covariates, and commonly is less than the fraction of observations missing when there are covariates that predict Y_i" (Rubin, 1987, p. 114).

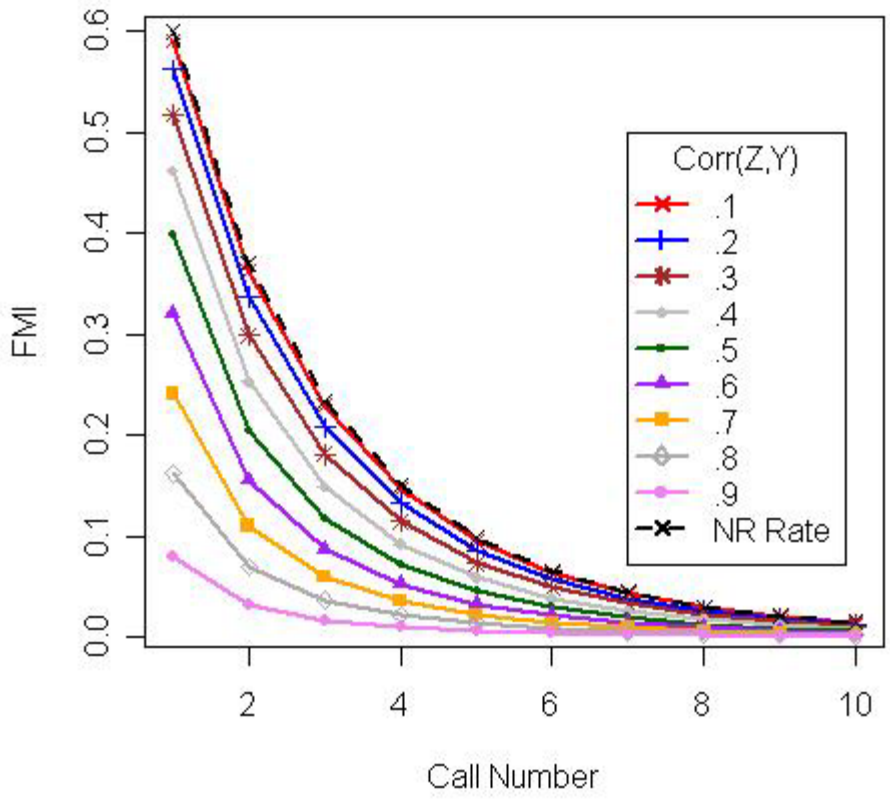
Rubin (1987) shows that the uncertainty about imputed values can be characterized by two components — between-imputation variance and average within-imputation variance. The latter is the variance component due to sampling error. It assumes that the missing values are known. The former is the variance component that results from creating an approximate distribution of imputed values. Little and Rubin (2002) denote these components as follows. With *D* multiple imputations, the estimate of the parameter θ is: $\overline{\theta}_D = \sum_{d=1}^{D} \hat{\theta}_d / D$. The within-imputation variance is the average of *D*

variances (W_d) computed from the D multiply-imputed datasets, $\overline{W}_D = \sum_{d=1}^{D} W_d / D$ The between-imputation component is $B_D = \sum_{d=1}^{D} (\hat{\theta}_d - \overline{\theta}_D)^2 / (D-1)$. These components are combined, incorporating an adjustment for finite D, to form the total variability associated with $\overline{\theta}_D$, $T_D = \overline{W}_D + (D+1)D^{-1}B_D$. The between-imputation variance component is a measure of the uncertainty about the value we would impute for a case. The quantity used to assess the impact of nonresponse is the fraction of missing information $\hat{\gamma}_D = (1+1/D)B_D / T_D$, defined as the ratio of the between-imputation variability to the total variance of the survey estimates.

If we conceptualize the sample as a data matrix, with current nonresponders as missing values, then monitoring between-imputation variance seems a natural choice for monitoring our uncertainty about nonresponse. Others have suggested using this measure as a diagnostic tool for latent class analysis (Harel, 2007). In the context of survey nonresponse, the fraction of missing information is a measure of the proportion of variance due to uncertainty about the values we have imputed for the current nonresponders. The imputations can be done by conditioning on the complete data on the sampling frame and the paradata (Couper, 1998; Couper, 2000) developed from the data collection process – for example, call records, interviewer observations, and key-stroke data generated by Computer-Assisted Personal Interviewing (CAPI) software. To the extent that these data are correlated with the survey variables, the fraction of missing information will be reduced below the nonresponse rate. In this case, we will have a better sense of the quality of our data after any nonresponse adjustments based on the frame data have been made.

Using this metric can be thought of as conditioning the response rate on the quality of frame data and paradata. For example, in the absurd case, if the survey variable was on the frame, a 0% response rate provides the same data quality as a 100% response rate. On the other hand, if there are no variables on the frame that correlate with the survey variable, then the response rate is the best information about the quality of the data (Rubin, 1987, p. 114). It is likely that most surveys fall somewhere on the continuum between these two extremes. Figure 2-1 shows simulation results meant to demonstrate the impact of the level of correlation between the frame data and the survey variable on the fraction of missing information. In the simulation, 200 normally distributed random draws were created to simulate the survey variable (y). Each random draw had a propensity to respond drawn from a beta distribution with a mean of 0.4. Using these probabilities, missing data on the survey variable (y) were created over a set of simulated "calls" by comparing a random draw from a uniform distribution to each case's propensity to respond. A frame variable (z) was generated with different correlations with y ranging from 0.1 to 0.9. This variable (z) was complete. I then did 1,000 simulations with this setup and calculated the nonresponse rate and fraction of missing information after each "call."

Figure 2-1: FMI by Call Number



While the nonresponse rate is invariant to the correlation between *z* and *y*, the fraction of missing information is not. With a very strong correlation (0.9) between *z* and *y*, much of the lost information on *y* has been recovered due to the correlation between *z* and *y*.

In addition, creating imputations allows us to assess the potential bias of our estimates of a parameter θ by comparing the complete case estimate $\hat{\theta}_{CC}$ (i.e. an estimate based on responding units only) to the estimate derived from the multiply imputed *D* datasets: $\overline{\theta}_D = \sum_{d=1}^{D} \hat{\theta}_d / D$. The difference $\overline{\theta}_D - \hat{\theta}_{CC}$ is the nonresponse bias – under the assumption that our imputation model is correct.

An obvious weakness of this approach is that it is model dependent. If the model is incorrectly specified, then the estimate of the fraction of missing information can be

biased. This could be a problem if the data collection agency is basing its actions on the estimate of the fraction of missing information. Even worse, a bad model can lead to a biased estimate. The fully imputed mean can be wrong. To the extent that the models used to develop the imputations are used in the adjustment strategy employed after data collection has been completed, these biased estimates can be propagated into the final results.

Given the importance of model choice, model diagnostics are a crucial step in preparing estimates of the fraction of missing information. There are multiple methods for developing imputations – regression, predictive mean matching, hot deck, and others (Little and Rubin, 2002). The model diagnostics appropriate for regression model building are also appropriate in the context of imputation methods based on regression. More general approaches compare the distribution of imputed values to the distribution of observed values. Graphical displays often can be used to determine if the regression methods are producing implausible or highly skewed values relative to the observed data.

Monitoring the fraction of missing information over the course of a survey data collection field period creates a unique environment for model checking. In this environment, the set of cases with missing data is constantly being reduced. Each day or call, there are new complete cases. The observed values for these new completes can be compared to the most recent imputed value. Graphical displays of these imputed and observed values offer a unique opportunity to evaluate the utility of the imputation model for preparing a distribution of plausible imputations. In general, this sequential environment creates unique opportunities to explore model diagnostics.

3. Applications

We have implemented this measure with two surveys. One survey, an area probability survey, has a rich set of frame data and paradata. The other survey, an RDD survey, has relatively limited frame data and paradata.

3.1 Application One: An Area Probability Sample

3.1.1 The Survey: NSFG

We implemented the fraction of missing information monitoring approach with a large area probability sample – the National Survey of Family Growth (NSFG). The NSFG collects data about pregnancies, births, and families. This survey screens households for persons aged 15-44. About 57% of households have an eligible person, so a large proportion of cases are not eligible.

The NSFG operates on a continuous basis with a new sample released every quarter. The sample is worked to completion in twelve weeks. The first ten weeks are phase one. In phase two, a subsample of nonresponding cases is selected and worked to finalization. Each quarter, approximately 1200-1500 interviews are taken with response rates ranging from 70-77%.

3.1.2 Data

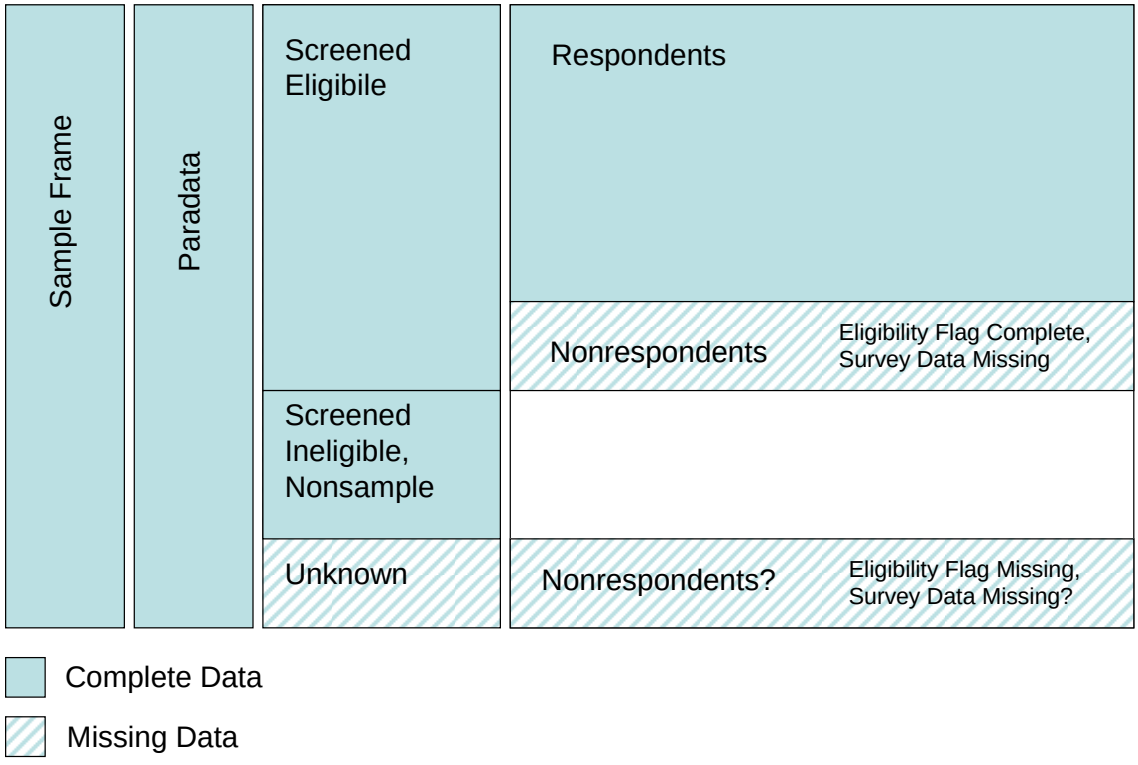
The goal is to evaluate the uncertainty about the values for the survey variables collected by the NSFG. This survey has a set of "key statistics" that are closely monitored while in production. The data can be divided into four main components:

- Data on the sampling frame
- Paradata regarding effort, including interviewer observations

- Household and respondent demographics, including eligibility, from a screening interview
- Questionnaire data supplied by respondents

These data are not always complete. The frame data and paradata are complete. The screening data are complete for a subset of the sample. The survey data are complete for a subset of the screened cases. The missing data are mainly the selected respondents who refuse or are unable to participate, and the unscreened housing units who may or may not be eligible. If there are eligible persons in the unscreened housing units, then we are also missing their survey data. This data structure is pictured in Figure 2-2. The proportion of cases that are unknown at the end of the survey period is generally small (5-7%), and the response rates are generally high (75-79%).

Figure 2-2: Schematic of NSFG Data Structure



The sampling frame data that I have include Census data characterizing the local area of the selected housing unit, such as the proportion of persons in the Census Tract who have never been married. I have also included the proportion of the population in the Zip Code Tabulation Area (ZCTA – a unit of Census geography that closely corresponds to US Postal Service ZIP codes) who are age eligible for the NSFG. We also have paradata – that is, the history of attempts to interview the case including number of calls, contacts, and indications of respondent reluctance.

Interviewers also make observations on sampled units. The NSFG tailored the observations hoping that they would be predictive of the survey variables. For example, the interviewer is asked to predict whether the selected respondent is involved in a sexual relationship. These predictions have been accurate enough to produce fairly strong correlations (some in the range of 0.2-0.3) with several key variables in the NSFG (Groves, Peytcheva, Wagner, 2007). Similar methods have been employed on other surveys (Copas and Farewell, 1998; Couper, 1997). Finally, for cases that are screened, we also have a set of demographic variables (age, sex, race, ethnicity, and household size) for the selected respondent. The data used in the imputation model are summarized in Table 2-1.

Table 2-1: Variables Used in NSFG Imputation Models

Eligibility Imputation Model
ZIP code-level Census data: % aged 15-44
Urbanicity
Census Block-level occupancy rate
Interviewer Observations
Access problems
Residential/Commercial neighborhood composition
Evidence of non-English speakers in neighborhood
Safety concerns in neighborhood
Housing unit in a multi-unit structure
All household members over age 45 (interviewer estimate)
Children under age 15 present in household (interviewer estimate)
Survey Variable Imputation Model
Variables from Eligibility Model
Census Region
ZIP code-level Census data: % in eligible age group never married
Selected respondent sex
Selected respondent age
Selected respondent lives in single person household
Selected respondent Hispanic status
Selected respondent in an active sexual relationship (interviewer estimate)
Selected respondent cohabiting (interviewer estimate)

In the imputation context, it is generally better to condition the imputations on all the observed data. This is the more conservative approach, as analyses based on the imputed data will tend to be more efficient than anticipated (Rubin and Schenker, 1987; Rubin, 1996). Collins, Schafer, and Kam (2001) consider the problem from a practitioner's standpoint. They simulated results from "restrictive" and "inclusive" strategies. Their simulations confirm the theoretical result from Rubin and Schenker – the inclusive strategy that includes more variables in the imputation than are planned for use in any given analysis is much preferred.

3.1.3 Methods and Results

The NSFG monitors the fraction of missing information on a daily basis. The survey has been monitoring several key variables. One of the key statistics is the proportion of persons who have never been married. Figure 2-3 shows the complete case estimate and the estimate for the fully imputed dataset by quarter and day.

For cases with unknown eligibility status, this variable was imputed. Then, conditional on the imputed eligibility status, the survey variables were imputed. This was done for the dataset as it stood after each day of the field period.

The imputations were created using IVEware, a software package which implements Sequential Regression Multiple Imputations (Raghunathan et al., 2001). There were 100 imputations done for each call attempt. Graham et al. (2007) note that when the fraction of missing information is high, a large number of imputations are needed in order to reliably estimate this quantity.

The models were fit using all available frame data and paradata (including interviewer observations). We subset variables until we found a set that worked well in producing imputations in the sequential regression setting. We found that different predictors were needed for the eligibility models than from the survey variables, so we did the imputations in the two-step procedure described above.

Figure 2-3: Proportion Never Married by Day

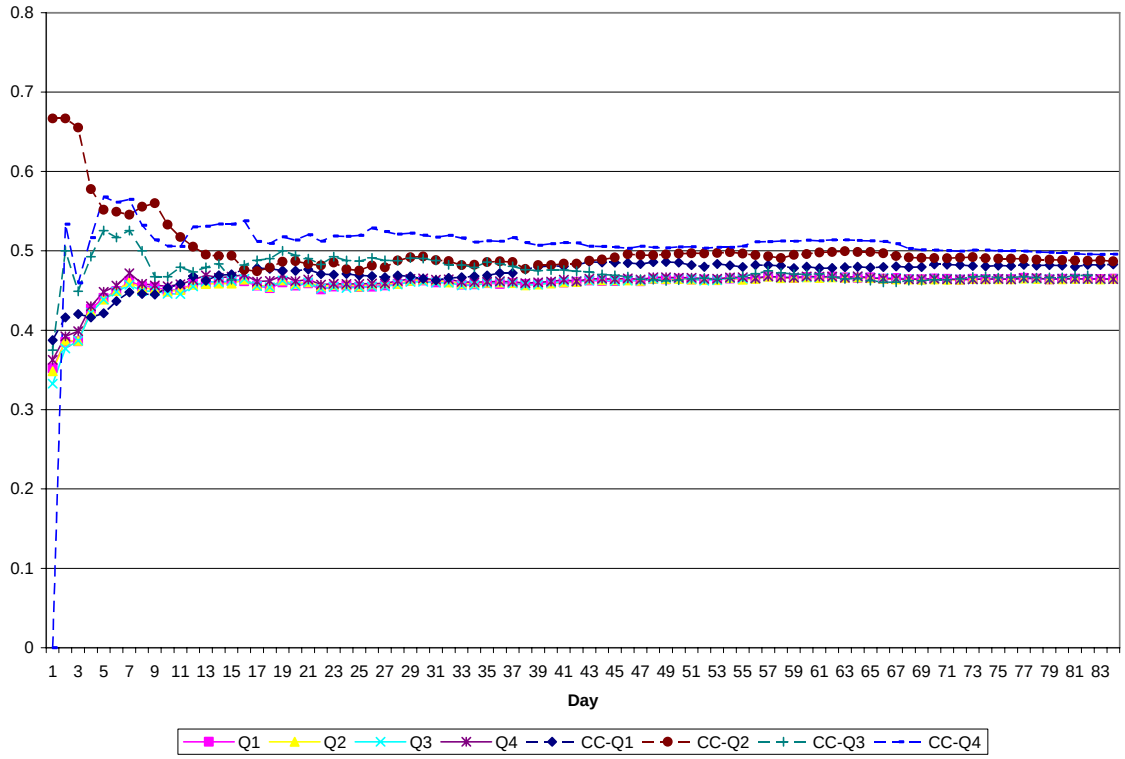
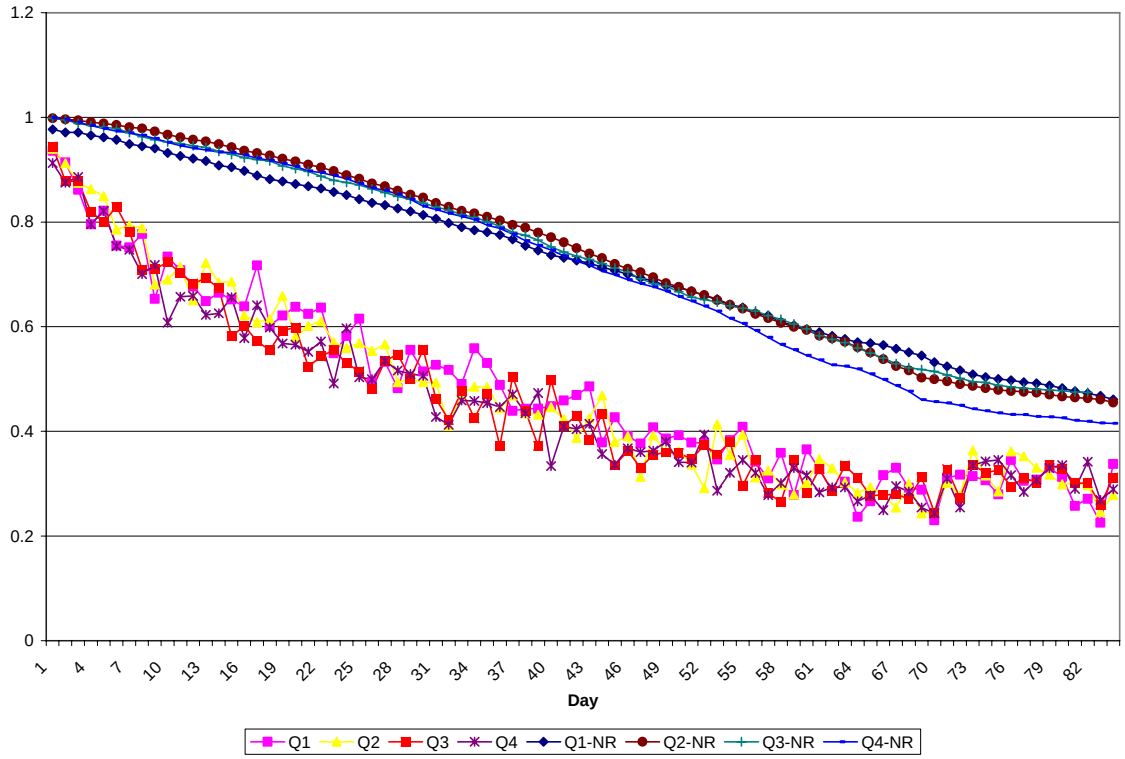


Figure 2-4 shows the fraction of missing information by day for the first four quarters (June 2006 to June 2007) of the NSFG data collection.

Figure 2-4: Fraction of Missing Information by Day: Proportion Never Married



The upper lines represent the nonresponse rate. These rates are calculated in a very conservative manner that make no assumptions about missing data. In particular, no adjustment is made for estimated eligibility among unscreened cases. This corresponds to the complement of the AAPOR Response Rate 2. This nonresponse rate was calculated in the following manner:

$$1 - \frac{Interviews}{TotalSample - Nonsample - Ineligible}.$$

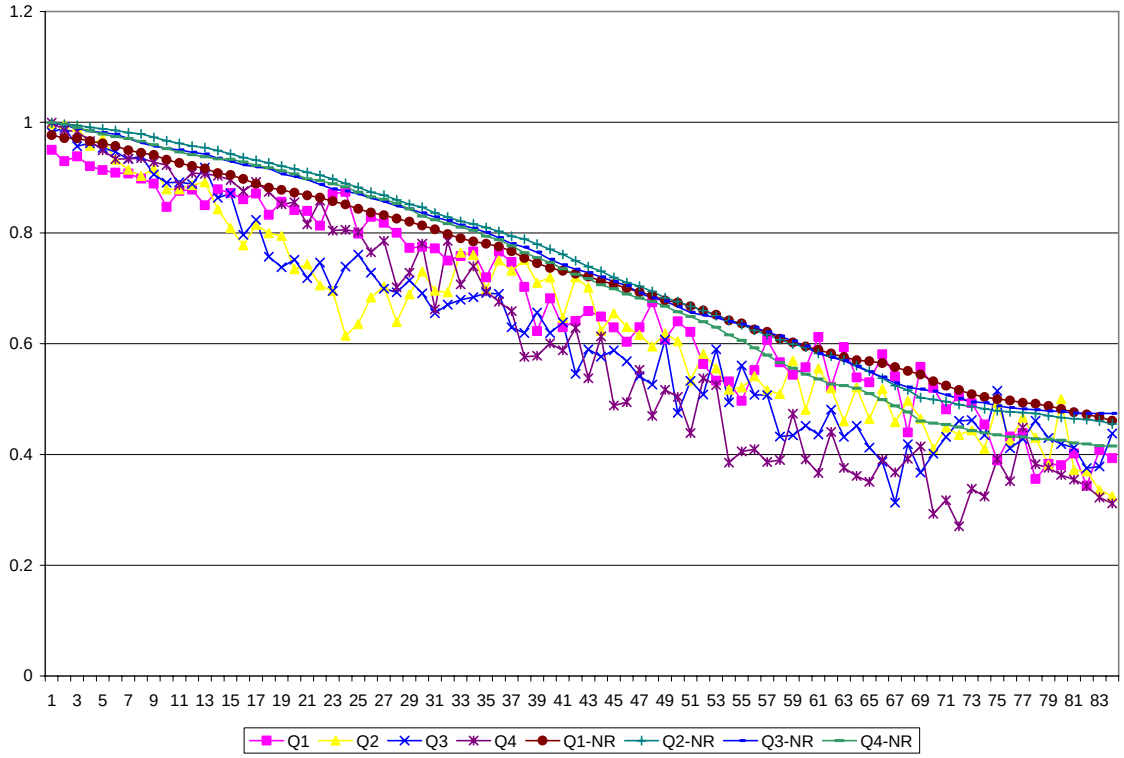
There are many cases where the eligibility status of the case is not known. Some response rate calculations assume that only a portion of these undetermined cases are eligible. In our analysis, however, we are involving this uncertainty about the status of the case. We impute the eligibility status of cases that are undetermined. As a result, this calculation of the response rate is the fairest comparison. In addition, we ignore the sampling (and

consequent weights) entailed by the two-phase subsampling procedure. Instead, we impute all the cases – including the phase one cases not sampled for interview in phase two.

It is clear from this figure that the frame data are reducing the uncertainty about the nonresponding cases. In the case of this variable, the data on the frame are correlated with the variable and, hence, the fraction of missing information is below the nonresponse rate.

In order to clearly demonstrate that the fraction of missing information is a variable-level statistic (as opposed to a survey-level), we will examine another key variable from NSFG. In this case, we find less reduction in our uncertainty. Figure 2-5 shows the fraction of missing information by day for the variable "number of sexual partners in the last 12 months."

Figure 2-5: Fraction of Missing Information by Day: Number of Sexual Partners in Past 12 Months



In this case, from the same survey, it seems that the frame variables and paradata are less correlated with the survey item. This information is also useful. It may help survey designers focus on developing paradata tailored to this item. Perhaps new interviewer observations can be created that are better correlated with this variable.

3.2 Application Two: An RDD Survey

3.2.1 The Survey: SCA

We implemented this monitoring approach with data from a completed monthly random-digit dial (RDD) survey – the Survey of Consumer Attitudes (SCA). The survey collects 300 RDD interviews each month. It typically attains an AAPOR Response Rate 2 of between 42% and 45%.

SCA asks respondents for their views on the state of their personal finances and the economy in general. A key statistic produced by this survey is the Index of Consumer Sentiment (ICS). This Index is widely reported and has been found to be highly predictive of economic trends (Curtin, 2007). There are two other key indices reported by this survey. Both are based on subsets of the ICS variables: the Index of Current Economic Conditions (ICC) and the Index of Consumer Expectations (ICE).

3.2.2 Data

The objective is to evaluate our uncertainty about the values for the three indices (ICS, ICC, and ICE) for the nonresponders using imputation models and calculating the fraction of missing information. These data have a structure similar to that of the NSFG. As with the NSFG, the frame data and paradata are complete. Additionally, the screening data are complete for a subset of the sample. The survey data are complete for a subset of the screened cases. The missing data are mainly the selected respondents who refuse or are unable to participate, as well as the noncontacted telephone numbers who may or may not be households. If they are households with eligible persons, then we are also missing their survey data.

In addition to the indices from the survey, there are a set of variables on the sampling frame. In an RDD survey, since the telephone numbers are randomly generated, we know nothing more than the telephone number. We can, however, estimate the geographic location (e.g. Census Tract, ZIP code, or county) for each telephone number. We can then attach data from the Census and other sources that describe the geographic context of the telephone number. For example, we can describe the urbanicity of the estimated geographic area in which the telephone number is located. Table 2-2 lists the

context variables used in the imputation models. In addition to Census data, I added data from surveys conducted by the Bureau of Labor Statistics: monthly average wages, monthly unemployment rates, and quarterly Consumer Price Index (CPI) estimates. Curtin (2007) suggests that these variables, as well as measures of month-to-month or quarter-to-quarter change are predictive of consumer sentiment. Curtin also mentions interest rates as important predictors of consumer sentiment; however, these data are more difficult to obtain at a local level. The unemployment and average wage data are summarized at the county level and attached to the sample record via the expected county of the telephone number. In the case of the CPI, these are estimated at the Core Based Statistical Area (CBSA) level for larger cities and at the Census Region/Urbanicity level elsewhere.

Table 2-2. Context Variables Used in Imputation Models

Census Data
Percent Listed
Household Density
Median Years Education
Log(Median Income)
Census Region
Proportion Aged 18-24
Proportion Aged 25-34
Proportion Aged 35-44
Proportion Aged 45-54
Proportion Aged 55-64
Proportion Aged 65+
Proportion White
Proportion African-American
Prop. Asian/Pacific Islander
Proportion Other Race
Proportion Hispanic
Proportion Owner-Occupied
Proportion Renter
BLS Data
Monthly Average Wages
Monthly Unemployment Rates
Quarterly Consumer Price Indices

Paradata, i.e. information from the call records (e.g. number of calls, ever a refusal), were also used in the imputation models. Every call attempt is recorded electronically. These records include date and time of call, interviewer ID, and a result code. This information is coded into the variables presented in Table 2-3.

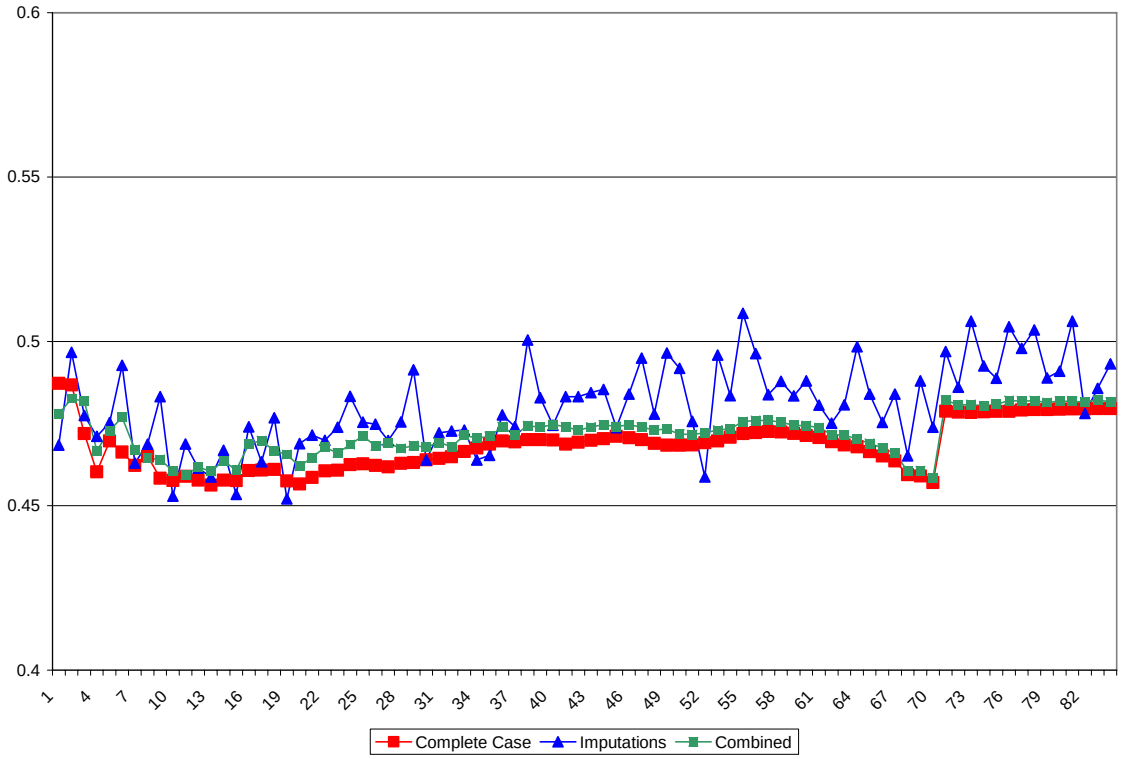
Table 2-3: Variables Created from Call Records (Paradata)

Number of calls
Ever resistant
Number of contacts
First contact
Days between attempts

3.2.3 Methods and Results

In an RDD survey, we are uncertain about which cases belong to the set of nonresponders. For some cases, we identify their eligibility early. For other cases, it may take several or even many calls to determine eligibility. For another substantial portion of each month's sample (7-15%), no contact is ever made (even with an answering machine) and we cannot determine whether the number is even assigned to a household. In order to deal with this problem, I imputed the eligibility status of each case for which it was not known. The eligibility flag was imputed using the information that existed for the case prior to each call. The imputations were updated for each call number. These imputations led to the eligibility rates in Figure 2-6.

Figure 2-6: Complete Case, Imputation, and Combined Eligibility Rates by Day

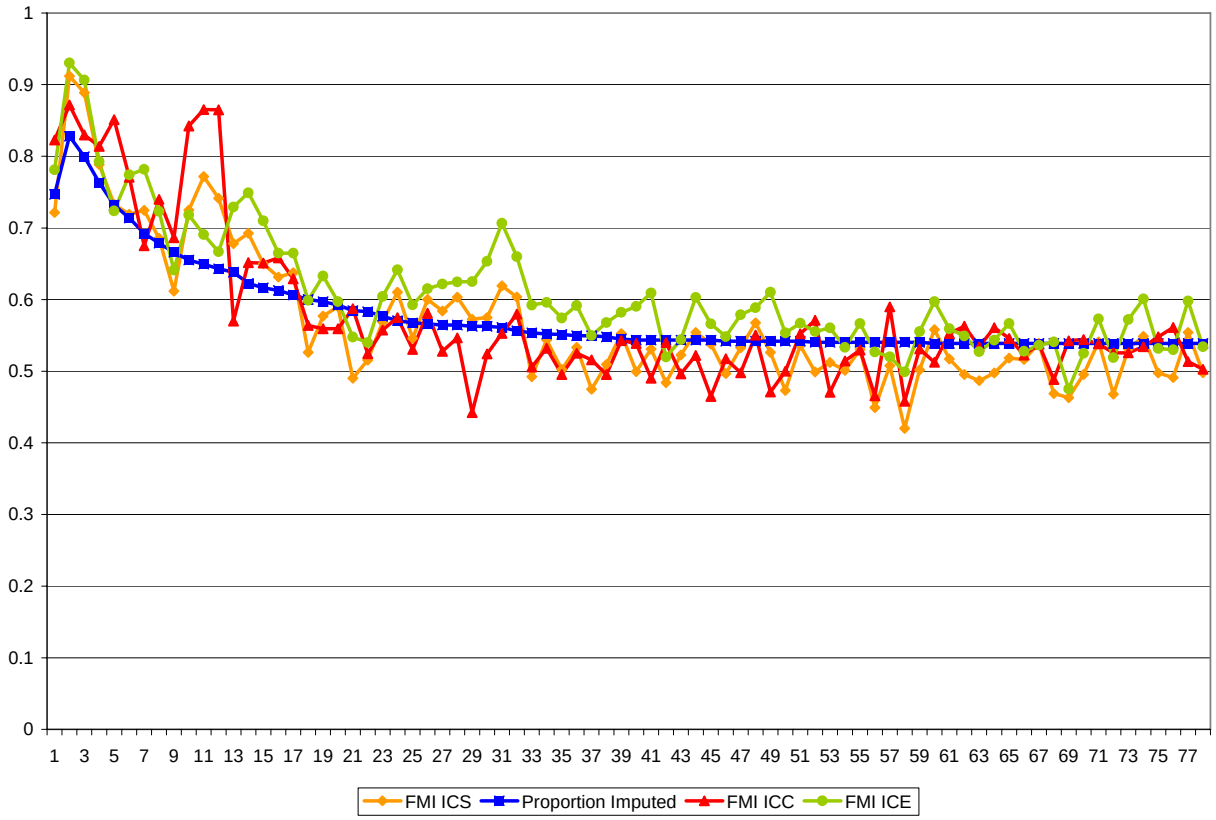


Conditional on the imputed eligibility flags, I imputed the survey variables that make up the components of the ICS. I did this for the dataset as it stood after each call. In other words, the imputations were made for the dataset as it stood after one call had been made on all cases, two calls on all cases, and so on. The imputations were created using IVEware. There were 100 imputations done for each call attempt.

Figure 2-7 shows the fraction of missing information as it existed after each call for the January 2006 dataset. The blue line with triangles represents the "proportion imputed," or the nonresponse rate. This rate includes all the cases for which eligibility flag unknown. This nonresponse rate is, therefore, the complement to AAPOR Response Rate 1.

We note, again, that our goal ought to be moving this fraction below the "proportion imputed" line through better paradata or by collecting cases that have a relatively larger impact on the information in the dataset. In this case, there appears to be some modest improvement over the response rate when we condition on the frame information, especially for the ICC.

Figure 2-7: Fraction of Missing Information and Proportion Imputed for ICS, ICC, and ICE



From this chart, conditioning on the available data does not reduce the fraction of missing information very much below the proportion imputed. In this case, the approach does not provide us much more guidance than monitoring the response rate – at least at an overall level. It is possible that monitoring the fraction of missing information for subgroups might lead to a different result.

4. Discussion

A strength of this measure – relative to response rates – is that it allows us to look at the relationship between nonresponse and the survey estimates, albeit under an

assumed model. This, however, is the inherent nature of the nonresponse problem. Any analysis that deals with missing data must make some model assumption – either implicitly or explicitly. In fact, the use of models is already ubiquitous in the presence of nonresponse. For example, nonresponse weighting class adjustments are an implicit model about the relationship between the attributes of responders and nonresponders. Since these models are presumed reasonable for statistical adjustment purposes, why not bring them into the data collection process? This may improve the quality of the final data by reducing our uncertainty about the model fit for certain cases. In other words, we may want to identify the subsets of nonrespondents who contribute the most to the between imputation component of the variance. It would make sense to use relatively more resources, if needed, to include those cases. Thus, the monitoring of the fraction of missing information can lead to tailoring of the protocol to minimize the uncertainty associated with nonresponse. Cases associated with more uncertainty could be targeted for a more costly and more effective strategy. Conversely, we could stop collecting data on cases for which our uncertainty about the value we would impute for the case is within certain bounds.

One simple approach would be to stratify all sampled cases using the propensity score. Then, calculate the fraction of missing information within propensity stratum. The stratum with the highest fraction of missing information could then be targeted for some special effort aimed at bringing in more cases from that stratum. Depending on the response rate and the quality of the frame information, this might not be the lowest propensity stratum. If the information does not allow us to discriminate between the

strata, then at least we can produce the salutary effect of equalizing response rates across subgroups.

Another advantage of this measure is that it may redirect attention away from the response rate and toward improving the quality of the data. Monitoring the response rate leads survey methodologists to focus on approaches that bring in the most responsive case at each step. Research organizations focus on designs that bring in "low-hanging fruit." Monitoring the fraction of missing information has the possibility of leading survey data collection organizations to focus on interviewing cases that maximize the information in the sample. This might result in a smaller data set (for a fixed cost) or a data set collected with a lower response rate than one which was collected by maximizing the response rate. But by maximizing the information, this approach should produce more efficient estimates.

A secondary benefit of using these model measures to monitor surveys is that it should encourage the collection and testing of new types of frame data and paradata. These data should be related to the survey measures of interest – proxy survey measures are preferred. This is needed to build useful imputation models. These data should be more useful for adjustment as well (Little and Vartivarian, 2005) as they will be tailored in order to be correlated with the specific variables of a particular survey and not with nonresponse in general.

A weakness of this approach is that it is item specific. Response rates are generally thought of as measures for the entire survey. Although it is true that item-level nonresponse can create different item-level response rates, it is also the case that the preponderance of nonresponse occurs at the unit level for most survey items. On the other

hand, the fraction of missing information, as we have seen, can vary a great deal across survey items. For some survey variables, the frame data may be highly predictive while for other survey variables the frame data may not be predictive at all. These differences will lead to relatively large differences in the fraction of missing information across items.

On the other hand, this weakness can also be seen as a strength. In fact, adjustment strategies face this same issue. A particular adjustment strategy may be good for one survey item and bad for another item in the same survey. Monitoring data quality at the item level may help surveys face this issue more directly. In practice, surveys can choose "key" items for monitoring. These key items can then be used as the basis for developing a post-survey adjustment strategy. In general, methods developed for designing multi-purpose surveys should apply when designing an item-level monitoring strategy.

Nonresponse has become a fact of life for surveys. Much research in survey methodology has been focused on methods for improving response rates, without evaluating the impact on nonresponse bias. Model-based methods offer a way to move forward in evaluating the risk of nonresponse bias. The fraction of missing information is one model-based method for evaluating data quality that uses all the data we have – not just the complete data on the frame. The use of this method for monitoring surveys should have a salutary effect on the design of surveys. First, more emphasis should be placed on identifying useful complete data – whether interviewer observations or new database sources of information – since better correlations between these data and the survey variable tend to lower the fraction of missing information. Second, monitoring the

fraction of missing information should lead to concern over which cases are interviewed – as opposed to being concerned with the proportion of cases interviewed (i.e. the response rate). The goal of data collection should become maximizing the information content of the produced datasets. In the presence of nonresponse, this should lead to more efficient datasets. Indicators such as the fraction of missing information are needed if we are to begin developing data collection strategies that are aimed at producing the highest quality data.

Chapter 3

Adaptive Survey Design Protocols

1. Introduction

Survey research organizations have long been struggling against the decline of response rates. In an effort to maintain response rates, these organizations have increased the amount of effort applied to each sample – additional calls, higher incentives, and more refusal conversions (Curtin et al, 2000; Rogers, 2004; McCarty, 2006; Lynn, 2003; Kessler et al., 1995). Since survey organizations have long used the response rate as a performance target, this was the natural response. However, this increased effort has not been able to stop the decline in response rates (Curtin, et al., 2005; de Leeuw ad de Heer, 2002; Atrostic et al., 2001; Petroni, et al., 2004). As a result, concerns over the risk of nonresponse bias have been growing. This has led to the situation where declining response rates are seen by some as a threat to the ability to conduct surveys of human populations.

This threat has led survey methodologists to pursue research into methods that will improve response rates. All facets of survey design have been considered. For example, there is a large literature on the impact of incentives on response rates (see Singer et al., 1999 and Singer, 2002 for reviews). This research often is done using the

classic randomized, controlled design with at least two treatments randomly assigned to independent samples. The treatment effect is measured as a difference in response rates between these two groups. Although some studies have looked at the impact on survey statistics, the effect on response rates has been the major focus. As a result, there is much more emphasis on the proportion that respond rather than on who those respondents are.

Unfortunately, these types of experiments are ill-suited to important features of the survey process. Surveys are by their nature multi-stage processes. First, we attempt to contact households. Once we have established contact, we attempt to screen households. We may need to convert households that refuse to be screened. Finally, we interview the sampled respondents. If we fail to establish contact after the first call, we try again. If the current design is failing to meet study objectives, then ad hoc changes are made in order to improve outcomes. The standard randomized controlled experiments typically focus on one design feature – incentives, for example. They do not consider what next step to take after this design feature fails. Nor do these experiments consider which combination of design features may work better than others. In addition, these experiments do not frequently enough consider how these design features may interact with the characteristics of the sampled persons to produce differing effects.

If these methods are ill-suited to the problem at hand, what options do we have? In the field of clinical trials there has been consideration of other types of designs. There is emerging research into what have been called “dynamic treatment regimes” (Collins and Murphy, 2007). A dynamic treatment regime is a multi-course treatment. At each step, the patient is assigned a next treatment based on their characteristics, including the history of previous treatments and their response (or lack of response) to those

treatments. The sequencing of treatments is regulated by a set of decision rules for each step of treatment. These rules have been designed to maximize response – in this context, positive outcomes for the treatment such as remission in cancer treatments. This approach has been used most famously to develop a multi-course treatment for depression in the STAR*D clinical trials (Wisniewski et al., 2004).

Survey methodology may find these methods useful for developing new "treatment regimes" aimed at inducing survey response. In this chapter, I will suggest that current survey practices can be improved by incorporating experimental and statistical methods developed under the rubric of dynamic treatment regimes. These methods can be used to identify strategies or protocols that are tailored to the characteristics of the sampled person and are aimed at increasing each sampled persons probability of response. These characteristics include the fixed characteristics of sampled persons, but also the history of previous attempts to interview those persons. Even though such an approach could be used simply to raise response rates, I will contend that this approach would be necessary if survey data collections are to have a goal of maximizing the information in the dataset (as opposed to maximizing the response rate). In other words, such an approach will be necessary if we care about who responds as much as we care about the response rate. The dynamic treatment regimes model will help us move forward with this agenda. In Section 2, I will review the relevant clinical trials and survey methods literature. In Section 3, I will introduce the notation and proposed methods. In Section 4, I will present two applications of the method. In section 5, I will propose the use of computerized decision support models for field surveys as a means to implement dynamic treatment regimes. Section 6 will offer a conclusion and discussion.

2. Background

Dynamic treatment regimes consider sequences of treatments. At each step in these sequences, the treatments are adapted to the characteristics of the patient, including the history of previous treatments. In this section, I will contrast dynamic treatment regimes with much of the research being conducted in survey methodology. I will then propose Leverage-Saliency as a theoretical justification for applying the dynamic treatment regimes approach to surveys. Finally, I will provide examples from survey methodology of tailored designs.

2.1 Dynamic Treatment Regimes

In the field of clinical trials, there has been recognition of the need for new methods. In practice, when a treatment fails the caregiver does not give up. Instead, another treatment is tried. Although randomized controlled trials do not consider sequences of treatments, this is the practice. There are, however, a set of new techniques developing in this field that address this gap between research and practice. These new techniques have been called adaptive or dynamic treatment regimes (Murphy, 2003; Murphy, 2005; Thall, Millikan, Sung, 2000; Thall et al., 2002; Thall and Wathen, 2005; Lavori and Dawson, 2000; Lavori and Dawson, 2004; Lavori and Dawson, 2008; Collins et al., 2004; Collins et al., 2005; Collins and Murphy, 2007). Under this emerging model, treatments are considered to be multi-course and the object is to identify the optimal combination and sequence of treatments for inducing the desired response. These regimes are often tailored not only to the outcomes of previous treatments, but to the characteristics of the patient, as well. Outcomes are measured both by the efficacy of a treatment sequence and by potential adverse outcomes, such as toxicity. Such regimes

allow the dosage level and type of treatment to vary with time, using rules specified before the beginning of treatment; these rules are based on time-varying measurements of subject-specific need.

A relevant example of this approach is provided by Thall, Millikan, and Sung (2000). They consider competing, multi-course treatments for prostate cancer. In order to determine the best sequence of treatments, they estimate the salvage probabilities that one treatment has when another treatment has already failed. They also estimate the cross-resistance between consecutive treatments. Murphy (2003), in an investigation of interventions to improve reading skills, also considers how these optimal regimes might vary for demographic subgroups.

These complicated, multi-course treatment protocols may require new experimental methods for development and testing. With even just a few components, these treatment regimes would require many arms in a standard randomized controlled trial. For example, with just three treatments, there are six possible sequences. Testing all of these sequences against a control seems impractical. Murphy, Lynch et al. (2007) propose an alternative experimental method for elucidating dynamic treatment regimes. Their approach, which they call Sequential Multiple Assignment Randomized Trials or SMART, involves multiple randomizations of patients to treatments. For example, nonresponders to an initial treatment may be re-randomized to a second treatment. A fractional factorial design may be used to allow efficient allocation of sample sizes. This experimental design may be implemented within a rubric suggested by Collins and Murphy (2007) – the Multiphase Optimization Strategy, or MOST. This approach attempts to identify effective treatments by using a three phase process. In the first phase

– called the screening phase – effective components of the potential treatment regimes are identified. A second phase is used to refine the list of components identified in the screening phase and their optimal dosages. The final stage is a confirming phase in which a standard randomized controlled trial is used to evaluate the proposed treatment regime against the current standard of care.

The dynamic treatment regimes approach offers a model for considering optimal methods for treatments that require multiple stages. This allows the researcher to consider interactions between the components that may strengthen or weaken their separate effects. It also allows consideration of how the order of the treatments might matter.

2.2 Current Practice in Survey Methodology

There is a stark contrast between this emerging research into methods for dynamic treatment methods and the current practice in survey methodology. The standard model for research into survey methods has been the randomized controlled trial. Two or more treatments are compared to each other. The outcome is a difference in response rates. Some of this research will compare the estimates of the survey variable for the two groups. Very little research is aimed at identifying who responds to which methods (see Groves, 2005 for a review). In addition, very little of this research is focused on which sequence of treatments if most effective.

Research into methods for dealing with nonresponse has been largely concerned with improving overall response rates. There is very little literature on methods for improving response rates among various subgroups. There are a large number of studies on various design features (incentives, calling strategies, advance notification, etc. -- see Singer et al., 1999; Link and Mokdad, 2005; de Leeuw et al., 2005; Greenberg and

Stokes, 1990; Weeks et al., 1987; Houtkoup-Steenstra and van den Bergh, 2000 for reviews) aimed at reducing nonresponse. In some of these studies, estimates of the survey variable are also compared. These studies sometimes produce contradictory results. Link and Mokdad (2005), for instance, review disparate results on the impact of prenotification letters in Random-Digit Dial (RDD) surveys. There are exceptions to this generalization (I will review these exceptions later in the chapter), but this has been the dominant approach.

The virtue of this general protocol approach is that is avoids potential biases resulting from arbitrarily applying methods to some cases and not others. For example, Kennickell (2003, 2004) worries that allowing interviewers to decide when they will call each case will lead them to be selective in where they place their effort. He presents statistical models that successfully predict which cases will receive more calls when interviewers decide where to place their effort as evidence that this has happened on the Survey of Consumer Finances. The end result is a difference in response rates between groups that received more calls and those that did not.

The problem with this approach is that using equal protocols with unequal sampled units does not necessarily produce equal response propensities. In other words, different respondents will react differently to the same protocol. Putting aside the question of whether we can really define what "equal effort" would mean, it should be clear that for some cases more calls will not lead to response. In the extreme, offering a sampled case an incentive might increase response probabilities while for another case this same offer might decrease response probabilities. In general, even if a design feature

works on average to increase response rates, it might actually decrease them for small subgroups.

The general correlates of nonresponse across multiple surveys are well known (Groves and Couper, 1998). The inverse of this research is that characteristics of responders are also well known. Bickart and Schmittlein (1999), in a premonition of concerns over online panel surveys, go so far as to argue that a small proportion of the population is completing the majority of surveys. To the extent that these generalizations about the characteristics of nonrespondents are true, survey researchers should be concerned that they are good at bringing in certain kinds of people, but not others. New methods are needed that will allow us to increase the response propensities of persons who are in the class of persons who generally respond at lower rates. If we continue to focus on methods that produce the highest overall response rate, we run the risk of recruiting more of the "same kind" of person. Survey designs that are adapted to the characteristics of the sampled person are needed if we are to maximize the information in the sample.

2.3 Leverage-Saliency

Attempting to tailor the design to the specifics of the case, including the history of previous treatments, might be more effective than applying a single set of rules to all cases in producing the highest response propensity possible for each case. In fact, Groves, Singer, and Corning (2000) present a theory of survey participation they call "Leverage-Saliency" which supports this viewpoint. They define the "leverage" of each survey design feature as the "importance the sample person assigns to the attribute in the decision to participate" (p. 300). Saliency is the emphasis which is placed on the attribute

by the interviewer or survey organization in the request to complete the survey. When an attribute — for example, an incentive — is important to a sampled person, then placing emphasis on that attribute will lead to increasing the probability that the sampled person will agree to do the interview. This model is appealing in that it emphasizes that the decision to participate is a function not only of the design features of the survey but also the characteristics of each individual respondent.

The theory provides a useful heuristic for describing the decision to participate. From this viewpoint, it makes sense to search for survey design protocols that have the most leverage with each, specific respondent and make them salient. These protocols may be identified statistically using information available on the sampling frame and paradata, i.e. data about the data collection process such as call records and interviewer observations (Couper, 2000; see Couper, 1998 for a full discussion of the range of paradata).

2.4 Tailored Design

Research into tailored methods has faced at least two major impediments. First, tailored strategies would require that recommendations based on the current status of each case be made available to interviewers in real-time. In the past, paper and pencil materials would have made this impossible. Interviewers were asked to make their own judgments based on very general training and their experience. Computerization makes new methods possible. It is now possible for interviewers to deliver data about their work to a central server on a daily basis – or even more frequently. These data could be used to deliver recommendations to interviewers on the same timely basis.

Another impediment to such research may be the use of the response rate as the key metric for measuring data quality. Under this metric, a survey organization's goal is to obtain the highest possible response rate. The least expensive way to do this is by interviewing the most responsive cases – the "low-hanging fruit." Marginal groups, for whom special design features might be more effective, are not necessarily a priority when the response rate is the measure of success.

In spite of these obstacles, there is a sparse literature on tailored methods for surveys. The research of Groves and Couper (1998) showed that allowing interviewers to tailor the introduction based on interactions with the respondents improved response rates relative to the procedure where interviewers are required to follow the same script for each introduction.

Greenberg and Stokes' (1990) article on call scheduling is another example of a tailored method. The timing of each call to a case was determined by the history of previous calls. They use a Markov decision process to minimize the expected number of calls required for making contact with households in an RDD sample. They used logistic regression to identify important predictors of contact. Their model differentiated day, evening and weekend calls. They also included the number of previous attempts, days between calls, and whether the last call was a busy signal or ring-no-answer. The result of their model is a proposed call scheduling algorithm that prescribes the timing of the call conditional on these characteristics of the case. Although the recommended strategy was never implemented, it remains an interesting piece of research into tailored methods. 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 context data as well as information about the result from previous attempts.

Another approach to minimizing effort aimed at contacting households was developed in the field of operations research. Bollapragada and Nair (2001) explore ways to improve contact rates for credit card collection call centers. In their model, they attempt to estimate household-level probabilities of contact for calls placed during various times of day and days of week (call windows). Their models start by assigning the overall average contact rate within any window to each new case. Then they modify that rate (upward or downward) by a fixed multiplicative factor depending on whether the previous attempt succeeds or fails to establish contact. They use simulation to determine the optimal multiplicative factor. Their results have been implemented and yielded a 10% reduction in effort to establish contact for a large credit card collection call center.

Trussell and Lavrakas (2004) have considered methods for tailoring incentives. In their study, they analyzed data from telephone recruitment to a "high-burden" mailed survey. They found that the effort required to bring about response to the telephone recruitment was a very useful piece of information for effectively tailoring the optimal incentive amount. They found that incentives had a much higher impact on response rates to the mail survey for households that refused the initial telephone screening survey. The next largest impact was for households that were not contacted by telephone. The smallest impact resulted from an incentive to households that had agreed to do the mail survey. Their conclusion is that "the notion that one 'optimal' level of incentive should be sent to all respondents makes no theoretical sense whatsoever" (Trussell and Lavrakas, p. 365).

Finally, Groves and Heeringa (2006) define a conceptual approach to surveys that involves tailoring. They call the approach "responsive design." The approach suggests that survey research staff identify design features that impact cost and error properties of key survey items, identify statistics related to these cost and error properties, monitor these statistics, and change the design based on the monitoring of these cost and error tradeoffs. Central to the approach is the notion of "phase capacity." Surveys are implemented in phases. When a phase ceases to bring in sampled units whose responses change current estimates of key survey statistics, then it has reached its phase capacity. In their words, " '[p]hase capacity' is the minimum bias condition for an estimate in a specific design phase; that is, the best outcome for a statistic that a particular set of design features can produce" (Groves and Heeringa, p. 11). When the phase capacity has been reached, the survey needs a new design protocol. In this sense, the design is tailored – but tailored to the set of currently active cases. The tailoring is based on either the amount of time the case has been in the field or the number of calls made without response under the current protocol. If a sampled unit reaches the end of the phase without having responded, this is taken as evidence that this unit (aggregated with all other units at the end of this phase) needs a different protocol.

In sum, the dynamic treatment regimes model offers a promising avenue for research into survey design methods. Surveys are inherently multi-stage processes and dynamic treatment regimes are designed for this situation. Leverage-Saliency provides theoretical justification for the approach. While there are some examples of research into "tailored" design, much more is needed.

3. Methods and Notation

The statistical methods and design approaches developed for adaptive treatment regimes can be adapted to the survey realm. If we consider survey design features as treatments and completing the interview as the desired response, then the model translates directly into the conceptualizations of adaptive treatment regimes. Conditional on fixed covariates and previous treatments, the task is to find the most efficient next step. I will develop statistical models that allow us to identify efficient means for establishing contact with sampled persons. These models will be adaptive in that they incorporate information from each previous step as they seek the best next step.

I will use the following notation to describe the approach. First, let R_{ij} be the response indicator for the i^{th} person at the j^{th} call. The definition of "response" depends upon the context. It could mean completing a survey. In both of my applications, response signifies successfully contacting a household. Let $\mathbf{X_{ij}}$ denote the vector of k covariates for the i^{th} person at the j^{th} call. These covariates can include information about the history of previous protocols used for case on the first $(j\text{-}1)^{th}$ calls. Let $\mathbf{S_{ij}}$ be the vector of p variables defining the protocols available for the i^{th} person at the j^{th} call. The problem is to define a statistical model (logistic regression, for example) that allows us to estimate the values of $\mathbf{S_{ij}}$ that maximize the $\Pr(R_{ij}=1)$. The $\mathbf{X_{ij}}$ can be used directly to match sampled persons to a protocol $\mathbf{S_{ij}}$, or they can be summarized in a propensity score p_{ij}, and then propensity strata can be matched to a protocol.

The first application will hypothesize that there may be interactions between demographic characteristics of sample members and the design features used which impact the probability of contact. Greenberg and Stokes (1990) and Brick et al. (1996)

considered this hypothesis. I also hypothesize that the sequence and combination of protocols may lead to differing contact probabilities. Given the high dimensionality of this problem (matching cases based on all the details of previous protocols as well the fixed covariates on the frame), I will reduce these elements to a propensity score. Theoretically, it is possible to tailor the protocol to the specifics of each case, however, this is likely to be difficult with a large number of covariates. The propensity stratification approach is meant to balance the covariate distributions across the strata. This is a new approach to this problem.

An alternative approach to this problem would be to develop a predictive mean model that predicts the survey outcome variable. One could then stratify the sample based on this predictive mean and then determine which protocol is most effective for each of these strata.

In order to assess these interactions, I adopt the following two-stage strategy. In the first stage, I use the history prior to the call attempt *j* and the context data, to match those who were contacted with those who were not. I do so by creating strata based on propensity scores (Rosenbaum and Rubin, 1983). Let $\mathbf{X}_{ij}$ denote a $k_j \times 1$ vector of covariates for subject *i* prior to the call attempt *j* or at the end of the previous call attempt. Let R_{ij} denote the contact status (1: contact, 0: no contact) on the corresponding subject for the call attempt *j*. The propensity scores were estimated using a logistic regression model of the following form: $p_{ij} = logit(\Pr(R_{ij} = 1 \mid \mathbf{X}_{ij}, \boldsymbol{\alpha}_j)) = \mathbf{X}_{ij}\boldsymbol{\alpha}_j$ where p_{ij} is the propensity score to match the respondents and nonrespondents prior to the next call attempt. The $\boldsymbol{\alpha}_j$ are the k_j coefficients in the logistic regression model.

In the second stage, I fit propensity models for contact at call attempt *j* conditional on the specific protocol used at this call attempt. Let S_{ij} denote a vector of variables describing the protocol (treatment) used at call attempt *j*. Here I define protocols as the set of design features that are combined together for each attempt at contact. I then estimate the contact probabilities at call *j*, within each propensity stratum conditional on the specific protocol used during call *j*. Let R_{ij} be the outcome at call *j* for subject *i*, taking the value 1 if the attempt was successful and 0 otherwise. I estimate the probability of contact using logistic regression, $logit(\Pr(R_{ij} = 1 \mid \mathbf{S}_{ij})) = \mathbf{S_{ij}}\boldsymbol{\beta}_j$, where the $\boldsymbol{\beta}_j$ are the coefficients in the logistic regression model.

I estimate these models for each of the propensity strata. Several months of previous surveys were used to estimate the posterior distribution of $\boldsymbol{\beta}_j$. A normal prior on $\boldsymbol{\beta}_j$ was assumed. This approach allows us to identify different highest probability strategies for cases with different fixed characteristics (including previous effort). In this approach, the previous effort is not incorporated in the S_{ij}, but in the $\mathbf{X}_{ij}$ used for the propensity stratification models previously described. The task is to identify the set of S_{ij} that maximize the probability of contact within each propensity stratum. This configuration is perhaps the most efficient protocol that would have maximized the propensity of contact. I use the posterior distribution of the coefficients from this second-stage model to assess the probability that a given strategy is the contact propensity maximizing strategy.

The second application will use data from a panel survey. Sampled household are interviewed at multiple points in time, or waves. The sample was developed from an area

probability sample of households. I will estimate the probability of contacting a household using a random intercept logistic model. This approach will allow me to develop household-specific estimates of the best times of day and days of week (i.e. "call windows") for establishing contact with the household. The demographic variables denoted as the (1 x *p*) vector $\mathbf{X}_i$ for household *i* are used as predictors in the model. These are treated as fixed effects since there is almost no change in these variables within any household at any given wave. Let R_{il} denote the contact status (1: contact, 0: no contact) on household *i* in window *l*. Each household is assumed to have its own intercept β_{0i} which is from a $N(0,\sigma_i^2)$ distribution. The time and date of each call is used to identify each call as having been placed in one of the six windows. Then the model is estimated as:

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

A separate model is estimated for each call window *l*. Table 3-1 shows how the each of the windows was defined.

Table 3-1: Definition of Call Windows

Window	Definition
1	Weekday Morning (Mon-Fri, 8am-12pm)
2	Weekday Afternoon (Mon-Fri, 12pm-5pm)
3	Weekday Evening (Sun-Thurs, 5pm-10pm)
4	Weekend Evening (Fri-Sat, 5pm-10pm)
5	Weekend Morning (Sat-Sun, 8am-12pm)
6	Weekend Afternoon (Sat-Sun, 12pm-5pm)

In the next section, I will implement these methods. The first application is implemented on a centralized, telephone survey and the second application uses data

from a panel survey with telephone and face-to-face interviewing. I will use observational data. Methods for observational data have been developed in other areas, including propensity score methods (Rosenbaum and Rubin, 1983). Winship and Morgan (1999) provide a detailed review of these methods. The problem is that these methods rely on a specified model to account for differences between those "assigned" to the treatment and control groups. If the model is misspecified, then biased estimates of treatment effects can still occur. Even with this limitation, these methods provide a sound starting place for research into tailored methods. There are vast amounts of data available to be analyzed. Eventually it will be desirable to implement experiments to verify that tailored protocols identified through analysis of observational data are, in fact, effective.

In addition, I will be focusing on one phase (which nevertheless involves several steps) of the survey process. One goal, suggested by the dynamic treatment regimes model, is to extend these methods to multiple phases and – ultimately – seek to identify optimal dynamic treatment regimes over all phases.

Finally, the methods proposed here are not "optimal dynamic treatment regimes" in the sense indicated by Murphy (2003). They are not shown to be the most efficient methods to establish contact or conduct an interview over multiple calls or multiple phases. To use the terminology of reinforcement learning (Sutton and Barto, 1998), the protocols that I identify as "best" in the two applications presented here are "greedy." That is, at each step the protocol that is identified as best has the maximum immediate gain. They do not consider long-term consequences of the current action. An optimal regime might determine, for example, that a protocol with a lower probability of success for the current call is part of a multi-phase strategy with an overall higher probability of

success than a strategy that includes for the current call a protocol with a higher probability of success. However, if my applications are successful, this suggests that a research program aimed at identifying optimal dynamic protocols for surveys could also be successful.

4. Applications

4.1 Application One: An RDD Survey

4.1.1 The Survey: Survey of Consumer Attitudes

I have undertaken the examination of observational data from the Survey of Consumer Attitudes (SCA), an ongoing RDD survey conducted each month by the Survey Research Center at the University of Michigan. SCA collects approximately 300 RDD interviews per month. The main statistic produced by the survey is the Index of Consumer Sentiment (ICS). This Index is widely reported and has been found to be highly predictive of economic trends (Curtin, 2007).

One of the most difficult tasks in an RDD survey is establishing contact with prospective respondents. Using data from SCA, Curtin, Presser, and Singer (2005) note that the noncontact rate grew an average of .63 percent per year between 1979 and 2003. This trend has led us to the point where noncontacts are now as large a component of the nonresponse to this survey as refusals. While others have considered the timing of the call and the lag time between calls (Stokes and Greenberg, 1990; Weeks, et al., 1987; Kulka and Weeks, 1988; Dennis et al., 1999; Brick et al., 1996), no one has considered all of the design features taken together as a single "protocol" simultaneously. The goal of this analysis is to locate the next protocol with the highest probability of contact

conditional on the fixed covariates of cases as well as previous protocols administered to those cases.

4.1.2 Data and Methods

The data I analyze here include the history of all calls made, including the time, date, offer of incentives, and the result obtained for the combined months of 2003 through 2005 (n=25,602 sampled telephone numbers). The sample was developed using the Genesys sampling system. Genesys links telephone exchanges to an estimated geography. Census data for these geographic areas provide "context" data for all sampled telephone numbers. These data include information about the age, income, and race distributions of the population associated with the exchange of the sampled telephone number as well as information about urbanicity, housing, and other characteristics of the geographic area in which each telephone number is estimated to be. These types of data have been used by others to estimate contact, screening, and interview probabilities for use in weighting adjustments (Lu, Hall, and Williams, 2002; Johnson et al., 2006).

One limitation of an analysis of observational data is that we can only consider the methods that were actually employed by the process that created these data. The variation in the data will be limited by the calling rules and design features that are the current practice. We cannot estimate outside the range of these data. Fortunately, very general rules have been employed for the collection of these data, and, hence, the variation in the data is quite large. An additional limitation is that the calling rules and design features employed may have been determined by variables that are not observed. The choice of when to leave an answering machine message could be influenced by characteristics of the household that are observed by the interviewer but not available in

the data – something about the outgoing message, for instance, that leads the interviewer to leave messages when they are more likely to be effective. Confounding factors such as these are possible in observational data. This seems less likely for the timing of the call -- particularly in the calls made before contact since there is very little information available to interviewers or the calling algorithm that could influence this decision. The data employed here, other than the history of previous calls, are not used by the algorithm nor revealed to the interviewers.

The list of predictor variables ($\mathbf{X}_{ij}$) in the first-stage model are listed in Table 3-2. The context variables encompass most of the data available through the Genesys sampling system. Previous research in this area suggests that the urbanicity and median income of the estimated geographic area (Dennis, 1998; Brick et al. 1996) are predictive of differences in contact rates using a different definition of contact windows. As part of the modeling fitting exercise, I tried different transformations on some of these variables and found that the natural logarithm of the median income produced 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 history of previously applied protocols is also included in these propensity models. In this manner, the next step is conditioning on the previous protocols. The protocol used on a case for the first call is used to estimate the propensity of contact on the second call (and, hence, is used in the propensity stratification); the protocol used on a case for the second call is used to estimate the propensity of contact on the third call;

and so on. In theory, we could match protocols to each particular combination of history of protocols and context variables. Unfortunately, this would require enormous sample sizes. Instead, we are incorporating this information into the best scalar summary of these variables – the propensity score (Rosenbaum and Rubin, 1983). The predicted probabilities were divided into quintiles to create matched strata.

Table 3-2: Contact Propensity Predictor Variables (X_{ij})

CONTEXT VARIABLES	**HISTORY OF PREVIOUS PROTOCOLS**
Listed/Letter Sent	Call 1: Weekday Day
% Exchange Listed	Call 1: Weekend
Household Density (households per 1000Sq ft.)	Call 1: Answering Machine
Median Yrs Education	Call 1: Incentive Offered in Answering Machine Message
Log(Median Income)	Call 1: Left Message on Answering Machine
Census Region	Call 2+: Weekday Day
% 18-24	Call 2+: Weekend
% 25-34	Call 2+: Answering Machine
% 35-44	Call 2+: 1 Day after Previous Call
% 45-54	Call 2+: 2 Days after Previous Call
% 55-64	Call 2+: 3 Days after Previous Call
% 65+	Call 2+: 4 Days after Previous Call
% White	Call 2+: 5+ Days after Previous Call
% Black	Call 2+: Incentive Offered in Answering Machine Message
% Hispanic	Call 2+: Left Message on Answering Machine
% Owner Occupied	

In the second-stage model, the components of the protocol are used to predict the probability of contact. Table 3-3**Error! Reference source not found.** shows the predictor variables S_{ij} that were used in the second-stage model to define the protocols in these models.

Table 3-3: Variables Defining the Protocols (S_{ij})

VARIABLE	DESCRIPTION
Timing of Call	Day of week and time of day; 3 different windows
Time between calls	Same day, 1, 2, 3, 4, or 5 or more days
Answering Machine/Incentive	No Message, message, message with incentive offer

There are three call windows, six lag times between calls, and three approaches to answering machines giving a total of 54 (3x6x3=54) possible protocols. Due to sample size limitations, I needed to limit the overall number of protocols. I did this by creating only three call windows. Other authors have considered more than three (Dennis et al., 1997; Kulka and Weeks, 1988; Weeks et al., 1980). In an effort to define the most useful call windows, I broke down contact rates for each day of the week and each hour of the day using three years of data from the RDD survey. I then grouped these hours back into useful windows by "merging" hours with similar contact rates until I had three windows left. For the lagged number of days between calls, I applied a similar approach. I used the same three years of data to calculate contact rates by number of days after the previous call. These rates leveled off after five days, so I grouped all calls that were lagged five or more days after the previous call together. I decided to include answering machine

messages as part of the protocol as there is contradictory evidence on the utility of these messages. Xu et al. (1993) conclude that these messages help increase participation rates, while Link et al. (2003) conclude that these messages are not helpful. A message left on a previous call is considered to be part of the strategy for the next call since that is the call most likely to be affected by a message. For example, a message left on the first call is a part of the protocol for the second call.

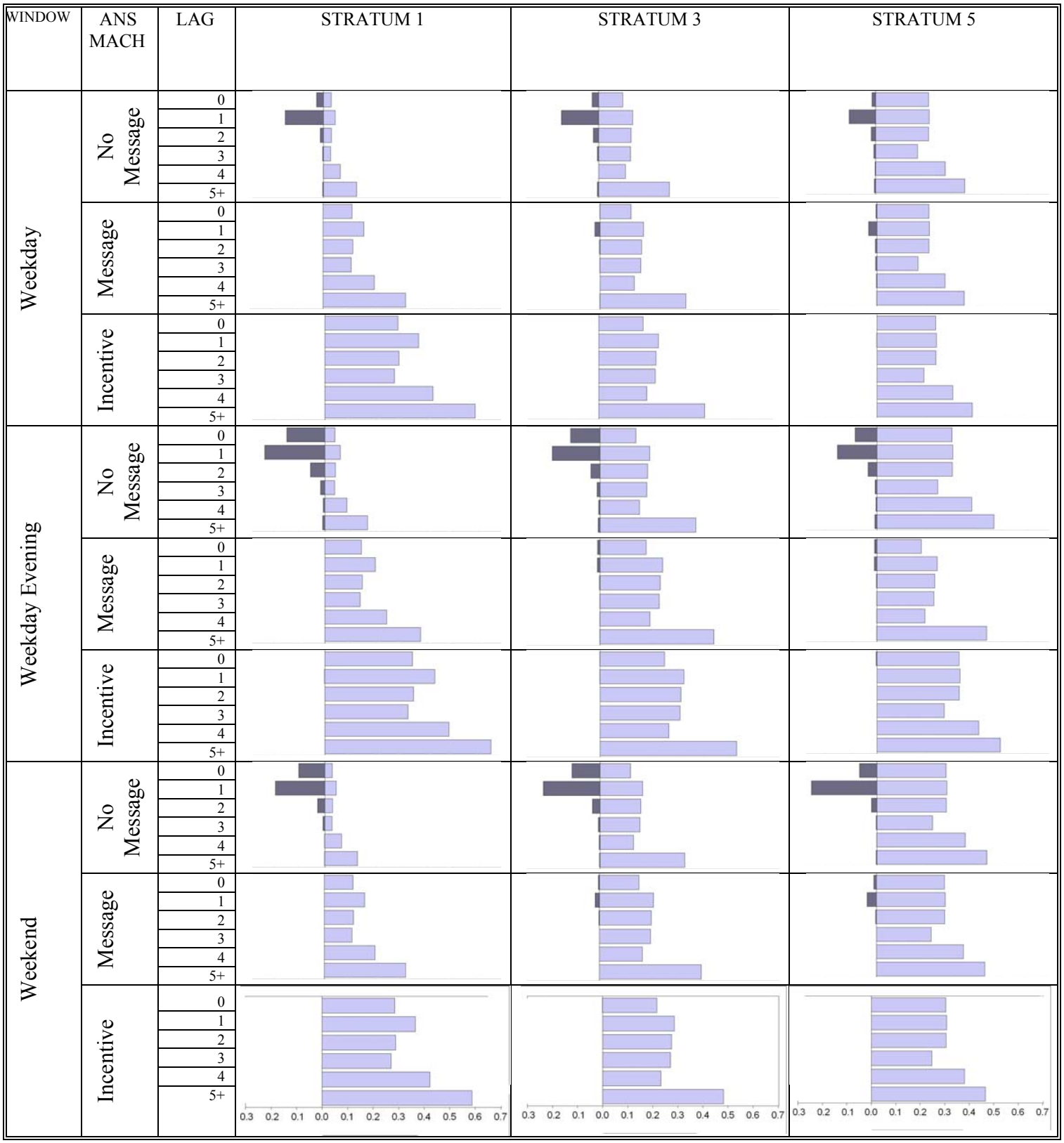
4.1.3 Results

The posterior distribution of the coefficients in the logistic model was simulated using the Metropolis algorithm (Gelfand and Smith, 1990; Tierney, 1994). Three chains of 10,000 iterations were run with 1,000 burn-in iterations. I used the Gelman-Rubin statistic to judge the convergence of these chains (Gelman and Rubin, 1992). Convergence is generally considered to have been achieved if this statistic is close to 1.0. In this analysis, the Gelman-Rubin statistic was 1.05 at its highest and was generally less than or equal to 1.01.

Figure 3-1 display the results for the analysis of the second call attempt. The light-colored bars to the right of the axis show the average posterior probability of contact for each of these protocols. The protocols are defined in the columns on the left. The first column with a graph shows the average posterior probability of contact for the lowest propensity stratum (stratum 1). In this stratum, there is a great deal of variation in the probabilities, suggesting that the choice of protocol does have an impact on the probability of contact. The second column with a graph shows the average posterior probability of contact for the middle propensity stratum (stratum 3). The third column with a graph shows the results for the highest propensity stratum (stratum 5). In the

highest propensity stratum, there is very little variation in the success of the various protocols. This result implies that in this stratum, the choice of protocol has less impact on the probability of contact.

Figure 3-1: Average Posterior Probability of Contact for Each Second Call Protocol by Propensity Stratum



The dark blue bars to the left of the axis show the proportion of cases that actually received each protocol. The actual protocols do not vary much across the strata. Most of the calls were placed on weekday evenings and weekends; no answering machine messages were left; and this call (the second call) was placed 1 to 2 days after the first call. In the medium and highest propensity strata (strata 3 and 5), these protocols are about as effective as any protocol that does not include the offer of an incentive. In other words, the answering machine message or time between calls seems to have little impact on contact rates. In the highest propensity stratum, even the choice of call window seems to have little impact on the probability of contact. In sum, it seems that the protocols that were actually used are tailored to the average or highest propensity cases. For these cases, an answering machine message or longer lag time between calls seems to have little impact on the probability of contact. However, these protocols are less effective in the lowest propensity stratum. For instance, less frequent calling and leaving messages on answering machines would do better in the lowest propensity stratum. This finding provides some confirmation of the Leverage-Saliency theory. In practice, a strategy that is equally effective for the high and medium strata is applied to the lowest propensity stratum. Thus, a tailored approach should improve the efficiency of the effort to contact sampled units.

In order to assess the increased efficiency of these strategies tailored to each of the propensity strata, I compared the empirical contact rates to the estimated contact rate of the protocol with the highest probability of contact. The protocol with the highest probability of contact varied across the propensity strata. However, the protocol with the

highest probability of contact often included an offer of an incentive or a long lag time (5 or more days) between the first call and the second call. The incentive may not be cost effective for establishing contact. Therefore, we may want to consider the impact of not offering it. In addition, waiting five days between calls is operationally difficult as it requires a long field period with low staffing levels. Therefore, Table 3-4 includes the estimated probability of contact under the protocol with the highest probability of contact, and also the same protocol with no incentive offer, the same protocol with a shorter lag between calls (the length of the lag is in parentheses), and the same protocol with no incentive and a shorter lag time between calls. Each of these protocols is estimated to achieve a higher contact rate than was achieved empirically.

Table 3-4: Achieved and Estimated Maximum Contact Rates by Propensity Stratum

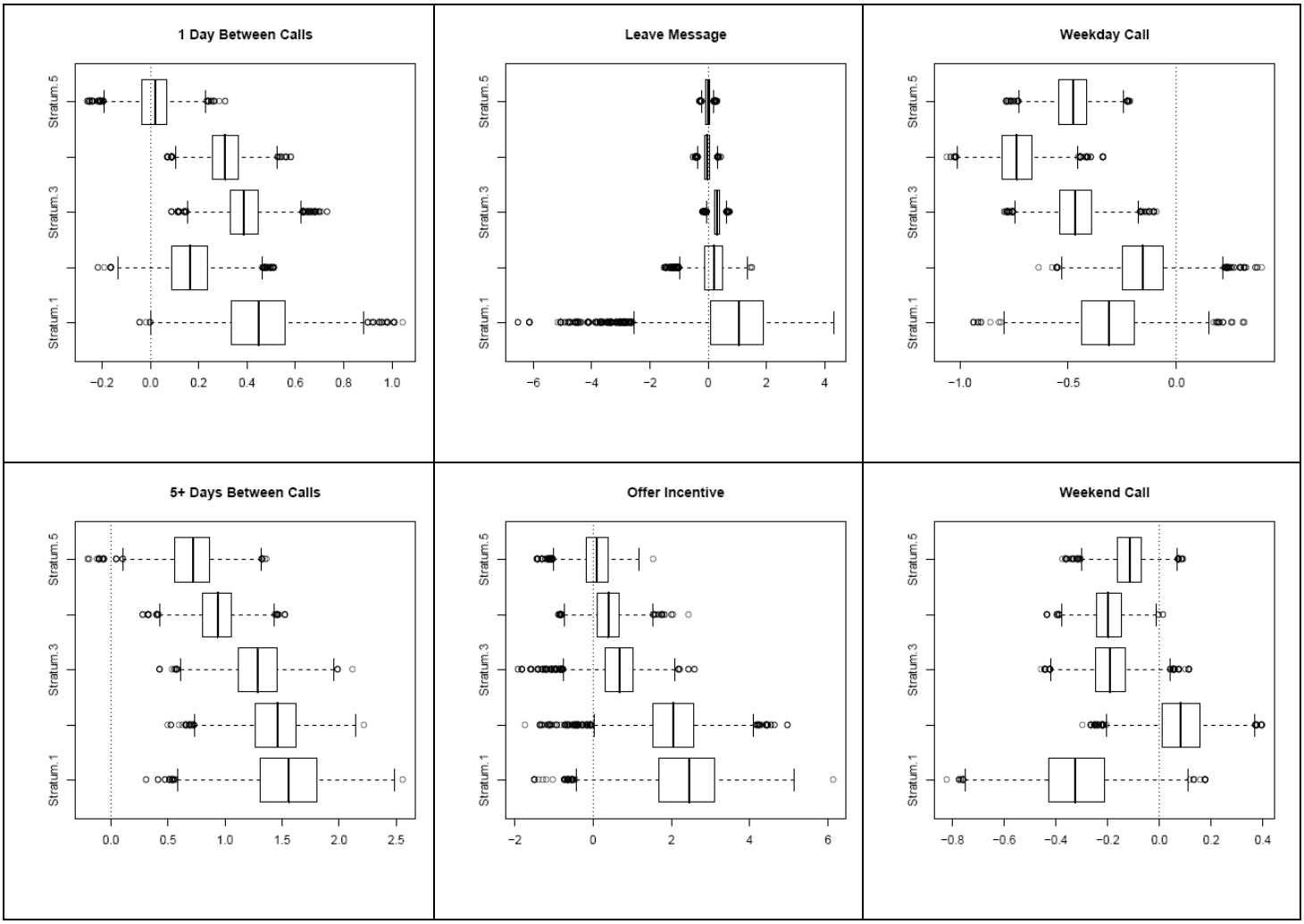
PROPENSITY STRATUM	ACHIEVED CONTACT RATE	MAXIMUM CONTACT RATE			
		Best Strategy	No Incentive	No 5+ Day Lags (Best Lag)	No Incentive/No 5+ Day Lags (Best lag)
1 Low	0.05	0.17	0.16	0.06 (1)	0.06 (1)
2	0.08	0.28	0.26	0.12 (2)	0.10 (2)
3	0.16	0.48	0.40	0.25 (1)	0.22 (1)
4	0.24	0.47	0.43	0.35 (2)	0.32 (2)
5 High	0.28	0.49	0.47	0.40 (4)	0.38 (4)
TOTAL	0.20	0.47	0.43	0.30	0.27

It is encouraging to note that there are simple changes that could be made to produce efficiencies in establishing contact. These changes involve changing the timing of calls, the delays between calls, and whether a message is left on an answering machine. Since these changes sometimes go in the opposite direction for cases in different propensity strata, there is hope that they would not require impossible staffing

plans in order to be implemented. A rule that said call every case on the first day and then call all nonfinal cases back on the second day, for example, would be difficult to implement for just this reason.

Figure 3-1 does not clearly illustrate whether the best strategies for achieving contact differ across the strata. Figure 3-2 does show the differences in strategy by stratum. In this figure, a dotted line indicates where zero is. Coefficients greater than zero represent protocols that tend to increase the probability of contact; those less than zero decrease the probability of contact. Each box-and-whisker plot represents the coefficient for a stratum. Stratum 1 is the lowest propensity stratum while stratum 5 is the highest propensity stratum. Looking at the posterior distribution of the coefficients for a weekend call (bottom right graph), it appears that for strata 1, 3, 4, and 5 most of the posterior distribution of the coefficients is less than zero, indicating that for these strata, a weekend call may be less effective than a call at other times. On the other hand, in stratum 2, most of the posterior distribution of the coefficient is positive, indicating that a call in this window may increase the probability of contact for cases in this stratum.

Figure 3-2: Posterior Distribution of Coefficients for Selected Components of the Second Call Protocols by Propensity Stratum



It seems that leaving a message has a positive impact on the probability of contact in the lowest propensity stratum (the stratum 1 posterior distribution of this coefficient is largely positive) and a negative or no impact in the other strata. The posterior probability that a message left on an answering machine will increase the probability of contact for a case in stratum 1 – the lowest probability stratum – (i.e., $\Pr(\beta_{Str1,AM} > 0)$) on the next call is 0.76 in stratum 1. The posterior probability for this protocol is 0.41 in stratum 5, the highest probability stratum. A similar result can be seen for placing a weekend call. This

is a better strategy for lower propensity cases (propensity stratum 2). We might speculate that persons in this stratum are at home more on the weekends than during the week, or that they screen their calls less on the weekend than during the week.

Table 3-5 summarizes the posterior probabilities that a coefficient is positive for the second call protocols. The table shows the probability that the impact of this protocol in this stratum will be to increase the probability of contact. For example, there is a 3% chance that if the second call to a stratum 1 case (lowest propensity stratum) is placed on a weekday day that this will increase the probability of contact relative to placing the call in a weekday evening window. For some protocols, we do not see as marked a differentiation in best strategies across the strata. In general, waiting longer between calls seems to have a higher probability of contact.

Table 3-5: Posterior Probability that a Coefficient for a Component of the Second Call Protocol is Positive

Stratum	Weekday	Weekend	Message	Incentive	Days Between Calls				
					1	2	3	4	5+
1	0.03	0.02	0.76	0.99	1.00	0.54	0.42	0.90	1.00
2	0.14	0.78	0.66	0.99	0.94	0.99	0.59	0.95	1.00
3	0.00	0.02	0.99	0.88	1.00	1.00	0.90	0.58	1.00
4	0.00	0.00	0.42	0.83	1.00	1.00	0.94	0.93	1.00
5	0.00	0.04	0.41	0.61	0.60	0.52	0.08	0.90	1.00

These results may be refined by matching strategies to specific covariates rather than the propensity score, which is a scalar summarization of all the covariates. Such refinements may help reduce the variability in the estimated coefficients seen in Figure 3-2. Unfortunately, even with our relatively large sample size, I was not able to do this. It is the case that some protocols are rarely used.

These analyses indicate that there are efficiencies to be gained. My hypothesis that different protocols will interact with demographic characteristics of potential

respondents as well as with previous treatments were at least partially confirmed. In addition, given the variety of best treatments and the nearness (in terms of probability) of next best treatments, these data suggest that a partial set of rules based on the results of these analyses would be operationally feasible.

This is all based on the relatively "thin" data available about RDD telephone numbers before contact. It seems likely that the data available after contact has been established, or in face-to-face surveys, would be even more powerfully applied under this approach.

This application has been limited to the sequence of protocols applied as we attempt to gain contact with sampled households. However, contact is only one stage in the survey process. There are also screening, refusal conversion, and interviewing steps to be completed. My approach has been myopic in that I only attempt to maximize contact without regard for how this might impact later stages of interviewing. For example, we can surmise that calling at 3am might be very successful at gaining contact, but that this protocol would greatly decrease the probability of interview. Carton and Loosveldt (1999) consider how the quality of the initial contact influences final response rates. I have considered the sequence of steps aimed at achieving contact. Further research is needed into the whole sequence of the survey process and how the protocols at each stage (e.g. screening) interact with those applied on other stages (e.g. refusal conversion or interviewing) of the process. The dynamic treatment regimes approach offers a roadmap for this research might be conducted. The results developed here suggest that such a research program could be successful.

4.1.4 A Proposed Learning Algorithm

My analyses are based on observational data. These analyses do, however, suggest a useful experimental strategy. I propose using the posterior densities of the model coefficients to randomly assign strategies to cases. Strategies would be assigned with probability proportional to the probability that the strategy is the one with the highest probability of achieving contact. Sutton and Barto (1998) term this a "softmax" strategy. For example, if in propensity stratum 2, there is a 78% chance that a weekend call would do better than a call placed at another time, then 78% of the cases in this stratum would be randomized to a weekend call. If the results are confirmed by the experiment, then when the posterior distributions are updated with the new data, the proportion of sample assigned to this protocol in the next iteration of the survey will grow.

This approach could be used even when very little or no data are available prior to fielding a survey. Weak priors could be used in order to give higher probability to strategies that are assumed to be better (e.g., weekend and evening calling). A weak prior would allow the experiment to explore strategies. As data accrue, successful strategies would have larger proportions of cases assigned to them.

This approach implies learning or adaptation at two levels. First, in the short term, it could be used to experimentally identify strategies that are tailored to the specifics of the case. The strategy would hone in on the best strategies as data accumulate. Second, in the long term – and assuming that some exploration is always allowed to happen – the strategy could adapt to a changing social environment in which the best strategy of today may not be the best strategy of tomorrow.

4.2 Application Two: Mixed-Mode Panel Survey

4.2.1 The Survey: The Health and Retirement Survey

A second application of this approach is considered for a mixed-mode panel survey. The Health and Retirement Survey (HRS) is a national sample of adults who have retired or are approaching retirement. The survey collects a large amount of financial and health information on each respondent. HRS attempts to interview panel members every two years. In 2006, the year I will be looking at, a random half of all panel members were selected for a face-to-face interview. Interviewing in this mode was necessary in order to allow the collection of physical measures on respondents – for example, height, weight and blood spots. For cases that are scheduled to be interviewed face-to-face, interviewers are encouraged to call by telephone and set an appointment for the interview in order to save travel costs. The panel is updated with new cohorts as they approach retirement age.

The goal of the following analysis is to evaluate strategies for establishing contact with households. However, in this case, I will only consider the timing of the call. Most previous research has focused on the average best time to call (Weeks, 1980), the best three-call sequence (Weeks, 1987), or the best nine-call sequence (Cunningham et al., 2003). The articles by Greenberg and Stokes (1990) and Brick et al. (1996) are important exceptions. Greenberg and Stokes consider the best time to call, conditional on the timing and outcome of previous calls. Brick et al. (1996) consider the best timing of a call conditional on the outcome of previous call as well as Census data from the estimated geography of the telephone number. Several studies have considered interviewer behavior in the context of face-to-face surveys (Snijkers, 1999; Purdon, 1999; Bates,

2003; Wang et al., 2005) where the timing of the call is determined by the interviewer. These studies note that the timing of the call is important in establishing contact. They also note that interviewers have different levels of ability in terms of determining when the best time to call is. Potthoff et al. (1993) attempt to model contactibility, but for the purpose of developing a weighting scheme that corrects for bias due to noncontact.

In fact, we are not interested in the average best time to call. When faced with the decision about when to time the next call for a case, the average is simply an estimate that does not use any information about the specific household. What we really want to know is the household-specific probability that someone will be at home (and answer the door or telephone) at different times of the day and days of the week.

In the area of marketing research, Rossi, McCulloch, and Allenby (1996) consider a similar problem. Their goal is to customize or tailor the face value of a coupon to a specific household. They attempt to estimate household-level parameters using demographic and purchase history data. Other researchers in the area of marketing are considering a similar problem in the context of advertising on the internet (Gooley and Lattin, 2000; Simester, Sun, and Tsitsiklis, 2006; Bertsimas and Mersereau, 2007). Bult et al. (1997) consider a method for estimating interactions between the target population and the characteristics of a fund-raising mailing that can be used to optimize these mailings.

Bollapragada and Nair (2001) consider the problem of improving "right party contact" rates at credit card collection calling centers. They conceive of the problem in a manner similar to that of Rossi and his colleagues. In other words, their goal is to estimate contact probabilities for each household the call center is attempting to reach.

They develop an algorithm that begins from the average contact rates and adjusts these starting values upward for each household when a call attempt is successful and downward when the attempt fails.

4.2.2 Data and Methods

In contrast with the situation with an RDD survey, this survey has a rich set of data on panel members. These data include both the survey items and the paradata collected from previous waves. HRS attempts to collect an interview every two years with panel members. These interviews include demographic data, health and financial measures. In addition, a record for every call is kept. This record includes the time and date of the call as well as a result code indicating the outcome of the call. I used data from the HRS (persons born between 1931 and 1941), CODA (Children of the Depression, persons born 1923-1930), and War Baby (1942-1947) panels (Leacock, 2006). There were 12,368 households included in the analysis. Table 3-6 shows the variables selected for use in the models. These variables are available on the "Tracker" file provided by the HRS. This file summarizes these demographic variables for each panel member for each wave of the survey. Of course, many more variables would be available if additional data from the survey itself were included. However, the variables in Table 3-6 are predictive of our ability to contact households (Groves and Couper, 1998; Brick et al., 1996; Dennis et al., 1998; Nicolleti and Peracchi, 2005). In particular, in the panel survey setting, information from previous waves of the panel is highly predictive of outcomes in subsequent waves (Nicolleti and Peracchi, 2005; Lynn et al., 2002). The call records are used to determine the timing of the call and whether contact was made.

Table 3-6: Variables Used in Models Predicting Contact in 2006

Demographic Variables (X_i)	**Call Record Variables**
Age	Date of Call
Race	Time of Call
Ethnicity	Result Code
Marital Status	Mode of Call
Living in Nursing Home	
Interview in 2004	
Final Refusal in 2004	
Education	
Total Calls in 2004	
Couple Present in Household	
Sex of Panel Member	

The goal is to determine the best timing of the next call. Therefore, I am simulating the estimates we would have had after each call of the 2006 survey. I estimate the probability of contact in each of six windows using multi-level logistic regression models. These models were estimated using the call records and demographic data described above. The models were first estimated using just the call records from 2004. In some households, not all of the six windows were called. For those cases, the estimate of the probability of contact was based on the average intercept and covariates.

We have call records dating back to 1992 for many of these households. I could have used these data, but would either have had to assume that the contact probabilities were stable over that time period or would have assumed some time series model. Instead, I assumed that the most recent wave (2004) would be the most informative and used only those data. I also ignored the mode of the call. To include it in the models would have led me to be prescriptive regarding the mode. I estimated the models with the mode variable and got similar results. Therefore, I decided to leave it out.

In addition, I assumed that the calls were independent, random draws from a Bernoulli distribution of the probability that the household could be contacted. In order to make this assumption more plausible, I deleted any calls that were set as appointments. Since I am assuming that these were independent trials, I did not enter the call number as a predictor into these models. This assumption enables me to estimate the probability of being at home for any call in that window. I did not want to estimate, for example, the probability of being at home after eight calls of a particular sequence.

After estimating the models using the 2004 data, I added the first call from 2006 and re-estimated the models. I continued this iterative process through the first ten calls of the 2006 survey. Many cases are finalized before the first ten calls have been placed. For cases with multiple calls, the successful or unsuccessful call attempts add data that can change our estimate of the probability of contact for each household and window combination.

The method proposed here extends and develops the approaches of Rossi et al. (1996) and Bollapragada and Nair (2001). First, rather than adjust estimated probabilities after each call by a fixed amount, I model the contact probability using a multi-level model patterned after Rossi et al. This approach allows us to assess our uncertainty about these estimates. This measure of uncertainty may be useful as we develop policies for timing of the call. Second, I involve demographic covariates in the model. Third, I propose learning algorithms that may help prevent getting stuck in a "dead end." Finally, in Section 5 I suggest how the policies might be implemented in a decentralized situation where interviewers make decisions about the timing of the call as opposed to having those decisions be made by a centralized software.

4.2.3 Results

The estimated and actual contact rates are presented in Table 3-7. The models (updated after each call placed in 2006) estimated that 52.6% of the calls would lead to contact. In reality, 53.8% of calls did lead to contact. However, if we pose the hypothetical, what would the contact rate have been had the interviewer always called the window with the estimated maximum probability of contact, the estimate is 69.8%. This indicates that there is room for increased efficiency in this survey.

Table 3-7: Actual, Estimated, and Estimated Maximum Contact Rate

Actual Contact Rate	Estimated Probability of Contact	Estimated Maximum Probability
0.526	0.538	0.698

Unless our estimates change from call to call, this method has very little chance of improving contact rates over a strategy that always called the window with the highest probability of contact estimated from a previous wave. If the same window is always estimated to be the window with the maximum probability of contact, then the added information is not leading us to adaptively change our strategy. Therefore, we hope to see changes in the estimate of the window with the highest probability. Table 3-8 shows the percentage of cases having the highest probability of contact after each call. The first column, labeled "Pre-Call 1" is the percentage of cases that were estimated to have each window as the window with the highest probability of contact. For example, 10.7% of cases prior to the first call had the maximum probability of contact for a call placed in Window one (Weekday Morning). However, after the first call placed in 2006, the

percentage of cases that have the maximum probability of contact in this window has grown to 16.6%. This indicates that our estimates are "adapting" to the new information.

Table 3-8: Estimated Percentages of Households that have the Maximum Probability of Contact in Each Window after Each Call

Call Window	Percentage of Cases Estimated with Maximum Prior to Each Call										
	Pre-Call 1	Pre-Call 2	Pre-Call 3	Pre-Call 4	Pre-Call 5	Pre-Call 6	Pre-Call 7	Pre-Call 8	Pre-Call 9	Pre-Call 10	Pre-Call 11
1: Weekday Morning	10.7	16.6	16.1	15.9	16.4	16.5	16.4	16.5	16.5	16.4	16.5
2: Weekday Afternoon	13.7	24.3	23.5	23.8	24.2	24.5	24.5	24.6	24.5	24.6	24.8
3: Weekday Evening	64.0	41.2	43.4	43.8	43.3	43.1	43.1	42.9	42.9	42.9	42.6
4: Weekend Evening	6.7	7.9	7.8	7.7	7.4	7.3	7.3	7.3	7.6	7.5	7.4
5: Weekend Morning	2.2	4.6	4.5	4.4	4.4	4.4	4.4	4.3	4.3	4.3	4.4
6: Weekend Afternoon	2.7	5.4	4.7	4.4	4.3	4.2	4.3	4.3	4.3	4.3	4.3

Although Table 3-8 presents the net changes, it is clear that as we update the data, our estimate of the window with the maximum probability is changing. The most dramatic change occurs after the first call. More than 20% fewer cases have window three (weekday evening) as the maximum probability window. Windows one (weekday morning) and two (weekday afternoon), on the other hand, have experienced large jumps. Again, the models that estimate these probabilities condition on the fixed characteristics described in Table 3-6. However, after each call, they add the success or failure of that call to the set of information included in the estimates for the next call. In this way, this method is somewhat akin to the method of Bollapragada and Nair (2001). They increase or decrease the probability of contact in each window for each household by a fixed multiplicative factor after each successful or failed contact attempt. In my method, the

hierarchical models re-estimate this "multiplicative" quantity after each call over the entire dataset.

In order to assess these changes at a household level, I looked at the proportion of cases that had changes in the estimate of the window with the maximum probability of contact. Table 3-9 shows the proportion of cases that had at least one change over the course of the first 10 calls. About half the cases never experienced a change in the estimated maximum probability window. The next largest proportion, about 41%, had one change over the course of 10 calls. The remaining 10% had more than one change. In sum, a slight majority of cases have a changed estimate of the maximum probability window that results from the accrual of data. This means that there is adaptation occurring. This adaptation will allow the search for contacts to be more efficient.

Table 3-9: Number of Changes in "Maximum Probability Window" over the First Ten Calls

Number of Changes	Count	Percent
0	6122	49.50
1	5051	40.84
2	674	5.45
3	381	3.08
4	82	0.66
5	44	0.36
6	11	0.09
7	2	0.02
9	1	0.01

Most of these changes are occurring quickly in the data collection process. Table 3-10 shows the proportion of cases that change the estimate of the maximum probability window after each call.

Table 3-10: Proportion of Cases Changing "Maximum Probability Window" by Call Transition

Call Transition	Proportion Change Maximum Prob. Window
1 to 2	0.471
2 to 3	0.057
3 to 4	0.038
4 to 5	0.029
5 to 6	0.021
6 to 7	0.025
7 to 8	0.017
8 to 9	0.018
9 to 10	0.015
10 to 11	0.013

This rapid change might be explained by several factors. First, there is an indicator variable in the model so that different waves will be allowed to have different effects. The impact of this fixed effect is averaged over all the cases. Second, life changes might result in changed at-home patterns that are picked up as attempts occur. Finally, the first call attempt is received by all sample members and quite often is successful (in terms of contact and interview). This may lead to a shift in estimate of the maximum probability window for cases that actually have high probabilities of contact in all windows.

The model did not include call number. The assumption of the model is that each call is an independent trial of a Bernoulli process. Under this assumption, the contact probability should be the same at each attempt. If we look at the distribution of contact rates and estimated contact rates by call number shown in Table 3-11, we notice that the actual contact rates are decreasing. In fact, those households that require more attempts are generally more difficult to contact regardless of the timing of the call.

Table 3-11: Actual, Estimated, and Estimated Maximum Probability by Call Number

Call Number	Actual Contact Rate	Estimated Probability of Contact	Estimated Maximum Probability
1	0.524	0.593	0.805
2	0.572	0.511	0.677
3	0.545	0.518	0.671
4	0.528	0.524	0.665
5	0.525	0.524	0.661
6	0.505	0.528	0.657
7	0.506	0.526	0.653
8	0.496	0.527	0.650
9	0.460	0.515	0.645
10	0.440	0.522	0.643
11	0.439	0.522	0.641

It seems clear that the model fits the data well since the model estimates correspond well to the empirically observed contact rates. In practice, this procedure provides household-level estimates of contact probabilities that can be used to guide data collection strategies. The evidence indicates that contact rates can be improved if the model estimates are used to guide data collection. This should improve the efficiency of

the survey since fewer calls that do not result in contact will be made. These decisions are currently left in the hands of interviewers. They must use their best judgment and experience when deciding which window is the best time to call for each household. Interviewers also see each housing unit and have, in this way, more "data" on each household with which to shape their judgments. However, interviewers only have data on the cases that are part of their workload. They do not see the full sample. A statistical rule such as the one proposed here has the benefit of being able to use data from the entire sample. These data are not available to the interviewers. The statistical model can average over all the cases in estimating the parameters. These reasons may explain why the statistical model does better at estimating the best time to call compare the interviewer decisions.

However, these data were gathered by interviewers who used their own judgment about when to place the call. If those decisions had been guided by a statistical rule, then a different set of calls would have been made. Would these decisions have led to the relatively quick estimation of the best time to call that we see in this analysis of observational data? It is possible that the interviewers are using a learning algorithm that is somewhat less efficient at first, but gives us the data we need in order to more quickly identify the best time to call. This raises the question about how to make these decisions adaptively such that over time the optimal strategy (in this case, timing of call) is identified quickly. The next section deals with this question.

4.2.4 Potential Learning Algorithms

The challenge that interviewers face is achieving contact in as few calls as possible. They are not trying to determine each household's probability of contact for

each of the different windows. If they establish contact on the first call, they will not need to attempt the other windows to see if they have a higher probability of contact. In the context of a learning algorithm, this is an important distinction. Given that we generally are not certain of when people will be home, establishing contact is a process of trial and error. The question is which process achieves our goal of minimizing the number of calls required to achieve contact.

Several algorithms for learning have been proposed (see Sutton and Barto, 1998 for a review). The tension is between the competing goals of exploitation and exploration. "Exploitation" strategies seek to maximize short-term gains without regard for long-term consequences. "Exploration" strategies will sacrifice short-term gains in the interest of gaining information which may increase long-term gains.

I did simulations of two algorithms to see how they performed in terms of learning to establish contact. The first algorithm is to always call the window with the estimated maximum probability of contact. This is an "exploitation" algorithm which may, in the end, explore very few options. The second algorithm uses the 95% confidence limits around the estimate of the probability of contact. The second algorithm always calls the window with the highest upper confidence limit on the estimate of the probability of contact. Although this is a deterministic strategy (i.e. there is no randomization in the selection of a next strategy) it is still an "exploration" algorithm since it encourages the use of protocols about which we are uncertain but which have some chance of being the best action. In this case, the latter algorithm will call windows that have relatively high estimated probabilities of contact (but not necessarily the highest) and relatively high variance. In other words, our uncertainty about these cases is

greater and prevents us from ruling out the hypothesis that this window has the highest probability of contact. This strategy will call a window until our uncertainty about its estimated probability of contact is reduced to the point where other strategies have higher upper confidence limits.

The second algorithm has shown to be optimal in long-run situations (Lai, 1987). The shortest run that Lai considers is 100 trials. However, our situation is generally a much shorter run. The average number of calls to completion in HRS is 7.0. Table 3-12 shows the percentage of cases completed at each call number in the 2006 survey. The median number of calls to finalization is four calls. About 83% of cases are finalized in 10 or fewer calls. For most cases, the number of calls required to finalize the case is a very short run relative to the "short-runs" considered by most learning approaches. In the extreme, if we only planned to make one call, we should always choose an exploitation strategy as there will be no ability to use anything learned from the call.

Table 3-12: Actual Distribution of Number of Calls Required to Finalize a Case (HRS 2006)

Calls to Finalize	Frequency	Percent	Cumulative Percent
1	438	1.95	1.95
2	4100	18.26	20.21
3	3913	17.43	37.64
4	2901	12.92	50.56
5	2115	9.42	59.98
6	1630	7.26	67.24
7	1221	5.44	72.68
8	1049	4.67	77.36
9	743	3.31	80.66
10	620	2.76	83.43
11	497	2.21	85.64
12	423	1.88	87.52
13	318	1.42	88.94
14	294	1.31	90.25
15	229	1.02	91.27
16	184	0.82	92.09
17	182	0.81	92.90
18	151	0.67	93.57
19	130	0.58	94.15
20-29	766	3.41	97.56
30-39	308	1.37	98.94
40-49	116	0.52	99.45
50+	123	0.55	100.00

In order to compare learning methods, I have developed Monte Carlo simulations for three approaches. Two of the approaches were described above: the exploitation approach of always calling in the window with maximum estimated probability, and the

exploration approach of always calling the window with the highest upper confidence limit. The third approach is to always randomly choose a window for the next call.

The Monte Carlo simulations were set up in the following manner. First, "true" contact probabilities were generated for each of six call windows for 1,000 simulated households (these probabilities are denoted π_{ij} for the i^{th} window and j^{th} household) using a beta distribution. This led to 6,000 independent draws. Second, the average probabilities for each window ($\bar{\pi}_i$) were calculated across the 1,000 households. These averages were used as starting values for our estimate of the contact probability of each household. In other words, each household had the same starting estimate. This value was the true population average. These starting values were used to specify the parameters of a beta distribution. The average was used as the estimate of the mean of the distribution. Since the beta distribution variance is a function of the "sample size," I tried different prior sample sizes for the beta distribution. In the end, the initial α and β parameters for the i^{th} window and j^{th} household were set in the following manner:

$$\alpha_{ij} = 3 * \bar{\pi}_i$$
$$\beta_{ij} = 3 * (1 - \bar{\pi}_i)$$

where $\bar{\pi}_i$ is the population average contact probability in window i.

These prior assumptions were then updated with data by simulating the first call following the three algorithms specified above and updating the α and β parameters depending upon the success or failure of the call. The result of the call was determined by comparing a draw from a uniform distribution to the "true" value (π_{ij}) for the household and window combination. The beta distribution was updated with the result of the call. The new estimate of the contact probability is then the mean of the updated distribution:

$$\hat{\pi}_{ij} = \frac{\alpha}{\alpha + \beta}$$

The process was iterated through ten calls. The simulation was repeated 1,000 times with the same setup.

The first simulation used the contact probabilities from HRS 2006 as parameters for the beta distribution from which the true values (π_{ij}) were drawn. These contact probabilities are displayed in Table 3-13. The average probabilities across the six windows do not vary much.

Table 3-13: HRS 2006 Empirical Contact Rates by Window

Window	Contact Rate
1	.471
2	.498
3	.488
4	.440
5	.423
6	.439

The results of the simulation are displayed in Figure 3-3. The random strategy shows the average probability of contact. A strategy informed with knowledge of the true probabilities should do better than this line.

Figure 3-3: Simulation of Learning Algorithms Using HRS 2006 Empirical Contact Rates

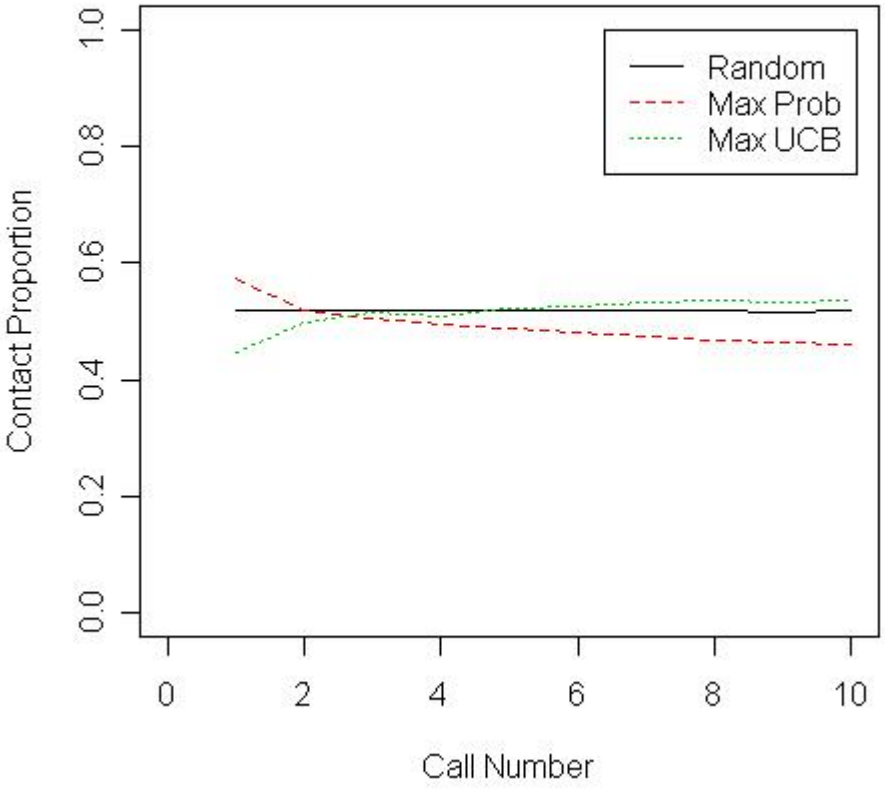


Table 3-14 shows the distribution of calls by window for each call number. From this table, it seems that the maximum estimate strategy moves quickly from the starting strategy of Window 2 (weekday afternoon), which is the maximum average strategy, into other strategies. The maximum upper confidence bound strategy, on the other hand, moves relatively slowly from Window 6 (weekend afternoon) into other windows. It stabilizes more quickly and has less movement between windows. It seems that the maximum estimate strategy quits trying an option too quickly and takes longer to settle on a particular window. This can be seen in Table 3-14. The "maximum probability" algorithm is very quickly spreading its effort over all the windows. Since it is relying on

overall averages for the starting point, it may start moving away from a window in a household that has a lower than average probability for that window too soon. The "maximum upper confidence bound" algorithm, on the other hand, stays with the current estimate longer. This reflects that it is not selecting another window to try until it has more certainty that the current window is not the best window.

Table 3-14: Percentage of Calls Placed in Each Window by Call Number and Algorithm

	"Maximum Estimate" Algorithm						Maximum Upper Confidence Bound" Algorithm					
	Call Windows						Call Windows					
Call	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
1	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
2	12.3%	12.4%	67.1%	4.0%	4.2%	0.0%	0.0%	0.0%	0.0%	3.3%	3.4%	93.4%
3	17.0%	17.1%	52.3%	6.3%	6.5%	0.9%	0.0%	0.0%	0.0%	16.9%	16.4%	66.6%
4	18.8%	19.0%	44.3%	8.0%	8.1%	1.8%	0.0%	0.0%	0.0%	17.3%	17.3%	65.3%
5	19.6%	19.8%	39.5%	9.2%	9.3%	2.7%	0.0%	0.0%	0.0%	13.4%	13.3%	73.3%
6	20.0%	20.1%	36.2%	10.1%	10.2%	3.4%	0.5%	0.5%	0.0%	15.5%	15.3%	68.2%
7	20.1%	20.2%	33.7%	10.8%	10.9%	4.2%	1.0%	0.8%	0.0%	17.6%	17.4%	63.1%
8	20.2%	20.3%	31.7%	11.5%	11.6%	4.8%	1.1%	1.1%	0.0%	17.5%	17.3%	63.1%
9	20.2%	20.3%	29.9%	12.0%	12.1%	5.5%	1.2%	1.1%	0.0%	16.5%	16.4%	64.8%
10	20.2%	20.2%	28.5%	12.5%	12.6%	6.1%	1.1%	1.0%	0.0%	16.7%	16.5%	64.7%

The question remains, over a short run, what calling strategy should we use to minimize total effort? In fact, in the panel survey example, we have prior information that may put us into the long-run situation. So, a long-term strategy may be appropriate. Otherwise, a combined strategy, of trying the maximum probability window early and then, once the "easy" cases have been contacted, moving into an exploratory strategy may work better for this type of problem. Figure 3-3 suggests a combined strategy that

goes from the maximum estimate strategy to the maximum upper confidence bound strategy at about the third call might work better.

In this section, I have shown two applications where conditioning the choice of the next treatment on the fixed characteristics of the case, as well as the history of previous treatments can increase efficiency. In addition, for each application I have suggested learning or adaptive models that may improve our ability to identify strategies matched to the characteristics of each case. These learning models go beyond the analysis of observational data conducted here. Although both of these applications are focused on contact, they deal with the process of establishing contact as a multiple step procedure. Future research could expand on this by treating additional steps or phases in the survey process.

While a statistical model may be useful for predicting the best time to attempt contact for each household, the decision about when to call the household is in the hands of a field interviewer. Interviewers are provided a general training that includes discussion of "best times" for contacting households. After training, interviewers are left to summarize their own experience and determine when to call each household. My analysis indicates that a statistical summary may be more efficient than this decentralized procedure. However, how will centralized "recommendations" of when to call be delivered to field interviewers? In the next section, I will propose a model for how statistically derived recommendations about which protocol to use can be delivered to decentralized staff.

5. Decision Support Model

In a telephone survey conducted using a centralized sample management system, it should be fairly simple to implement the learning approaches outlined above. The algorithm can be developed and updated in statistical software outside of the sample management system. The selected strategies can then be loaded into the sample management system either daily or continuously.

In a face-to-face survey, it is less clear how these statistically generated rules can be implemented. In the clinical trial realm, there is an analogous problem. Dynamic treatment regimes are often implemented as a complicated set of rules. Trivedi and Daly (2007) propose the use of "computerized decision support" to aid the implementation of these algorithms by physicians. The physician enters information about the patients current status and treatments into a computer using an interface to an electronic medical records database. This information is updated at each visit so that a history of treatments is developed over time. Based on this history and the patients fixed characteristics, a treatment recommendation is delivered to the physician.

There is a well-developed literature on decision support, particularly in medical applications (see Hersh and Hickam, 1998; Trivedi and Daly, 2007; Kawamoto et al., 2005; Sim et al., 2001; and Trivedi et al., 2002 for reviews). There have been numerous qualitative and quantitative evaluations of this approach. Kawamoto et al. (2005), in particular, identify traits of computerized decision support systems that are associated with implementations that improve health outcomes for patients. These factors include that the support system is part of the workflow, the users are involved in the design of the system, and that recommendations are delivered rather than just information. In addition,

Trivedi et al. (2002) note that it is useful to record whether the recommendation was followed and, if the recommendation was not followed, why not. They report that the recommendations were followed 85% of the time in the STAR*D trial for which they implemented their web-based decision support system. They also note that based on the notes recorded by the physicians that these deviations were often justified (Trivedi et al., 2007).

The computerized decision support model may provide a model for aiding interviewers in face-to-face surveys. Traditionally, interviewers in the field have been given very general recommendations in training. The decisions about timing of call, etc. are left to their discretion. A decision support model would offer treatment prescriptions to the interviewer as a recommendation. These recommendations should be delivered as part of the workflow in order to be used. The HRS currently uses "profiles" of respondents. These profiles are web pages accessed by interviewers on demand (as opposed to being part of the workflow). In addition, they only provide information, not recommendations. Only about 65% of cases ever have their profiles accessed and no impact on response rates or effort is discernible (Guyer, Wagner, Cheung, 2007). On the other hand, a flag indicating that a case should be prioritized that is part of the workflow and is a sort of recommendation does show improved response rates for the flagged cases (Groves, et al. 2008) as compared to an emailed recommendation of which cases should receive a higher priority. These results parallel those from a clinical trial comparing the impact of electronic alerts to those of an alert that is delivered on-demand (van Wyk, 2008).

Further work needs to be done testing a computerized decision support system, but such a system would help survey organizations direct interviewer effort in field surveys. Linked with adaptive survey design protocols, this might lead to improved efficiency, response rates, and reduction in the fraction of missing information. But even without such adaptive regimes, the decision support model could help survey organizations standardize interviewer behaviors.

6. Discussion

Tailoring is a widely accepted strategy for convincing sampled persons to respond to the survey request. Most frequently, this tailoring is done on the introduction, but survey organizations frequently tailor effort in other ways (e.g., increasing or decreasing incentives at different points) on an ad hoc basis. Unfortunately, the randomized controlled trial most often employed for exploring new methods does not fit well with the fact that surveys are inherently multi-stage processes. This research model does not easily allow the development of methods that involve sequences or combinations of treatments.

I have suggested that the dynamic treatment regimes approach developed in the framework of clinical trials may be useful for evaluating survey design protocols. The dynamic treatment regimes model allows us to consider the problem as a multi-stage problem. Further, this model gives us experimental and statistical means to explore "tailored" designs.

The two applications developed here suggest that these approaches can be successful. I have shown that tailoring the design to the characteristics of the case, including previous treatments can improve efficiency. In the case of an RDD survey, I matched protocols – sets of design features – to the characteristics of the cases, including

the history of previous treatments. These characteristics were summarized using the propensity score. The results showed that persons with different sets of characteristic respond differently to various protocols. The contact maximizing protocol was defined for each of the first few steps of the screening process.

In the second application, I showed that we can successfully tailor our protocol to the household. The data we had from previous waves of a panel survey, when combined with the accumulating data on treatments from the current wave, was able to learn which call window had the highest probability of contact for each household. This technique adapted the proposed policy (i.e. timing of the call) to the incoming data as well as the fixed demographic characteristics of each household.

These results are suggestive. Future research will need experimental validation of the results for each of these applications. In addition, I have considered only parts of the survey process. This approach can also be extended to all stages of the survey process. For example, there may be interactions between the protocol used for screening a case and the protocol used for interviewing a case. A thoroughgoing analysis of how the various protocols used for establishing contact impact the probability of an interview is needed. The dynamic treatment regimes approach gives a method for approaching these questions.

Finally, the question remains, how do we use these techniques to address nonresponse bias? In Chapter 2, I argued against using the response rate as a key metric. It is probably true that these tailored strategies could be used for the purpose of simply raising response rates. However, they become necessary if we want to do something beyond that. If we wanted to pursue cases other than the easiest-to-respond, then we

would need methods that allowed us to alter our protocol such that those cases were more likely to respond. Such a shift offers the possibility of survey research becoming much more focused on minimizing nonresponse bias, as opposed to simply minimizing nonresponse.

Chapter 4

Stopping Rules for Surveys

1. Introduction

Current sampling designs are conceived of as being fixed. After the sample size and (usually) the target response rate are set by the design, then the data collection takes these as targets. Short of a cost overrun, these targets remain fixed. The target response rate may be more difficult to achieve than expected and then a slightly lower response rate might be expected, but the number of completed interviews is usually considered to be a contractual obligation. There might be some plans for altering the size of the final released sample if the eligibility or response rates are lower than expected, but information from the accumulating data is rarely used to modify these sample designs. This information includes the survey data, the relationship between the survey data and the complete data on the frame, and characteristics of the nonresponders using the frame variables. In effect, this approach assumes that the risk of nonresponse bias is low at the specified response rate and that sound adjustment strategies will be available using the complete data on the frame. However, these assumptions are often not checked until after all the interviews are completed.

If we were monitoring these data as they accumulate, then we might use them to help answer the question, when should we stop collecting data? These data could be used to modify our original assumptions about the variances of key survey statistics. Further, these data could also be used to explore the possibility that it does matter which set of respondents we get. The current "stopping rule" is to stop data collection at a specified number of interviews and response rate. This rule creates a situation in which it makes sense to collect data from those easiest to interview, without regard to whether they contribute anything new to our estimates. As we have become more concerned with the possibility of nonresponse bias, new stopping rules are needed that rely on information about the current set of responders and nonresponders. In this chapter, I will suggest a rule for stopping that attempts to account for the risk of nonresponse bias. The decision to stop will be based on the data we are collecting as well as the complete data on the frame. The rule will suggest stopping when the probability that further data will change our adjusted estimates is very low. Such a rule would provide a new context for survey data collections. Instead of seeking to add the easiest cases, such a rule would encourage data collection agencies to interview cases that reduce the risk of nonresponse bias most quickly. In this chapter, I will propose a stopping rule for surveys that attempts to account for the risk of nonresponse bias.

1.1 Current Sampling Practice

Much of sampling theory was developed as if nonresponse were a tangential problem. Sampling theory in its early development assumed 100% response rates. This was a very reasonable approach. It is certainly much simpler to design samples under this assumption. Dealing with nonresponse requires some sort of model assumption. Early

sampling statisticians were very wary of using models (Hansen et al., 1983). They were mainly concerned with establishing the principle that probability sampling should form the basis for inference. An additional barrier was that data about empirical response rates, eligibility rates, and the survey variables themselves were not generally available until after the survey was completed. As a result there was no possibility of incorporating this information into an adaptive design.

Today, we confront a new situation. First, nonresponse is an ubiquitous problem. Although sample designs are typically built using an assumed response rate and are often adjusted in size when the empirical rates differ from the expectation, the sample design does not usually account for the characteristics of the empirically achieved set of responders. In other words, the data collection process is not guided by the characteristics of the nonrespondents. The data collection occurs as if it does not matter who responds. Designs adapted to information about the current set of nonresponders are now needed. Second, the computerization of survey data collection allows that information about the current set of responders and nonresponders is available to survey designers on a real-time basis. This information can be used to adapt the sample design – not just to achieve the targeted sample size, but to achieve specified goals for the precision of estimates in the presence of nonresponse.

Of course, sample designs need to make some prior assumptions about key parameters in order to create an initial sample design. However, these assumptions can be compared to incoming data during the field period. The design assumptions can then be updated as the data accumulate. In addition, we can involve variables from the sampling frame to estimate the impact of statistical adjustments. For example, given the data from

the current set of responders, does our model well predict the values of nonresponders? If not, then perhaps we should continue collecting data. This would be an adaptive approach to sample design.

1.2 Chapter Preview

In this chapter, I will propose stopping rules for surveys that rely on imputation of missing data. These rules will be adaptive to the survey data — including the response status of each case — as they are collected and will produce a decision to stop collecting data once the survey objectives have been achieved. In Section 2, I will provide some background on stopping rules from the clinical trials context and discuss a recent article (Rao, Glickman, and Glynn, 2007) that proposes stopping rules for surveys. In Section 3, I will propose a new stopping rule and compare it to the rule of Rao, Glickman and Glynn (RGG). In Section 4, I will present the results of a simulation study comparing my new rule to that of RGG. Finally, In Section 5, I will implement the proposed stopping rule using data from an actual survey.

2. Background

Stopping rules have long been a key feature of monitoring clinical trials. The National Institutes of Health "Policy for Data and Safety Monitoring" (1998) suggests that every clinical trial has the responsibility to "recommend conclusion of the trial when significant benefits or risks have developed or the trial is unlikely to be concluded successfully" (p. 2). Ethical concerns have led researchers to propose clinical trials only when there is a state of equipoise between two treatments. That is, a trial should only be conducted when there is no clear preference between two treatments. As soon as a

treatment is determined to be clearly better, then the ethical requirement is that the trial should be stopped and all patients be given the preferred treatment. Statisticians have responded to this ethical demand by developing efficient rules for determining when this requirement has been met. Although there is no consensus on which methods are preferred, there are several viable alternatives.

2.1 Stopping Rules for Clinical Trials

An early stopping rule was developed by Pocock (1977, 1983). This rule was developed from a frequentist perspective. Under the frequentist perspective, the type 1 error rate (false positive) can increase above the nominal rate of a test if multiple tests are performed. That is, the more often a test is repeated, the higher the probability that a false positive will result. In a clinical trial, several analysis points are usually planned. The Pocock approach adjusts the type 1 error rate at each planned test such that the overall probability of a false positive is equal to or less than a specified value. The adjusted type 1 error rate is equivalent at each analysis under this approach.

Another approach was developed by O'Brien and Fleming (1979). This approach is similar to Pocock's method. The O'Brien and Fleming rule adjusts the type 1 error rate at each analysis to achieve the specified overall probability of a type 1 error. It does so, however, by being more conservative in early analyses and less conservative in later analyses. As a result, the probability of an early stopping is lower than under Pocock's approach.

Bayesian approaches to the problem of stopping rules have been discussed by Spiegelhalter, Abrams, and Myles (2004). They suggest stopping rules based on careful elicitation of priors and monitoring of posterior probabilities. Evidence for stopping a

trial with the conclusion that the new treatment is better than the current standard of care should be strong enough to overcome a "sceptical" prior – that is, a prior distribution that prefers the current treatment. On the other hand, evidence for stopping a trial with the conclusion that the current standard of care is as good or better than the new treatment should be strong enough to overcome an "enthusiastic" prior (Spiegelhalter et al., 2004, p. 205). They also recommend the use of predictions to monitor the probability that a significant result will be found, conditional on the current data. All of these methods have been employed on clinical trials. There is still a great deal of controversy in this area; however, there is also a great deal of experience that may be useful for surveys.

2.2 Rao, Glickman, and Glynn's Stopping Rules for Surveys

A recent article has suggested the use of stopping rules with survey data collections (Rao, Glickman, and Glynn, 2007). It is not surprising that this article originates from the context of survey data collection for a clinical trial. The authors propose four rules to be used in determining when data collection should be stopped. The type of survey for which they have designed their rules is a mailed survey. They consider "waves" of data collection; that is, each wave is a new mailing to nonresponders from the previous waves. The problem, as they define it, is to determine after which wave to stop collecting data.

The first two rules they propose are quite simple and take no account of nonresponse. The first rule is based on a likelihood ratio statistic that is meant to determine whether the survey statistic of interest is associated with the wave of response (i.e. the propensity to respond). If the wave of response is an important predictor of the survey variable, after controlling for the frame data, then continue data collection. Their

rules are defined for binary survey outcomes, Y_i. Therefore, they define two logistic models for predicting the survey variable. The first model involves only covariates on the frame predicting the survey outcome variable. In their notation, $\pi_i = P(Y_i = 1)$ where *i* is the i^{th} subject. The first model is:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + X_i'\beta,$$

where X_i' is a vector of *m* covariates available on all cases. The second model involves the covariates, the wave of response, and the interaction of the two.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + X_i'\beta + U_i'\delta + U_i'\Gamma X_i,$$

where U_i is a vector with k elements. The element of this vector u_{iw} takes on the value 1 if response occurred for subject *i* in wave *w* and 0 otherwise. Γ is a *k* x *m* matrix of interaction terms for wave and the frame covariates. Rao, Glickman and Glynn's first stopping rule (Rule 1) tests the significance of the ratio of the likelihood of the second model to that of the first model. If this ratio is larger than a percentile of a χ^2 distribution, then data collection is continued.

The second rule (Rule 2) is based on a comparison of proportions from two independent binomial samples. These two proportions are estimated as:

$$\hat{p}_k = \frac{\sum_{i=1}^{N} u_{ik} y_i}{n_k}$$

where u_{ik} is an indicator for whether subject *i* responded at wave *k* and n_k is the sum of u_{ik}, i.e. the number of responders at wave *k*. The second proportion is:

$$\hat{P}_{(k-1)} = \frac{\sum_{i=1}^{N}\sum_{l=1}^{k-1} u_{il} y_i}{N_{(k-1)}}$$

where $N_{(k-1)}$ is the number of respondents through wave *k-1*. In other words, the first proportion is the proportion estimated from respondents from wave *k* and the second proportion is estimated from respondents from wave 1 to *k-1*.

If the most recent responders change the estimate of the proportion, then continue collecting data. In this case, the first sample is respondents in waves 1 to *k*-1. The second sample is respondents to the current wave, *k*. They compute the standardized difference of the proportions estimated from these two samples and continue to follow nonrespondents if this standardized difference is larger than a percentile of the standard normal distribution.

Rule 2 ignores any covariates and does not attempt to adjust for nonresponse when estimating the proportions with Y_i=1. RGG also propose two rules that attempt to account for nonresponse. The two rules are similar. For both rules, they use a logistic regression model to develop imputations for the missing data. In the first rule, which they call Rule 3.1, one set of imputations is prepared using the data from all waves prior to the current wave (1 to *k-1*). The estimate of the proportion with $Y_i = 1$ based on these imputed datasets they call $\hat{P}_{k-1}$. A second set of imputations is prepared using the data from all waves including the current wave (waves 1 to *k*). The proportion estimated using these imputed datasets they call $\hat{P}_k$. The rule (Rule 3.1) is that if the two sets of imputations provide different estimates of the proportion, then data collection should continue. RGG propose two different tests to see if there is a difference between these

two sets of imputations. Rule 3.1a is the first test. For Rule 3.1a, the difference in the estimated proportions $\left(\hat{P}_k - \hat{P}_{k-1}\right)$ is standardized and data collection is continued if the standardized difference is larger than a percentile of the standard normal distribution. They also suggest another test of this rule (Rule 3.1b) which would stop data collection if the difference in the estimated proportions $\left(\hat{P}_k - \hat{P}_{k-1}\right)$ is less than a specified proportion (.01, for example). A second rule (Rule 3.2) of this type is proposed in which the first set of imputations is based on the previous wave only (wave *k*-1), and not all previous waves. The second set of imputations is based only on the current wave (wave *k*). In their simulations, the last two rules performed equally well with relatively low root mean squared error.

Rao, Glickman and Glynn's rules are flexible enough to be applied to surveys other than mailed ones. Although the rules were developed for a mailed survey conducted in waves, it could easily be applied to either a telephone or a face-to-face survey. Instead of comparing waves of response, the rule could be implemented after each call or other specified points in production. The rules can also easily be generalized to situations other than binary variables.

A weakness of the approach is that it requires at least two waves of data collection. For the estimate from prior waves, we only include part of the data (waves 1 to *k*-1). This may lead to some loss of efficiency. In this sense, the method is retrospective. The rule stops data collection only after the latest "wave" fails to contribute anything new to the estimate. In other words, we would have had the same estimate had we stopped at least one wave earlier.

Another weakness is that the rule assumes that a similar relationship among covariates and the survey variable obtains for waves beyond the current wave. Of course, this might not be the case and stopping might leave us with biased estimates. However, such an argument can be used against any stopping point short of a 100% response rate. Very few surveys have any possibility of achieving 100% response. The current approach is to specify a response rates at which to stop. This approach ignores all data except for the response indicator. It is also vulnerable to the same criticism – that the relationship between the covariates and the survey variable is different for nonresponders. The model for nonresponse offered by RGG's rules 3.1 and 3.2 at least attempts to place the decision to stop on a reasoned and statistical basis that uses all the available data by modeling the impact of nonresponse and creating a statistical rule for when to stop. A response rate, when used as a stopping rule, does neither.

In the next section, I will propose an alternative stopping rule. My rule will make use of all the data. I will conduct simulations to understand how this rule works, compare the results to rules suggested by RGG, and then implement a real-world application.

3. A Proposed Stopping Rule for Surveys

3.1 The "Stop and Impute" Rule

The rule that I will propose is based on imputation methods. It involves the comparison of two estimates. The first estimate is the one we would have if we were to stop collecting data right now and impute the missing values. The second estimate is the one we would have if we were to continue collecting data until we had achieved a specified number of additional interviews and then we were to impute the missing data. If

these two estimates are likely to produce the same results, then we should stop collecting data. These two estimates can be denoted:

$$e_1 = (\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n} \hat{y}_i) / n$$

and

$$e_2 = (\sum_{i=1}^{n_1+n_2 p} y_i + \sum_{i=n_1+n_2 p+1}^{n} \hat{y}_i) / n,$$

where y_i is the observed survey variable for case i, $\hat{y}_i$ is the predicted value derived from the model $\hat{y}_i = \hat{\beta} z_i$, z_i is a covariate available for all sampled cases, $\hat{\beta}$ is an estimated coefficient for a regression-through-the-origin model with y regressed on z, n_1 is the current set of responders, n ($n = n_1 + n_2$) is the total sample, and p is the proportion of the remaining n_2 interviews that is expected to be collected with continued data collection (e.g. interviews expected on the next call or next few calls). The error term ε_i in the model $y_i = \beta z_i + \varepsilon_i$ is distributed $N(0, \sigma_e^2)$.

We then evaluate the probability that these two estimates are the same. The investigator must specify a value δ that is acceptably small such that a difference between e_1 and e_2 that is less than this value is not meaningful. If the probability

$$\Pr(|e_1 - e_2| < \delta \mid z, p, \beta, \sigma_e)$$

is sufficiently large, then option e_1 – where we stop data collection now and impute the remaining missing values – is preferred.

In order to determine this probability, the variance of $e_1 - e_2$ is needed. The difference between e_1 and e_2 can be expressed as

$$\frac{\sum_{i=n_1+1}^{n_1+n_2 p} \hat{y}_i - \sum_{i=n_1+1}^{n_1+n_2 p} y_i}{n}.$$

This is the difference between the imputed and observed values for the n_2p cases we are considering adding to our data. If we substitute the regression equation for the predictions and the mean value for the cases that would be interviewed, then we have the following reformulation of the previous expression:

$$\frac{\sum_{i=n_1+1}^{n_1+n_2 p} \hat{\beta} z_i - n_2 p \overline{y}}{n}$$

The next step is to note that the expected value of this quantity is 0. If we can derive the variance of this quantity, then we can estimate the probability that this quantity is close to zero. If we denote the cases that would respond (i.e. the n_2p cases) with further effort using the subscript *wr*, then this can be shown to be equivalent to the following:

$$\frac{n_2 p}{n}\left(\hat{\beta}_R \overline{z}_{wr} - \overline{y}_{wr}\right).$$

The variance of this difference is:

$$Var\left(\frac{n_2 p}{n}(\hat{\beta}_r \overline{z}_{wr} - \overline{y}_{wr})\right).$$

Using the subscript *r* to indicate the cases that have already responded, this can be rewritten as:

$$\left(\frac{n_2 p}{n}\right)^2 \left[\overline{z}_{wr}^2 Var(\hat{\beta}_r) + Var(\overline{y}_{wr})\right].$$

Since $Var(\hat{\beta}_r) = \sigma_e^2 \Big/ \sum_{i=1}^{n_1} z_i^2$ and $Var(\bar{y}_{wr} \,|\, z) = \sigma_e^2 \Big/ n_2 p$, the variance of the difference between e_1 and e_2 is:

$$\hat{\sigma}_e^2 \left(\frac{n_2 p}{n} \right)^2 \left(\frac{\bar{z}_{wr}^2}{\sum_{i=1}^{n_1} z_i^2} + \frac{1}{n_2 p} \right).$$

where $\hat{\sigma}_e^2$ is the residual variance from the model which regresses *y* on *z*:

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{n_1} (y_i - \hat{y}_i)^2}{n_1 - 1}.$$

This variance estimate for the difference between e_1 and e_2 can be used to standardize the results for comparison with the standard normal distribution. If the probability

$$\Pr(|\, e_1 - e_2 \,| < \delta \,|\, z, p, \beta, \sigma_e)$$

is large enough, then data collection can be stopped. In other words, if the probability that additional data (n_2p cases) will not substantially change our current estimate – conditional on our model, the covariates, and the residual variance – is very high, we should stop collecting data.

A multivariate extension of this very simple case (univariate *z*) is needed. In order to do this extension, we can rewrite the variance of e_1-e_2 in the same manner as used above:

$$Var\left(\frac{\sum_{i=n_1+1}^{n_1+n_2 p} \hat{y}_i - \sum_{i=n_1+1}^{n_1+n_2 p} y_i}{n} \right),$$

where $\hat{y}_i = \mathbf{z}_i'\boldsymbol{\beta}$, where $\mathbf{z}$ is a (1 x m+1) matrix of m covariates and a vector of 1's for the i^{th} sampled unit. This can be rewritten using the regression model used to predict $\hat{y}_i$ and the expected value for the sum of the y_i:

$$Var\left(\frac{\sum_{i=n_1+1}^{n_1+n_2 p}\left(\beta_0+\sum_{j=1}^{m} z_{ij}\beta_j\right)-n_2 p\bar{y}}{n}\right)$$

If we denote the vector of means of the covariates $\mathbf{z}$ (m+1 covariates) for the n_2p cases that would respond as $\bar{\mathbf{z}}_{wr}$ and the matrix of observed covariates (respondents) as $\mathbf{Z_r}$, then this variance can be written in the following manner:

$$\hat{\sigma}_e^2\left(\frac{n_2 p}{n}\right)^2\left[\left(1+\bar{\mathbf{z}}_{wr}'(\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\bar{\mathbf{z}}_{wr}\right)+\frac{1}{n_2 p}\right].$$

This can be described as the prediction variance for the cases that would be collected plus the conditional variance of the mean of y.

This rule is cost efficient since it allows us to stop data collection when the imputation model is precise enough that we can have nearly the same certainty in our estimate as we would if we were to collect a specified number of additional cases. The rule does, however, assume a model that relates the covariate z to y. If this model is incorrect, or if the estimated coefficient $\hat{\beta}$ is different among responders than among nonresponders, then it is possible that we will stop too early.

One practical issue is determining which n_2p cases you believe will be collected. This can be done by taking a random sample of the remaining n_2 cases, or by sampling using an estimated probability of response to select a sample of cases likely to respond given a specified protocol (this protocol could be tailored to the case). Another practical

issue is how to set p. This can be set either by empirical observation from other surveys about the expected number of interviews to be completed with additional effort, or as a number that is meaningful for continued data collection.

This rule is certainly more focused on the risk of bias than on meeting targets for sampling error. While it is true that the rule requires some minimum size in order to decide when to stop, this minimum can vary quite a lot depending on the specific interrelationships among y, z, and the propensity to respond (this will be seen in the simulations in the next section).

One simple solution is to have a second rule that the sampling error (estimated using multiple imputation-appropriate methods) must be within a specified limit before stopping. Then both this rule about sampling error and the "Stop and Impute" rule must be met before stopping.

It is also possible to build in protection against early stopping using rules developed for stopping in clinical trials. For instance, an approach similar to the O'Brien-Fleming rule can be used to require stronger evidence to justify stopping early. Another approach is suggested by Spiegelhalter et al. (2004). They suggest a specifically Bayesian approach where a "sceptical" prior is specified that might represent the prior opinion of someone who believes that the current standard of care is as good or better than the proposed new treatment. Such a skeptic would require stronger evidence before concluding that the new treatment was superior. As a result, early stopping is less likely and a larger sample size is required. In the application proposed here for surveys, a sceptic would be one who believes that the risk of nonresponse is higher than what the

optimist believes. As a result, a prior specified in this manner would require additional evidence (i.e. interviews) before stopping would be justified.

Working out the required sample size before collecting any data is surely more difficult in this circumstance. Simulations may be helpful in this regard. Simulations could be used to determine what sample sizes might result under various assumptions about the interrelationships of *y*, *z*, and the propensity to respond. These simulations could help project a range of outcomes. Then the investigators can vary the value of the parameters (δ and the cutoff probability for defining a "high probability" that e_1- e_2 is small). These parameters can be set to help tune the expected sampling error.

4. Simulations

There are two key conditions that impact our estimates of the survey variable (*y*). The first is the relationship between the propensity to respond (here denoted *r*) and *y*. If the two are correlated, this can lead to bias. Data collected at different waves are likely to lead to different estimates of $\bar{y}$. The second condition is the relationship of covariates (*z*) on the frame to the survey variable *y*. If the propensity to respond (*r*) and *y* are correlated, but we still have a frame variable (*z*) that predicts well the survey variable, then we may still be able to adjust our estimates (either through imputation or weighting) with this variable and still produce unbiased estimates. If, however, the correlation of *z* and *y* is confounded with *r*, then we are back in the difficult situation where *z* is not useful for adjustment purposes.

Following Rao, Glickman and Glynn (2007), I developed a simulation study to test the performance of this rule while varying these two relationships — the correlation

between the *z* and *y* variable and the relationship between *z* and the propensity to respond *r*. I generated 1,000 data sets under the cross-classification of the following conditions:

- correlations of *z* and *y* either does not or does depend on *r*,
- wave of response (*r*) either does not or does depends on *z*.

Crossing these two conditions produces the four simulations detailed in Table 4-1.

Table 4-1: Parameters of Simulations

SIMULATION	WAVE OF RESPONSE (R)	CORRELATION Z,Y
1	Does not depend on Z	Does not depend on R
2	Depends on Z	Does not depend on R
3	Does not depend on Z	Depends on R
4	Depends on Z	Depends on R

I have attempted to follow the key features of the simulation study of Rao, Glickman, and Glynn (2007) in order to be able to compare the results of their stopping rules to those of the rule proposed here. I also generalized RGG's rules to the normally distributed variable to which the "Stop and Impute" rule applies. In these simulations, *z* is normally distributed with a mean of 10 and a standard deviation of 1. The wave of response *r* is modeled as a Poisson distribution with a mean of 1 when *r* does not depend on *z*. In the condition where *r* does depend on *z*, *r* has a mean of 1 when $z<10$ and a mean of 5 when $z>=10$. When the *z*-*y* correlation does not depend on *r*, it is fixed at a single value for all waves. When the correlation does depend on *r*, the correlations increase with the wave of response. In addition, in order to demonstrate the impact of the correlation, I also simulate *z* with different standard deviations to see what impact this has. Since the "Stop and Impute" rule depends on the choice of a suitably small δ, I chose δ to be 1% of

the mean. In addition, if the probability that the difference between e_1 and e_2 is less than δ is greater than .95, I stopped collecting data.

In the first simulation, *r* does not depend on *z*, and the correlation of *z* and *y* is independent of *r*. This corresponds to the Missing Completely at Random assumption described by Little and Rubin (2002). Under these conditions, the responders are effectively a random sample of the total sample (including nonresponders). This is shown by Figure 4-1, which plots the observed $\bar{y}$, the fully imputed $\bar{y}$, and the true $\bar{y}$ by wave. The three points are virtually identical at each wave. The wave two respondents are equivalent to the wave one respondents, and so on.

Figure 4-1: Observed, Imputed, Actual Mean Y for Simulation 1

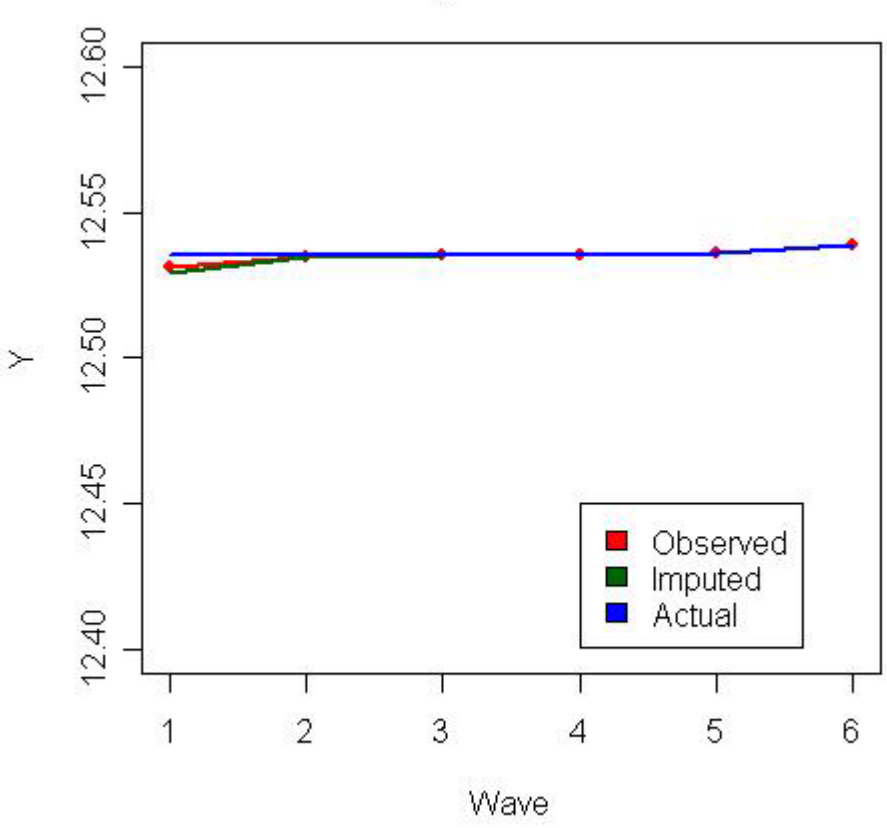


Table 4-2 presents the results of 1,000 simulations of these conditions (y depends only on z, r does not depend on z). I created 1,000 datasets at each of nine different correlations between z and y. The column "Mean Stopping Wave" reports the mean number of waves completed before stopping. The next column is the standard deviation of the stopping wave. The next column is the variance of $\hat{\bar{y}}$. This variance is calculated using methods for combining multiple imputations (Little and Rubin, 2002). The bias of the estimate is multiplied by 100, as is the root mean squared error (RMSE). The final column is the proportion of the time that the 95% confidence intervals attain the nominal coverage of the population mean.

Table 4-2: Simulation 1 Results

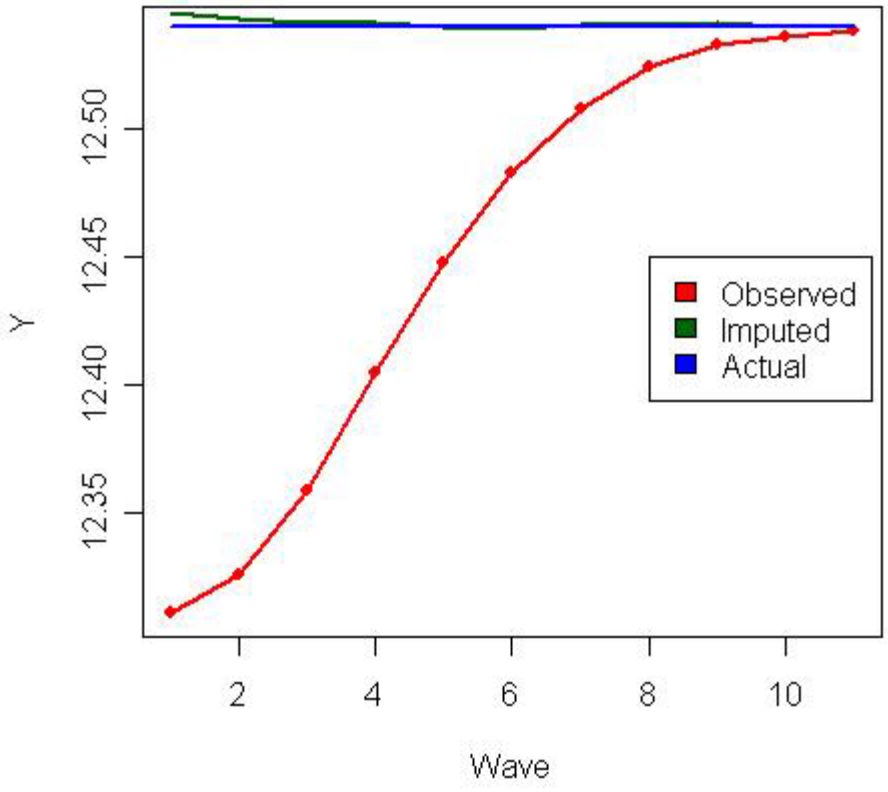
	RGG 3.1a					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)* 100**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	2.10	0.29	0.007	0.28	8.39	1.00
0.2	2.10	0.31	0.007	-0.02	8.37	1.00
0.3	2.10	0.30	0.007	-0.03	8.29	1.00
0.4	2.09	0.30	0.007	-0.10	8.22	1.00
0.5	2.12	0.33	0.006	0.02	8.05	1.00
0.6	2.10	0.32	0.006	-0.05	8.00	1.00
0.7	2.10	0.31	0.006	-0.11	7.78	1.00
0.8	2.08	0.28	0.006	-0.14	7.55	1.00
0.9	2.10	0.30	0.005	0.09	7.33	1.00
	RGG 3.1b					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)*10 0**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	2.28	0.45	0.007	0.21	8.17	1.00
0.2	2.22	0.42	0.007	-0.06	8.22	1.00
0.3	2.18	0.38	0.007	0.00	8.19	1.00
0.4	2.12	0.33	0.007	-0.12	8.19	1.00
0.5	2.11	0.31	0.006	0.03	8.06	1.00
0.6	2.07	0.25	0.006	-0.05	8.02	1.00
0.7	2.03	0.18	0.006	-0.10	7.82	1.00
0.8	2.02	0.12	0.006	-0.11	7.57	1.00
0.9	2.00	0.04	0.005	0.08	7.35	1.00
	"Stop and Impute" Rule					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)* 100**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	1.85	0.36	0.008	0.29	8.93	0.99
0.2	1.82	0.38	0.008	-0.03	9.03	0.99
0.3	1.73	0.44	0.008	0.05	9.17	0.99
0.4	1.59	0.49	0.009	-0.13	9.50	0.99
0.5	1.37	0.48	0.010	0.19	9.87	0.99
0.6	1.15	0.36	0.011	0.02	10.45	0.99
0.7	1.01	0.11	0.010	0.14	10.18	0.98
0.8	1.00	0.00	0.009	-0.10	9.48	0.99
0.9	1.00	0.00	0.007	0.15	8.40	1.00

In this situation, it should be clear that stopping earlier is preferred since the outcome variable is uncorrelated with the response propensity and the *z* variable is uncorrelated with the wave. Under these conditions, the "Stop and Impute" rule performs better than either rule of RGG. Mainly, this is because RGG's rules require at least two

waves of data collection, whereas the "Stop and Impute" rule can stop after the first wave. As the correlation between *z* and *y* increases, this early stopping after the first wave is more likely to happen.

In the second simulation, the wave of response (*r*) is a function of *z*. In the simulations, this effect was implemented by having *r* drawn from a Poisson distribution with mean 1 when *z*<10. When *z*>=10, then *r* was drawn from a Poisson distribution with mean 5. In this case, there is a risk of bias. This can be seen from Figure 4-2. The observed $\bar{y}$ are different than the true $\bar{y}$. Fortunately, the *z* variable allows us to correct for this bias through the imputation model that correctly relates *z* and *y* even at wave 1. This is why the actual and impute lines are nearly identical. These conditions correspond to the Missing at Random assumption described by Little and Rubin (2002).

Figure 4-2: Observed, Imputed, and Actual Mean Y for Simulation 2



I created 1,000 datasets under these conditions (r depends on z, the correlation of z and y does not depend on r). The results are presented in Table 4-3.

Table 4-3: Simulation 2 Results

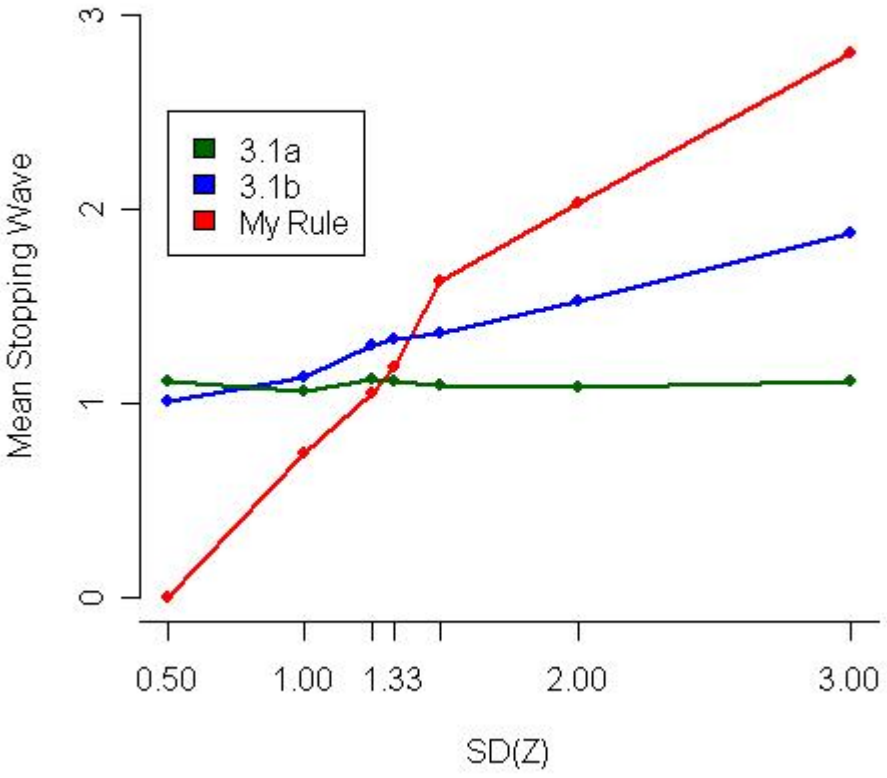
	RGG 3.1a					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)* 100**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	2.54	0.70	0.026	-0.06	16.04	0.94
0.2	2.52	0.71	0.025	0.18	15.93	0.95
0.3	2.55	0.74	0.024	0.08	15.57	0.95
0.4	2.56	0.74	0.023	-0.12	15.12	0.96
0.5	2.58	0.73	0.020	0.00	14.16	0.96
0.6	2.58	0.73	0.018	-0.05	13.39	0.97
0.7	2.60	0.73	0.015	-0.27	12.26	0.97
0.8	2.53	0.70	0.013	-0.08	11.22	0.98
0.9	2.53	0.72	0.009	-0.07	9.52	1.00
	RGG 3.1b					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)* 100**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	2.93	0.89	0.022	-0.11	14.74	0.95
0.2	2.86	0.86	0.021	0.71	14.55	0.96
0.3	2.82	0.83	0.021	-0.04	14.55	0.95
0.4	2.72	0.77	0.021	0.19	14.56	0.96
0.5	2.66	0.73	0.019	-0.08	13.73	0.95
0.6	2.59	0.68	0.018	0.05	13.24	0.97
0.7	2.51	0.63	0.015	-0.34	12.35	0.97
0.8	2.37	0.55	0.013	-0.11	11.51	0.98
0.9	2.20	0.41	0.010	-0.13	9.87	0.99
	"Stop and Impute" Rule					
Corr(Z,Y)	**Mean Stopping Wave**	**Std Dev Stopping Wave**	**Var ($\hat{\bar{y}}$)**	**Bias($\hat{\bar{y}}$)* 100**	**RMSE($\hat{\bar{y}}$)* 100**	**Coverage**
0.1	1.88	0.52	0.036	0.60	19.01	0.92
0.2	1.84	0.51	0.036	-0.12	18.89	0.92
0.3	1.79	0.50	0.037	0.04	19.14	0.93
0.4	1.66	0.51	0.040	-0.26	19.98	0.92
0.5	1.53	0.51	0.037	-0.21	19.33	0.92
0.6	1.34	0.47	0.042	-0.60	20.44	0.92
0.7	1.10	0.30	0.041	0.28	20.13	0.92
0.8	1.01	0.08	0.035	0.16	18.59	0.93
0.9	1.00	0.00	0.021	-0.76	14.35	0.94

In this situation, RGG's rules are more conservative. They generally result in a smaller bias of the survey statistic than the "Stop and Impute" rule, but the bias is still relatively small for the "Stop and Impute" rule, but it is large enough that the nominal

coverage for the 95% confidence interval is not quite attained. Again, the "Stop and Impute" rule stops much earlier than either of RGG's rules.

Since the "Stop and Impute" rule is more directly dependent on the variance of *z*, this parameter setting can make a large difference in when stopping will occur. Figure 4-3 shows the impact of different standard deviations of *z* on the mean wave of stopping. With a mean of 10, a standard deviation of 1 still represents a large variation relative to the mean.

Figure 4-3: Mean Stopping Wave by SD(Z)

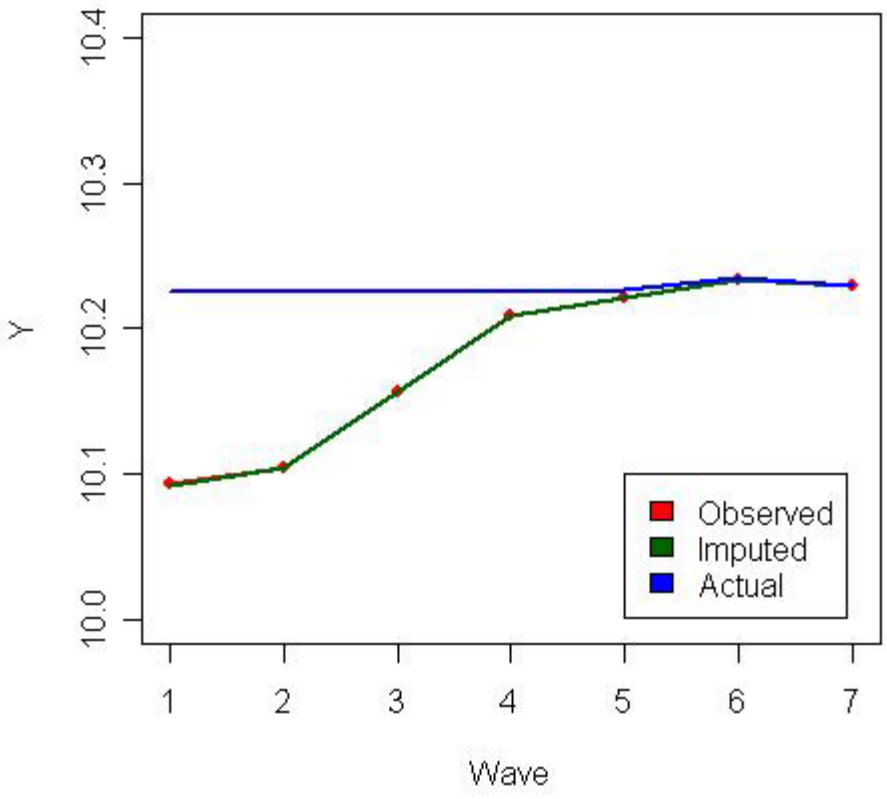


It is clear from the figure that the "Stop and Impute" rule is much more susceptible to variation in *z*. This should be expected as the rule directly incorporates this variation.

RGG Rule 3.1a involves the difference in the imputed means between two waves. This difference is very insensitive to variation in *z*. RGG Rule 3.1b, on the other hand, is somewhat sensitive to variation in *z*. In sum, high variances on the covariate may reduce its utility for the "Stop and Impute" rule.

Simulation 3 is the situation where *r* does not depend on *z*, but the correlation of *z* and *y* is a function of *r*. This simulation was implemented by specifying the correlation for each value of *r* (0.01, 0.01, 0.04, 0.1, 0.1, 0.2, 0.2, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3). The correlation of *z* and *y* is an increasing function of *r*. Figure 4-4 depicts this situation. The imputed and observed $\bar{y}$ depart from the true value until the response rate becomes relatively high. This is the Not Missing at Random assumption (at least for the first 4 waves) described by Little and Rubin (2002).

Figure 4-4: Observed, Imputed and Actual Mean Y for Simulation 3



The results of this simulation are presented in Table 4-4. Since the correlation of *z* and *y* varies with *r*, I did not try different correlations (0.1 to 0.9) as I did with simulations 1 and 2.

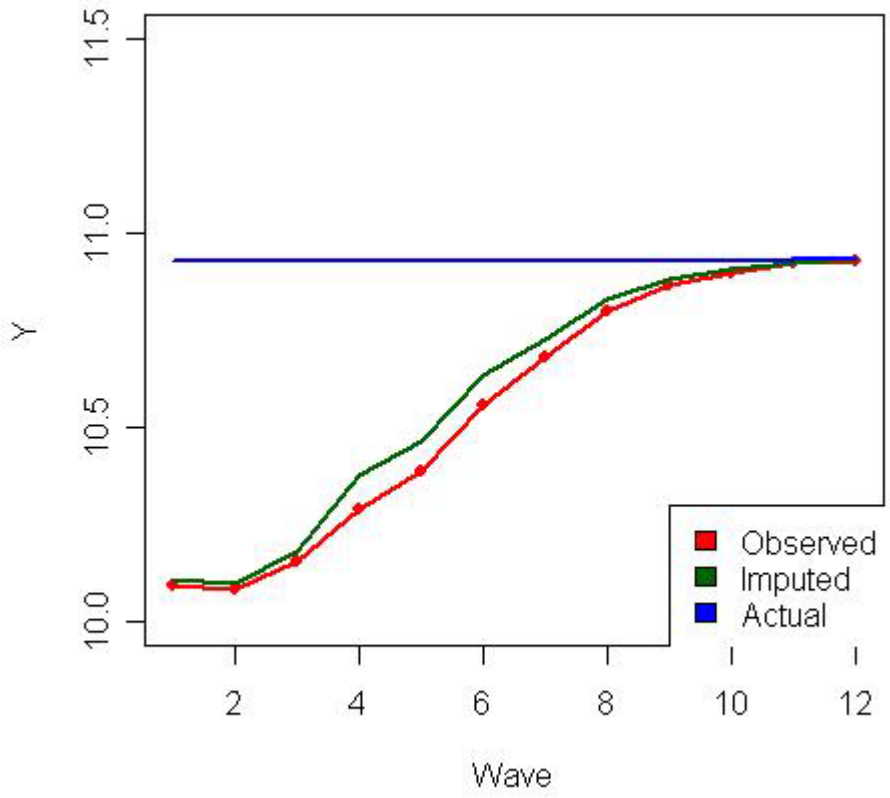
Table 4-4: Simulation 3 Results

	RGG 3.1a	RGG 3.1b	"Stop and Impute" Rule
Stopping Wave Mean	2.10	3.14	3.90
Stopping Wave Std Dev	0.31	0.90	0.32
Bias($\hat{\bar{y}}$)*100	-11.60	-6.37	-2.07
RMSE($\hat{\bar{y}}$)*100	27.67	23.74	21.64
Coverage	99.6	99.9	100.0

In this situation, it appears that the "Stop and Impute" rule and RGG 3.1b are more conservative and produce less biased estimates of $\bar{y}$. The RMSE of the estimate using the "Stop and Impute" rule is lower than either of the other rules. If limiting bias is more important than the overall RMSE, the more conservative rule might be preferred for this situation. However, it is also much more costly, requiring on average 1.8 more calls than RGG rule 3.1a.

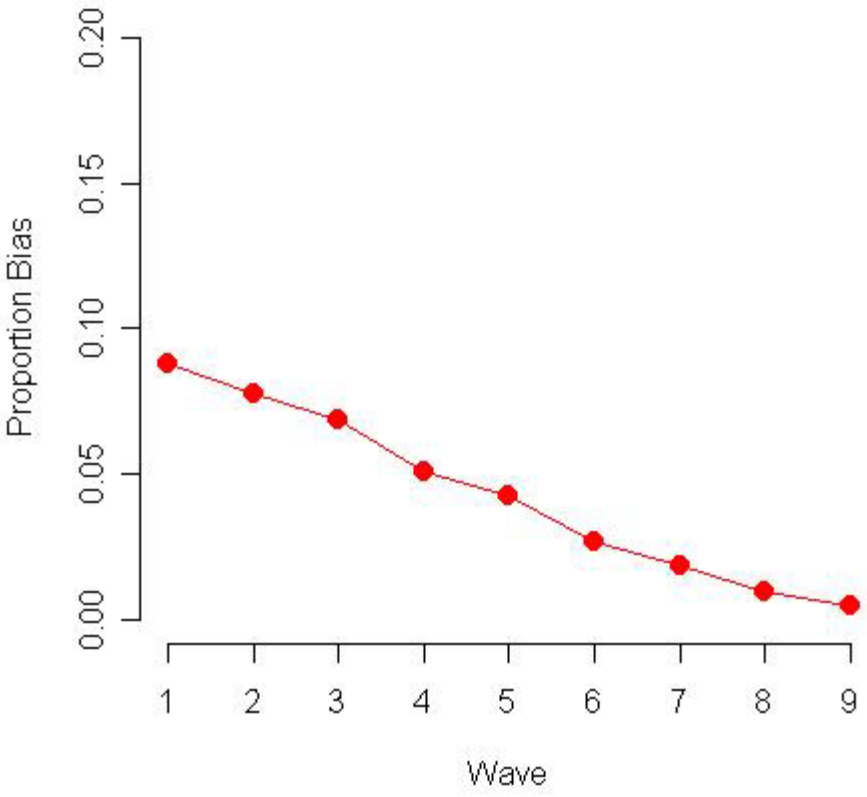
In the final simulation, the wave of response (*r*) depends on *z* and the correlation of *z* and *y* is a function of *r*. Figure 4-5 shows the observed, imputed, and true $\bar{y}$ by wave for this simulation. Again, these conditions produce the Not Missing at Random situation.

Figure 4-5: Observed, Imputed, and Actual Mean Y for Simulation 4



The bias is a function of the wave. Figure 4-6 shows the average proportionate bias over the 1,000 simulations at each wave for the imputed mean (as opposed to the mean derived from the observed data alone).

Figure 4-6: Simulation 4 Proportion Bias By Wave



From Figure 4-6 it can be seen that the bias is still relatively high until after the 6th wave (wave 5). The results for the three stopping rules under simulation 4 are presented in Table 4-5.

Table 4-5: Simulation 4 Results

	RGG 3.1a	RGG 3.1b	"Stop and Impute" Rule
Stopping Wave Mean	1.58	3.95	7.22
Stopping Wave Std Dev	0.77	2.03	0.75
Bias($\hat{\bar{y}}$)*100	-77.56	-47.43	-8.51
RMSE($\hat{\bar{y}}$)*100	90.94	57.84	24.23
Coverage	57.6	70.8	94.5

In this situation, the bias begins to dominate the RMSE for RGG rule 3.1a. The "Stop and Impute" rule does much better in terms of the bias and is the only rule to attain anything close to the nominal coverage. Of course, much greater effort is required.

In general, in comparing the "Stop and Impute" rule to RGG Rules 3.1a and 3.1b, the rule proposed here is more efficient in the situations that are MCAR. In the MCAR situation, the "Stop and Impute" rule can be very efficient and truncate effort earlier than either of RGG's rules considered here. When the situation is not MCAR, the "Stop and Impute" rule appears to be conservative. It may be overly conservative, depending on the relative importance of the bias and variance components of the mean squared error. The "Stop and Impute" rule is more protective against potential bias, but generally requires more effort when bias is a risk.

5. Implementation

This approach was implemented with data from the Survey of Consumers (SCA). The survey collects 300 random-digit dial (RDD) interviews each month. It typically attains an AAPOR Response Rate 2 of between 42% and 45%. SCA asks respondents for their views on the state of their personal finances and the economy in general. A key statistic produced by this survey is the Index of Consumer Sentiment (ICS). This Index is widely reported and has been found to be highly predictive of economic trends (Curtin, 2007).

The Index of Consumer Sentiment is the survey variable (y). The mean ICS for the month analyzed here is 91.46. The frame variable (z) is the change in the CPI for the associated Core Based Statistical Area (CBSA) or MSA Status and Region from the prior quarter to the current quarter (see below for a complete description of frame data). This

variable was chosen since it had the largest correlation with the ICS of any of the available variables. The correlation between *z* and *y* is -0.15. The *r* variable in this case is the call number of the completion. The correlation between the ICS and the call number is practically zero (-0.01).

An important difference between these data and the simulation data has to do with the frame quality. Since the SCA uses an RDD frame, there are numerous blanks (non-household, non-working numbers) on the frame. Some of these nonsample cases will never be identified. This is because telephone companies often allow unassigned numbers to ring through as if they were working. These numbers will ring instead of playing a message indicating that the number is not working. Other numbers will be identified as nonworking or nonhousehold only after repeated attempts. Therefore, the total sample size changes at each call.

The survey designer has two important choices to make when implementing this stopping rule. The first decision is the probability at which e_1 and e_2 are judged to produce the same estimate. For these analyses, I have stopped if the probability that the difference between e_1 and e_2 is less than δ is greater than 0.95. The second decision involves δ. A value needs to be chosen that is small enough such that as long as the difference between e_1 and e_2 is smaller than this value it can be considered to be unimportant. Table 4-6 shows the results for different choices of δ. For comparison, a 1% δ would be 0.91.

Table 4-6: Stopping Rule Results Using Various Delta Values

DELTA	STOPPING CALL	EST. ICS	VAR(ICS)	NONRESPONSE RATE	COMPLETES
1	78	90.93	4.95	0.51	297
2	78	91.64	4.74	0.51	297
3	78	91.87	4.90	0.51	297
4	78	90.77	4.13	0.51	297
5	78	91.07	4.78	0.51	297
6	21	92.28	6.24	0.57	266
7	21	90.47	7.21	0.57	266
8	18	91.12	6.00	0.59	256
9	12	90.65	3.87	0.64	228
10	10	91.88	6.43	0.65	220

With these data, the stopping rule does not stop before the point where the survey actually stopped (i.e. n=297) until δ reaches 6. This is 6.7% of the estimate and is probably too large of a difference to be considered "reasonably small." It would seem that in this situation, with a reasonably small δ, the rule would not lead us to stop before the current rule. In order to demonstrate the functioning of the rule, Table 4-7 includes — for each call number of the survey — the probability that $|e_1 - e_2| < \delta$, the estimated ICS, the nonresponse rate, and the number of completes.

Table 4-7: Stopping Rule Probability by Call Number, Delta=6

Call Number	Test Probability	Est. ICS	Nonresponse Rate	Completes
1	0.33	87.20	0.90	74
2	0.47	89.92	0.83	112
3	0.41	90.38	0.80	130
4	0.44	91.22	0.77	153
5	0.56	91.36	0.73	173
6	0.48	91.92	0.71	183
7	0.62	92.10	0.69	198
8	0.59	92.61	0.68	206
9	0.73	91.94	0.66	215
10	0.77	88.94	0.65	220
11	0.77	92.99	0.64	224
12	0.83	92.01	0.64	228
13	0.55	90.31	0.63	231
14	0.76	90.88	0.62	241
15	0.82	89.94	0.61	245
16	0.76	90.60	0.60	248
17	0.75	91.78	0.60	252
18	0.90	92.56	0.59	256
19	0.82	91.12	0.59	258
20	0.70	91.13	0.58	261
21	0.98	90.28	0.57	266

If an investigator were more willing to tolerate the risk of bias, then stopping earlier under this rule might be effective. It is also important to note that additional effort is not producing nearly the results of earlier effort in this case. While the first call yielded 74 completes, the last call (in this table) only yielded 5 completes. This makes it difficult for the stopping rule to be implemented. In this case, grouping calls to create larger sample sizes at each "grouped" call does not significantly change the results. For example, grouping calls 23-78 creates a group with 30 completed interviews. However, this does not change the result of the stopping rules. It is not until δ becomes relatively

large that stopping occurs, and this stopping occurs at the "grouped" call that includes the call number identified above (i.e. call 21).

The multivariate model proves to be more useful in this example. There are several variables on the frame that are related to the ICS. The RDD sample for SCA is generated using the Genesys sampling system. This system associates every telephone exchange with a geographic location using telephone listings. Then Census data of the associated geographic location can then be attached to the telephone number. Of course, with the portability of telephone numbers, this estimate of the geographic location of the number may be wrong. However, only a small proportion of numbers have been ported and these estimates of geography continue to be fairly accurate (Johnson, et al., 2006). In addition to the variables supplied by Genesys from Census data, I added data from the Bureau of Labor Statistics on employment, wages, and prices. These data included monthly unemployment rates at the county level, quarterly reports of average weekly wages at the county level, and the monthly reports of the Consumer Price Index (CPI) at the Core Based Statistical Area (CBSA) for larger areas or region and MSA status for smaller areas. Since the ICS is a measure of consumers' views on the economy, these economic measures were expected to be related to the ICS. Curtin (2007) presents evidence that this is the case. Curtin suggests that interest rates are also important predictors. These are not included in these models as it is more difficult to obtain these data at a local level. In addition, Curtin notes that the change from one time period to the next in these rates is also predictive of consumer sentiment. Therefore, changes in unemployment rates, average weekly wages, and the CPI from the quarter or month prior

to the month of the interview to the current month or quarter were also included. These variables are listed in Table 4-8.

Table 4-8: Available Frame Data

VARIABLE	SOURCE
Listed/Letter Sent	Genesys
Percent Exchange Listed	
Percent Age 18-24	Genesys (Census)
Percent Age 25-34	
Percent Age 35-44	
Percent Age 45-54	
Percent Age 55-64	
Percent Age 65+	
Percent Income <$10,000	
Percent Income $10-15,000	
Percent Income $15-25,000	
Percent Income $25-35,000	
Percent Income $35-50,000	
Percent Income $50-75,000	
Percent Income $75-100,000	
Percent Income $100,000+	
Percent Owner Occupied	
Percent Black	
Percent Hispanic	
Percent White	
Percent Asian/Pacific Islander	
Log(Median HH Income)	
Household Density (Households per 1000 sq ft)	
Unemployment Rate (County)	BLS
Change in Unemployment Rate (County)	
CPI (CBSA/MSA Status by Region)	
Change in CPI (CBSA/MSA Status by Region)	
Average Weekly Wages (County)	
Change in Avg. Weekly Wages (County)	

A subset of these variables (Table 4-9) was selected using a stepwise regression modeling approach. Variables were allowed to enter the model if they had a p-value less than 0.3. The model had an R-squared value of 0.105. I wanted to find a parsimonious model since it would need to be used with continually updated datasets.

Table 4-9: Variables Used in Impute Model

VARIABLE
Percent Exchange Listed
Percent Age 18-24
Percent Age 65+
Percent Income $100,000+
Percent Owner Occupied
Percent Black
Percent Hispanic
Log(Median HH Income)
Household Density (Households per 1000 sq ft)
Unemployment Rate (County)
CPI (CBSA/MSA Status by Region)
Change in CPI (CBSA/MSA Status by Region)
Average Weekly Wages (County)

The results are reported in Table 4-10. With δ=1, the rule would stop data collection after 33 calls and collects 286 interviews. The estimate of the ICS would be 90.82 – very close to the estimate from all 297 interviews (91.46). The difference between the two estimates is within δ (91.46-90.82=0.64, which is less than 0.91).

Table 4-10: Stopping Rule Results Using Various Delta Values

DELTA	STOPPING CALL	EST. ICS	VAR(ICS)	NONRESPONSE RATE	COMPLETES
1	33	90.82	7.0	0.54	286
2	15	90.17	10.5	0.61	245
3	10	88.77	17.7	0.65	220
4	9	89.05	10.2	0.66	215
5	8	90.84	4.5	0.68	206
6	5	91.36	22.8	0.73	173
7	5	94.21	26.2	0.73	173
8	5	92.65	15.3	0.73	173
9	5	93.37	15.2	0.73	173
10	5	91.80	18.1	0.73	173

Although these savings are moderate, the important result is that in this application the stopping rule identified a stopping point that was not arbitrary. It was

based on the data. In fact, it was based on the data from all the interviews and the complete data on the frame.

6. Conclusion

Stopping rules have long been used in clinical trials where there is an ethical demand that a trial be stopped when one treatment is clearly better than another. Data monitoring committees are charged with analyzing the accumulating data to determine when this happens. Surveys, on the other hand, have relied on fixed sample designs built on assumptions about the variance of the statistics of interest. These assumptions are infrequently tested once the survey is in the field. Surveys almost never analyze their accumulating data in order to assess their ability to achieve targeted levels of precision. These estimates are complicated by the problem of nonresponse. This nonresponse can potentially bias survey estimates.

If we are to build adaptive designs for surveys, we will need some means for determining when to stop collecting data. Current practice stops at a pre-specified response rate or sample size. This current rule does not consider the data available on the frame and their relationships to the accumulating survey data. These data may help inform us about the risk of nonresponse bias. A stopping rule that monitors these frame data and the accumulating survey data is preferred. A rule that attempts to account for risk of bias due to nonresponse would perform better across a variety of situations.

In this chapter, I have proposed such a rule. This rule is based on imputation methods for dealing with nonresponse bias. In the presence of variables which are correlated with the survey variable, it is possible for this rule to be protective against bias. In contrast to rules proposed by other authors, the “Stop and Impute” rule makes use of

all the available data to determine whether to stop. This increases the efficiency of the rule. In simulations, this rule proves to be quite robust. In the most favorable situations (i.e. when data are missing completely at random), it also tends to be efficient compared to other rules proposed for surveys. An application to a real survey shows that this rule is feasible. This rule if focused on the risk of bias. However, I have also suggested simple modifications that will accommodate setting targets for sampling error as well.

Chapter 5

Conclusion

1. Introduction

Nonresponse is a growing problem. While this phenomenon has been examined in specialized studies conducted for the sole purpose of studying nonresponse bias (for example, Keeter et al. 2000), much of the research into methods for dealing with this bias have focused on ways to increase response rates. Although the response rate may establish an upper limit on the potential nonresponse bias, this is not very informative as this limiting condition is often very from the estimated mean, even at a relatively high response rate. Given that the relationship between nonresponse rates and nonresponse bias is not necessarily linear, it makes sense to be concerned about the composition of the final set of respondents. In the field of survey methodology, there has been very little research in this direction.

In order to address this gap, I have proposed a new measure of the risk of nonresponse bias that includes more data than just the response indicator. This measure is the fraction of missing information. Under such a measure, new methods would be needed in order to maximize the information in the sample. I have suggested that methods tailored to the characteristics of the case are needed and that the experimental

and statistical methods developed for dynamic treatment regimes may be helpful in this regard. Surveys are inherently sequential and the dynamic treatment regimes approach is designed for identifying optimal solutions in just this sort of environment. I have demonstrated that this approach can be effective at identifying more efficient means for establishing contact with sampled households. Future research can extend these methods to the whole survey process.

2. Summary

It seems natural to turn to methods developed for analysis in the presence of missing data when looking for indicators for nonresponse bias. The fraction of missing information is a measure long used with imputation methods to judge the quality and efficiency of these imputations. As opposed to the response rate, this measure involves all the data that we have – the response indicator, the frame data, and the incomplete survey data. Since it uses more data than the response rate, it should allow us to make more informed decisions about the risk of nonresponse bias.

The fraction of missing information does involve a model assumption. The imputation model is specified by the user. An incorrect model can lead to further errors. If the fraction of missing information was being used to help target cases, then incorrect model specification could increase the nonresponse bias. However, every approach to survey data collection in the presence of nonresponse relies upon some modeling assumption. The response rate as a monitoring tool implies that the nonresponse bias is a linear function of the nonresponse rate. In fact, there is evidence that this is not always true. Groves (2006) indicates that for the specialized studies reviewed there is very little correlation between response rate and nonresponse bias. Perhaps, we can model the risk

of nonresponse bias better if we involve more data and make explicit our model assumptions. In addition, a model that relies on more of the data is likely to be more robust. In this way, the fraction of missing information is a better indicator than the response rate of the risk of nonresponse bias.

If the response rate is the key metric, then the rational approach is to prioritize cases by their propensity to respond. Cases with the highest propensity are targeted for interview first. Unfortunately, this approach may not create data with the lowest risk of nonresponse bias. A new metric, such as the fraction of missing information, might create an environment in which other approaches could flourish. The FMI, as a metric, might lead to the development of data collection strategies aimed at maximizing the information content of the incomplete survey data. More "information" would be recovered under data collection guided by this strategy than one guided by the response rate.

Under this measure, new methods would be needed. We may want to target specific cases based on something other than the propensity. This creates the need to tailor our methods, and to continue this tailoring throughout the data collection process. For example, suppose that leaving an answering machine message may be beneficial for establishing contact in most cases. However, for an important subgroup a message left on an answering machine is a "warning" that we are calling that may lead to call screening – increasing the difficulty of the case. Or worse, the answering machine message gives the potential respondent time to prepare a "script" with reasons why they cannot do the survey. In this case, a method tailored to the specifics of the case is needed. No message would be left for those cases for whom the answering machine message is actually harmful. This sort of tailoring can be generalized over many design features.

Unfortunately, standard randomized controlled trials are not amenable to investigating these sorts of questions. The dynamic treatment regimes approach, as a set of experimental and statistical tools, may offer a way forward. In fact, the survey process is inherently sequential. We move from attempting contact, to screening, to possible refusal conversion, to interviewing. The dynamic treatment regimes approach allows us to expand our focus from one stage (or even one feature of a stage) to the whole sequence of stages. The approach allows us to more broadly consider interactions between the sampled person's characteristics (including their history of previous "treatments") and the proposed set of design features, or protocols. These interactions can include how early treatments may change the effect of later treatments (for example, the answering machine message impact – positive or negative – on later ability to contact or interview a case). I have presented two case studies that show efficient methods can be identified using this approach.

Finally, if the response rate is not the goal, then how do we know when to stop collecting data? The current rule is often to stop collecting data when a target response rate has been reached. Again, a rule that is based on all the data should perform better than one that uses only the response indicator to determine when to stop. In addition, a rule that conditions on all the data is truly adaptive to the data that are being collected. Several stopping rules have been proposed for surveys. Rao, Glickman and Glynn (2007) have proposed a set of rules. In Chapter 4, I proposed a new rule – the "Stop and Impute" rule. This rule would allow the decision about when to stop collecting data to be based on the data as they are collected. Such a rule attempts to account for the risk of nonresponse bias when determining whether to stop.

Taken together, these methods may allow the field to develop methods for dealing with nonresponse bias during data collection. Working to maximize the information in our data should produce higher quality data. Designing "dynamic treatment regimes" for surveys may help us meet this goal.

3. Future Work

3.1 Using FMI to Guide Effort

I have suggested that we might want to target cases based on something other than the propensity to respond. The fraction of missing information is a viable candidate. The problem is that this is not a case-level statistic. It is a feature of the dataset. However, there are at least two possible approaches. The first involves creating propensity strata and then estimating the fraction of missing information within each stratum. There are several issues to work through with this approach. The estimation of the propensity strata could be updated daily. This would allow information about previous treatments to be included in the model. We would then target cases within the stratum with the highest fraction of missing information. I have been monitoring FMI by propensity stratum (re-estimated daily) for NSFG.

We could also estimate the impact on the fraction of missing information that each case has using a sort of bootstrap approach. Denote $\hat{y}_{im}$ as the m^{th} imputed value for i^{th} case. Suppose that sampled units y_1 to y_{i-1} are complete and y_i to y_n are missing. Denote the FMI for the current set of data (and model) as FMI_{i-1}. Then we want to estimate the FMI if we were able to collect data for case i. This is dentoted FMI_i. It is estimated using the following bootstrap approach:

$$FMI_i = [\sum_{l=1}^{M} (\bar{V}_l^* + (1 + M^{-1}) B_l^*)] / M \ ,$$

where M is the number of imputations. In this case, $\bar{V}^*$ and B^* are estimated assuming that each of the M imputations for case i is actually known. Then the M estimates of the FMI are averaged to produce an FMI estimate for the situation where case i is added to the dataset. I might also consider using the range or the minimum (similar to my proposed learning model) of the M estimates of the FMI as an indicator. The cases that have the smallest predicted FMI will be targeted for interview.

3.2 Experiments in Adaptive Strategies

I proposed a learning model for surveys. The ideal situation would be to conduct an experiment using this approach on a sample while a sample that was given the standard approach was run simultaneously. The results could be compared over time. Ideally, the survey variable would be known for all cases to facilitate the assessment of nonresponse bias.

I would also like to explore the impact of answering machines. It appears that leaving message has a differential impact on potential respondents. SCA recently implied a standardized protocol for answering machines. This could provide a useful contrast for an experiment that used targeted answering machine messages that were matched to the respondent. After experimenting with targeting tailored to the fixed characteristics of the case, I would like to experiment with the placement of these messages in the sequence. Our preliminary analysis found that we most frequently left messages on the first attempt and that the frequency of messages declined with subsequent attempts. Does it make a

difference at what point in the sequence these messages are left? In particular, the rule currently in place leaves messages after the first contact. Could these be harmful?

3.3 Cost-Error Tradeoffs

I have ignored this problem in this dissertation. This is a very important problem. The results from the analysis of SCA data suggested that higher contact rates could be achieved if answering machine messages with an offer of an incentive were left. However, these incentives are expensive to implement. The question becomes, given a fixed budget, where should we allocate resources? It might be the case that those incentives are essential if we are to have any hope for some particular cases to be interviewed. The analysis problem is the question of whether a cheaper protocol might obtain nearly the same results.

One interesting subquestion is whether we can determine more quickly than current methods whether to change strategies. This might offer the possibility of saving money. However, my hypothesis would be that for a fixed budget, you might consider differentially allocating resources to cases to produce the highest information content. The problem then is to somehow maximize information content. This problem has to be handled sequentially. This makes it more difficult.

3.4 Computerized Decision Support Model

The implementation of adaptive designs in field surveys will require more centralized control of protocols that are administered by field interviewers. This is akin to the situation for adaptive treatment regimes in the medical field. These complex regimes are often implemented by physicians. There is some research in this field indicating that

training and information are not very effective means for disseminating new treatment recommendations (Lin et al., 1997; Garg et al., 2005). The previously cited literature on Computerized Decision Support Systems indicates that these methods can improve physicians' ability to adopt evidence-based treatment recommendations (Kawamoto, 2005, Garg et al., 2005; Hersh and Hickam, 1998; Hunt et al., 1998). There is also research into the features that make these software applications effective (Kawamoto, 2005; Trivedi et al., 2007; Sim et al., 2001). This research may help us design new software systems for field staff. Properly designed systems have been effective in the medical realm. I expect they will be helpful for field surveys as well.

The most obvious candidate for experimental testing is to deliver a recommendation to each interviewer about the best time to attempt contacting any given household. I would expect a period of time in which interviewers would need to learn the effectiveness of these recommendations before they would begin to follow them. They would need to be convinced that the recommendations are sound and learn when they are not. Currently, we notice that even with training, it takes some time for interviewers to learn which times are generally best for calling attempts.

3.5 Stopping Rules

The "stop and impute" rule appears to be useful. The model developed in this dissertation is for normally distributed variables. I would like to extend this rule to other types of variables as well – poisson, binomial, etc. This should be quite easily done.

In the current climate, surveys may be reluctant to implement such a rule. Most surveys have specified targets for response rate and sample size. I would propose running the stopping rule analysis with an ongoing survey. Determine each month when data

collection would have stopped under the "stop and impute" rule as well key survey estimates (adjusted) and their variances at that point. These could be compared to the actual stopping point and the final achieved key survey estimates and variances at that point. The problem would arise in those cases when the "stop and impute" rule did not recommend stopping before the current rule (sample size or response rate). Over time, this approach could help develop confidence in the method until it might be accepted for use. At that point, the interesting problem would be to see what happens when data collection continues beyond what would have been carried out under the response rate/sample size rule. No observational data exist for this scenario.

4. Conclusion

Certainly, nonresponse bias is a risk that surveys must face. As with all missing data problems, the best strategy is to get complete data. Unfortunately, for most surveys, this is simply not possible. The next best thing is to use some reasonable model to account for the missing data. While the response rate may be based on complete data (the response indicator), the nonresponse bias is a function of the response rate and data that are usually missing (the nonresponders survey values). The response rate as an indicator of data quality is important only insofar as it affects the behavior of survey organizations. Unfortunately, the use of the response rate has led to behaviors that may be unhealthy for surveys. The focus is on the status of the response indicator, and not on the total set of data – complete frame data with a response indicator and incomplete survey data. Survey methods have largely been focused on increasing the number of respondents without much regard for who is responding. If the field of survey methodology is to develop

methods for dealing with nonresponse bias, then new indicators of nonresponse bias are needed. These indicators should perform better than the response rate.

The fraction of missing information is the indicator that I have proposed. This measure is model-based. However, the model assumption allows us to use more of the data – it uses the incomplete survey data. If our models are reasonable, then this measure should be informative of the risk of nonresponse bias. The assumption that our models are reasonable will be used again in the post-survey adjustment procedures. If we make this assumption during data collection, then the proposed indicator may guide effort toward collecting data that is more information relative to data collected under another guiding indicator.

I have also proposed that surveys may benefit from using techniques developed in the field of clinical trials. These techniques have been called "dynamic treatment regimes." The goal of these regimes is to identify treatment courses that adapt to the fixed characteristics and history of previous treatments for each case in order to identify optimal methods. This approach allows us to consider interactions not only between the characteristics of potential respondents and the proposed protocols, but also interactions between the different protocols used at different phases of the survey process. Placing the various design features under consideration in their context may help us produce more robust results and develop more controlled experiments.

This is an exciting time for the field of survey methods. The specter of nonresponse bias has created an opportunity for research that delves into the dual components of this bias – the nonresponse rate and the difference between responders

and nonresponders on the survey value. New indicators for this risk are needed. Armed with new indicators, we may begin exploring new methods for reducing this risk.

References

Atrostic, B. K., N. Bates, et al. (2001). "Nonresponse in US Government Household Surveys: Consistent Measures, Recent Trends, and New Insights." Journal of Official Statistics **17**(2): 209-226.

Babbie, E. R. (2007). The practice of social research. Belmont, CA, Thomson Wadsworth.

Bates, N. (2003). "Contact Histories in Personal Visit Surveys: The Survey of income and Program Participation (SIPP) Methods Panel." Presentation at the Annual Conference of the American Association for Public Opinion Research, Nashville, Tennessee.

Bertsimas, D. and A. J. Mersereau (2007). "A Learning Approach for Interactive Marketing to a Customer Segment." Operations Research **55**(6): 1120-1135.

Bethlehem, J. G. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. Survey Nonresponse. R. M. Groves. New York, Wiley.

Bickart, B. and D. Schmittlein (1999). "The Distribution of Survey Contact and Participation in the United States: Constructing a Survey-Based Estimate." Journal of Marketing Research **36**(2): 286-294.

Biemer, P. P. and L. Lyberg (2003). Introduction to survey quality. Hoboken, N.J. ; Chichester, Wiley.

Bollapragada, S. and S. Nair. (2001). "Improving Right Party Contact Rates at Outbound Call Centers." General Electric Technical Information Series, from http://www.crd.ge.com/cooltechnologies/pdf/2001crd106.pdf.

Brick, J. M., B. Allen, et al. (1996). "Outcomes of a Calling Protocol in a Telephone Survey." Proceedings of the Survey Research Methods Section of the American Statistical Association: 142–149.

Bult, J. R., H. van der Scheer, et al. (1997). "Interaction between target and mailing characteristics in direct marketing, with an application to health care fund raising." International Journal of Research in Marketing **14**(4): 301-308.

Carton, A. and G. Loosveldt. (1999). "How the Quality of the Initial Contact can be Determinant for the Final Response Rate in Face to Face Surveys." from http://www.jpsm.umd.edu/icsn/papers/carton.htm

Cobben, F. and B. Schouten (2007). An empirical validation of R-indicators, Statistics Netherlands, Division of Technology and Methodology. **2007:** Manuscript under preparation.

Collins, L. M., S. A. Murphy, et al. (2004). "A conceptual framework for adaptive preventive interventions." Prevention Science **5**(3): 185-96.

Collins, L. M., S. A. Murphy, et al. (2005). "A strategy for optimizing and evaluating behavioral interventions." Annals of Behavioral Medicine **30**(1): 65-73.

Collins, L. M., S. A. Murphy, et al. (2007). "The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent eHealth Interventions." American Journal of Preventive Medicine **32**(5, Supplement 1): S112-S118.

Collins, L. M., J. L. Schafer, et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." Psychological Methods 6(4): 330-351.

Copas, A. J. and V. T. Farewell (1998). "Dealing with Non-Ignorable Non-Response by Using an `Enthusiasm-To-Respond' Variable." Journal of the Royal Statistical Society. Series A (Statistics in Society) **161**(3): 385-396.

Couper, M. P. (1997). "Survey Introductions and Data Quality." Public Opinion Quarterly **61**(2): 317-338.

Couper, M. P. (1998). "Measuring Survey Quality in a CASIC Environment." Proceedings of the Survey Research Methods Section of the American Statistical Association: 41-49.

Couper, M. P. (2000). "Usability Evaluation of Computer-Assisted Survey Instruments." Social Science Computer Review **18**(4): 384-396.

Cunningham, P., D. Martin, et al. (2003). "An Experiment in Call Scheduling." Proceedings of the Section on Survey Research Methods, American Statistical Association: 59-66.

Curtin, R. (2007). "Consumer Sentiment Surveys: Worldwide Review and Assessment." Journal of Business Cycle Measurement and Analysis **2007**(1): 9-45.

Curtin, R., S. Presser, et al. (2000). "The Effects of Response Rate Changes on the Index of Consumer Sentiment." Public Opinion Quarterly **64**(4): 413-428.

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

de Leeuw, E. and W. de Heer (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Survey Nonresponse. R. M. Groves.

New York, John Wiley & Sons**:** 41-54.

de Leeuw, E., J. Hox, et al. (2005). "The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis." Working Paper series of the Program in Survey Research and Methodology(10): 1-26.

Dempster, A. P., N. M. Laird, et al. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society. Series B (Methodological) **39**(1): 1-38.

Dennis, J. M., C. Saulsberry, 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.

Fisher, R. A. (1925). "Theory of statistical estimation." Proceedings of the Cambridge Philosophical Society **22**(5): 700-725.

Garg, A. X., N. K. J. Adhikari, et al. (2005). "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review." JAMA **293**(10): 1223-1238.

Gelfand, A. E. and A. F. M. Smith (1990). "Sampling-Based Approaches to Calculating Marginal Densities." Journal of the American Statistical Association **85**(410): 398-409.

Gelman, A. and D. B. Rubin (1992). "Inference from iterative simulation using multiple sequences." Statistical Science **7**(4): 457-472.

Gooley, C. G. and J. M. Lattin (2000). Dynamic Customization of Marketing Messages in Interactive Media, Graduate School of Business, Stanford University.

Graham, J., A. Olchowski, et al. (2007). "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." Prevention Science **8**(3): 206-213.

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

Groves, R. M. (2005). "Research synthesis: nonresponse rates and nonresponse error in household surveys." 16th International Workshop on Household Survey Nonresponse, Tällberg, Sweden: 28-31.

Groves, R. M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys." Public Opinion Quarterly **70**(5): 646-675.

Groves, R. M., J. M. Brick, T.W. Smith, J. Wagner. (2008). “Alternative Practical Measures of Representativeness of Survey Respondent Pools.” Preseneted at the Annual Conference of the American Association for Public Opinion Research.

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

Groves, R. M. and S. G. Heeringa (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs." Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(3): 439-457.

Groves, R. M., N. Kirgis, et al. (2008). Responsive Design for Household Surveys: Illustration of Management Interventions Based on Survey Paradata.

Groves, R. M., E. Peytcheva, et al. (2007). Use of Interviewer Judgments About Attributes of Selected Respondents in Postsurvey Adjustment for Unit Nonresponse: An Illustration with the National Survey of Family Growth. Presented at Joint Statistical Meetings

Groves, R. M., S. Presser, et al. (2004). "The Role of Topic Interest in Survey Participation Decisions." Public Opinion Quarterly **68**(1): 2-31.

Groves, R. M., E. Singer, et al. (2000). "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." Public Opinion Quarterly **64**(3): 299-308.

Guyer, H., J. Wagner, et al. (2007). Impact of the Use of Respondent Profiles on Response Rates and Efficiency. Preseneted at the Annual Conference of the American Association for Public Opinion Research.

Hansen, M. H., W. G. Madow, et al. (1983). "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys." Journal of the American Statistical Association **78**(384): 776-793.

Harel, O. and D. Miglioretti (2007). "Missing Information as a Diagnostic Tool for Latent Class Analysis." Journal of Data Science **5**: 269-288.

Heerwegh, D., K. Abts, et al. (2007). "Minimizing survey refusal and noncontact rates: do our efforts pay off?" Survey Research Methods **1**(1).

Hersh, W. R. and D. H. Hickam (1998). "How Well Do Physicians Use Electronic Information Retrieval Systems?: A Framework for Investigation and Systematic Review." JAMA **280**(15): 1347-1352.

Houtkoop-Steenstra, H. and H. van den Bergh (2000). "Effects of Introductions in Large Scale Telephone Survey Interviews." Sociological Methods Research **28**(3): 281-300.

Hunt, D. L., R. B. Haynes, et al. (1998). "Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review." JAMA **280**(15): 1339-1346.

Johnson, T. and L. Owens (2003). "Survey response rate reporting in the professional literature." Annual Conference of the American Association for Public Opinion Research, Nashville, Tenn., May 15.

Johnson, T. P., Y. I. K. Cho, et al. (2006). "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey." Public Opinion Quarterly **70**(5): 704-719.

Kawamoto, K., C. A. Houlihan, et al. (2005). "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success." British Medical Journal **330**(7494): 765.

Keeter, S., C. Kennedy, et al. (2006). "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey." Public Opinion Quarterly **70**(5): 759-779.

Keeter, S., C. Miller, et al. (2000). "Consequences of Reducing Nonresponse in a National Telephone Survey." Public Opinion Quarterly **64**(2): 125-148.

Kennickell, A. B. (2003). "Reordering the Darkness: Application of Effort and Unit Nonresponse in the Survey of Consumer Finances." Proceedings of the Section on Survey Research Methods, American Statistical Association: 2119-2126.

Kennickell, A. B. (2004). "Action at a Distance: Interviewer Effort and Nonresponse in the SCF." from http://www.federalreserve.gov/Pubs/OSS/oss2/papers/asa2004.5.pdf.

Kessler, R. C., R. J. A. Little, et al. (1995). "Advances in strategies for minimizing and adjusting for survey nonresponse." Epidemiologic Reviews **17**(1): 192-204.

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

Lai, T. L. (1987). "Adaptive Treatment Allocation and the Multi-Armed Bandit Problem." The Annals of Statistics **15**(3): 1091-1114.

Lavori, P. W. and R. Dawson (2000). "A design for testing clinical strategies: biased adaptive within-subject randomization." Journal of the Royal Statistical Society. Series A (Statistics in Society) **163**(1): 29-38.

Lavori, P. W. and R. Dawson (2004). "Dynamic treatment regimes: practical design considerations." Clinical Trials **1**(1): 9-20.

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

Leacock, C. P. (2006). "Getting Started with the Health and Retirement Study." from

http://hrsonline.isr.umich.edu/docs/dmgt/IntroUserGuide.pdf.

Lin, E. H. B., W. J. Katon, et al. (1997). "Achieving Guidelines for the Treatment of Depression in Primary Care: Is Physician Education Enough?" Medical Care **35**(8): 831-842.

Link, M. W. and A. Mokdad (2005). "Advance Letters as a Means of Improving Respondent Cooperation in Random Digit Dial Studies: A Multistate Experiment." Public Opinion Quarterly **69**(4): 572-587.

Link, M. W., A. Mokdad, et al. (2003). "Improving Response Rates for the BRFSS: Use of Lead Letters and Answering Machine Messages." annual meeting of the American Association for Public Opinion Research, Nashville.

Little, R. J. A. and D. B. Rubin (2002). Statistical Analysis with Missing Data. Hoboken, N.J. :, Wiley.

Little, R. J. A. and S. Vartivarian (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" Survey Methodology **31**(2): 161-168.

Lu, R. C., J. Hall, et al. (2002). "Resolvability, Screening, and Response Models in RDD Surveys: Utilizing Genesys Telephone-Exchange Data." JSM-SRMS Conference Proceedings **2002**: 2198-2202.

Lynn, P. (2003). "PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents." Quality and Quantity **37**(3): 239-261.

Lynn, P., P. Clarke, et al. (2002). The Effects of Extended Interviewer Efforts on Nonresponse Bias. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons.

McCarty, C., M. House, et al. (2006). "Effort in Phone Survey Response Rates: The Effects of Vendor and Client-Controlled Factors." Field Methods **18**(2): 172-188.

Merkle, D., M. Edelman, et al. (1998). "An Experimental Study of Ways to Increase Exit Poll Response Rates and Reduce Survey Error." annual conference of the American Association for Public Opinion Research, St. Louis, Missouri.

Merkle, D. M. and M. Edelman (2002). Nonresponse in Exit Polls: A Comprehensive Analysis. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons**:** 243-257.

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): 1455-81.

Murphy, S. A., K. G. Lynch, et al. (2007). "Developing adaptive treatment strategies in substance abuse research." Drug and Alcohol Dependence **88**(Supplement 2): S24-S30.

National Center for Educational Statistics. (2002). "NCES Statistical Standards." from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003601.

National Institutes of Health. (1998). "NIH Policy for Data and Safety Monitoring." from http://grants.nih.gov/grants/guide/notice-files/not98-084.html.

Nicoletti, C. and F. Peracchi (2005). "Survey response and survey characteristics: microlevel evidence from the European Community Household Panel." Journal of the Royal Statistical Society: Series A (Statistics in Society) **168**(4): 763-781.

O'Brien, P. C. and T. R. Fleming (1979). "A multiple testing procedure for clinical trials." Biometrics **35**(3): 549-556.

Office of Management and Budget. (2006). "Standards and Guidelines for Statistical Surveys." from http://www.whitehouse.gov/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Orchard, T. and M. A. Woodbury (1972). "A Missing Information Principle: Theory and Applications." Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability **1**: 697-715.

Petroni, R., R. Sigman, et al. (2004). "Response Rates and Nonresponse in Establishment Surveys BLS and Census Bureau." Presented to the Federal Economic Statistics Advisory Committee: 1-50.

Pocock, S. J. (1977). "Group sequential methods in the design and analysis of clinical trials." Biometrika **64**(2): 191-199.

Pocock, S. J. (1983). Clinical trials: a practical approach, Wiley, Chichester.

Potthoff, R. F., K. G. Manton, et al. (1993). "Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks." Journal of the American Statistical Association **88**(424): 1197-1207.

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

Raghunathan, T., J. M. Lepkowski, et al. (2001). "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." Survey Methodology **27**(1).

Rancourt, E. (2002). Using Variance Components to Measure and Evaluate the Quality of Editing Practices. U. N. S. C. a. E. C. f. Europe.

Rao, R. S., M. E. Glickman, et al. (2007). "Stopping rules for surveys with multiple waves of nonrespondent follow-up." Statistics in Medicine(Early epublication).

Rogers, A., M. A. Murtaugh, et al. (2004). "Contacting Controls: Are We Working Harder for Similar Response Rates, and Does It Make a Difference?" American Journal of Epidemiology **160**(1): 85-90.

Roose, H., J. Lievens, et al. (2007). "The Joint Effect of Topic Interest and Follow-Up Procedures on the Response in a Mail Questionnaire: An Empirical Test of the Leverage-Saliency Theory in Audience Research." Sociological Methods & Research **35**(3): 410.

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika **70**(1): 41-55.

Rossi, P. E., R. E. McCulloch, et al. (1996). "The Value of Purchase History Data in Target Marketing." Marketing Science **15**(4): 321-340.

Rubin, D. B. (1987). Multiple imputation for Nonresponse in Surveys. New York ;, Wiley.

Rubin, D. B. (1996). "Multiple Imputation After 18+ Years." Journal of the American Statistical Association 91(434): 473-489.

Rubin, D. B. and N. Schenker (1987). "Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes." Journal of Official Statistics 3(4): 375-387.

Schouten, B. and F. Cobben. (2007). "R-Indexes for the Comparison of Different Fieldwork Strategies and Data Collection Modes." Retrieved June 2007.

Sim, I., P. Gorman, et al. (2001). "Clinical Decision Support Systems for the Practice of Evidence-based Medicine." Journal of the American Medical Informatics Association **8**(6): 527-534.

Simester, D. I., P. Sun, et al. (2006). "Dynamic Catalog Mailing Policies." Management Science **52**(5): 683-696.

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons**:** 163-178.

Singer, E., J. V. Hoewyk, et al. (1999). "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." Journal of Official Statistics **15**(2): 217-230.

Snijkers, G., J. Hox, et al. (1999). "Interviewers' Tactics for Fighting Survey Nonresponse." Journal of Official Statistics **15**(2): 185-198.

Spiegelhalter, D. J., K. R. Abrams, et al. (2004). Bayesian Approaches to Clinical Trials

and Health-Care Evaluation. Chichester ; Hoboken, NJ, John Wiley & Sons.

Sutton, R. S. and A. G. Barto (1998). Reinforcement Learning: An Introduction. Cambridge, Mass., MIT Press.

Thall, P. F., R. E. Millikan, et al. (2000). "Evaluating multiple treatment courses in clinical trials." Statistics in Medicine **19**(8): 1011-28.

Thall, P. F., H. G. Sung, et al. (2002). "Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials." Journal of the American Statistical Association **97**: 29-39.

Thall, P. F. and J. K. Wathen (2005). "Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments." Statistics in Medicine **24**(13): 1947-64.

Tierney, L. (1994). "Markov chains for exploring posterior distributions." Annals of Statistics **22**(4): 1701–1762.

Trivedi, M. H. and E. J. Daly (2007). "Measurement-based care for refractory depression: A clinical decision support model for clinical research and practice." Drug and Alcohol Dependence **88**(Supplement 2): S61-S71.

Trivedi, M. H., J. K. Kern, et al. (2002). "Development and implementation of computerized clinical guidelines: Barriers and solutions." Methods of Information in Medicine **41**(5): 435-442.

Trussell, N. and P. J. Lavrakas (2004). "The Influence of Incremental Increases in Token Cash Incentives on Mail Survey Response: Is There an Optimal Amount?" Public Opinion Quarterly **68**(3): 349-367.

van der Grijn, F., B. Schouten, et al. (2006). Balancing Representativity, Costs and Response Rates in a Call Scheduling Strategy. Occasional Paper produced by Statistics Netherlands**:** 1-20.

van Wyk, J. T., M. A. van Wijk, et al. (2008). "Electronic Alerts Versus On-Demand Decision Support to Improve Dyslipidemia Treatment. A Cluster Randomized Controlled Trial." Circulation.

Wang, K., J. Murphy, et al. (2005). Are Two Feet in the Door Better than One? Using Process Data to Examine Interviewer Effort and Nonresponse Bias. Federal Committee on Statistical Methodology Research Papers.

Weeks, M. F., B. L. Jones, et al. (1980). "Optimal Times to Contact Sample Households." Public Opinion Quarterly **44**(1): 101-114.

Weeks, M. F., R. A. Kulka, et al. (1987). "Optimal Call Scheduling for a Telephone Survey." Public Opinion Quarterly **51**(4): 540-549.

Winship, C. and S. L. Morgan (1999). "The estimation of causal effects from observational data." Annual Review of Sociology **25**: 659-707.

Wisniewski, S. R., D. Stegman, et al. (2004). "Methods of testing feasibility for sequenced treatment alternatives to relieve depression (STAR*D)." Journal of Psychiatric Research **38**(3): 241-8.

Xu, M., B. J. Bates, et al. (1993). "The Impact of Messages on Survey Participation in Answering Machine Households." Public Opinion Quarterly **57**(2): 232-237.