ABSTRACT

Title of Dissertation:	BAYESIAN METHODS FOR PREDICTION OF SURVEY DATA COLLECTION PARAMETERS IN ADAPTIVE AND RESPONSIVE DESIGNS
	Stephanie Michelle Coffey Doctor of Philosophy, 2020
Dissertation Directed By:	Professor Michael R. Elliott Joint Program in Survey Methodology Michigan Program in Survey Methodology

Adaptive and responsive survey designs rely on estimates of survey data collection parameters (SDCPs), such as response propensity, to make intervention decisions during data collection. These interventions are made with some data collection goal in mind, such as maximizing data quality for a fixed cost or minimizing costs for a fixed measure of data quality. Data quality may be defined by response rate, sample representativeness, or error in survey estimates. Therefore, the predictions of SDCPs are extremely important.

Predictions within a data collection period are most commonly generated using fixed information about sample cases, and accumulating paradata and survey response data. Interventions occur during the data collection period, however, meaning they are applied based on predictions from incomplete accumulating data. There is evidence that the incomplete accumulating data can lead to biased and unstable predictions, particularly early in data collection.

This dissertation explores the use of Bayesian methods to improve predictions of SDCPs during data collection, by providing a mathematical framework for combining priors, based on external data about covariates in the prediction models, with the current accumulating data to generate posterior predictions of SDCPs for use in intervention decisions.

This dissertation includes three self-contained papers, each focused on the use of Bayesian methods to improve predictions of SDCPs for use in adaptive and responsive survey designs. The first paper predicts time to first contact, where priors are generated from historical survey data. The second paper implements expert elicitation, a method for prior construction when historical data is not available. The last paper describes a data collection experiment conducted using a Bayesian framework, which attempts to minimize data collection costs without reducing the quality of a key survey estimate. In all three papers, the use of Bayesian methods introduces modest improvements in the predictions of SDCPs, especially early in data collection, when interventions would have the largest effect on survey outcomes. Additionally, the experiment in the last paper resulted in significant data collection cost savings without having a significant effect on a key survey estimate. This work suggests that Bayesian methods can improve predictions of SDCPs that are critical for adaptive and responsive data collection interventions.

BAYESIAN METHODS FOR PREDICTION OF SURVEY DATA COLLECTION PARAMETERS IN ADAPTIVE AND RESPONSIVE DESIGNS

by

Stephanie Michelle Coffey

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Michael R. Elliott, Chair
Professor Mei-Ling Ting Lee
Professor Yan Li
Associate Professor James Wagner
Associate Professor Brady T. West

Acknowledgements

Thank you to my committee members for your patience and support throughout my dissertation. Thank you also for your generosity with your time and expertise - I am a better researcher because of this experience. In particular, thank you to my chair, Mike Elliott – your expertise was invaluable, as was all the time you spent helping me formulate my research, reviewing my work, and providing feedback. Also, thank you for taking on a student several states away, and being willing to travel (and answer emails and phone calls) to make sure I always felt like my advisor was available.

Research reported in this dissertation was supported by the National Center for Science and Engineering Statistics of the National Science Foundation under contract number 0099-FY16-IPA-0002, and the National Institute on Aging of the National Institutes of Health under Award Number R01AG058599. I am grateful for this financial support, without which this research would not have been possible. Also, thank you to James Wagner for giving me the opportunity to author a paper supported by your grant.

Thank you to my friends, family and colleagues for always encouraging me to keep at it. Thank you to Ed Binkowski for setting me on this path in the first place – you are missed. And thank you to Marc. I could not have done this without your encouragement, understanding, and support.

Table of Contents

List of Tables

List of Figures

1. Summary Introduction

Survey organizations concern themselves with metrics of data collection progress, cost and quality when designing surveys and managing survey data collection operations. Particularly, as data collection costs increase alongside survey nonresponse, there is a clear need to implement data collection designs that are both cost-effective and high quality to meet the continuing needs of policy makers and researchers that rely on survey data.

Metrics of data collection progress, cost and quality are created based on a specific survey's data collection goals, and are monitored in an effort to determine whether data collection is staying "on track" or not. If not, survey managers may respond in a variety of ways to help the survey meet expectations. Providing feedback to survey interviewers, reassigning workload, or extending data collection by a short time are common *ad hoc* changes to data collection procedures, implemented as needed to help reach survey data collection goals in interviewer-assisted surveys. Case prioritization, tailored data collection, and mode switching have been tested in a variety of settings in order to evaluate the potential of a variety of data collection features to help improve survey outcomes. More recently, adaptive and responsive survey designs have emerged as a more structured way of intervening in data collection based on cost and quality metrics.

Regardless of the type of monitoring or intervention under consideration for implementation, all of these strategies require survey managers to develop expectations of what "should" happen during data collection, or the expected change in data collection outcomes if a particular intervention is carried out. These expectations allow survey managers to determine if data collection is progressing as expected. If not, and if an intervention is determined to be necessary, expectations help evaluate whether the intervention worked as expected.

This dissertation focuses on how the expectations of progress, cost and quality metrics are created, specifically during the data collection process. Currently, it is very common for data collection organizations to generate expectations using historical or current round survey data. For example, a simple expectation of progress based on historical data may be that 80% of sample cases are attempted 20% of the way through data collection. That expectation could be based on just the prior implementation, or an average of several prior implementations.

Historical data is also used in a more nuanced way than simple averages. For example, survey managers could build a predictive model for expected data collection costs, based on covariates available prior to data collection. After building the model on the prior round of data collection, the estimated coefficients can be applied to the current round of data collection to generate predicted costs at the case level. Data collection costs can then be monitored with respect to the expected cost. Response propensity models can be estimated in a similar way with historical data, and could be used early in data collection to identify cases that are less likely to respond. Those cases may be identified as requiring an intervention. Finally, historical data may be used to create predictive models for survey responses. By applying these models to current round data, expectations of survey estimates could be created. Then, as responses are collected, metrics of quality like mean squared error of the actual versus the predicted estimates, or the stability of estimates over time could be monitored throughout data collection, leading to interventions that may include stopping data collection for particular cases.

Current data can be used similarly, but in this case, the accumulating current round data is used to estimate predictive model coefficients, and those coefficients are used to estimate predictions for the remainder of the current sample. For example, response propensity can be calculated at a point in data collection, based on the response status for all cases at that time. Estimating response propensity at different times throughout a data collection period will result in different predictions, based on the amount of data accumulated.

While these approaches are common in practice, there are drawbacks to using only historical data or only current data to develop these expectations. Using only historical data to develop expectations requires several assumptions. Most significantly, this process assumes that the current implementation of the survey will proceed very similarly to past implementations. If there are time lags between implementations or changes in response behavior over time, this may not be a reasonable assumption. Additionally, using only historical data to set expectations means that if the current implementation does not follow the past implementations, there is no systematic way to update expectations with the new information from the current implementation. Ignoring historical data and focusing only on the current implementation to develop these expectations assumes that partial data collected during the early part of a data collection implementation is representative of data that will be collected later in that same implementation. If the survey implementations are fairly similar over time, this may ignore valuable predictive information.

This dissertation explores the potential of a Bayesian framework for generating predictions of parameters of interest during a survey data collection. Bayesian models for prediction utilize both historical and current data, in an attempt to overcome the limitations of each of the two sources used independently. For any data collection parameter of interest (e.g., time to contact with a respondent, response propensity, etc.), prior beliefs are constructed from information external to the current round of data collection. The point estimates and standard errors generated from that external information reflect the fact that external data is being used as an initial assumption for the coefficient estimates. Then, those priors are updated as the current data is accumulated, resulting in posterior predictions that take both sources of data into account. The strength of the prior coefficient, determined by its standard error, is a reflection of confidence in the prior being representative of the current round of data collection.

Due to its relative novelty in survey data collection, the use of a Bayesian framework for generating these predictions during data collection requires validation of its usefulness, from both simulations and experimental implementations. The research contained in this dissertation contributes both types of evidence, using large national surveys that have characteristics common to many other surveys. This research is part of a broader research agenda in Bayesian methods for improving survey outcomes.

This dissertation includes three self-contained papers, each focused on the use of Bayesian methods to improve predictions of survey data collection parameters for use in adaptive and responsive survey designs. The first paper incorporates Bayesian methods when predicting time to first contact, where priors are generated from historical survey data. The second paper implements and evaluates expert elicitation, a method for prior construction when historical data is not available. The last paper describes a data collection experiment conducted using a Bayesian framework, which attempts to minimize data collection costs without reducing the quality of a key survey estimate. In all three papers, the use of Bayesian methods introduces modest improvements in the predictions of survey data collection parameters, especially early in data collection, when interventions would have the largest effect on survey outcomes. Additionally, the experiment in the third paper resulted in significant data collection cost savings without having a significant effect on a key survey estimate. While areas for improvement are discussed throughout these papers, this dissertation suggests that Bayesian methods can improve predictions of survey data collection parameters that are critical for adaptive and responsive data collection interventions.

2. Predicting Time to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach

Stephanie Coffey[1], Michael R. Elliott[2,3,4], James Dahlhamer[5]

[1] Joint Program in Survey Methodology, University of Maryland
United States Census Bureau, Washington, DC
[2] Survey Research Center, Institute for Social Research, University of Michigan
[3] Michigan Program in Survey Methodology, University of Michigan
[4] Department of Biostatistics, University of Michigan
[5] National Center for Health Statistics, Hyattsville, Maryland

Abstract

Monitoring and reducing time to respondent contact can help reduce nonresponse due to noncontact in surveys. In order to intervene based on data monitoring, survey managers need an estimate of expected time to respondent contact. Surveys currently estimate these types of expectations, often using historical means, models built from historical data, or models built using the current, partial, round of data collection. We propose a new method that, under a Bayesian framework, utilizes historical data in the form of priors on model coefficients and combines those with current accumulating data to estimate posterior predictions of the expected time to contact. Results demonstrate that the Bayesian method results in lower root mean squared error of predictions of time to contact than the other, more commonly used, methods.

2.1 Introduction

Making contact with a survey respondent is an important part of the data collection process. Noncontact makes up a significant portion of nonresponse in face-to-face surveys. Durrant and Steele (2013) discuss six large, government-sponsored, face-to-face household surveys carried out in the UK that have noncontact rates ranging from eight percent to 40 percent of sample cases. Williams and Brick (2017) report that for nine government-sponsored, primarily face-to-face household surveys in the United States, noncontact rates as a portion of nonresponse ranged from four percent to 40 percent in 2014, the most recent year included in the paper. Additionally, Willimack and Dalzell (2006) found that noncontact was even a significant component of nonresponse in establishment surveys, though their focus is on nonresponse follow-up by telephone.

Groves and Couper (1998) discuss methods for dealing with noncontact, including increasing the number of contact attempts in face-to-face surveys. Without making successful contact, it is not only impossible to obtain cooperation and a completed interview, but it may also be impossible to determine the eligibility of a case, affecting response rates, nonresponse adjustments, and ultimately, variances of estimates. Additionally, if noncontact is related to the survey items of interest, nonresponse due to noncontact may also introduce or increase nonresponse bias in survey estimates.

Because a successful contact on the first attempt is not guaranteed, interviewers must make attempts early enough in the data collection period to make contact with all sample units. Waiting too long into the data collection period to begin making attempts could lead to nonresponse that is due to insufficient progress rather than post-contact reluctance or refusals. Further, the data collection resources that are expended making these attempts that do not result in a contact can increase overall survey costs without improving response rates, and increase measures of costs used for evaluations, such as costs-per-case or costs-per-complete.

A lag between first attempt and first contact means there is a time gap between first attempt and the ultimate resolution of a case. If an interviewer never makes contact, it is difficult to resolve the case correctly with respect to eligibility or refusal status. Additionally, if long lags are accompanied by several unsuccessful attempts, the interviewer may not be optimizing their attempts to obtain a contact, while increasing survey costs. Finally, as the end of data collection approaches, long lags mean that cases that have not been worked might never be completed, affecting response rates.

In order to improve survey outcomes, survey managers might want to reduce the lag between a first attempt and first contact with a sample member by intervening with respect to individual cases shortly after a first contact attempt is made. Those interventions could reflect different passive or active levels of management at different points during data collection. A more *passive* intervention might simply involve field supervisors informing interviewers of the cases that are expected to be difficult to make contact with once the first attempt is made. Alternatively, an interviewer whose cases are approaching (or exceeding) expected lag times might prompt feedback from a supervisor, and the earlier that intervention can occur, the faster an improvement might be made (Maitland, Hubbard and Edwards 2016). An *active* intervention could involve reassigning cases from an interviewer who is taking longer than expected to make contact with a sample unit to an interviewer with a smaller remaining workload, or one who is making contact with their sample units within the expected timeframe.

In order to implement any of these interventions during the data collection period, it is necessary to first determine how long it “should” take to make contact with a particular sample unit. Once estimates of expected lags are generated, supervisors then can use those expectations to provide input, feedback, or implement other interventions. Ideally, these estimates of expected lag would be produced during the data collection period itself, and as close to the time of the first contact attempt as possible, in order to facilitate interventions while the interviewers have time to alter their behavior. Therefore, these estimates are actually predictions of the expected lag between the first contact attempt and the first contact with a respondent.

Currently, interim predictions of estimators of progress, cost or quality are generated during survey data collection, and the most common prediction made is response propensity. Groves and Heeringa (2006) developed discrete time logistic regression models (Singer and Willett 2003) for the National Survey of Family Growth (NSFG) to predict the propensity to respond at the next contact attempt in order to determine the phase capacity of a data collection operation. In Schouten, Cobben and Bethlehem (2009) and Schouten, Shlomo and Skinner (2011), daily estimates of response propensity were used to generate representativeness (R-) indicators, which are then used to identify over- and under-represented groups of cases, with respect to the overall selected sample. Chesnut (2013) used discrete time logistic regression to predict the daily propensity to respond to the American Community Survey (ACS) by web, in order to determine the optimal time to switch nonresponding cases to the mail mode. West and Groves (2013) used discrete time logistic regression models to estimate the probability of responding at the next contact. These probabilities were used to compute difficulty-adjustments in interviewer evaluation metrics. Coffey, Reist and Miller (2019) generated predictions for propensities in order to identify cases for adaptive interventions.

In all of these examples, only the current round of data collection was used to generate predictions, including sample information, administrative data known that could be linked to the sample, and paradata (Couper 2000; 2017) accumulated up to each prediction time. Wagner and Hubbard (2014) demonstrated that, by relying only on the partial current data as it accumulates, predictions generated from this partial data might be biased. As a result, intervention decisions made based on these predictions can be inefficient or even harmful. When only the current round of data is utilized, one is effectively extrapolating the relationships found between covariates and outcomes using partial paradata, collected during the early part of fieldwork. The implicit assumption, then, is that the data collected in the early part of data collection is representative of data that will be collected later. Wagner and Hubbard (2014) demonstrated that this assumption does not necessarily hold. Further, relying only on current data does not leverage potentially valuable historic or external information.

There are some examples in the literature of utilizing historical data. Historical data does not necessarily mean prior contact information for *the same case*, as in a longitudinal survey. Prior data could also include a prior implementation of the same survey. Peytchev, Rosen, Riley, Murphy and Lindblad (2010) and Roberts, Vandenplas and Stahli (2014) used prior survey wave data to classify cases into response propensity strata which were then used for data collection tailoring. Calinescu, Bhulai and Schouten (2013) simulated a static adaptive design to minimize mode effects for a given cost using historical response data and population register data to assign cases to particular data collection protocols. In these applications, the implicit assumption is that the relationships between covariates and outcomes found in prior rounds of data collection would remain the same in the current round of data collection, meaning that the current round of the survey will behave very similarly to the data on which the estimates of the model coefficients are based. Errors or acceptable ranges around these expectations might be created through the use of historical deviations or *ad hoc* heuristics for acceptable tolerances, but do not take into account the information being gathered during the current fieldwork period.

Peytchev (2014) provides an example of using both historical data and current data to implement a response propensity-based intervention in the National Intimate Partner and Sexual Violence Survey (NISVS). Historical data were used to generate estimates of response propensities and contact attempts that would be saved by stopping work on cases below a particular response propensity, in order to generate a threshold for cutting off contact attempts in the current round of data collection. While this leverages historical data in the current survey implementation, the author states that this method also has drawbacks, including the fact that the thresholds set are not “responsive to current data collection outcomes”.

In the present context, we are not developing a model to predict response propensity, but are instead predicting the length of time between the first contact attempt and the first successful contact for a case. However, the assumptions required when using only current data, or only historic data, to generate predictions are the same and so suffer from the same disadvantages. This manuscript develops a Bayesian method of combining historical datasets and current information to make a more accurate prediction of the lag between first contact attempt and first successful contact, which could be useful information for minimizing nonresponse due to non-contact. While this is not a commonly used monitoring metric in survey management, clinical trials predict and monitor similar types of progress indicators – subject accrual rates and times. Zhang and Long (2012) provide a review of methods used for predicting accrual rates. In clinical trials, as in surveys, insufficient recruitment can lead to poor data quality, and an inability to make statistical statements. As a result, monitoring and reacting to shortfalls in recruitment can be an important part of clinical trial management. Kim, Han, and Youngblood (2018) propose a method of monitoring managers of clinical trials that provides ongoing feedback based on the existing recruitment data that can alert managers if recruitment needs are unlikely to be met, allowing for potential reactions or interventions, similar to our purpose here.

A logical approach to leveraging information from past rounds of data collection, while also using information about the current round of data collection is Bayesian modeling with informative priors. Wagner and Hubbard (2014) and Schouten et al. (2018) discuss the potential benefits of using Bayesian methods to improve predictions during data collection to support responsive and adaptive designs, respectively. Here, we find this method is particularly valuable for inference about contact lag, since occurrence of the first contact attempt is an important time-varying covariate within a data collection period that is not well-estimated by early data. We apply our methodology to predict contact lags in the National Health Interview Survey, using monthly survey data from July, 2014 through June, 2016.

The remainder of this manuscript is organized as follows. Section 2.2 describes the National Health Interview Survey and relevant auxiliary data sources. Section 2.3 describes the time to event models used to predict contact lag, and the construction of the prior distributions necessary to implement the Bayesian methods. Section 2.4 provides results, comparing the proposed Bayesian methods with other standard alternatives, including using only current data, and using only previous data. Section 2.5 concludes with a discussion of the results and directions for future work.

2.2 Description of Data

2.2.1 The National Health Interview Survey

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the United States and is one of the major data collection programs of the National Center for Health Statistics (NCHS). The main objective of the NHIS is to monitor the health of the United States population through the collection and analysis of data on a broad range of health topics.

The NHIS is a cross-sectional household interview survey that is carried out monthly. The sample design follows a multistage area probability design that permits the representative sampling of households and non-institutional group quarters (e.g., college dormitories). First, the entire United States is divided into approximately 1,700 primary sampling units (PSUs), which can consist of a county, a small group of adjacent counties, or a metropolitan statistical area. A selection of PSUs is made, with some metropolitan areas being selected with certainty (self-representing PSUs), while others are sampled probabilistically (non-self-representing PSUs). Within those selected PSUs, clusters of households and non-institutional group quarters are selected for interview. Additionally, an oversample is taken for geographies with higher expected populations of particular age and race/ethnicity groups. Interviewing for the NHIS is conducted continuously throughout each calendar year. Each month is its own self-contained data collection period, and each quarter, interim estimates are produced from the prior three months, which make up a representative subset of the overall annual NHIS sample.

The U.S. Census Bureau is the data collection agent for the NHIS. Survey data are collected continuously throughout the year by Census interviewers. The NHIS consists primarily of face-to-face interviews conducted in respondents' homes, but follow-ups to complete interviews may be conducted over the telephone. A telephone interview may also be conducted when the respondent requests one or when road conditions or travel distances would make it difficult to schedule a personal visit before the required completion date (NCHS 2018). We used two years' worth of monthly data collection periods, spanning the months from July, 2014 to June, 2016 for this work.

2.2.2 Auxiliary Data Sources

Four data sources were used for this evaluation, in addition to the NHIS sample itself. The first is the Census Bureau Planning Database (PDB) (Census 2016), a detailed dataset including sociodemographic information at the block group level that is produced using data from 5-year ACS estimates (Census 2008) and the Decennial Census. This dataset is created every year, and for this evaluation, we used the 2016 version of the PDB. Second, we obtained a dataset of some basic employment information about interviewers, including which regional office an interviewer belongs to and their experience level on the NHIS.

We also utilized two sources of paradata, the Neighborhood Observation Instrument (NOI) and the Contact History Instrument (CHI). In the NOI, interviewers are asked to record information about the housing unit and neighborhood from their own observations.

These neighborhood observations require no contact with the respondent, and ideally are reported prior to contact with the respondent. For the CHI, on the other hand, interviewers record the date, time and outcome of each contact attempt and information about interactions with sample persons. Additionally, the CHI includes information about the field management structure, including which interviewers are assigned to work each case and whether a case was reassigned during the field period.

These data sources were selected for this application partially because of their availability, but also because they include the types of data identified in Groves and Couper (1998) for predicting outcomes like household contactability and survey cooperation or response propensity. Here, we are predicting the time between first attempt and first contact, however, many of the same predictors could be useful explanatory covariates for our purposes, as well. The NHIS was selected for this application due to the collection of both the NOI and the CHI, as this leads to an expanded set of paradata that can be used to predict the lag between first attempt and first contact.

2.3 Model Selection

Our first consideration was for the category of model used for our predictions. In the NHIS, contact is made with the sample unit at the first contact attempt approximately 60% of the time. This results in a lag of zero days (“zero lag”) between the day of the first attempt and the day of first contact. The remainder of sample units have a lag of at least one day, and while contact is eventually made with most sample units, some sample units never have a successful contact, meaning the true observed lag length is censored.

Additionally, as Table 1 shows, for cases with positive lags, the variance of the lags is much larger than the mean lag, across all months covered by this study.

Table 1. Mean and Variance of Lag in Days between First Attempt and First Contact by Interview Month

Interview Month	Number of Cases	% Cases with Lag > 0	Mean(Lag) if Lag > 0	Var(Lag) if Lag > 0
07/2014	5200	0.4171	8.890	48.67
08/2014	5100	0.3990	8.216	41.90
09/2014	4900	0.4131	7.941	42.54
10/2014	5300	0.4236	8.009	43.89
11/2014	5000	0.3895	7.837	40.78
12/2014	5100	0.4058	9.192	48.30
01/2015	5600	0.3861	8.178	39.93
02/2015	5400	0.4181	8.464	44.49
03/2015	5400	0.3984	8.658	49.41
04/2015	5800	0.4165	8.396	41.84
05/2015	5400	0.4026	8.424	45.74
06/2015	5600	0.3842	8.072	43.88
07/2015	5400	0.3938	8.589	40.90
08/2015	5200	0.3929	8.385	46.09
09/2015	5100	0.4003	8.449	43.12
10/2015	5300	0.4015	8.847	49.48
11/2015	5000	0.3754	8.062	42.60
12/2015	4800	0.3915	9.684	59.51
01/2016	5200	0.3876	7.797	35.65
02/2016	6000	0.3952	8.657	42.47
03/2016	6600	0.3893	8.861	51.00
04/2016	6000	0.3947	8.665	43.58
05/2016	5700	0.3951	8.448	44.40
06/2016	5800	0.4009	8.724	47.89

Additionally, Table 1 shows that, for the 40% percent of cases where contact is not made on the first attempt, the average and variance of the lag in days is similar across months of data collection. Below, Table 2 shows that within a month, however, the average lag and the variance of the lag differs by the week in which the first contact attempt was made. As the time remaining in data collection decreases, the lag between the first attempt and first contact also decreases. This makes some intuitive sense. As a fixed data collection period progresses, there is less remaining time for cases to be contacted, and so interviewers may make contact attempts closer together, or at higher frequency in order to make contact with the sample member. Additionally, interviewers may have fewer remaining cases in their workloads, and so more attention can be paid to those remaining cases. Lags were averaged over all 24 months of data collection that were used in our application to generate Table 2.

Table 2. Mean and Variance of Lag in Days between 1st Attempt and 1st Contact by Week of 1st Attempt

Week of Data Collection	Mean(lag) if Lag > 0	Var(lag) if Lag > 0
1	8.891	48.81
2	7.763	33.56
3	5.770	16.94
4	3.716	9.94

Table 2 suggests that the time point during data collection when the first attempt is made (either day or week) is an important predictor for the lag between that first attempt and first successful contact. In order to accommodate this time-varying parameter and retain in the analysis the cases whose lag was censored (because a contact was never made), we made the decision to model the time until first contact using a survival function. One benefit of using a survival model is that the model can account for censoring in the outcome variable, meaning that cases that have a first attempt but no contact can still be included in the model. This would not be possible with a negative binomial model, or other inflated count models. However, two characteristics of our data meant that a survival model alone was not sufficient for modeling the time to first contact. First, survival models are used to estimate the time, $t > 0$, until an event of interest happens. Therefore, a survival model would not provide predictions for the 60% of cases where contact was made on the first attempt, who effectively have a time to first contact $t = 0$ (Klein and Moeschberger 2003).

The second complication is that many of the variables available for prediction are only available *after* the first attempt is made. Again, in a production setting where the goal is to prospectively predict the time to first contact, these variables would not be available prior to data collection. Therefore, they could not be used to aid prediction of a contact on the first attempt (a zero lag), but they *could* be used to predict the length of the lag *given* the fact that contact was not made on the first attempt. Table 3 below shows the five data sources, sample data items from that source, when that source would be available during a typical data collection period, and whether the data items from the source are fixed or time-varying.

Table 3. Data Sources and their Availability for Use in Prediction of Lag

Data Source	**Sample Items**	**Availability for Inclusion in Models**	**Fixed or Time-Varying**
Sample File	• Block-Level Geography • Census Management Region	Prior to Data Collection	Fixed
Planning Database	• % of Housing Units with No Health Insurance Plans • % of Housing Units Vacant	Prior to Data Collection	Fixed
Interviewer Information	• Experience on the NHIS	Prior to Data Collection	Fixed
Interviewer Observations	• Evidence of Children • Evidence of Smoking	After the First Contact Attempt	Fixed
Contact History Instrument	• Count of Contact Attempts • Activities Completed on Attempts • Case Reassignment Indicator	After the First Contact Attempt	Time-Varying

Table 3 shows that while all of these sources would be available for use to predict the lag, provided it was greater than zero, only the first three data sources would be available prior to data collection in order to predict whether contact would be made on the first attempt or whether there would be a positive lag. As a result, we employed a hurdle model in order to combine two different processes – one that would predict the likelihood of making contact on the first attempt, and a second that would predict the lag between the first attempt and first contact, given that contact was not made on the first attempt. An alternative to a hurdle model would be to simply wait until a first attempt was made for a case and, if the first attempt did not result in a contact, use the paradata collected from that first attempt to estimate *only* a time-to-first-contact model. However, that would mean cases would need to be attempted prior to modeling, and so there would be no way to predict the length of the lag prior to the first attempt. This could ignore information that might be useful for interviewers; if they knew which cases were likely to have longer lags, they may plan their workdays differently than they would without that information.

2.3.1 Hurdle Models

A hurdle model (Mullahy 1986) is useful when one believes there are two separate processes at work – one that generates zeros (in our case, “zero lags”, where the interviewer makes contact on the first attempt) and the other that generates some positive lag (Ma et al. 2015). Additionally, hurdle models are similar to negative binomial models (Rose et al. 2006) in the fact that they are appropriate for count data with overdispersion, where the variance is larger than the mean. However, where negative binomial models assume that a single underlying process governs both the generation of excess zeroes and the nonzero values, hurdle models allow the data generation processes to be different. There is a “hurdle” that must be crossed before the outcome is a nonzero value. Then, for those that cross the hurdle, a different distribution predicts the outcome variable, the lag between first attempt and contact. This was a useful model structure for our application. As shown in Table 1, approximately 60% of our sample units have a lag of zero days, and the variance of the non-zero lags is much larger than the mean. Additionally, this model would allow us to incorporate different predictors in each of the portions of the model, which would not have been possible without the hurdle model.

We chose to use a logistic regression model to predict the hurdle portion of the model, which determine whether contact would be made on the first attempt. Then, for each of the 24 months of data collection, we used all records with a positive lag to determine which parametric distribution for the survival portion of the model was most appropriate. We evaluated the empirical distribution of positive lags (without controlling for any factors) against Poisson, Weibull, Gamma, and Lognormal distributions both visually and with goodness-of-fit statistics, using the `fitdistrplus` package in R. Figure 1 below shows panels for each of the four distributions evaluated for one data collection month to illustrate how different time-to-event distributions compare with the theoretical distribution for a given month.

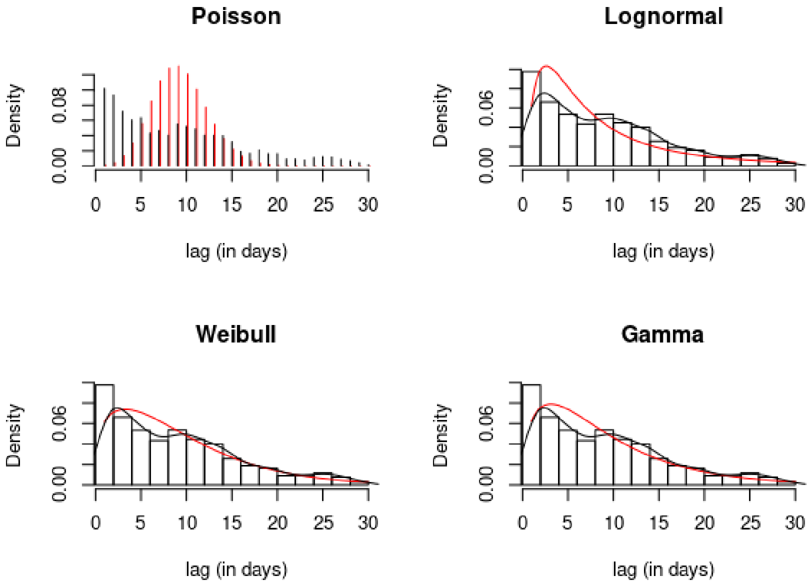


Figure 1. Four Parametric Survival Distributions Compared to Empirical Distribution of Lag (in Days)

The red curves (or bars in the case of the Poisson distribution) show the theoretical distribution of the four proposed distributions for modeling the lag between first attempt and first contact. The Poisson distribution displays histogram bars at discrete values, rather than a smooth curve, because it is not a continuous distribution. The black bars, which are the same in all four plots, display the empirical distribution of the actual lag between first attempt and first contact, based on the actual data. The black line shows the smoothed distribution. The Poisson distribution provided the worst fit, followed by the lognormal distribution. The Weibull and Gamma distributions were similar in their fits, and so to make a determination, we also looked at Chi-Square tests, AIC and BIC. For this particular month, the AIC and BIC were both smaller for the Weibull, and the Chi-Square test was larger. Looking at all 24 months, the Weibull had smaller AIC and BIC scores more often than the Gamma distribution, so we elected to use a Weibull distribution for model fitting and prediction. There was little difference between it and the Gamma distribution, however, so either could be a reasonable choice.

2.3.2 Survival Models

Survival models, or time-to-event models, attempt to estimate the time until some event happens. Both non-parametric and parametric survival modeling are common, though here we focus on parametric survival models, in order to take advantage of the shape of the distribution to help us predict the lag between the first attempt and first contact.

Several related functions are important for survival modeling (Klein and Moeschberger 2003, Ch.2). First, the *survival* function provides probability that an event occurs after time t, and is written:

$$S(t) = \Pr(T > t) = \int_t^\infty f(t)dt \quad ,$$

where the *probability density* function, $f(t)$, is integrated from the time point of interest, t, to infinity to calculate the probability of an event happening *after* time t. The cumulative distribution function, which represents the probability that an event occurs by time t, is written:

$$F(t) = 1 - S(t) = \Pr(T \leq t) = \int_0^t f(t)dt \quad .$$

These two expressions are helpful for explaining how likely an event is to occur before or after a certain point in time; however, it is often of interest whether an event will happen at a particular instant in time. This quantity is explained by the *hazard* function, which is the instantaneous risk of an event happening in the next moment. The hazard function is defined as:

$$\lambda(t) = \lim_{dt \to 0} \left\{ \frac{\Pr(t \leq T < (t + dt) | T \geq t)}{dt} \right\} = \frac{f(t)}{S(t)} \ ,$$

where the hazard is the probability of an event happening in the next small window of time, divided by the size of that window of time, dt, as that window shrinks to zero. The *hazard* function can also be written as the quotient of the probability distribution function and the survival function.

These relationships are important for estimating the expected time-to-event, but particularly so when the event of interest is *right-censored* for a particular case, that is, when the event has not yet occurred in the observation time window (Klein and Moeschberger 2003, Ch.3). Using the *probability density* function, the *survival* function and the *hazard* function, we can explain how each individual case contributes to the overall likelihood of the distribution.

If the event has occurred at time t, the likelihood can be written: $L = f(t) = S(t)\lambda(t)$. This expression represents the fact that the case survived up until time t, represented by $S(t)$, and then had the event occur in the instant, $\lambda(t)$, of time t. For *right-censored* cases where the event has not occurred by time t, only the *survival* function contributes to the likelihood, which is written: $L = S(t)$.

As a result, the full likelihood, where δ_i is an indicator for whether the event has occurred for the i^{th} case, can be written:

$$L = \prod_{i=1}^{n} f(t)^{\delta_i} S(t)^{(1-\delta_i)} = \prod_{i=1}^{n} S(t)\lambda(t)^{\delta_i} \ .$$

We will use these relationships in the next section when defining the likelihood for the Weibull hurdle model.

2.3.3 The Weibull Hurdle Model

In this setting, the prediction of interest, y_i, is the length of time in days that would elapse between the first contact attempt and the first contact with a household sample member for the i^{th} case. The hurdle model estimates two processes. For each observation, with probability $(1-\pi_i)$, the outcome variable, lag, is zero, and with probability (π_i), a non-zero outcome is estimated. Given these two factors, the probability distribution function for an observation can be written as:

$$f(y_i) = I(y_i = 0)\ (1-\pi_i)(0) + I(y_i > 0)\ (\pi_i) g(y_i; \alpha, \beta\ ) = \ (\pi_i) g(y_i; \alpha, \beta\ )$$

where $g(y_i; \alpha,\ \beta)$ is the probability distribution function (p.d.f.) of the Weibull distribution, $\pi_i = \ln\left(\frac{\rho_i}{1+\rho_i}\right)$ is the link function with a logistic distribution; $\rho_i = \exp(\boldsymbol{\gamma}'\boldsymbol{z_i})$ based on the individual case's covariate(s), $\boldsymbol{z_i}$; $\hat{y}_i = (\pi_i) g(y_i)$ is the estimate of the outcome using the estimated regression parameters; α is the scale parameter of the Weibull distribution; and β is the shape parameter in the Weibull distribution. (Gelman et al. 2013).

As in section 2.3.2, the p.d.f., and therefore the likelihood, of the Weibull distribution incorporates both the survival function, $S(t_i)$, for all cases, and the hazard function, $\lambda(t_i)$, for those cases where the event of interest, contact with the respondent, has occurred. The likelihood function for the Weibull model can be written:

$$L = \prod_{i=1}^{n} S(t_i)\big(\lambda(t_i)\big)^{\delta_i} = \underbrace{e^{-\left(\frac{y_i}{\alpha}\right)^{\beta}}}_{S(t_i)} \underbrace{\left(\frac{\beta}{\alpha}\left(\frac{y_i}{\alpha}\right)^{\beta-1}\right)^{\delta_i}}_{\lambda(t_i)} \quad .$$

This formulation provides the flexibility to include cases in the model for which contact has not yet occurred (right censored cases). In the event that there is no censoring, which would happen in a retrospective analysis where the event occurred for all cases, both the survival function and hazard function would contribute for all cases.

Incorporating the likelihood for the Weibull distribution into the hurdle model formulation, we obtain the following full p.d.f. for the Weibull hurdle model:

$$f(y_i) = (\pi_i)\left(e^{-\left(\frac{y_i}{\alpha}\right)^{\beta}}\left(\frac{\beta}{\alpha}\left(\frac{y_i}{\alpha}\right)^{\beta-1}\right)^{\delta_i}\right) .$$

The full likelihood can then be written as follows:

$$L(\gamma,\alpha,\beta) = \prod_{i=1}^{n}\left(\frac{1}{1+e^{z_i'\gamma}}\right)^{(1-d_i)} \prod_{y_i>0}\left(\frac{e^{z_i'\gamma}}{1+e^{z_i'\gamma}}\right)^{d_i} \prod_{y_i>0}\left(e^{-\left(\frac{y_i}{\alpha}\right)^{\beta}}\left(\frac{\beta}{\alpha}\left(\frac{y_i}{\alpha}\right)^{\beta-1}\right)^{\delta_i}\right) \ ,$$

where d_i is an indicator specifying there *will be* a lag between the first attempt and first contact, and δ_i is an indicator noting whether contact has occurred by the time of model estimation, or whether the case should be considered censored. The first term in the likelihood represents the probability that there will *not* be a lag between the first event and the first contact; that is, contact will be made on the first attempt. The second term represents the probability that there will be a positive lag, and the last term in the likelihood represents the survival portion of the model, accounting for censoring.

While this likelihood is complicated, and includes parameters for both the binary process and the time-to-event process, maximum likelihood estimation simplifies the expression somewhat. More importantly is that MLE shows that the different contributing terms can be maximized separately allowing the model to be estimated in parts, computationally. The log-likelihood can initially be written :

$$lL(\gamma,\alpha,\beta) = \sum_{i=1}^{n}(1-d_i)(-log\Big(1+e^{z_i'\gamma}\Big) + \sum_{y_i>0}(d_i)z_i'\gamma(-log\Big(1+e^{z_i'\gamma}\Big)$$

$$+\sum_{y_i>0}\left(\frac{y_i}{\alpha}\right)^{\beta} + \delta_i\big(\beta-\alpha+(\beta-1)(y_i-\alpha)\big)\ .$$

As a result of this construction, the portion of the model that predicts the binary outcome (lag/no lag), and the portion of the model that predicts the lag length, can be estimated separately (Smithson and Merkle 2013, Chapter 5). As a result, throughout this application, we estimate the two parts of the model separately, as shown in Appendix G. However, we refer to this as a single model, as the goal of this prediction is to determine the expected lag between first attempt and first contact, and both portions of this model are needed to arrive at that prediction.

2.3.4 Variable Reduction and Model Building

We combined the datasets described in Table 3 to create two sets of analytic files for each monthly NHIS sample from July 2014, through June, 2016. The first set were *summary* files and consisted of one record per sampled household, with all fixed covariates, and the final status of all time-varying covariates. The second set were *attempt-level* files and consisted of one record per *attempt* per sampled household, with all fixed covariates, and the most-recent status of all time-varying covariates.

Using the summary files, we first attempted to reduce the set of variables through both factor analysis and latent class analysis, in order to reduce potential collinearity as well as obtain a parsimonious model. These efforts were not successful in yielding data reduction. We then executed backwards stepwise regression (for both the logistic and Weibull portions of the model) on each monthly summary file, and only retained variables that were significant at the $p < 0.05$ level in over 10% of those models (e.g., at least 3 months out of 24). The logistic regression model was estimated using the `glm()` function in `base` R, while the Weibull model was estimated with the `survival` package. The final list of explanatory variables used to predict the lag between first attempt and first contact is in Table 4 below.

Table 4. Covariates Included in Model to Predict Lag

Variable	**Description**	**Logistic**	**Weibull**
Lag (in Days) Between Initial Attempt and Initial Contact	Dependent Variable		X
Non-Zero Lag Indicator	Dependent Variable (Derived from Lag (in Days) Dependent Variable)	X	
Day of First Attempt	Integer (1 – 31) noting the day of month when the first attempt was made	X	X
Regional Office (RO)	Highest Level of Field Organization at Census – 6 Levels for the US		X
Interviewer Experience Measure	Indicator identifying interviewers who have worked on NHIS less than 1 year.	X	
Reassignment Indicator	Indicator to identify cases that have experienced one or more reassignments during data collection		X
% Mobile Homes	Percentage (0-100) of housing units with these characteristics in a Census Block Group from PDB	X	
% Vacant Units		X	
% College Graduates	Percentage (0-100) of population with these characteristics in a Census Block Group from PDB	X	
% Not HS Graduates		X	
% Without Health Insurance		X	
% Urbanized Population		X	
% Vacant Units * Day of First Attempt	Interaction Variable from PDB	X	
Existence of Bars on Windows	Indicator based on interviewer observations		X
Evidence of a Wheelchair at HU			X
Evidence of Children at HU			X
Below Average Condition of HU			X
Barriers to Accessing HU			X

The variables that were retained as significant in each of the models, shown in Table 4, make some intuitive sense. The NHIS is a cross-sectional survey, and so the interviewer may not know very much about their specific cases before visiting. However, general geographic information or interviewer characteristics, like their experience with the survey in general, may be correlated with making a successful contact on the first attempt, which results, mathematically, in a zero lag. Perhaps the interviewer has cases in an area (described by PDB variables) that is overall more responsive to survey requests, and a respondent is more likely to answer the door, or where individuals are home more often so that contact is more likely to be made. It is reasonable then that variables from the PDB would appear in the logistic regression portion of this model, as this is information known before attempts on the household are made. On the other hand, the non-zero (positive) lag process, which represents the lag between attempt and contact, may have something to do with best practices within their supervisory structure, once contact is not made on the first attempt, or using interviewer reactions to observations they make about the sample unit itself and its members. Therefore, the NOI variables appeared in the Weibull portion of the model.

It is important to note that this variable selection process could only be undertaken because we were conducting a retrospective analysis, and therefore had the true lag length available to us. In a situation where historical data are not available, one would not be able to identify *a priori* the most significant variables across a large number of months. Instead, a broad array of variables might be included in the prediction process until the estimation stabilized, or until there was enough historical data to conduct variable selection. We discuss the benefits and limitations of this process in Section 5.

2.4 Predicting Time to First Contact: Methods

For this application, we used 24 months of data from the NHIS, covering the time period from July 2014 through June 2016. For each month, the prediction of interest was the length of time in days that would elapse between the first contact attempt and the first contact with a household sample member. We generated these predictions using either current data only, historical data only, or the combination of both through the use of priors generated from historical data which are then updated with current accumulating data.

1) The first method uses accumulating data throughout the current round only to estimate a Weibull hurdle model, and then uses those parameters to predict the expected lag for each open case in the current month. We will refer to this as the *current method.*

2) The second method uses historical data to estimate the mean expected lag. The expected lag for all cases in the current month, then, is just the overall average lag of the three prior months, ignoring any additional information. We will refer to this method as the *mean method.*

3) The third method estimates the parameters for a Weibull hurdle model from historical data, and then uses the point estimates of those parameters to predict the expected lag for each open case in in the current month. We will refer to this as the *historical method.*

4) The fourth method combines the *current* method and the *historical* method statistically. Estimates of the parameters for a Weibull hurdle model from

historical data and their standard errors are incorporated as priors in a Bayesian Weibull hurdle model that also leverages accumulating data for the current month. Then, posterior predicted values of the lag based on the priors and current data are estimated for each open case in the current month. We will refer to this as the *Bayesian method.*

For methods that leverage historical data, three consecutive months were used as the historical data, and the next month was considered the current month, the period of predictive interest. For example, if October 2014 was the predictive period of interest, July, August, and September of 2014 would be used to generate the mean for Method 2, coefficients and standard errors for model covariates for Method 3, and point estimates and standard errors for incorporation as priors in Method 4.

The *mean* method is clearly the simplest to implement. However, using a single number from the end of data collection is only useful if progress throughout a self-contained data collection period is constant during that period. If not, the *mean* method may work well some of the time (e.g., early in data collection), but not at other times, like the end of data collection. Table 2 suggests that the mean method may not be particularly useful for this reason. Additionally, because it uses historical data only, this method is useful is only if data collection periods behave similarly, which may not be true for a variety of reasons, including seasonality effects, changes to the data collection instrument, or external factors that affect data collection progress, like severe weather.

The *historical* and *current* methods include time-varying covariates, such as the day of first attempt and case reassignment status, so that more information about the data collection process can be incorporated into the predictions. However, the *historical* method only uses historical data, again requiring the current data collection period to be nearly identical to historical data collection period to be useful. The *current* method only uses current round data, suffers from the right censoring of open cases, and ignores potentially useful historical information that could help create meaningful expectations.

The *Bayesian method* leverages information from past rounds of data collection, while also using information about the current round of data collection is to take advantage of a Bayesian modeling approach with informative priors. We accomplish this by first fitting a Weibull hurdle model using three months of historical data to obtain parameter estimates $(\hat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$, and the associated variances, $V(\hat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$. These parameters capture the time-varying nature of some of the covariates within a single data collection period in a survival model framework.

Assuming approximate normality by the properties of maximum likelihood estimates, we form priors, $P(\boldsymbol{\gamma}, \boldsymbol{\beta}) \sim N\left(\binom{\hat{\boldsymbol{\gamma}}}{\widehat{\boldsymbol{\beta}}}, c\hat{V}(\hat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})\right)$ where c is a constant that controls the degree to which the prior information is used in the daily estimation procedure. We vary the value of c to demonstrate how c can be used to (relatively) weight historical and current data in the posterior predictions. The standard errors were based on 3 months of data, so we inflated them by a factor of $\sqrt{3}$ to represent one month of data, or a factor of 3 to represent 1/3 of a month of data. This resulted in five values for c, and therefore the standard errors around the coefficients, representing 1/3 month, 1/2 month, 1 month, 2 months, and three months. Once these priors were obtained, a Weibull hurdle model is estimated each day using the data available up to that day in the current month, combined with the previously obtained prior. The `brms` package was used to conduct resampling and estimation for Bayesian inference of parameters and estimation of posterior predictions. The code for generating predictions via all methods for a given day and month are provided in Appendix Table 5 summarizes the methods and their main characteristics.

Table 5. Summary of Prediction Methods by Data Types Used

Data Used	**Method 1: Current Method**	**Method 2: Mean Method**	**Method 3: Historical Method**	**Method 4: Bayesian Method**
Historical Data		X	X	X (as priors)
Accumulated Current Data	X			X

We repeat these predictions for twenty-one sets of data collection periods, which are made up of a historical period (3 months) and current period (one month). The four methods will be compared primarily using measures of mean prediction bias (MPB) and root mean square prediction error (RMSE). The MPB is defined as:

$$MPB^m = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^m{}_i - y_i\right)$$

and the RMSE for the m^{th} method is defined as:

$$RMSE^m = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^m{}_i - y_i\right)^2}$$

where n is the number of cases in the accumulated current month dataset, $\hat{y}_i$ is the predicted value of the lag, y, for the i^{th} case, and y_i is the true value of the lag, y. The percent difference between the RMSE for the m^{th} method and the method selected as the baseline is calculated as:

$$PCHG^{m_1 m_b} = 100 * \left(\frac{RMSE^m - RMSE^b}{RMSE^b}\right)$$

Percent change in the MPB can be calculated similarly. The expectation is that the *Bayesian* method will produce predictions of the expected lag closer to that of the actual lag, resulting in a smaller overall MPB and RMSE, as the modeling procedure is effectively borrowing strength across the historical data and current data to make a prediction. For this evaluation, we do not use design-adjusted variance estimates. We are concerned primarily with prediction of outcomes within a single survey sample, and therefore focused on internal validity of the predictions, rather than attempting to create predictions or estimates for the full target population.

2.5 Results

2.5.1 Comparison of Four Predictive Methods

In order to compare the four discussed methods for predicting lag (*current (C), mean (M), historical (H),* and *Bayesian (B)*), we first compare the MPB and RMSE of the prediction of the lag for each of the four methods, and then plot improvements of the Bayesian method over its closest competing method. Predictions of the expected lag and the resulting MPB and RMSE of those predictions depend not only on the prediction method used, but also on when the prediction is made during data collection, which we refer to as the cut point. This is because each day, new cases are attempted for the first time, and contact is being made in other cases. Therefore, we are attempting to predict the expected lag for a pool of cases that is changing on a daily basis. In order to evaluate the quality of predictions, we evaluate RMSE at several cut points in data collection, after days 2, 4, 6, 8, 10, 15, 20 and 25. As data accumulate over time, the effectiveness of different methods may change. All figures are generated using `ggplot2` in the R programming language.

Figure 2 and Figure 3 below display estimates of MPB and RMSE for predictions of the lag. For each cut point during each of the 21 time periods, we used all open, uncontacted cases to generate estimates of bias and RMSE. The boxplots were then generated using the MPB^m and RMSE^m for all *m* time periods. Here, we initially discuss results for the *Bayesian* method where the variance is scaled to represent 1/3 of a month of data collection. Later, we discuss the impact of increasing the contribution of the prior.

Figure 2 shows that all methods underestimate the actual lag between first attempt and first contact throughout the data collection process. Additionally, as data collection goes on, the underestimation MPB increases. It is also evident that, until day 15, all methods that use historical data in some form (*M, H, and B)* outperform the c*urrent* method with respect to MPB. While the three methods that use historical data all perform similarly late in data collection, the *historical* and *Bayesian* methods perform better than the *mean* method until day 10. The *Bayesian* method appears to provide small improvements over the *historical* method over the 21 time periods, as evidenced by central tendencies of estimates of the mean being closer to zero.

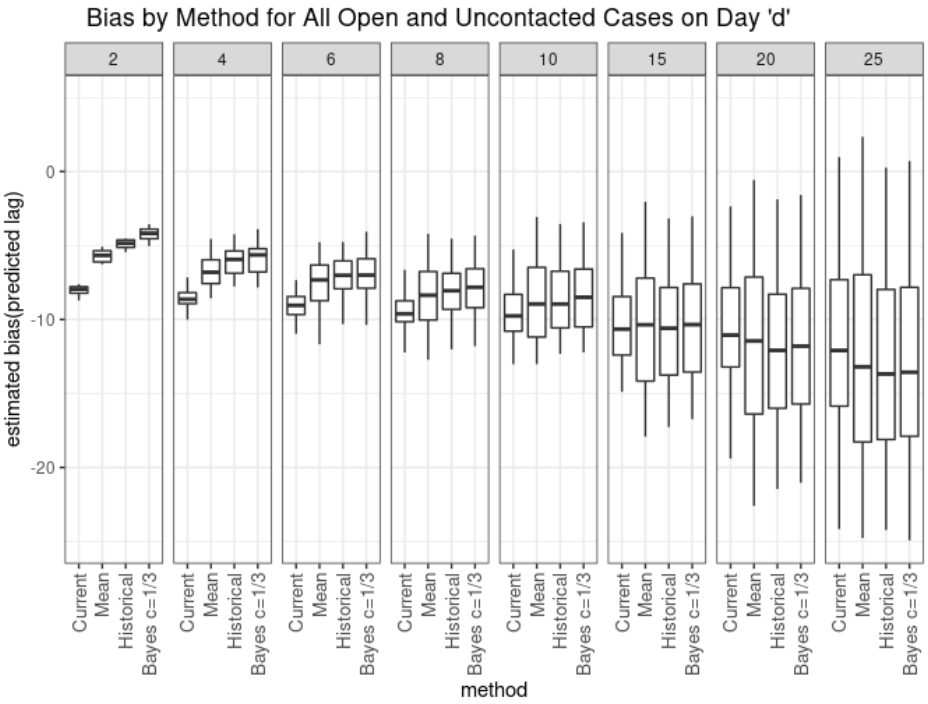


Figure 2. Bias by Prediction Method for All Open and Uncontacted Cases on Cutpoint Day 'd'

We see the same pattern in the RMSE of the lag predictions in Figure 3. The *current* method performs worse than the other methods until mid-way through the data collection period, and the other three methods are competitive with each other. Again, the *Bayesian* method appears to provide small improvements over the *historical* method, as the central tendencies and intraquartile ranges are smaller in the *Bayesian* method than in the *historical* method.

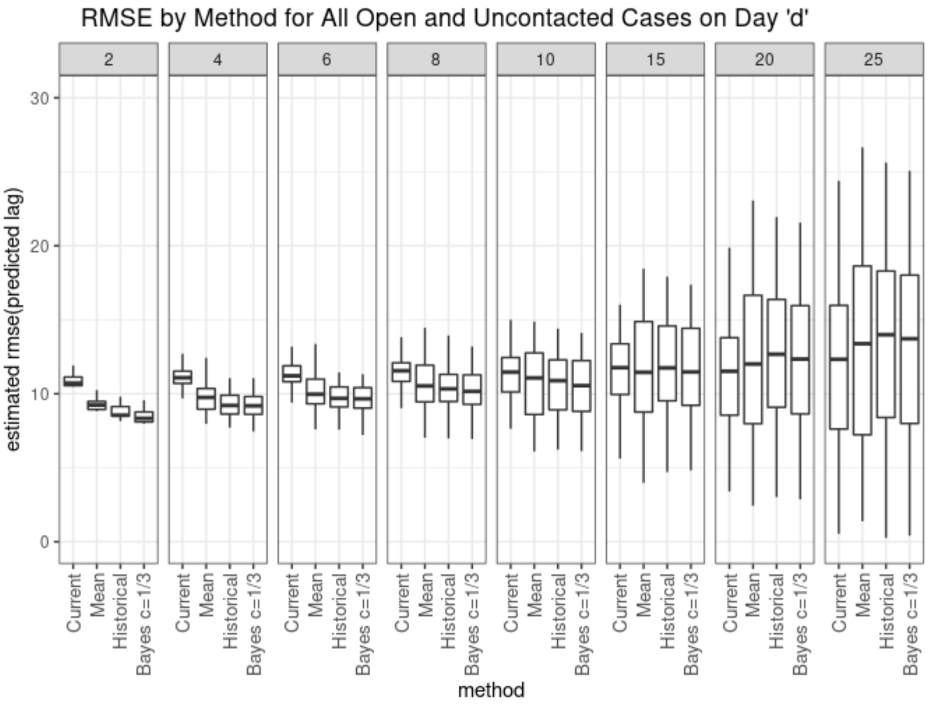


Figure 3. RMSE by Prediction Method for All Open and Uncontacted Cases on Cutpoint Day 'd'

In addition to the overall predictive abilities of the four different methods, we are particularly interested in the temporal effectiveness of the predictions. In other words, we are interested in how well we can predict the expected lag near the point in time of the first attempt, when we are close to the entry point of a given case into the dataset. Just after the initial contact attempt, we have the least data and the most time to implement an intervention if needed. In order to explore this, we generated Figure 2 and Figure 3 again, but limited the cases included to those worked within two days of the cutpoint. So, when $d = 4$, only cases that were first attempted on day 2 or 3 are included in the estimates of MPB or RMSE. On day 25, only those cases first attempted on days 23 and 24 are included.

Figure 4 below displays the MPB in prediction of lag for recent cases. While following the same general pattern, there are some differences from Figure 2. Most notable is that as time progresses through the data collection period, the MPB in predicting lag for recent cases is much smaller (the bias is closer to zero) for all cases, whereas in Figure 2, the under-estimation MPB increases over time. Again, the *current* data method performs the worst until late in data collection, but here, it never really outperforms the *historical* or *Bayesian* methods. Late in data collection, the *mean* method begins outperforming other methods, but this is not consistent throughout data collection. Again, we see that the central tendencies of the bias in the *Bayesian* method are closer to zero than in the *historical* method, demonstrating a small improvement.

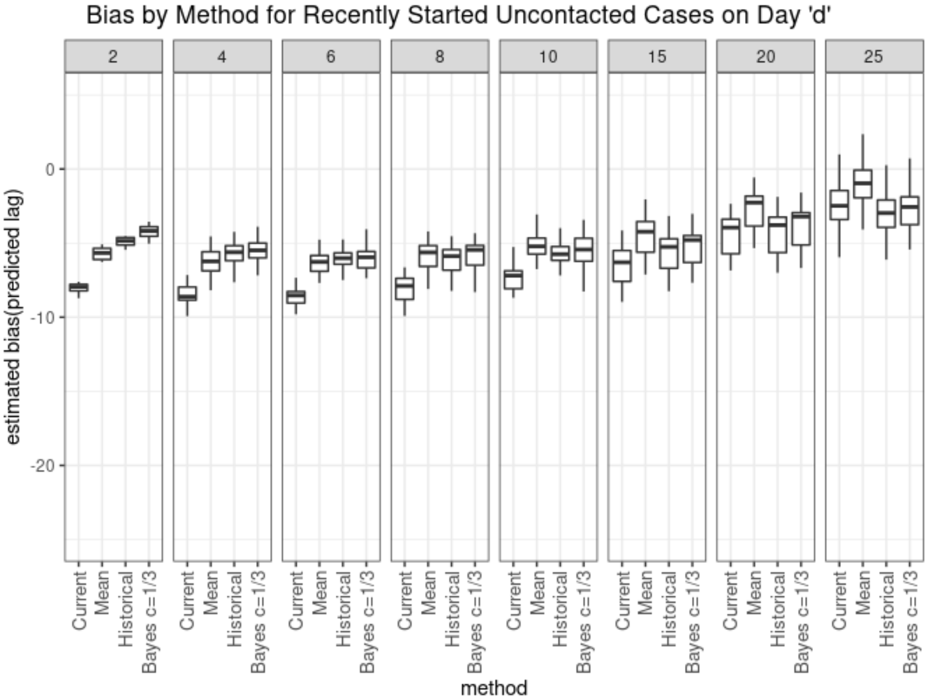


Figure 4. Bias by Prediction Method for Recent Open and Uncontacted Cases on Cutpoint Day 'd'

Figure 5 examines the RMSE of predictions of lag for recent cases, and continues to demonstrate that the use of historical information is generally helpful for improving predictions. Additionally, throughout data collection, the *Bayesian* method provides small improvements in the central tendencies of RMSE over the *historical* method.

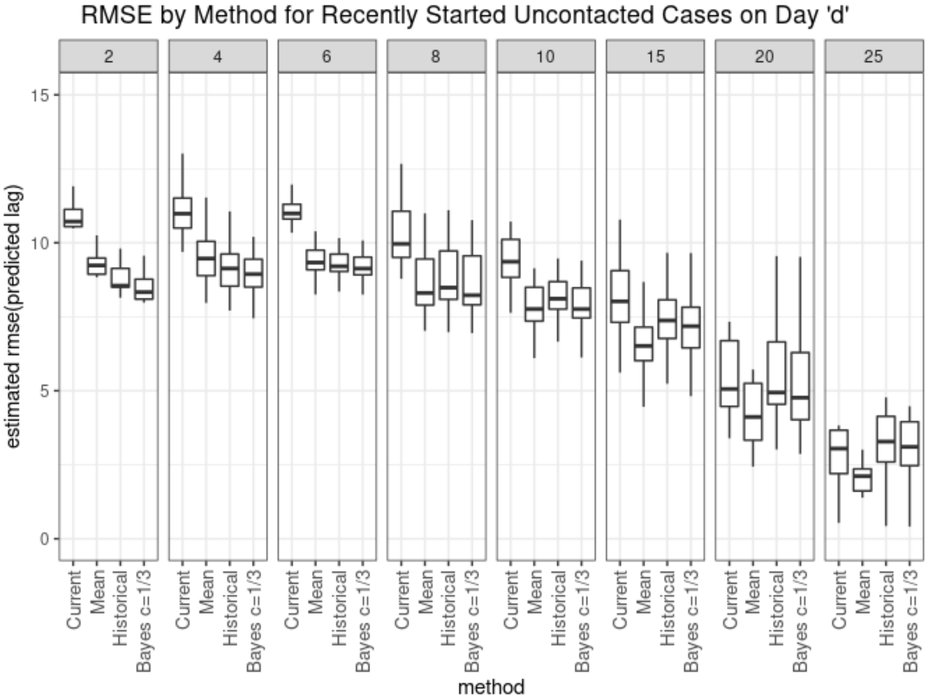


Figure 5. RMSE by Prediction Method for Recent Open and Uncontacted Cases on Cutpoint Day 'd'

Figures 2 through 5 display either the MPB or RMSE of predictions, focusing on when during data collection a prediction was made. Figure 6 plots the percent change in RMSE and MPB of the *Bayesian* method where the prior is equivalent to 1/3 of a month from the *historical* method. Each point on the scatter plot is for a time point within a data collection period, and recent case and older cases are plotted separately. For example, one data point in the plot of recent cases would be for the measures of MPB and RMSE on Day 4, for cases first attempted on Day 2, in a given data collection period.

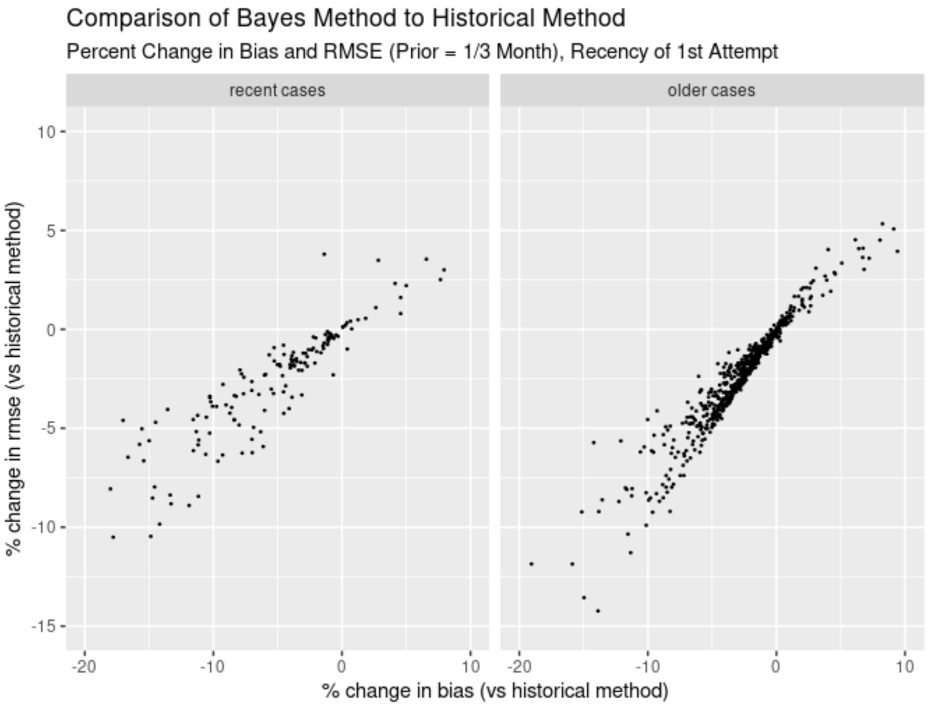


Figure 6. % Change in RMSE and Bias of Bayesian vs. Historical Method by Recency of 1st Attempt

Most of the data points fall in the southwest quadrant, representing a reduction of both RMSE and MPB in the *Bayesian* method when compared to the *historical* method. This is true for both recent and older cases. Additionally, when there are improvements in MPB and RMSE, those improvements have a larger range (reductions in the MPB and RMSE reaching 20% and 10% respectively) than the situations where the MPB or RMSE increase when the *Bayesian* method is used (increases generally limited to 10% increases in MPB and 5% increases in RMSE). Taken together, this suggests the *Bayesian* approach can improve predictions, even if the improvements are small.

2.5.2 Effect of Varying the Strength of the Prior

Thus far, we have only examined the benefits of the *Bayesian* method when a prior equivalent to one-third of one month of data is used for prediction. However, we can vary the prior in order to weight the posterior prediction more or less towards the current data.

In Figure 7 below, we replicate Figure 6 and compare the percent change in RMSE and bias of the *Bayesian* method when compared to the *historical* method varying the strength of the prior from 1/3 of a month to three months. Again, we split the cases into recent cases (top row) and older cases (bottom row).

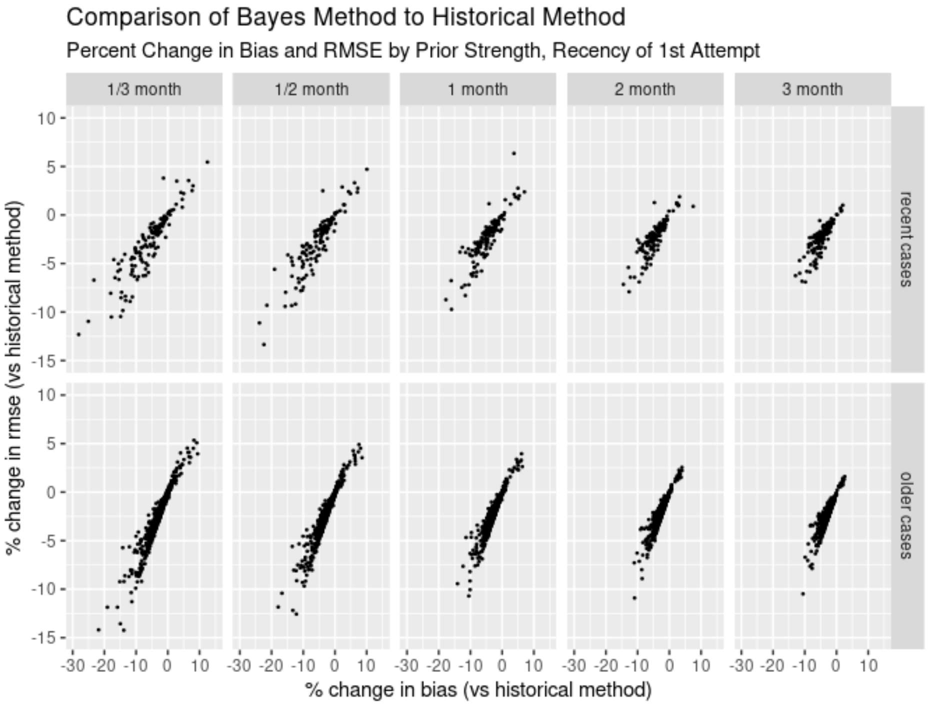


Figure 7. % Change in RMSE and Bias of Bayesian vs. Historical Method by Recency of 1st Attempt and Strength of Prior

Moving from left to right, as the prior increases in strength, the posterior is weighted more strongly to the prior, generated from historical data. This is visible in the scatterplots, as the percent differences between the *Bayesian* method and the *historical* method shrink toward zero. The weaker priors weight the likelihood, or the current accumulating data, more strongly in the posterior, resulting in larger differences from the *historical* method. Here, the weaker priors lead to larger reductions (in the lower left quadrant) in the bias and RMSE than increases (in the upper-right quadrant) than stronger priors. Still, for most prediction points shown above, the Bayesian approach provides benefits over only using the historical data.

2.5.3 Application

We also considered a simple application of these models to demonstrate a potential implementation of this method during data collection. By obtaining more accurate predictions of the expected lag soon after the first attempt, we maximize the opportunity for identifying cases likely to have a longer than desirable lag time and for intervening in data collection to improve outcomes. If it were possible to identify cases likely to have excessive lags, survey managers might consider interventions at the interviewer level to reduce the lag until the first contact, ranging from simply informing the interviewer that the case may be more difficult to contact than usual to reassignment of the case to a different interviewer.

We defined an “excessive lag” to be any lag predicted to be over four days. While this is an arbitrary number, at the Census Bureau, field supervisors monitor casework to ensure that there is not a gap of more than three days between contact attempts for monthly surveys. Time between contact attempts is not the same concept as time between an attempt and a successful contact; however, there is no current definition for what would constitute an excessive lag. If interviewers are expected to make contact attempts at least once every three days, setting an excessive lag at four days can be considered to mean a case where contact is not expected to be made in the next one or two contact attempts.

We then used the predicted lags for cases that have not been contacted at a particular point during data collection to classify cases into those with acceptable or excessive predicted lags. We also classify the true lags as normal or excessive. This allows us to calculate sensitivity and specificity (Altman and Bland 1994) for the different prediction methods. In order to focus on the early to middle portion of data collection when interventions are more likely to have an effect, we identified five points during data collection – days 4, 6, 8, 10, and 15 – to carry out this classification exercise.

Sensitivity is the proportion of true excessive lags, as measured by the actual lag, that are correctly classified using the different prediction methods. Specificity is proportion of true acceptable lags, as measured by the actual lag, that are correctly classified using the different prediction methods. These measures of prediction quality have operational implications, as well. Classifying all cases as having excessive lags means that the resources available to spend on interventions are being spread across all cases, making those interventions impossible or at least less effective. Similarly, classifying all cases as having acceptable lags means no intervention is carried out. As a result, there needs to be a balance between sensitivity and specificity. Figure 8 and Figure 9 below provide ranges of sensitivity (Figure 8) and specificity (Figure 9) across the 21 time periods for six different cut points in the first half of data collection, and comparing the two illustrates that balance. In these figures, the current method is labeled “C”, the mean method is labeled “M”, the historical method is labeled “H”, and the Bayesian methods are referred to by the strength of the prior: B_0033 uses c = 1/3, or one-third of a month; B_0050 uses c = 1/2 , or one-half of a month; and B_0100, B_0200, and B_0300 use c = 1, 2, and 3, respectively, or one, two and three full months of data collection. As c increases, the contribution of the prior (relative to the likelihood) increases.

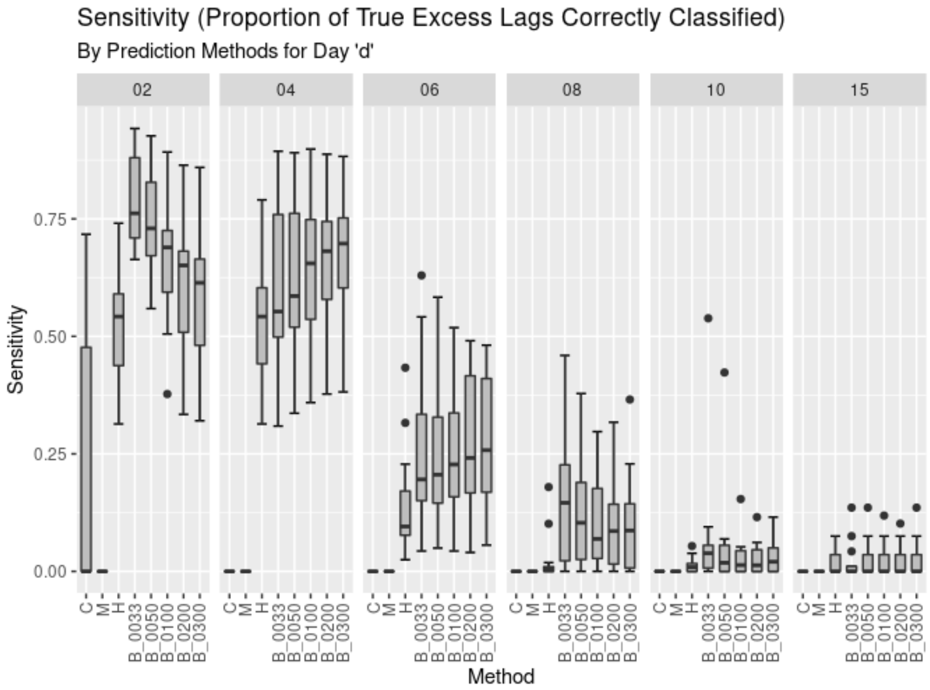


Figure 8. Sensitivity of Classification of Lags as Excessive Using a Cutoff of Four Days

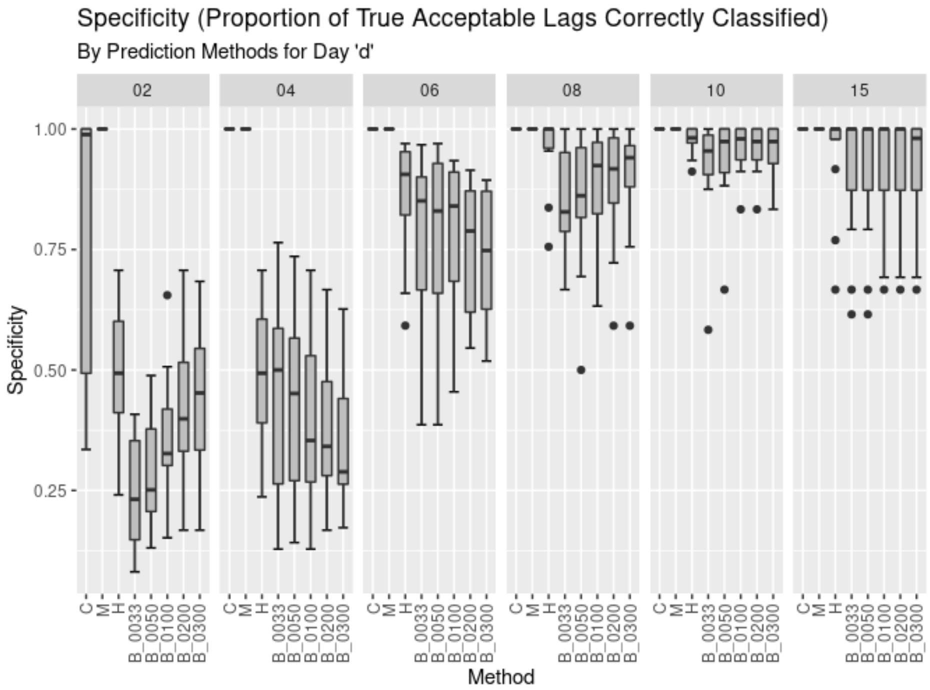


Figure 9. Specificity of Classification of Lags as Excessive Using a Cutoff of Four Days

Both the *current* and *mean* methods have extreme values for the mean of sensitivity and specificity across all time periods. Conceptually, this means that very few cases with excessive lags were categorized as such, and nearly all cases with acceptable lags were correctly classified. In an intervention situation, we would intervene on very few cases given these classifications. For the *mean* method, this occurs simply because the historical average lag time was smaller than four days, and so no case would be predicted to have an excessive lag. If we had made the threshold for an excessive lag less (e.g., 2 days), all cases would be predicted to have excessive lags because the historical average lag is greater than 2 days. This is a function of the fact that the *mean* method is an overly simple method for predicting lags in future data collection periods. The *current* method, on the other hand, does generate lags based on models that are updated with accumulating paradata. However, the coefficients that are estimated and then applied to open cases are only based on the subset of paradata that has been accumulated so far. Those coefficients may not reflect the true relationships between covariates and true lag when only partial data has been collected, and in this case, biases the predicted lags downward, resulting in low sensitivity and high specificity.

The *historical* method performs more similarly to the *Bayesian* method, generally with lower sensitivities and higher specificities. While the *historical* method relies only on historical data, similar to the *mean* method, it outperforms the *mean* method because it is model-based and can therefore account for differences in both fixed and time-varying effects. This results in a distribution of predicted lags that is closer to the distribution of the true lags, rather than a single point estimate for all cases, as occurs when the *mean* method is used. The *Bayesian* method provides a range of values for sensitivity and specificity. With the exception of day 2, stronger priors result in higher sensitivity in predictions, but lower specificity. This effectively means that, while stronger priors correctly identify cases that will have excessive lags, they also cause more cases with acceptable true lags to be incorrectly classified. Depending on the nature of potential interventions, sensitivity may be prioritized over specificity, or vice versa.

A receiver-operator characteristic (ROC) analysis can be used to evaluate the different methods by comparing the tradeoffs between sensitivity and specificity (Zou, O'Malley and Mauri 2007). Figure 10 shows a generic ROC curve, where sensitivity is plotted against (1-specificity). The line marked "C" displays the ROC curve for random chance predictions, and has an area under the curve (AUC) of 0.50. The point marked "A" is the theoretical best ROC, when both sensitivity and specificity equal 1, meaning no cases are misclassified. In this case, the ROC curve would extend up the y-axis and then across to the x-axis value of 1.0, leading to an AUC of 1.0. Generally, ROC curves look like the line marked "B", falling between "A" and "C".

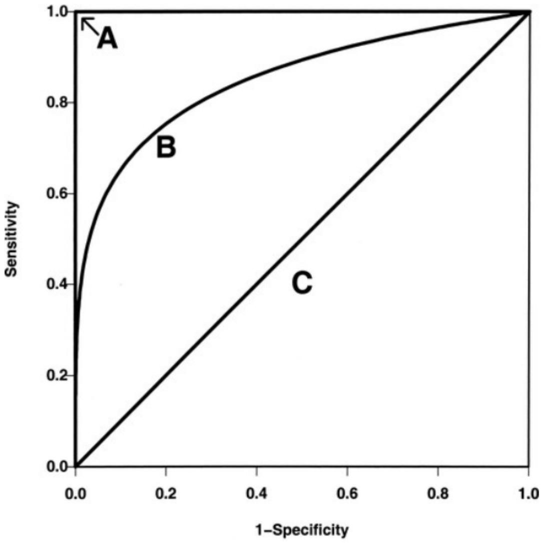


Figure 10. Example Receiver-Operator Characteristic Curve

We can use this ROC space to compare our five predictive methods at the different cutpoints shown in Figures 8 and 9 by calculating the Cartesian distance of each of the sensitivity-specificity pairs from perfect discrimination (x,y)=(0,1). The shorter the distance, the more successful the prediction method.

In order to summarize our results from Figures 8 and 9 above, we use the 25th, 50th, and 75th percentiles of sensitivity and specificity to come up with a below average, average, and above average estimate of the distance from perfect discrimination for the eight prediction methods over the 21 time periods in this application. We carry out these calculations for each of the eight time periods shown in Figures 8 and 9. Results are summarized in Table 6 below.

Table 6. Summary of Prediction Methods by Data Types Used

Day	**Percentile Sensitivity Specificity**	**Distance from Perfect Discrimination by Prediction Method**							
		Current	**Mean**	**Historical**	**Bayesian Methods**				
					c = ⅓	**c = ½**	**c = 1**	**c = 2**	**c = 3**
2	p25	1.121	1.000	0.814	0.900	0.859	***0.808***	0.830	0.845
	p50	1.000	1.000	0.683	0.804	0.796	0.742	0.695	***0.670***
	p75	***0.523***	1.000	0.572	0.658	0.646	0.643	0.580	0.566
4	p25	1.000	1.000	***0.827***	0.891	0.874	0.867	0.834	0.837
	p50	1.000	1.000	0.683	***0.671***	0.688	0.732	0.732	0.773
	p75	1.000	1.000	0.559	***0.479***	0.495	0.533	0.583	0.612
6	p25	1.000	1.000	0.811	***0.810***	0.822	0.845	0.854	0.843
	p50	1.000	1.000	***0.675***	0.684	0.680	0.733	0.761	0.769
	p75	1.000	1.000	0.554	***0.526***	0.557	0.574	0.620	0.659
8	p25	0.872	1.000	***0.810***	0.821	0.834	0.837	0.871	0.874
	p50	0.809	1.000	0.672	***0.671***	0.702	0.734	0.769	0.793
	p75	0.697	1.000	***0.551***	0.581	0.597	0.637	0.633	0.652
10	p25	0.865	1.000	***0.795***	0.867	0.862	0.883	0.891	0.907
	p50	0.814	1.000	***0.666***	0.675	0.711	0.736	0.787	0.795
	p75	0.761	1.000	0.546	***0.542***	0.549	0.552	0.597	0.634
15	p25	1.000	1.000	***0.785***	0.859	0.868	0.879	0.894	0.895

p50	1.000	1.000	0.645	***0.644***	0.695	0.736	0.762	0.806
p75	0.992	1.000	0.547	0.550	***0.536***	0.562	0.593	0.595

Consistent with the analyses reported earlier in this paper, the *historical* method and the *Bayesian* method where c = 1/3 provide the best combinations of sensitivity and specificity. Early in data collection, particularly days 4 and 6, the *Bayesian* method is superior to the *historical* method, and on days 8 and 10, the *historical* method is superior. Additionally, on day 2, the Bayesian method with stronger priors (c = 2 and c = 3) perform better than either the *historical* method or the *Bayesian* method with a weaker prior (c = 1/3). Given the need to obtain higher quality predictions early in data collection, the *Bayesian* method provides a slight advantage when considering equally balanced sensitivity and specificity.

2.6 Discussion and Future Work

This paper discussed several methods for generating predictions for the estimated lag (in days) between first attempt and first contact for a case in the NHIS. The difficulty in estimating these lags arises from the fact that we are making predictions as data collection progresses, meaning we are using partial data that is not necessarily representative of the full data collection. We evaluated four methods that utilize current data and historical data to different degrees: using current data only, using historical data to estimate a mean expected lag, using historical data to estimate model coefficients that were then used to score the current dataset, and using historical data to estimate priors that are then combined with current data to create posterior predictions of estimated lag.

The results showed that utilizing data external to the current data collection, in this case historical survey data, can be useful for improving predictions in the current month. If there is more variation in the lag between first attempt and contact within a data collection period than there is variation across data collection periods, historical data can be a useful way to complement the partial data in the current data collection period, filling in the gaps in what we know about the relationship of certain covariates with the expected lag.

Additionally, both of the model-based methods that incorporate historical data were superior to the historical mean method, particularly earlier in the data collection period. Because of the time-varying nature of the day of the first attempt covariate (cases worked earlier generally have longer lags than cases first attempted very late), the *historical method* and the *Bayesian method* were able to capture that variability, resulting in better predictions, particularly near the point of first attempt. Further, the model-based methods improved estimates with relatively little additional information – there is some low level geographic information, minimal interviewer information, and the day of first attempt. Improvements could be more significant with more information, either for individual cases in a longitudinal setting, or just richer auxiliary frame data even in a cross-sectional setting.

When considering the *historical method* versus the *Bayesian method*, the Bayesian model provided modest improvements over the historical data model, and those gains were observed throughout the data collection period. Early in data collection the benefit of the Bayesian model was that external data can be incorporated into the prediction process as priors. Those priors can help improve the stability of predictions when working with the partially accumulated, sparse data that exist early in the data collection period. Later in data collection, the current data takes over, reducing bias that might have been introduced by considering historical data only, although the prior continues to reduce the variance of the predictive coefficients, stabilizing the posterior predictions of expected lag.

The Bayesian method also offers flexibility in how much influence the prior has on the posterior prediction, in the form of the constant, *c*. By increasing the value of *c*, the strength of the prior increases, and for this application, a stronger prior resulted in smaller RMSE than a weak prior or using the historical predictors as fixed coefficients. A strong prior may be particularly valuable early in data collection, when the current data would have been missing many cases, and have many cases that were attempted, but not contacted.

There are limitations to the model-based predictions, though these are not all specific to the Bayesian method. From a data quality perspective, the NOI and the CHI data are self-reported by interviewers, who are expected to record outcomes of contact attempts immediately after they occur. Some research has considered potential errors in this source of data. West and Kreuter (2013) discussed how neighborhood observations may constitute guesses on the part of interviewers, and erroneous observations decrease the predictive power of these covariates. Separately, Biemer, Chen and Wang (2013) raised concerns about interviewers under-reporting contact attempts. The CHI could be used to identify undesirable interviewer behaviors (Bates et al. 2010), such as high numbers of contact attempts or the repetitive use of contact strategies, causing a reduction in the recording of contact attempts that may draw attention. However, the two events used for our predictions – first contact attempt and first successful contact – are desirable behaviors, hopefully reducing the likelihood these go unreported.

Additionally, as illustrated in the Results section, the predictive models did not perform consistently across all time periods included in this analysis. This is likely a function of our decision to fix the set of covariates across all time periods, rather than allowing them to vary based on which variables were most predictive for the prior three months, potentially ignoring a seasonal effect of specific variables, or simply natural variation in which variables were most predictive. We required any variable retained after the backwards stepwise regression discussed in the model selection discussion to be a significant predictor in at least seven of the 21 time periods. In doing this, we were able to achieve a smaller, more parsimonious model, and simplify the coding required to extract priors from the data. However, we may also have excluded variables that were highly predictive during selected time periods. This limitation could be mitigated by either including a larger set of model covariates throughout all time periods, or allowing the variables included in the model-based methods to vary over time.

Generally, the variables included in the models, and how the priors are developed, are complexities without straightforward solutions. Here, because the NHIS is an ongoing survey with available paradata, we were able to obtain several sources of data from the frame and interviewer observations. Additionally, as this application was based on analyzing retrospective data, we were able to build an analytic file and conduct variable selection and build a model using information over the 24 months in order to reduce the set of covariates to those that were predictive of our outcomes in at least a third of the included time periods. However, in a true prediction situation, models would have to be built from existing (historical) data only. Even more difficult, priors for a new survey may need to be generated from somewhat dissimilar surveys, or other sources of information. Despite these limitations, the *Bayesian* method generally resulted in lower RMSE of predictions, which translated into the identification of cases with excessive lags at a higher sensitivity than the other methods. When sensitivity and specificity were considered equally in prediction evaluation, the *Bayesian* method performed slightly better early in data collection, though the *historical* method was certainly competitive. .

Future work could extend into several different areas. First, more work could be done to append more useful auxiliary variables onto the survey data, and potentially include a flexible set of model covariates over time, in order to observe whether larger improvements can be found from the Bayesian model when the underlying model has a better fit. For example, the planning database information that was appended to the sample represented low-level aggregate information about the block group in which a sample case was located. If information about the specific household were available, the model may have more predictive power. Additionally, it could be useful to have information about past response behavior for the sampled household (in the same or different surveys), or more information about past contact rates of the interviewers making contact attempts. While the *Bayesian* methods showed improvement over the *historical* and other methods, predicted lags were still generally underestimating the true lag, sometimes by several days. Therefore, better external covariates that can be appended to the frame could lead to even larger reductions in the MPB and RMSE of predicted lags. Additionally, this method should be replicated for different estimators in different data collection settings. Ongoing surveys would benefit from the ability to leverage their historical data to improve survey outcomes.

Research might also focus on identifying an appropriate value of *c*. For this application, we varied the value of *c*, allowing the priors to represent a range of 1/3 of a month of data to 3 months of data for the results in order to demonstrate the effect of prior data strength on the posterior predictions of lag. However, the optimal value for *c* may be a prediction-specific issue. Additionally, tradeoffs between sensitivity and specificity may be considered when interventions are potentially resource intensive. It may be worth selecting a prior that provides prioritizes sensitivity over specificity if the intervention is lower risk or requires fewer resources. On the other hand, for an intensive or expensive intervention, prioritizing specificity, in order to avoid applying the intervention to misclassified cases, may be more important.

Finally, it is important to consider how to utilize these improved predictions during data collection. This lag, for example, could be used to identify cases that are at risk of non-completion because of the predicted time it would take after making a first attempt to make contact with a particular sample member, as demonstrated in the hypothetical application. Interviewers working on those cases could be alerted to this risk and coached or given different instructions for how to work those cases. Additionally, if at-risk cases were clustered within specific interviewers, additional training could be provided to mitigate the risk of excessive lags on survey outcomes. Obtaining more accurate estimates for progress metrics such as the one examined here could be useful for monitoring, or intervening in, data collection operations.

3. What Do You Think? Using Expert Opinion to Improve Predictions of Response Propensity Under a Bayesian Framework

Stephanie Coffey[1], Brady T. West[2], James Wagner[2], Michael R. Elliott[2,3]

[1] Joint Program in Survey Methodology, University of Maryland
United States Census Bureau, Washington, DC
[2] Survey Research Center, Institute for Social Research, University of Michigan
[3] Department of Biostatistics, University of Michigan

Abstract

Responsive survey designs introduce protocol changes to survey operations based on accumulating paradata. Case-level predictions, including response propensity, can be used to tailor data collection features in pursuit of cost or quality goals. Unfortunately, predictions based only on partial data from the current round of data collection can be biased, leading to ineffective tailoring. Bayesian approaches can provide protection against this bias. Prior beliefs, which are generated from data external to the current survey implementation, contribute information that may be lacking from the partial current data. Those priors are then updated with the accumulating paradata. The elicitation of the prior beliefs, then, is an important characteristic of these approaches. While historical data for the same or a similar survey may be the most natural source for generating priors, eliciting prior beliefs from experienced survey managers may be a reasonable choice for new surveys, or when historical data are not available. Here, we fielded a questionnaire to survey managers, asking about expected attempt-level response rates for different subgroups of cases, and developed prior distributions for attempt-level response propensity model coefficients based on the mean and standard error of their responses. Then, using respondent data from a real survey, we compared the predictions of response propensity when the expert knowledge is incorporated into a prior to those

based on a standard method that considers accumulating paradata only, as well as a method that incorporates historical survey data.

3.1 Introduction

Responsive Survey Design (RSD; Groves and Heeringa, 2006) relies on accumulating paradata (i.e. data about the process of collecting survey data, see Couper 2000, 2017) and response data in order to introduce changes to data collection protocols or tailor data collection features to specific cases. These changes are made in pursuit of a survey goal, such as quality improvement or cost control. Unfortunately, by relying only on the partial current data as it accumulates, predictions generated from this partial data may be biased (Wagner and Hubbard 2014) and, as a result, decisions made based on these predictions can be inefficient or even harmful.

Recently, survey researchers have introduced Bayesian approaches (Schouten et al. 2018) to mitigate this bias by supplementing the current accumulating data with prior beliefs, generated from external data such as past implementations of the same survey or the survey methodological literature (West, Wagner, Coffey and Elliott 2019). While priors generated from past implementations of the same survey may be the most informative for a particular survey, that solution is not always an option. New surveys, or surveys whose designs have changed dramatically, may need to develop priors from different data sources. West et al. (2019) explored using a literature review to source prior information for response propensity models in the National Survey of Family Growth (NSFG). While priors from the literature review did not perform as well as priors from historical NSFG data, they outperformed model predictions made only using current accumulating paradata, particularly in the middle portion of the data collection period.

The present study evaluates another potential source of prior information. Here, expert knowledge was elicited from survey managers (“experts”), through a self-response questionnaire designed to collect their predictions of attempt-level response rates, or changes in those expected response rates, for various types of sample members. Given those survey responses, pooled priors were created from expert respondent data. The structure of the items in the questionnaire completed by the experts mimicked that of the existing response propensity model. We then evaluated these priors’ ability to improve predictions of response propensity in the National Survey of Family Growth (NSFG) relative to only using partial data from the current round or using historical data as an alternative source for the development of priors. This manuscript discusses the content of the questionnaire, the identification of experts, the method for generating priors, and an evaluation of how the information from expert elicitation affects the bias and root mean squared error (RMSE) of the daily predictions of response propensity. We found that priors based on expert opinion led to modest improvements in prediction during the middle and late portions of data collection when compared to using only current round data. Additionally, we found that priors based on expert opinion were sometimes competitive with, though generally did not outperform, an approach that used historical data evaluated in West et al. (2019). We also identified several ways to improve upon our elicitation process that may lead to further improvements in predictions based on expert opinion over methods more commonly used in RSDs.

3.2 Background

3.2.1 Responsive Survey Design

Responsive survey design (RSD; Groves and Heeringa 2006) has emerged as a framework for maintaining or improving survey outcomes in an increasingly difficult survey climate. Increasing data collection costs, and decreasing cooperation and response rates, have caused survey methodologists and managers to explore alternatives to the prevailing “one path fits all sample members” approach to data collection operations (Axinn, Link and Groves 2011). Instead, RSD uses accumulating paradata and response data to make changes to later data collection protocols. These changes attempt to increase data quality in some specified way or control costs, relative to continuing with the standard data collection protocol. Types of protocol changes may include introducing another mode (Coffey, Reist and Miller 2019), changing the effort spent on specific cases (Rosen et al. 2014), or a change in tokens of appreciation combined with subsampling (Wagner et al. 2012).

In an RSD, one of the most common ways to tailor data collection features to specific cases is with predicted propensity scores. Based on frame data and accumulated paradata, these predictions can be used to alter data collection operations. Various surveys have utilized propensity scores to differentially implement a variety of data collection features, including protocol assignment (Peytchev, Rosen, Riley, Murphy and Lindblad 2010; Roberts, Vandenplas and Stahli 2014), incentives (Chapman 2014), and allocation to nonresponse follow-up (Laflamme and Karaganis 2010; Thompson and Kaputa 2017) in hopes of improving survey outcomes.

Paradata from the current round of data collection provide useful predictors of survey outcomes, such as response propensity, for the sampled cases currently receiving recruitment effort. In an RSD, targeted interventions are applied to cases during the data collection period in order to shift response propensities in pursuit of a cost- or quality-related survey goal, necessitating high quality predictions of these propensities. However, during the survey period when an RSD would be implemented, the accumulating paradata are “incomplete” relative to the final data, in that completed cases and incoming data from early in the data collection period may not be representative of that which will be collected later in data collection. As a result, only using the accumulating data from the current round of data collection could result in biased predictions of response propensity (Wagner and Hubbard 2014) or reduced prediction performance when predicted propensities are classified into response categories, either of which could lead to inefficient decisions. In this paper, we focus on the error in the predictions of response propensity scores, as opposed to the secondary step of classification error.

In order to improve predictions, survey practitioners often use external data that may be more representative of a full data collection period. It is relatively common to estimate the coefficients of a predictive model using historical data, such as a prior implementation of the survey, and then apply those coefficients to the current round of data collection (Schouten, Calinescu and Luiten 2013; Schouten, Wagner and Peytchev 2017; Schouten, Mushkudiani, Shlomo, Durrant, Lundquist and Wagner 2018). While this method provides data that might be representative of an entire data collection, it ignores current data in the prediction process.

More recently, survey researchers have begun exploring Bayesian approaches that utilize both external and current data in the prediction process. Prior beliefs are generated from external data, most commonly historical data from the same survey, and those priors are then updated as the current data accumulates. Schouten et al. (2018) discuss using Bayesian methods for predicting response and cost under different scenarios. Through simulation, they demonstrate value in the Bayesian methods in terms of reduced RMSE of predictions, while stressing that misspecification of the priors with respect to the true data should be relatively small. Empirical evidence is also emerging (West et al. 2019) that combining published estimates or historical information and current round information in a Bayesian setting can improve prediction.

3.2.2 Empirical Evidence and Sources of Prior Information

West et al. (2019) compared the performance of predictions of response propensity in the NSFG, a nationally representative quarterly survey in the U.S., when Bayesian methods are used versus when only current data is used. The Bayesian methods incorporated external information in the form of priors, either from past implementations of the NSFG or from published research on propensity models found through a literature review. Results demonstrated that the Bayesian approaches consistently reduced both the bias and the mean squared error (MSE) of predicted response propensities, particularly in the middle of data collection, when an RSD may be implemented. This was true for either source of prior information -- the historical data or the literature review.

The quality of the prior information is directly related to its ability to improve predictions of interest, and so the source of prior information is an important consideration. It seems reasonable that historical data from the same survey would result in the most informative priors for the prediction of interest; however, there may be cases where this information is not available. New surveys, for example, would not have access to historical information. Additionally, surveys that have undergone significant redesign, such as introducing a new mode, changing an incentive amount, or dropping a screening interview, may find that priors based on historical paradata are no longer available.

There may be cases where even a literature review produces limited or no useful external information. In the case where a survey has an unusual or unique target population, or the prediction of interest is not as common as response propensity, there may not be sufficient information in the literature from which to develop priors. In these cases, where there is an absence of objective information, expert opinion may be the only option for generating the necessary information for prior construction. Expert opinion is often used implicitly in survey planning – experienced survey managers may provide input into expected response rates to help determine sample sizes, or for estimating budgets. Additionally, they may help explain variation progress or response rates during data collection. Transforming expert opinion into priors explicitly incorporates this information into the prediction model.

3.2.3 Expert Elicitation

Clinical trials and health care evaluations often rely on prior beliefs for a variety of reasons. Dallow, Best and Montague (2018) describe a protocol for eliciting expert opinion in order to improve the drug development process. Mason et al. (2017) propose a practice for leveraging expert opinion in the analysis of randomized controlled trials when there are missing observations for patients. Additionally, Boulet et al. (2019) demonstrate the use of expert opinion in a variable selection process for personalized medicine. When novel treatments are tested, or prior trials have very small sample sizes or are otherwise not comparable, expert opinion can be relied upon for developing priors (Hampson, Whitehead, Eleftheriou and Brogan 2014)**.**

Spiegelhalter et al. (2004, Ch. 5) as well as O'Hagan (2019) provide overviews of the expert elicitation process, and the potential biases that may arise in priors elicited from individuals. *Availability bias* may arise when experts are asked about easily recalled events – they may estimate a higher or lower probability than is accurate. For example, if survey experts have recently seen frequent reports of language barriers along with increasing non-interview rates, the experts may inflate the effect that a language barrier has on overall response rate or response propensity, even if there are other contributing factors to increasing non-interview rates. *Anchoring bias* may lead experts to shrink intervals between different categories or groups based on a provided piece of information or their initial elicited quantity or probability. Once an expert learns from the elicitation instrument, or offers through the elicitation process, that the expected response rate for one group is 45%, future answers about different subgroups may be biased towards 45%.

Overconfidence bias may lead to distributions of the priors with insufficient variance. This may occur when elicitation happens in small groups and some strongly opinionated experts convince others of their opinion, a behavior also known as groupthink. Alternatively, in individual elicitation, overconfidence bias may arise because of the expectation of experts that they have, in fact, a greater amount of expertise than they actually do, resulting in under-reported uncertainty. *Conjunction fallacy bias* may arise when a particular event is given a higher estimated probability when it is the subset of another event. For example, on any given contact attempt, the probability that any open case will have had a callback request and response is necessarily smaller than the probability that any open case will respond. However, an expert may suggest the opposite, thinking that having a callback request makes response much more likely. This bias is often due to the rarity of one of the two events, which in this case would be the callback request. Finally, *hindsight bias* may arise if the expert is asked to provide a prior expectation after looking at the current data. Awareness of all of these types of bias is useful in the design of the expert elicitation process.

Spiegelhalter et al. (2004, Ch. 5) also discuss four common methods for elicitation: informal discussion, structured interviewing, structured questionnaires, and computer-based elicitation. Each of these methods requires different amounts of interaction with experts, and allows for different levels of complexity of prior development. Additionally, these authors discuss three methods for combining information when multiple experts are utilized: arriving at a consensus value among all experts, arithmetic pooling, or retaining individual priors. O'Hagan (2019), whose elicitation method elicits distributions from experts, discusses the combination of those distributions to generate a pooled empirical distribution for the prior.

Here, we adapted the concept of expert elicitation of priors from the clinical trials literature. Our goal was to evaluate whether expert opinion can be helpful when little objective data is available for generating priors for the coefficients in a logistic regression model used to estimate propensity of response. In this application, we elicited opinion from experts independently through an internet questionnaire, and used arithmetic pooling to combine the elicited information into priors for models used to generate daily predictions of response propensity in the NSFG.

3.3 Data and Methods

3.3.1 Overview of the National Survey of Family Growth

The NSFG is conducted by the National Center for Health Statistics, under contract with the Institute for Social Research (ISR) at the University of Michigan. The NSFG, in its current iteration, is a cross-sectional survey for which data were collected continuously throughout the calendar year from 2011-2019. In a given year, four data collection operations are conducted, with data being collected from four independent, nationally representative samples. The field operations for each sample last three months, or one quarter (e.g., January to March, April to June). The survey selects a national sample of U.S. housing unit addresses each quarter of the year. The target population from which the NSFG selects these four independent national samples is 15 – 49 year old persons living in the U.S. (Lepkowski, Mosher, Groves, West, Wagner and Gu 2013). The NSFG is a two-stage survey, meaning there is first a screener interview to determine eligibility, followed by the main interview. Interviewers first visit randomly sampled households and attempt to screen the households for eligibility. Within eligible households, one of the eligible individuals is randomly selected to complete the main survey interview, which usually takes 60-80 minutes and covers a variety of fertility-related topics.

NSFG paradata are aggregated on a daily basis and used to predict the probability that active households will respond to either the screening interview or the main interview. Survey managers might use these predictions for prioritization of active cases (e.g., Wagner et al. 2012) or for stratifying the sample when selecting a subsample of active cases for the new data collection protocol after 10 weeks (Wagner et al. 2017). At this point, managers may oversample high-propensity cases, or offer a higher token of appreciation to encourage response. Accurate model-based predictions are thus essential for maximizing the efficiency of the data collection effort in any given quarter. For purposes of this study, we focus on models for the probability of responding to the initial screening interview.

3.3.2 Response Propensity Models in the NSFG

For this application, we used data from five quarters of the NSFG (Quarters 16 – 20), covering the June 2015 to September 2016 time period. For each of the five quarters, our prediction of interest was the probability of response to the screening interview at the next contact attempt, using *either* the current accumulating paradata only, or the combination of priors generated from expert elicitation and the current accumulating paradata. We also compared these methods to the best performing method in West et al. (2019), which combined current accumulating paradata with priors based on historical data from the eight preceding quarters of data collection.

In order to compare predictions generated from our proposed method with those discussed in West et al (2019), we used the same predictive modeling approach (discrete time logistic regression), and the same set of predictors of screener response propensity. In that paper, eight quarters (or two years) of the NSFG (Quarters 13 – 20) were combined into a stacked dataset containing all contact attempt records and a binary outcome for each record that indicated whether the screener interview was completed on that particular attempt or not. The authors then fit a discrete time-to-event logistic regression model to this dataset to identify significant predictors. Available predictors included sampling frame information, linked commercially-available data, and NSFG paradata, all of which have been used to predict response propensity in the NSFG (West

2013; West and Groves2013; West et al. 2015). The authors used a backward selection approach to model-building, retaining all predictor variables that appeared in all eight quarters with a *p*-value less than 0.05 based on a Wald test for all regression parameters associated with a given variable.

They then included two predictor variables that were important for sampling and weighting in order to control for sampling domain in the response propensity model. The first was the sociodemographic domain of each housing unit, based on the percentage of the population in the Census Block Group containing the segment that is Black and/or Hispanic as reported in U.S. Census data. The second was a three-level categorical variable indicating whether a case was in a self-representing area, a non-self-representing metropolitan statistical area (MSA), or a non-MSA non-self-representing area. Self-representing sampling areas are geographic sampling domains that are large enough to be sampled with certainty in a probability proportionate-to-size sample, and, therefore, represent only themselves during weighting and estimation. These two variables were initially included in the backwards stepwise procedure, but were not found to be statistically significant, and so were not retained. However, after consultation with data collection managers, these two variables were added back into the response propensity model in order to control for sampling domain in the predictive model.

All retained predictors from the backward selection process carried out in West et al. (2019), including their estimated coefficients and standard errors, are listed in Appendix A. Several predictors came from each available data source: the sampling frame, commercially-available data, and paradata. By using the same list of predictors, and the same discrete-time logistic regression model specification, we are able to compare the effect that priors based on expert elicitation have on the predictions of response propensity, versus excluding prior information, or using priors from historical data. The focus of our analysis is on the relative performance of these methods given a particular model.

3.3.3 Design of Prior Elicitation Process

For this proof-of-concept study, we wanted our prior information to be based upon a relatively large group of experts $(n \cong 20)$ to generate a reasonable distribution from which to estimate priors. Our target sample size meant that elicitation methods requiring significant interaction with experts, including informal discussion and structured interviewing, were not feasible. As a result, we created and distributed a structured questionnaire to selected experts, who could then respond at their convenience. The questionnaire asked experts to provide their opinions on attempt-level response rates for subgroups with various types of characteristics, and, in some cases, opinions on changes to response rates based on certain characteristics.

The questionnaire included the significant predictors found in the retrospective analysis of the NSFG response propensity model, as described in Section 3.2. These predictors include items from the sampling frame, including geographic and sampling strata information, as well as time-varying attempt-level information, derived from accumulating paradata. Fixed characteristics include sampling frame or commercially available data, like the 9-level Census Division geographic variable. In the questionnaire, we asked experts their opinions on their expected response rates for each of the nine categories. Time-varying covariates were based on paradata and include indicators for past contact or instances of the sample member expressing questions, comments or concerns. In the questionnaire, we requested information about the expected *change* in response rate for characteristics like each additional contact attempt, or whether the sample member expressed comments on concerns on the most recent contact attempt. We also asked experts to provide their experience with survey data collection by selecting one of three categories: 0 to 4 years, 5 to 15 years, and 15 or more years.

We solicited feedback from two survey experts prior to distributing the questionnaire in order to get basic feedback about content, complexity, and readability. In some cases, edits resulting from this initial feedback changed the format of the questions to make them easier to understand and answer. This meant that the format of the questions did not always match the format of the predictor in the propensity model. The final version of the questionnaire can be found in Appendix B, and in the Center for Open Science repository (https://osf.io/3kxzb/) at the Open Science Framework (log-in required).

Given the target number of experts, we opted to develop priors through arithmetic pooling of all respondent information. At the same time, we wanted to avoid the biases mentioned by Spiegelhalter et al. (2004, Ch. 5). In order to avoid *anchoring bias* while still eliciting reasonable responses, we provided an overall expected attempt-level response rate (24%), but did not provide anchor points for any particular category in the survey, allowing the experts to provide input for all items and categories. To avoid *hindsight bias* (Schouten et al. 2018) arising from the fact that experts at ISR also conduct the NSFG, we recruited additional experts from the U.S. Census Bureau (Census). These additional experts have experience managing interviewer-administered data collections, but do not have experience with the NSFG or its data. By soliciting predictions from two geographically dispersed survey organizations with varying familiarity with the NSFG, we also hoped to protect against *overconfidence bias* (Schouten et al. 2018), which can lead to prior distributions that are too narrow and do not accurately reflect the uncertainty in the prior.

At both ISR and Census, we worked with senior survey managers to identify experienced interviewer supervisors, field directors, and survey methodologists who were knowledgeable about survey processes and reviewed progress data on a daily basis as part of their job responsibilities. We recruited eight individuals from ISR, and 12 from Census (two from each of the six regional offices). During March 2019, the recruited experts were asked to complete the questionnaire, and were encouraged to provide feedback, either directly or through a scheduled debriefing. We summarize the feedback received in the Results section.

3.3.4 Method for Deriving Priors

We obtained 20 sets of expert responses about the effects on attempt-level response rates of various characteristics of sample members and paradata items, subject to some item nonresponse. We used arithmetic pooling to combine the priors and generate an expected mean and standard error for a coefficient in an attempt-level response propensity model (Spiegelhalter et al. 2004, Ch. 5).

Before pooling, however, we had to convert the estimates of differences in response rates to model coefficients for use in a logistic regression model. When categorical variables are included as predictors in a logistic regression model, the estimated coefficients are generally interpreted with respect to a reference category. Therefore, the mathematical manipulation involved identifying a reference category, calculating odds ratios with respect to the reference category, and then taking the natural log of the odds ratio to obtain a logistic regression model coefficient, or *beta*. We first did this for each respondent's information individually.

Formula 1 below demonstrates how to calculate the coefficient for the k^{th} category of the j^{th} item for the i^{th} expert, $\hat{\beta}_{ijk}$, given the estimated probability of response for category *k* of interest, $\hat{p}_{ijk}$, and the estimated probability of response for a reference category *R*, $\hat{p}_{ijR}$.

$$\hat{\beta}_{ijk} = ln\left(\frac{\hat{p}_{ijk}/(1-\hat{p}_{ijk})}{\hat{p}_{ijR}/(1-\hat{p}_{ijR})}\right)$$

Using gender as an example (abbreviated G in the expression below), assume that the i^{th} respondent estimates the expected call-level response rate for female sample members to be 85% (as opposed to 70% for males), and male is the reference category. The *beta* for female sample members, for the i^{th} expert, would be:

$$\hat{\beta}_{iGF} = ln\left(\frac{\hat{p}_{iGF}/(1-\hat{p}_{iGF})}{\hat{p}_{iGM}/(1-\hat{p}_{iGM})}\right) = ln\left(\frac{0.85/(1-0.85)}{0.70/(1-0.70)}\right) = 0.8873$$

Continuous variables were converted to model parameters using the same formula but with a slightly different explanation. For these items in the questionnaire, expert opinion was elicited about the *change* in response propensity, given some unit change in the continuous variable. For example, survey managers were asked to provide their expected change in response rate for each additional contact attempt made on a sample member, and a survey manager might have responded saying they would expect a -10% change, or a 10% reduction, in response propensity for each additional contact attempt.

However, unlike standard linear regression, where there is linear change for every unit increase, logistic regression results in exponential change for each unit increase, meaning the change in response propensity is dependent on *which* unit increase is being considered (e.g. from 1 to 2 attempts, or from 8 to 9 attempts). In the case of continuous variables, we did not have a defined reference category, and so the reference is always to the average attempt-level response rate of 24%.

If the i^{th} expert believes increasing the number of contact attempts, j, by one would change the attempt-level response rate by some amount, we can adapt Equation (1) above for a continuous variable. While we do not have a defined reference category, we have the overall average attempt-level response rate, 24% and the expected change provided by the expert, 5%. This results in a model coefficient of:

$$\hat{\beta}_{ij} = \ln\left(\frac{odds\left(attempts = \ (n+1)\right)}{odds\left(attempts = \ (n)\right)}\right) = ln\left(\frac{0.29/0.71}{0.24/0.76}\right) = 0.2573\ .$$

We note at this point that, while we have elicited priors on a linear scale, linking these back to the logistic scale changes the interpretation. We provide more consideration of this issue in the Discussion section.

To pool the expert information, we then took an arithmetic mean, $\hat{\hat{\beta}}_{jk}$ (or $\hat{\hat{\beta}}_{j}$ for continuous items), of the coefficients from the expert respondents. The standard error of the prior, $SE\left(\hat{\hat{\beta}}_{jk}\right)$, was estimated by dividing the standard deviation of the coefficients from the respondents by the square root of the number of respondents.

$$\hat{\hat{\beta}}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_{ijk}$$

$$SE\left(\hat{\hat{\beta}}_{jk}\right) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\beta}_{ijk} - \hat{\hat{\beta}}_{jk}\right)^{2}}$$

We chose to transform each expert response into an odds ratio, take the log, and then pool the individual log-odds ratios for a few reasons. Mathematically, by first transforming each expert response into a log-odds ratio before pooling, we are working under the assumption that the log-odds are normally distributed, as opposed to the response rate or response propensity, which is how the experts provided their opinions. We felt this assumption was reasonable. First, response rates and response propensities are bounded at (0,1), and are not normally distributed, whereas the log-odds can take on any number on the real line. Additionally, the log-odds is a linear function, while the function for the odds (and for probabilities) are multiplicative and exponential, which suggests that the log-odds might converge to a normal distribution more quickly than the odds, given enough sample size.

Operationally, by generating a model coefficient for each expert, we were able to calculate a mean and standard error for each model coefficient. If we had first taken the mean of the expert response first, and then transformed that estimate to obtain our model coefficient, we would no longer be able to generate a variance, as we would have only one estimate.

For each covariate of interest, we used $\left(\hat{\bar{\beta}}_{jk}, SE\left(\hat{\bar{\beta}}_{jk}\right)\right)$ to define a normal prior distribution in our prediction models. Each prior was based on a maximum of 20 responses, but item-level nonresponse reduced the number of responses to varying degrees (see Table A2 for individual response counts). Due to the small sample sizes, we ignored the potential covariance between the coefficients, resulting in a variance-covariance matrix that is only non-zero on the diagonal. This is different from the methods evaluated in West et al. (2019) that utilize historical data to generate priors. For those methods, including the historical method replicated in our results, estimated covariances were generated from the existing historical data.

Appendix C provides the prior information, $\left(\hat{\bar{\beta}}_{jk}, SE\left(\hat{\bar{\beta}}_{jk}\right)\right)$, for each covariate included in the propensity models, provided that there were at least three contributing respondents. Further, an Excel spreadsheet available in the online supplementary material provides a template for estimating these priors for the survey items in the propensity model. For demonstration purposes, simulated data are included in the table, including missing cells, which would occur should an expert not respond to a particular question.

3.3.5 Methods for Predicting and Evaluating Response Propensities

Each of the five NSFG quarters of interest (Quarters 16 through 20, representing June 2015 – September 2016) were analyzed independently to introduce replication in our analysis. First, we used the expert opinions to generate the prior distributions for the response propensity model coefficients as described above. These priors were used for all five quarters.

We generated our “target” prediction at the case level for each of the five evaluation quarters by fitting a discrete time-to-event logistic regression model using the predictors identified in the backward selection model discussed in Section 3.2 to all contact attempt records from that quarter. This allowed us to estimate a “final” probability of responding to the screener interview at the last contact attempt for each case. Because this model uses all available information for a given quarter, we consider this the benchmark against which the prediction methods under evaluation will be compared. Table 7 below shows the ROC-AUC values when all contact attempt records were used to predict final response.

Table 7. Model Fit Statistics for In-Sample Predictions of Response, 5 Evaluation Quarters

	Q16	**Q17**	**Q18**	**Q19**	**Q20**
ROC-AUC	0.711	0.682	0.661	0.690	0.654
Nagelkerke-Pseudo R2	0.143	0.115	0.089	0.130	0.086

These model fit statistics are reflect the in-sample performance of the models demonstrate that the variable selection procedure from West et al. (2019), where these statistics are extracted from, yielded a reasonable list of predictors for our target response propensity. From that point, we are concerned with the case-level differences from the target propensity that the different methods produce.

Then, we generated daily predictions of response propensity based on contact history data accumulated prior to each day. Our baseline predictions came from the model using only accumulating current round paradata. Our proposed predictions came from the model that also incorporated prior information from expert opinion. Additionally, we included predictions that incorporate prior information from historical data, as presented in West et al. (2019). In that paper, the authors found that the historical data method performed the best in their application. We include the historical data method here so we can understand how well the expert elicitation method performs when compared to both the "current data only" method and one of the historical data methods evaluated in West et al. (2019).

Prediction of daily response propensity for each of these three methods is carried out just as it would have been if the approach were to be employed during data collection. For each of the five quarters of interest, we use the accumulated contact attempt record information (with a screener response indicator for each record) up to day d to estimate the coefficients for the discrete time logistic regression model for that data collection period. Then we use those coefficients to predict the response propensity at the next contact attempt for all cases who were nonrespondents on day d. We repeat this for each day of data collection from Day 7 to Day 84.

Using only the current quarter of paradata, the response propensity, $\hat{p}_{id}$, was modeled as follows:

$$\hat{p}_{id} = \hat{p}(y_{id} = 1|X_{id}) = \frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}$$

where y_{id} is the response status for the i^{th} case after a contact attempt on the d^{th} day, and X_{id} is the set of predictors v for the i^{th} case after the d^{th} day. These predictors may be fixed (e.g., geographic predictors) or time-varying (e.g., prior contact status). The $\hat{\beta}_v$ are estimated coefficients for the X_{idv} predictors. They are estimated from the likelihood in equation (5) based on the contact attempt records that have been accumulated through day d.

$$L(\hat{\beta}_0, \dots, \hat{\beta}_v) = \prod_{i=1}^{n} \prod_{j=1}^{d} \left(\frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right)^{y_{id}} \left(1 - \left(\frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right) \right)^{(1-y_{id})}$$

The only difference between the target prediction and the baseline, current-data only method is the time at which the prediction is made. For the target predictions, all contact attempt records from a given quarter are used (d is after the last contact attempt is made in a given quarter); for the baseline method, only data accumulated through day d are used.

In a Bayesian setting (Gelman et al. 2013), the likelihood matches the frequentist formulation. The only estimated parameters in this expression are the $\hat{\beta}_v$, and so these are the parameters for which priors are defined. As described in Section 3.4, we assumed a normal distribution, $\beta_v \sim N(\mu_v, \sigma_v^2)$, for our priors with the mean and variance based on our expert elicitation procedure. The posterior multiplies the prior over the parameters in the likelihood to combine the information, as shown in equation (6):

$$\begin{aligned} pos(\hat{\beta}_0, \dots, \hat{\beta}_v) = \prod_{i=1}^{n} \prod_{j=1}^{d} \Bigg[\left(\frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right)^{y_{id}} \Bigg(1 \\ - \left(\frac{\exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)}{1 + \exp\left(\sum_{v=0}^{V} \hat{\beta}_v X_{idv}\right)} \right) \Bigg)^{(1-y_{id})} \Bigg] \times \prod_{v=0}^{v} \frac{1}{\sqrt{2\pi\sigma_v^2}} exp \left(-\frac{1}{2} \left(\frac{\beta_v - \mu_v}{\sigma_v} \right)^2 \right) \end{aligned}$$

In the Bayesian version of the prediction, it is clear that the priors add additional information to the prediction. This can be beneficial when the likelihood is based on very sparse data, or partial data that are not representative of the full data collection process, both of which occur earlier in the data collection process. Code in the SAS 9.4 programming language that can be used to carry out these predictions is available in the online supplementary materials.

For each method, we will compare predictions for each contact attempt on each day of the data collection quarter to the “target” predictions (based on all cumulative data) in order to generate daily estimates of the bias and root mean squared error (RMSE) for the predictions. The mean daily bias for the m^{th} method is defined as:

$$B^{m} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{\rho}^{m}{}_{i} - \rho_{i}\right)$$

and the daily RMSE for the m^{th} method is defined as:

$$RMSE^{m} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{\rho}^{m}{}_{i} - \rho_{i}\right)^{2}}$$

We then summarized those estimates using boxplots for three different parts of data collection: early (day 7 – 30), middle (day 31 – 60), and late (day 61 – 84).

The end-of-data-collection response propensity is not the only possible target, but this choice does allow us to evaluate whether the use of Bayesian approaches with informative priors can reduce error in the predictions of response propensity at a given contact attempt versus using only current round paradata. Additionally, we will be able to evaluate whether the use of expert opinion (in the absence of historical data) can perform similarly to the historical data, were it available.

3.4 Results

3.4.1 Descriptive Statistics for Selected Priors

We first wanted to understand if ISR experts have different expectations than Census experts, potentially due to the varying familiarity with NSFG or simply being a part of a different survey organization. We also collected information about the experts' length of experience with survey data collection, thinking opinion may vary with length of experience and more experienced managers may provide more useful information. We then examined distributions of the individual experts' betas, generated using Equations (1) and (2) above, by organization and experience level. Here we provide examples of these distributions to illustrate similarities and differences in the provided opinions. Due to the small sample sizes, we do not provide tests of significance with respect to these differences. Instead, we are interested in the means and general trends of the expert opinion by category in order to understand, at a high level, if different types of experts provide different information.

We first examined distributions of coefficients related to two time-varying covariates, Contact Status and Concerns Status. Contact Status had three possible response categories: if there was ever contact with the sample member, contact on the previous attempt, or if there had never been contact with the respondent, which was used as the reference category. Concerns Status had four possible response categories: if concerns were ever expressed by the sample member, if concerns were expressed on the previous visit, if strong concerns were ever expressed, or if no concerns were ever expressed (the reference category). We looked at how responses differed by organization (Figure 11 and Figure 13) and level of experience (Figure 12 and Figure 14).

For both variables, we found largely the same results. There were no large differences found in the point estimate for the priors by survey organization, shown.

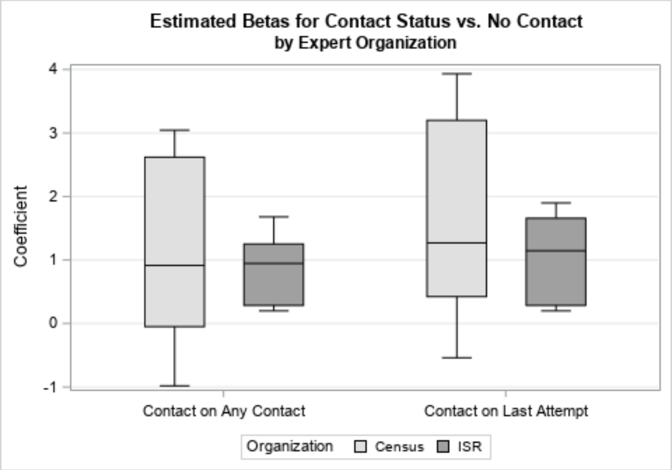


Figure 11. Coefficients for Contact by Organization

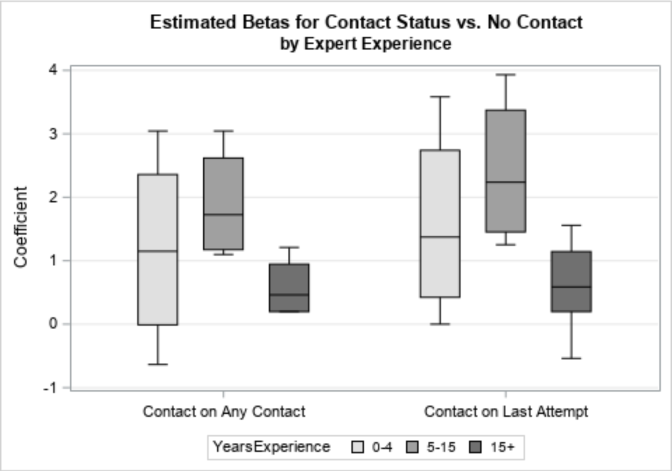


Figure 12. Coefficients for Contact by Experience

When examining the priors by level of experience (Figure 12 and Figure 14), interviewers with 0-4 or 5-10 years of experience generated similar point estimates for the betas, while experts with fifteen or more years of experience showed differences with respect to the point estimates. Specifically, experts with 15 or more years of experience appear to perceive, on average, that any one covariate has less of an impact on response propensity than do experts with less experience.

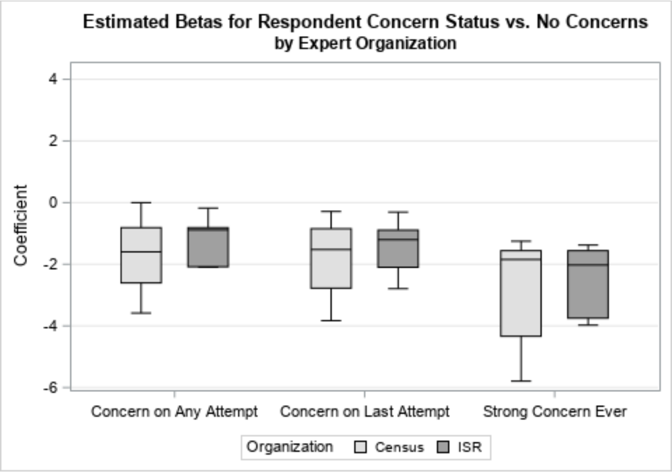


Figure 13. Coefficients for Concerns by Org

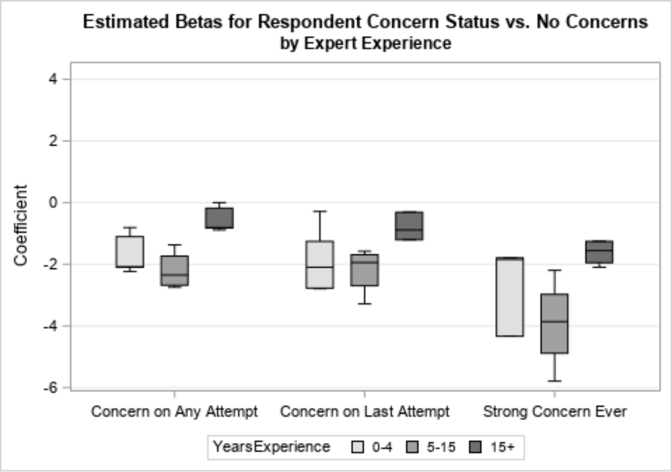


Figure 14. Coefficients for Concerns by Experience

Other questionnaire items showed more clear differences between the survey organizations. Figure 15 shows the effect of various types of listing procedures on response propensity, versus listing alone on foot. Here, there are not only differences in the means by survey organization, particularly for listing in a car with another person and on foot with another person, the means are in the opposite directions from the reference category, and the Census Bureau estimates are highly variable compared to estimates from ISR. In this particular case, feedback showed that Census Bureau experts did not see a link between listing method and response propensity, resulting in highly variable responses. We discuss the additional expert feedback that we received on the survey later in Section 3.5.

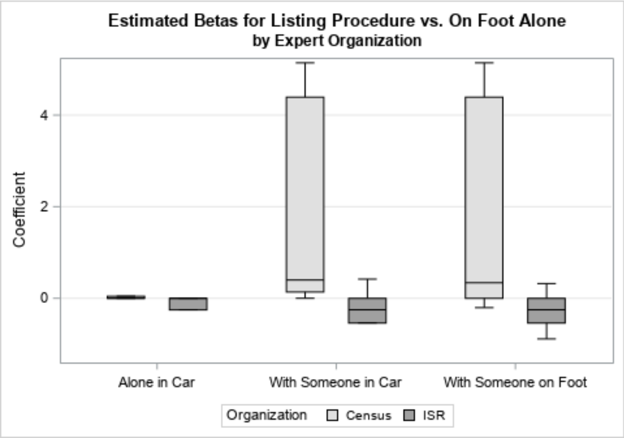


Figure 15. Estimated Betas for Listing Procedure by Organization

Figure 16 below displays the distributions of the betas by survey organization for the effect of evidence of a language other than English being spoken at home. Here, Census Bureau experts feel that evidence has a more negative effect on response propensity than ISR experts do. This may have to do with differences in the availability of bilingual interviewers or language specialists.

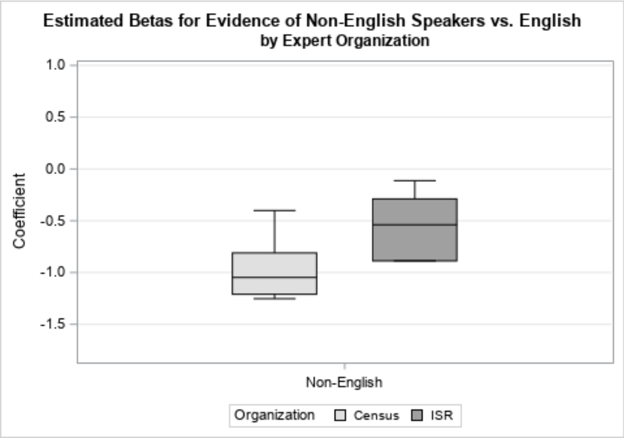


Figure 16. Estimated Betas for Likely Non-English Speaker by Organization

Understanding these similarities and differences is important for selecting the most appropriate experts to interview. Depending on the survey of interest, it might be more important to select interviewers with specific skill sets, such as language specialties. It may also affect which questions are included on the questionnaire, or which priors are actually used in the prediction model. In the case of listing procedure, the feedback obtained might suggest ignoring the prior information for some or all of the experts, and either using an uninformative prior or dropping the variable from the model.

3.4.2 Comparison of Methods

For each quarter, we treated the final prediction of response propensity, based on all accumulated contact data for the quarter, as the unbiased "target" prediction of response propensity. For each method, we then generate daily estimates of bias and RMSE with respect to the target prediction. Figure 17 through Figure 22display the performance of the Bayesian method using expert elicitation (EXPERT) to the current data-only method (Standard) and the precision-weighted prior Bayesian method (PWP) from West et al. (2019) that incorporates historical data. Our primary interest was to evaluate whether predictions generated using priors derived from expert opinion would be of higher quality than those generated using current data only, assuming historical data were not available for use. However, we were also interested in how the priors from expert opinion perform versus priors from historical data, which were evaluated in West et al. (2019). Because this was a retrospective analysis, we were able to examine both of these questions. Figure 17, Figure 19, and Figure 21 present the summarized distributions of estimated bias, while Figure 18, Figure 20, and Figure 22 present the summarized distributions of estimated RMSE.

Figure 17 and Figure 18 focus on the early portion of data collection, from day 7 through day 30 (24 days). For each quarter, the 24 daily estimates of bias (Figure 17) or RMSE (Figure 18) were summarized using box plots. Early in data collection, the expert elicitation (EXPERT) method has a small but inconsistent effect on the bias and RMSE versus the standard method. For example, in quarters 19 and 20, the EXPERT method results in mean, median, and intraquartile ranges of both the bias and RMSE of the predictions that are slightly closer to zero than the Standard method, signifying an improvement. However, in quarter 16, the EXPERT method performs worse than the Standard method with respect to the mean and median values of bias and RMSE, and delivers no improvement in quarter 17. Overall, however, neither the PWP nor the EXPERT method offer consistent improvement over the Standard method early in data collection.

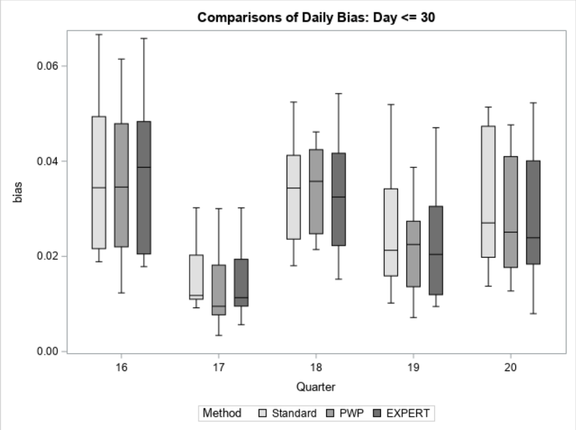


Figure 17. Bias in Propensities by Quarter (Early)

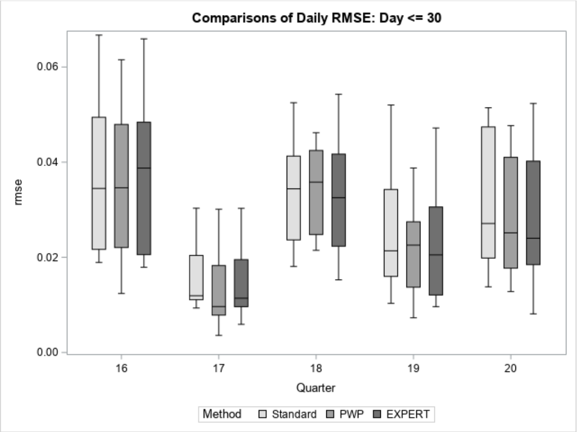


Figure 18. RMSE of Propensities by Quarter (Early)

Figure 19 and Figure 20 below represent the middle portion of data collection from day 31 to day 60. Beginning on day 31, there are noticeable reductions in the bias and RMSE of predictions for the EXPERT method. In all five quarters, the central tendencies of both the bias and the RMSE, as well as the intraquartile range, are shifted towards zero versus the Standard method. Further, in quarter 19, neither of the metrics have interquartile ranges that overlap between the Standard and EXPERT methods. For the most part, the PWP method continues to perform at least as well as the EXPERT method on measures of bias and RMSE, though the EXPERT method is certainly competitive, particularly in quarters 18 and 20. Here, unlike in the early portion of data collection, there is a clear benefit to using priors from expert elicitation if historical data are not available.

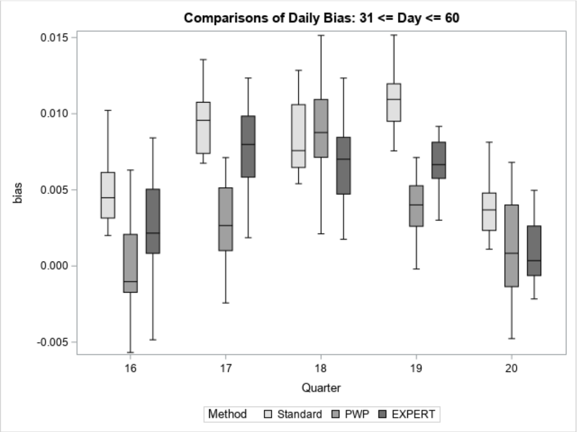


Figure 19. Bias in Propensities by Quarter (Mid)

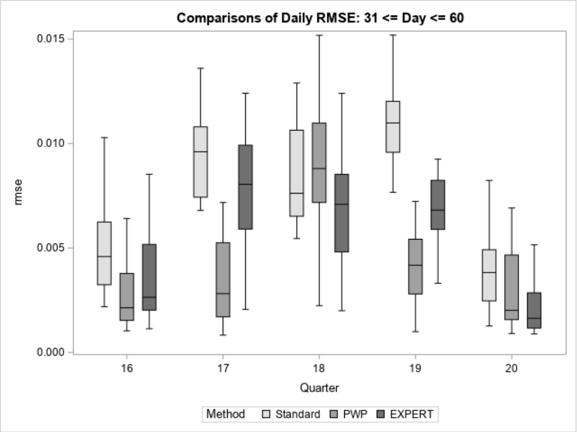


Figure 20. RMSE of Propensities by Quarter (Mid)

During the final third of data collection, shown below in Figure 21 and Figure 22, we continue to see that the EXPERT method leads to reduced measures of bias and RMSE versus the Standard method. These improvements are generally smaller than those found in Figure 19 and Figure 20. Over the course of data collection, as more data are accumulated, it is likely that the Standard method improves in its ability to predict response, leading to smaller differences between the Bayesian methods and the Standard method. Additionally, it is more mixed as to whether the historical method or the expert opinion method is superior.

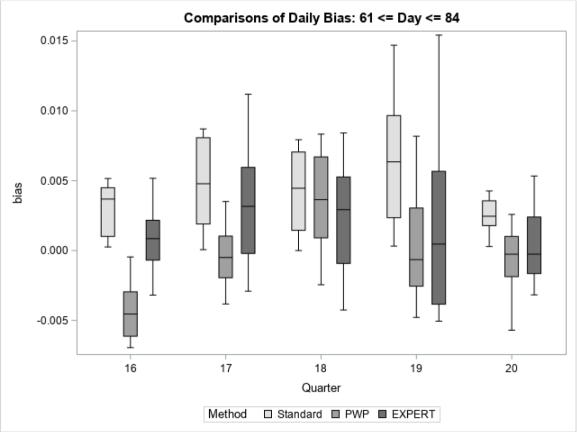


Figure 21. Bias in Propensities by Quarter (Late)

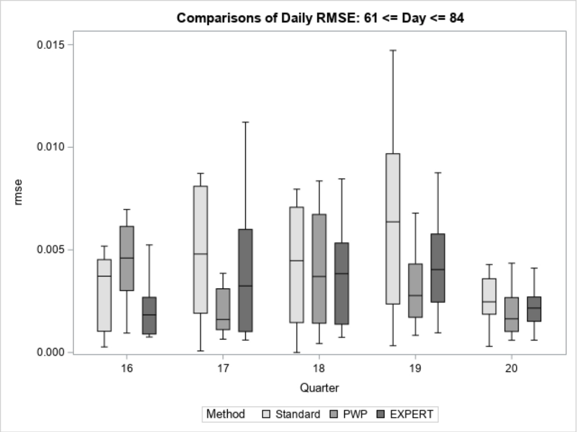


Figure 22. RMSE of Propensities by Quarter (Late)

These results show that for this application, the PWP method results in the most consistent improvements in bias and RMSE of predictions of response propensity. However, the results also show that, in the absence of historical information, predictions that incorporate expert opinion still generally outperform the standard method, and can be a useful way to improve predictions of response propensity during data collection for the purposes of an RSD.

3.5 Feedback from Survey Experts on Prior Questionnaire Development

Within two weeks of receiving questionnaire responses, we elicited feedback from experts in order to uncover issues with the questionnaire and identify potential areas for improvement. The experts had feedback in three main areas: the concepts identified in the questionnaire, how those concepts were translated into variables and categorical subgroups, and the lack of anchor points throughout the questionnaire.

The design of the questionnaire was driven by the variables available from the frame or from paradata. However, the concepts measured in the questionnaire did not always match concepts considered by the recruited experts. In our questionnaire, the experts provided two examples of this issue. In one instance, the predictive covariates from existing data sources were not meaningful concepts for survey managers. Mail Delivery Point Type is a categorical variable providing information on how mail is delivered to an address. This variable comes from the commercially available data and has several different categories that were significant in the variable selection model discussed in Section 3.2. However, when we included this variable (and all significant categories) on the expert questionnaire, only three out of 20 survey managers responded for any of the categories. During debriefing, survey managers explained that they did not have any experiential evidence that there was a relationship between response propensity and mail delivery. As a result, the survey managers generally declined to provide information for this concept.

On the other hand, survey managers explained that they do make use of concepts that were not included on the questionnaire. When providing feedback, one survey manager from the Census Bureau mentioned “perceived safety in a neighborhood” as a predictor of response propensity. In this case, this category was not included on the questionnaire because it was not a significant predictor in the response propensity model described in Section 3.2. It may be worthwhile to elicit information about predictors suggested by field experts, in order to capture information about predictors the experts find informative or predictive. This would allow confirmation that those particular items do not offer more explanatory power than the items retained from the propensity model.

In addition to defining meaningful concepts, it was also important to translate each concept into a variable that generated informative predictions, to the extent possible. This included determining whether a variable should be categorical or continuous, and, if categorical, how to define subgroups. Again, we found two clear examples of this issue. First, there were some instances where the categories that we provided in the expert questionnaire were not the same as those in the baseline model. As an example, age of householder, sourced from the sampling frame, was defined in the current model as having four categories: 18 - 44; 45 - 59; 60+; and Missing. In the questionnaire, we only included three categories to simplify the response options: Under 50; 50+; and Missing. Age of the householder is provided on the sampling frame as a continuous variable, so in this instance, the different classifications posed no issues for generating predictions of response propensity. However, if the questionnaire included categories that were not able to be derived from the existing frame or paradata, the priors derived from expert information would not easily translate to covariates in the existing data.

The survey experts also suggested that the functional form of some of our variables was not ideal. For example, on the questionnaire, we asked the experts to predict the change in attempt-level response rates for every $10,000 increase in household income over the median. At least one expert suggested that the relationship was likely not linear, and a better way to elicit opinion might be categorical, such as using quartiles of household income. This would better represent what the experts suggested, which was that the top and bottom quartiles of household income would have a lower attempt-level response rate than those in the middle two quartiles.

The experts also provided feedback regarding anchor points. In designing the questionnaire, we made a conscious decision to only include the overall attempt-level response rate, 24%, in the introduction, leaving it up to respondents to generate all subgroup level response rates. This was primarily to avoid generating anchoring bias among the survey expert responses. However, while survey managers were comfortable ordering different subgroups of a variable, from highest to lowest predicted response rates, and even defining relative differences, they were less comfortable defining an initial response rate for one category, in order to then provide response rates that reflected the subgroup ordering and relative differences. We found evidence of this in the response data itself. Survey managers provided responses for nearly all questions, but on occasion, the predicted response rate ranges varied significantly (e.g., one manager might have all subgroup response rates in a range of 20% to 40%, while another would provide responses in a range of 60% or 80%). One survey manager suggested providing an anchor point for one subgroup in the categorical variable, from which they could then provide the relative differences for the remainder of the subgroups. We provided an overall anchoring point in order to facilitate estimates of effect levels. The 24% value acts as an “intercept” attempt-level response rate, from which specific categories of the questionnaire deviate. However, we did not provide any category-level anchor points in an effort to avoid anchoring bias. There was a concern that if we provided the overall attempt level response rate (24%) *in addition to* an anchor point for one of the categories, the experts would focus on the relationships between categorical response rates and the overall response rates. For example, had we provided the 24% overall attempt-level response rate, and a response rate of 35% for female respondents, the expert may ignore their own expertise to provide a response rate around 13% in order to have the categorical response rates roughly match the overall attempt-level response rate. Our goal was to provide the minimum necessary amount of background information to allow the experts to use their own judgement to the fullest extent possible.

3.6 Discussion

We hypothesized that in the absence of historical survey data, survey researchers would be able to generate priors from the experiences of survey managers that lead to improved predictions of response propensity over those made from just the data available for the current round of data collection. The results of this study demonstrate that eliciting expert opinion is a useful way to generate priors and improve prediction of response propensities. Particularly after the first month of the NSFG data collection process, priors generated from expert opinion resulted in predictions of next-contact response propensity with both lower bias and RMSE than predictions based on only current round data. One potential explanation for why the Bayesian methods did not improve the predictions in the first month of data collection is that the early experience in any quarter is highly variable. That is, in Bayesian terms, the likelihood varies from quarter to quarter in the first few weeks. The observed data are somewhat more stable after 30 days, but do not normally align with the final model until near 60 days into the quarter. Hence, it is during that interval – i.e. after the first 30 days but before the 60th day of the quarter – that the prior information is most useful.

This prior elicitation process is significantly more involved than building models from existing historical data. Developing a questionnaire, conducting data collection with survey experts, aggregating and organizing the response data, and generating priors may be time consuming, particularly as the number of covariates increases. As a result, eliciting expert opinion for generating priors may not always be the ideal solution. In our experience, the large majority of the time and effort was spent on the initial development of the questionnaire. We would expect changes, adaptations, and future implementations to require much less effort. Experts themselves spent, on average, less than an hour on the actual survey. Assuming a pay rate of $50 per hour, the actual elicitation portion of the survey would cost roughly $1,000. We can imagine numerous applications where this type of expenditure would be worth this cost, as in the case where a new survey has a specific target population that may not have coefficients well-estimated by the published literature. Further, this method may be useful for mathematically incorporating expert opinion into predictions of response rates for budgetary purposes, sample sizes, and power calculations. Given the high costs of face-to-face data collection, improved response propensity predictions may help data collection managers make better decisions in an adaptive or responsive design framework. Evaluating of the ability of predictions based on such an approach to improve data collection outcome is an interesting direction for future research. We are currently pursuing experimental work in this area.

Through the process of designing and implementing the questionnaire, debriefing the survey managers, and analyzing the collected data, we identified four areas survey researchers should consider when developing and implementing expert elicitation surveys. These areas include the selection of concepts for inclusion into the survey; the translation of those concepts into covariates and/or categories; the potential need for anchor points for categorical covariates; and lastly, the selection of experts for the survey. Attention to these areas will lead to information from experts that is more helpful for generating priors, which are ultimately combined with current data to generate posterior predictions of response propensity.

For this particular questionnaire, through debriefings and response analysis, we observed several opportunities for improvement in the design process for expert surveys. Mindful selection of concepts and the subsequent translation of categorical variables will help experts provide more informative prior expectations. By working with experts to determine which data fields on the frame and in the paradata effectively translate to concepts used by survey managers, the value of the elicited information may increase. Additionally, it may uncover concepts used by survey managers when developing *ad hoc* expectations for response propensities that are not currently provided by data systems. There may be an opportunity then for expert opinion to motivate a modification of existing systems, either by appending an additional piece of information from the survey frame (if available), or capturing this concept in paradata, potentially through interviewer observations.

In order for experts to provide opinions on attempt level response rates for a survey, particularly when they are unfamiliar with the exact topic questionnaire, it may be helpful to provide context to the survey managers about general attempt-level response rates, or even provide an anchor point for one category of a variable. Providing an anchor point for a particular subgroup may be a reasonable solution to this issue, but it may increase anchoring bias in the remainder of the experts' responses. Additionally, in the case of categorical covariates in a logistic regression, it may not be absolutely critical. Generating priors requires constructing odds ratios, using one subgroup as a reference category. Because of this, odds ratios focus on the relative difference between a category of interest and a baseline category more than point estimates of response propensities provided by the survey managers. As a result, if the ordering and relative differences are accurate, that may be sufficient for generating relatively useful priors.

Associated with this is the fact that continuous variable were queried about on a linear scale, while the logistic regression modeling assumes a log-odds scale. For categorical variables this transformation is straightforward, since there is only a fixed set of options for the categorical variable to take; for continuous covariates, however, extrapolations outside of the specific values considered lead to different predictions. Thus, if an expert suggests that an additional contact attempt increasing the probability of a successful contact from 5% from a 24% baseline, this yields a beta parameter of 0.26; thus five contact attempt increases the odds of contact by $e^{(5*0.26)} = 3.67$, to 54%, instead of the 49% on the linear scale, and at 81% after transformation from the log-odds scale for 10 contact attempts, vs. 74% on the original linear scale. Hossack, Hayes and Barry (2017) have proposed eliciting priors at a series of quantiles of the continuous predictor values in order to better approximate the log-odds transformation; we leave this as a future extension.

An iterative process to address these issues is difficult to carry out without collaboration with the targeted experts and may not be possible in all situations. However, if it is possible to first validate a questionnaire with some experts, keeping in mind the potential biases like overconfidence and anchoring biases, the resulting questionnaire may have more predictive power. Similarly, the SHELF method, proposed by O’Hagan (2019) relies on a significant amount of interaction with the experts throughout the elicitation process in order to elicit a probability distribution form each expert. While this method can be highly informative, providing both a point estimate and a measure of uncertainty for each expert’s opinion, the number of items in our questionnaire would not have allowed for this level of individual interaction.

We also used the variability in the point estimates across our sample of experts to determine the variability in the prior distribution. This simplified the task of constructing the prior, since the experts were required only to supply point estimates, not estimates of uncertainty. This required a relatively large sample size of experts compared to many such elicitation studies. It also allowed us to take advantage of the Central Limit Theorem to utilize a normally-distributed prior, which in turn allowed more direct comparisons with West et al. (2019); alternatively, more heavy-tailed priors (e.g., t-distributions with small degrees of freedom) could be used. We did not rescale the prior to account for this sample size; one could construct a prior based on a “pseudo-sample size” of m by multiplying $SE\left(\hat{\bar{\beta}}_{jk}\right)$ in (4) by $\sqrt{n/m}$ (that is, standard deviation of the arithmetic mean by the square root of m rather than the square root of the actual number of respondents). Alternatively, one could elicit estimates of uncertainty as well as point estimates from the expert sample, and use information for both the direct elicitation and the sampling variability to construct the variance of the prior; we leave this to future research.

A limitation of our approach is that we used historical data to determine the key covariates to include in our survey of experts. We did this in order to make a fair comparison with historical data in our analysis, but in practice one might at best have data available from other studies with greater or lesser degrees of similarity. Indeed, one might have no historical data whatsoever from which to build a propensity model, in which case one would have to rely on experts' opinion about potentially predictive items to develop an effective model for response propensity. As noted in Section 5, querying experts for the key covariates may have advantages over model selection, even if historical data is available from similar studies.

Finally, it is important to elicit expert opinion from appropriate individuals, based on the survey characteristics. Experts at ISR were identified through discussions with survey managers to identify appropriate individuals. At the Census Bureau, we worked with senior leadership in the Field Directorate to identify the two "most knowledgeable" survey managers in each of the six regional offices. This provided geographic coverage over the entire country and, we hoped, significant experience in demographic surveys that could be translated into priors for response propensity prediction. We did not include any other requirements in our identification of survey managers for interview. After collecting responses, we found that survey experience ranged anywhere from '0-4 years' to '15 or more years', and we found potential correlations between experience and predictions of attempt-level response rates predictions for some covariates. Due to the small sample size, we cannot conclude that these correlations are meaningful. However, it is useful to consider whether additional requirements would be useful when identifying experts. Relevant experience, either with respect to survey topic (e.g., health, education, etc.), operations (e.g., multimode vs. in-person interviewer-administered), or other characteristics, may lead to more informative expert opinion for incorporating into priors.

4. Optimizing Data Collection Interventions to Balance Cost and Quality Under a Bayesian Framework

Stephanie Coffey[1], Michael R. Elliott[2,3]

[1] Joint Program in Survey Methodology, University of Maryland
United States Census Bureau, Washington, DC
[2] Survey Research Center, Institute for Social Research, University of Michigan
[3] Department of Biostatistics, University of Michigan

Abstract

All aspects of a survey design, from the length of the survey period, to the mode of data collection, to individual data collection features like incentives or mailings, will affect both who responds to a survey and how much it costs to obtain their response. In order to conduct data collection successfully in a budget-conscious environment, decisions related to survey design require balancing concepts of data quality and costs. Recently, responsive survey designs have emerged as a way to tailor data collection features to specific subgroups or specific cases within a data collection period in order to save costs and/or improve survey outcomes. For the most part, however, responsive designs in the survey methodological literature do not incorporate actual survey response data into their decision framework. Here, we report on a responsive design experiment in the National Survey of College Graduates that incorporates optimization as a way to minimize data collection costs for a small increase in the root mean squared error of a key survey estimate. We demonstrate both the benefits of Bayesian methodology for the optimization problem as well as the ability to incorporate optimization during live data collection. Results include a comparison of the data collection costs and RMSE of a survey estimate in the experimental treatment group to a control group that follows the standard NSCG data collection pathway.

4.1 Introduction

Survey design requires balancing concepts of data quality and costs, and understanding the cost and quality properties of different data collection features is critical for managing survey operations and for making improvements in future survey rounds. Information about how successful a particular data collection feature is at yielding response, or how the historical response behavior of sample cases predicts their future behavior, can be used to better estimate survey data collection outcomes, like expected costs and response rates. This, in turn, can offer insight on how to offer or mix modes, how to target incentives, and generally how to adapt data collection features to meet survey cost and quality goals.

Developing expectations regarding the performance of different response modes can help determine which modes to offer in a survey, and additionally, how to order those modes. Survey organizations choose to mix modes in order to balance competing priorities and attract different types of respondents. De Leeuw (2005) provides an overview of mixed- or multimode surveys, and describes their use as the “opportunity to compensate for the weaknesses of each individual mode at affordable cost”. Survey operations in a multimode survey may vary the types of contacts, the response options provided, or both, and the different modes may be offered simultaneously or in a sequential order. As de Leeuw (2005) points out, multimode surveys have most commonly been used to reduce nonresponse error. Cost savings gained by using less expensive modes may allow for a longer data collection period. Alternatively, different modes may appeal to different sample persons, and so offering multiple modes may reduce the selection bias that results from offering a single mode. Multimode surveys still follow a standardized data collection protocol for all cases, but offer a more diverse set of contact and response options within that standardized protocol.

Adaptive and responsive survey designs take this framework further by relaxing the traditional standardization in order to adapt data collection features to sample members in pursuit of some data collection goal. Schouten et al. (2017) discuss adaptation as a concept, pointing out that sample members are different. This results in different preferences for completing a survey request, and may require different features to elicit that cooperation. As a result, predictions of data collection characteristics such as response propensity (West et al. 2020), response mode (Schouten et al. 2018), and costs (Wagner, West and Elliott 2020) are critical for making interventions in adaptive and responsive designs.

Typically, metrics such as response rate and response propensity are used as proxies for data quality when making data collection interventions. This leads to interventions where the highest response propensity cases are prioritized to increase response rates (Peytchev et al. 2010); cases with the largest base weights are prioritized in order to decrease the sizes and variability of nonresponse adjustments (Wagner et al. 2012); or extra effort is applied to under-represented cases to improve balance (Coffey, Reist and Miller 2019) , among other examples. However, these methods ignore the impact each case might have on actual survey estimates.

Ideally, survey methodologists would know whether a particular sample member would respond to a particular data collection pathway or feature *(response propensity)*, the data collection resources associated with that pathway or feature *(cost)*, and the ability of the sample member to impact the quality of the information collected in the survey *(response data)*. If this information were known, sample members could be assigned to the optimal data collection pathway to balance costs and quality. For example, if some sample members would respond to inexpensive pathways, this allows other sample members to be assigned to more expensive pathways, potentially increasing response rates, and decreasing the variance of estimates generated from survey response data while staying within the survey budget. Alternatively, if the assignment of all sample members to their most effective pathways for data quality would exceed the survey budget, some sample units could be assigned to cheaper data collection modes based on how impactful a sample member would be on the survey data.

These parameters of interest are not known prior to data collection, however, so we must rely on predictions of these parameters to effectively allocate sample members to different data collection pathways. Schouten et al. (2018) simulate how the data collection pathway assignment for cases could be generated *prior* to the start of data collection, or in-between rounds of data collection for a survey that is in the field regularly. The authors utilize a Bayesian framework to incorporate historical information into predictions of response propensity, costs, and survey response data in order to determine whether cases should be assigned to a self-response internet mode, an interviewer-assisted face-to-face mode, or a combination. In the authors' examples, the predictive models based on historical data are used to define fixed business rules that assign cases to treatment paths.

It is also possible to make decisions *during* the data collection period, which would allow optimization based not only on the information obtained from prior rounds of data collection or external sources, as in Schouten et al. (2018), but also on updated historical expectations given accumulating information from the current round. Paiva and Reiter (2017) demonstrate, through simulation, the implementation of stopping rules based on accumulating survey response data. Adapting or optimizing data collection pathways during a data collection period can be useful when a survey has a long data collection period, has natural decision points like mode changes, or when there is enough time or change between implementations of a survey that the predictive power of models based only on historical data suffers. However, in order to intervene in an effective way, it is important that the predictions upon which the interventions are made are as accurate as possible. West, Wagner, Coffey and Elliott (2020) and Coffey, West, Wagner and Elliott (2020) show that, when data collection interventions are carried out during a data collection period in a responsive or adaptive design, a Bayesian framework can improve predictions by statistically leveraging both historical data and accumulating data from the current round of data collection.

This manuscript discusses a dynamic adaptive design experiment in the 2019 National Survey of College Graduates (NSCG), where data collection interventions were carried out in order to minimize costs while avoiding a large increase in the root mean squared error of a key survey estimate, the mean respondent-reported salary. Throughout the 2019 data collection period, we generated predictions for overall response propensity, response propensity by phase, data collection costs, and salary for all non-responding cases. We used these predictions at natural points in data collection, when more expensive modes were being introduced, in order identify a subset of nonresponding cases to not receive the new mode. The selection of cases was based on their overall response propensity and predicted value of salary. The goal of this experiment was to demonstrate that interventions could be applied during data collection that would reduce data collection costs without reducing the quality of survey response data.

The remainder of this manuscript is organized as follows. Section 4.2 describes the National Survey of College Graduates and includes information on the sample design, the data available for this application, data collection operations, and past experiences with adaptive and responsive designs. Section 4.3 discusses the predictive models for the survey parameters of interest, including overall response propensity, respondent-reported salary, and data collection costs. We demonstrate, using historical data, that generating predictions under a Bayesian framework results in less bias and error in predictions of survey parameters of interest, such as response propensities and survey responses, particularly in the early and middle parts of the NSCG, when interventions are available and more likely to have an impact on costs. These results add to the recent evidence that Bayesian methods for prediction perform better than methods relying on either only historical survey data, or only current, accumulating data (West et al. 2019; Coffey et al. 2020). Section 4.4 explains the structure of our data collection experiment, including the experimental design, the planned intervention points, the steps of optimizing the selection of cases for the intervention, and the evaluation methods for the experiment. Section 4.5 summarizes each of the three intervention points. Section 4.6 details the results of the experiment, and the manuscript ends with some discussion and directions for future work in Section 4.7.

4.2 Description of the Data

4.2.1 National Survey of College Graduates

The National Survey of College Graduates (NSCG) is a longitudinal survey that collects information on the employment, educational attainment, and demographic characteristics of the college-educated population in the United States with a focus on those educated or employed in a science or engineering field. The National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF) conducts the NSCG with the U.S. Census Bureau serving as the data collection contractor. In this data collection contractor role (and with active NCSES input), the U.S. Census Bureau designs and implements the data collection features and schedule for the survey, as well as any embedded experiments.

The target population of the NSCG covers non-institutionalized individuals under the age of 76 with at least a bachelor's degree who reside in the United States, as of the reference date for a given survey implementation. In 2010, the NSCG began implementing a rotating panel design with sample drawn from respondents to the American Community Survey (ACS), and past rounds of the NSCG. This means that for the 2015 NSCG, the new sample was drawn from the 2013 ACS respondents, and returning sample from the 2013 NSCG included respondents to the 2011 and 2009 ACS that met the above-defined eligibility requirements (Fecso, Frase and Kannankutty 2012).

The sample size for any given round of the NSCG is approximately 124,000 individuals, of which approximately 40% is new sample. The NSCG uses a stratified sampling design, and stratification cells are defined by demographic group, highest degree type, field of occupation, and field of degree. These classifications are derived from responses to the ACS for the new sample, and from responses to past NSCG interviews for returning sample. Within stratification cells, a combination of probability proportional to size or systematic random sampling (depending on the stratification cell) are used to select the NSCG sample.

As NSCG cases are selected from ACS respondents, they inherit the ACS final person-level weights as their initial "population weight" (Hall, Cohen, Finamore and Lan 2011; NCSES 2015). Those weights are then adjusted to account for sample selection in the NSCG to obtain the NSCG base weights. The variability of weights from the ACS, in part due to subsampling in the ACS (U.S. Census Bureau 2014, pp 14-15), coupled with differential sampling within stratification cells leads to more variable weights than a typical national survey, which typically use a two-stage self-weighting design. (U.S. Bureau of Labor Statistics and U.S. Census Bureau 2006, Chapter 3). Additionally, cases in the NSCG are assigned replicate weights prior to data collection that are then subject to adjustments including nonresponse and raking in order to obtain a set of final replicate weights that can be used for point and variance estimation (White and Opsomer 2011, 2012; Opsomer, Breidt, White and Li 2016). New cases, set to receive their first interview, are referred to throughout this document as "New Cohort". Returning cases, who have been interviewed at least once, are referred to as "Old Cohort". For example, the 2015 NSCG New Cohort was sampled out of the 2013 ACS, and will be considered Old Cohort during 2017, 2019 and 2021. New cohort nonrespondents are dropped after the first interview period and are not included in the old cohort. This means that all old cohort cases have responded at least once to the NSCG. The adaptive design experiment in this manuscript only includes the New Cohort.

4.2.2 NSCG Data Collection Schedule

The data collection schedule in the NSCG follows a sequential multimode design and runs for approximately 26 weeks every other year. Data are collected using one of three modes: self-administered web interviews, mail questionnaires, and telephone interviews. Prior research (Finamore and Dillman 2013) on mode ordering and response mode found that NSCG sample members are most likely to respond via the mode offered initially, which led to a mode sequencing most sensitive to cost.

A “web push” phase lasts for the first eight weeks of data collection. During this time, invitations are mailed with a username and password requesting the sample person to respond via the web instrument. In week 8, a paper questionnaire is introduced for nonrespondents who were unwilling or unable to respond by web. In week 12, remaining nonresponding cases are sent to a Computer Assisted Telephone Interviewing (CATI) operation for nonresponse follow-up (NRFU).

Pre-defined “influential” cases, or those that have a high weighted response influence (Sarndal and Lundstrom 2008; Coffey and Reist 2014), also receive a prepaid $30 incentive card with their Week 1 web invitation. There are other operations throughout data collection, including automated telephone reminders, email reminders, and postcard or letter mailed reminders. Additionally, there are exceptions to the traditional data collection pathway; for example, cases could call into the Telephone Questionnaire Assistance (TQA) line and request a paper questionnaire, or respond to the survey with an interviewer on the spot. Despite these exceptions, the vast majority of cases follow the traditional data collection pathway. Figure 23 displays the main characteristics of the typical NSCG data collection design (with reminder contacts noted with an “R”).

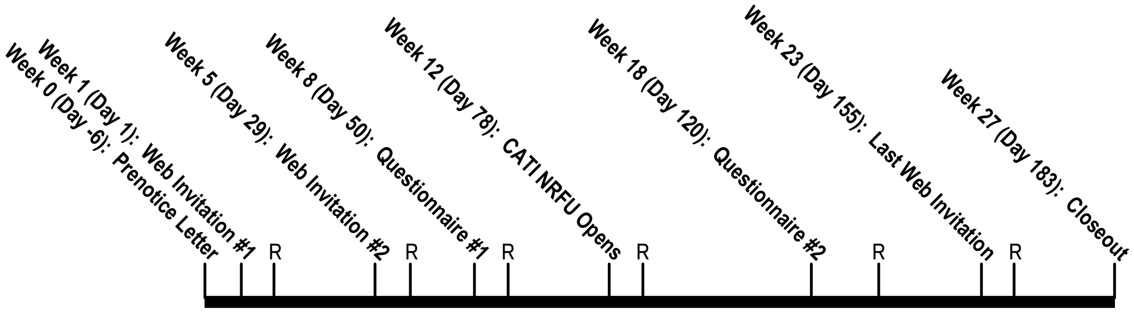

Week 0 (Day -6): Prenotice Letter
Week 1 (Day 1): Web Invitation #1
R
Week 5 (Day 29): Web Invitation #2
R
Week 8 (Day 50): Questionnaire #1
R
Week 12 (Day 78): CATI NRFU Opens
R
Week 18 (Day 120): Questionnaire #2
R
Week 23 (Day 155): Last Web Invitation
R
Week 27 (Day 183): Closeout

Figure 23. Major Scheduled Operations During the NSCG

For the purposes of this paper, we can classify these operations into four data collection phases, as shown in Table 8 below. As data collection progresses, the variety of modes available to the majority of the sample population is increased.

Table 8. Contact Phases in the NSCG

Phase	**Primary Modes**	**Weeks**	**Days**
1. Web Push Phase	Web	0 – 7	-6 – 49
2. Mail Questionnaire Phase	Web, Mail	8 - 11	50 – 77
3. Telephone Follow-up Phase	Web, Mail, CATI	12 – 17	78 – 119
4. Late Follow-Up Phase	Web, Mail, CATI	18 – 26	120 – 182

It is expected, then, that the distribution of remaining sample members responding by particular modes will change as modes like mail or telephone are more widely introduced. Additionally, due to both the cumulative contacts, and the fact that modes introduced later are generally more expensive, the cost of earlier responders will be different than that for later responders.

4.2.3 Data Sources

Four sources of data were used for this application, in addition to the NSCG sample itself. First, because the NSCG is sampled out of the ACS, we were able to obtain response data to the ACS (Census 2019) at a point two years before a case entered the

NSCG. For a 2015 NSCG new cohort case, then, we have 2013 ACS response data. Some of the responses are used to carry out sampling for the NSCG, but we explored a large number of additional covariates as potential predictors of parameters of interest. We were also able to incorporate one piece of ACS paradata – the mode of response. The ACS, like the NSCG, has a sequential multimode design, and so response mode is an indicator of both resistance of the survey request and cost of obtaining the survey response.

Second, we incorporated several types of paradata from the NSCG. The NSCG does not have an in-person interviewer component, so the majority of the paradata is operational, e.g., counts of log-ins to the web instrument; dates and types of outbound mailings; date, time and outcome of outbound telephone calls; and records of any incoming assistance requests coming from sample members. Operational paradata also includes undeliverable mail information from the US Postal Service (USPS 2009).

Third, we have available to us the data collection costs of nearly all data collection features, including incentives, printing and mailing letters and questionnaires, post-data collection processing for paper questionnaires, and average costs for unproductive calls and telephone responses. These specific costs are provided by the survey operations team and allow us to estimate a case-level cost-per-case, rather than an aggregate average cost-per-case. Table 9 below shows an example list of contact and response types and their costs in the NSCG.

Table 9. Sample Costs per Case for Data Collection Features in the 2015 NSCG

Week	Type	Cost ($)
N/A	Fixed Development Costs	3.21
0	Prenotice Letter	0.87
1	Phase 1 Web Invitation Letter	0.93
1	Phase 1 Questionnaire	3.09
2	Reminder Letter	0.87
5	Reminder Letter	0.98
6	Reminder Letter	0.98
8	Phase 2 Web Invitation	0.97
8	Phase 2 Questionnaire Mailing	3.13
9	Reminder Postcard	0.67
12	Phase 3 Web Invitation (Start CATI)	0.78
13	Reminder Letter	0.77
16	Reminder Letter	1.26
18	Phase 4 Web Invitation Letter	6.69
18	Phase 4 Questionnaire Mailing	7.63
20	Phase 4 Reminder Letter	1.26
23	Phase 4 Reminder Letter	0.80
24	Final Letter	0.79
Several	Automated Telephone Reminder	0.10
Several	Email Reminder	0.01
Any	Paper Response	24.72
Any	CATI Nonresponse Attempt	2.07
Any	CATI Response	45.91
Any	TQA Response	31.82

Finally, we use the accumulating NSCG responses for predicting survey outcomes, as well as for evaluating those predictions against actual survey outcomes.

4.2.4 Adaptive and Responsive Design in the NSCG

Starting in 2013, the NSCG has incorporated adaptive survey design experiments into their data collection operations. The experiments have focused on different aspects of adaptive survey designs, starting with a proof of concept, maturing to a study on whether data quality metrics could be improved through dynamic interventions, and finally focusing on automation of interventions. From the data quality perspective, past NSCG adaptive design experiments have focused on improving representativeness, a measure of data quality (Coffey, Reist and Miller 2019).

While interventions were applied in support of this goal, the interventions were not made under any formal decision rule framework, nor did we prospectively predict the expected change in response or costs by implementing an intervention on a particular subset of cases. In other words, the cost savings and relatively stable response rates have largely "just worked out". However, it is possible that we could have obtained better results by maintaining a neutral cost per case, and intervening on more cases (or reducing contact effort on fewer cases), which could have further improved response rates. Alternatively, we could have attempted to minimize cost without decreasing representativeness, as a way to find data collection resources available for reallocation without hurting the target data quality metric. Additionally, there is also no guarantee that, in the future, continuing to intervene as we have in the past will result in the same outcomes of cost savings and stable response rates. Further, while representativeness is often considered a proxy for nonresponse bias, the interventions made historically do not take into account any expected effect on key estimates, such as the root mean squared error or the size of nonresponse adjustments.

In order to evaluate the effect of adaptive design on key survey estimates of the NSCG, while applying cost-efficient decision rules, we propose a dynamic adaptive design implementation under a Bayesian framework (Elliott 2017).

4.3 Predicting Survey Data Collection Parameters of Interest

In order to make effective intervention decisions, we would ideally like to know, for a given sample member in a data collection pathway, if they will respond, and if so, when during the survey they will respond, and how they will respond. This information would allow us to allocate sample members to different data collection pathways in order to minimize costs while obtaining a given measure of data quality (or alternatively, maximize a measure of data quality for a fixed cost). We do not know these pieces of information *a priori*, however, and so we must rely on predictions. Further, as we want to make interventions during the data collection period, we will be relying on interim predictions, meaning we are generating predictions of end-of-data-collection parameters with only partially accumulated data.

In this section, we define each survey data collection parameter necessary for our experiment. In the Introduction, we mentioned the need to predict *response propensity*, *response data*, and *cost*. We first discuss the methods by which we will evaluate the different prediction methods. For both response propensity and survey response data, we use historical NSCG data in order to compare three methods for estimating response propensity and the survey response data.

1) The first method generates predictions using only data from the current survey implementation as paradata and response data accumulate. Model coefficients used in a predictive model for a survey parameter of interest are estimated using the accumulated data up through time t, and those coefficients are used to generate predictions for non-responding cases at time t. This means that over time, as new data is accumulated for open cases, the estimates of the coefficients

(and the resulting estimated parameters) may change. We refer to this method as the *current method*.

2) The second method generates predictions using the complete set of paradata and response data from a prior implementation of the survey. Here, model coefficients are estimated using the full set of historical data, and are then applied to current round cases at each prediction time. In this case, the estimated coefficients are fixed, meaning that not only do the coefficients ignore any of the accumulating data during the current round, they also assume the values of the coeficients are constant. As a result, this method fails to account for uncertainty in the estimates of the coefficients. If all covariates in the predictive model are fixed, the predictions in the current round that are based on historical coefficients will never change. If on the other hand the predictive model includes time-varying covariates, a prediction for an individual case may change, despite the fixed coefficients, as the value of one of the covariates changes. We refer to this method as the *historical method*.

3) Finally, the third method is a statistical combination of the first two methods. First, coefficients used in a predictive model for a survey parameter of interest and their standard errors are estimated using a past implementation of the survey, as in the *historical* method. Those coefficients and their standard errors are then used to define prior distributions which are then updated by the accumulating data for the current month. The resulting posterior estimates of the model coefficients are then used to generate predictions of survey parameters of interest in the current round. We refer to this as the *Bayesian method*.

We used the 2017 NSCG as our evaluation period for these three methods. For the *historical* and *Bayesian* methods, historical model coefficients and standard errors were estimated from the 2015 NSCG. The cases eligible for inclusion in this evaluation were those from the 2015 and 2017 NSCG adaptive design experiments. We identified these cases for two main reasons. First, there are often several experiments embedded in the NSCG, and we wanted to estimate coefficients that were based on the standard NSCG data collection survey design. The adaptive design experiments included both treatment (NSCG Adaptive Design Treatment, or NADT) and control (NSCG Adaptive Design Control, or NADC) group. These samples were both selected to be representative of the full data collection population and are 8,000 cases each. Therefore, the NADC group would serve as the set of cases from which we could evaluate the predictive power of these three different methods on a sample of the NSCG that was not confounded with any other experiments.

For response propensity and survey response data, we used the 2015 NADC cases to conduct variable selection for a predictive model. We output and stored the estimated coefficients to use in both the *historical* and *Bayesian* prediction methods, as described above. Then, using the 2017 NADC, we generated predictions weekly or at the end of each phase for each of our parameters of interest during the 2017 data collection period, using the accumulating 2017 paradata and response data.

In order to evaluate the three methods, we compared the weekly case-level predictions generated by each method to the end-of-data-collection case-level “target estimate” for each parameter. That target estimate was generated using the complete set of 2017 paradata and response data at the end of data collection, after all information about the data collection period was known and available. In the case of response propensity, we used the final case-level estimates of response propensity in the 2017 NADC as the target estimate. Similarly, for the response data, we used the final case-level reported salary for respondents as the target estimate. We evaluated the three methods on measures of mean prediction bias (MPB) and root mean squared prediction error (RMSE) with respect to the target estimate.

The *historical method* only uses historical data, and for this method to be effective for prediction, the current data collection period needs to be very similar to the historical data collection period to be useful. This may be an unreasonable assumption for intermittent surveys, surveys where design changes may occur between rounds, or surveys that see large shifts in response behavior, such as a large decrease in contactability by phone, or a large increase in response by web. The *current method* only uses current round data, and therefore assumes that data accumulated early in data collection are representative of outcomes later in data collection. This could be especially problematic in a sequential multimode design, such as the NSCG, where some modes are not even available until late in data collection. Additionally, the *current method* ignores potentially useful historical information that could help create meaningful expectations.

The *Bayesian method* is, conceptually, a combination of the other two methods. A logical approach to leveraging information from past rounds of data collection, while also using information about the current round of data collection is to take advantage of a Bayesian modeling approach with informative priors. We can accomplish this by first fitting an appropriate predictive model using prior survey data to obtain parameter estimates and the associated variances. These parameters capture relationships between covariates and outcomes when accumulated data represents a full data collection period, though not the one of interest.

For response propensity and survey response data, we assume an approximately normal distribution for the prior parameters by the properties of maximum likelihood estimation. Means and variances for the prior parameters are estimated once from historical data, and so are fixed values. Once the priors for all coefficients are obtained, a predictive model is estimated periodically using the data available up to that day in the current data collection period, combined with the previously obtained prior. This way, historical information can be used to set baseline expectations for survey data collection outcomes, and current round information can be used to update those expectations. This ability to borrow strength from historical information but still reflect the current data collection reality is an important benefit when data collections are intermittent, and there may be changes in how sample members respond to the survey. Similar to the work carried out in Chapter 2 of this dissertation, we evaluated the difference in prediction quality when the priors were scaled to represent one-half of a data collection sample (by doubling the variance) versus a full data collection sample. The full data collection period was more successful at improving predictions of response propensity and survey data, and so we limited our discussion to the full data collection sample.

Our expectation is that the *Bayesian* method will produce predictions closer to the actual survey data collection outcomes, as the modeling procedure is effectively borrowing strength across the historical data and current data to make a prediction. The results in this section aim to support the use of the 2017 NADC group for generating priors for use in the 2019 NSCG adaptive design experiment.

4.3.1 End of Data Collection Response Propensity

In order to make effective interventions in data collection, it is important to understand whether a sample member will respond to a particular data collection pathway. For example, if one sample member is equally likely to respond to an inexpensive pathway and a more expensive pathway, assigning that sample member to the less expensive pathway would make resources available to assign a reluctant sample member to the more expensive pathway in order to obtain response. We therefore need to ensure that we are basing our interventions on reasonable estimates for response propensity.

We used the same methodology found in Wagner and Hubbard (2014), West et al. (2020), and Coffey et al. (2020) to estimate and evaluate response propensities. We started with the 2015 NADC cases, and fit a logistic regression model to the final binary response status (respondent or nonrespondent) for those cases, using a rich set of covariates from the frame and accumulated paradata. We used a backwards selection procedure, retaining all variables that had a p-value of at least 0.25. We took this inclusive view of predictors as we are concerned about out-of-sample prediction, and want to protect against eliminating variables from the model that may be predictive for some time periods but not others.

We removed three paradata items that were strongly associated with response but were not explanatory predictors so much as observations that are endogenous to the data collection process: the number of web logins, the number of mailings, and the number of outbound CATI calls. Nearly all sample members who access the web instrument finish the survey in the web. Additionally, they generally finish the survey on the first log-in. Therefore, while this variable was highly associated with response, it is not useful for prediction. The number of mailings and the number of outbound CATI attempts were also strongly associated with response. As data collection progresses, cases that remain non-responders are reluctant sample members, and generally have lower response propensities than cases that responded early. Additionally, in the NSCG, as data collection progresses, additional modes are introduced. Questionnaires are not available until at least week 8, and outbound telephone attempts in CATI do not start until week 12. The coefficients for number of mailings and number of outbound telephone attempts are negative, so as these contact types accumulate, response propensity decreases. However, it also means that prior to week 8, cases who *should* have lower estimated response propensities may have their propensities biased upward as these modes are not yet available. Table 10 displays the variables retained after this step.

Table 10. Retained Predictors for Response Propensity Model, 2015 NADC

Variable Name
Age Group
Demographic Group
Highest Degree Earned
Science and Engineering Degree Indicator
Full-Time/Part-Time Work
Veteran’s Status
Internet Access Type
Incentive Sent
ACS Response Mode
Contact Research Indicator
Refusal Indicator
Cumulative Web Logins
Cumulative Mailings
Cumulative Telephone Attempts

Both measures suggest a reasonable set of predictive covariates.Appendix D displays all predictors retained in the model, with their parameter estimates, standard errors, and significance. The in-sample ROC-AUC when including all binary predictors listed in Appendix D was 0.769, and the response rates by predicted probability percentile (using 2% wide cutoffs) are shown below in Figure 24. Both measures suggest a reasonable set of predictive covariates.

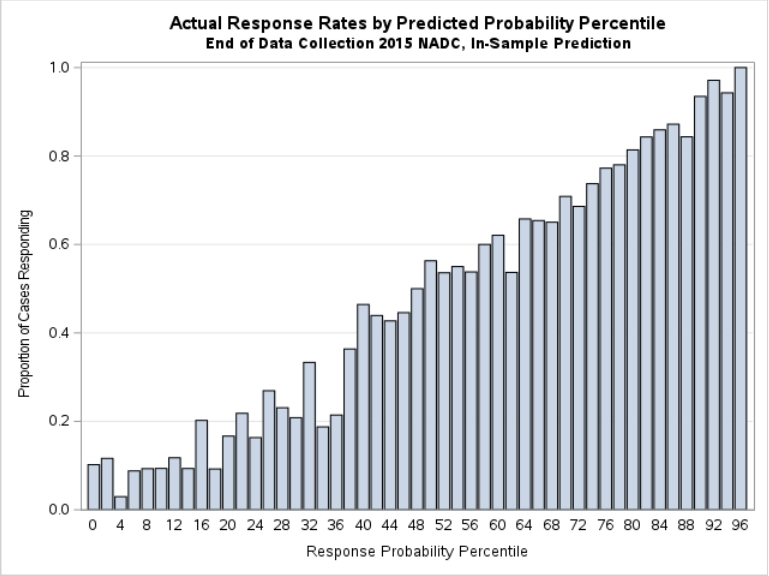


Figure 24. Response Rates by Predicted Probability Percentile, 2015 NADC, Simplified Model

We used the list of predictive covariates above as the list of predictors in the 2017 NADC. Each week, we used frame data, the most-up-to-date accumulated paradata, and response status to estimate coefficients in a logistic regression model. We then used these coefficients to estimate the likelihood that a particular case would respond to the NSCG by the end of data collection. These predictions were for the *current* method. While the list of covariates remained the same between 2015 and 2017, the models were refit weekly as new paradata accumulated. We used the last prediction from this method (after Week 26) as the "target" response propensity. Additionally, the parameters and standard errors from Appendix D were retained for use in the *historical* and *Bayesian* methods for predicting response in the 2017 NSCG.

In order to generate predictions using the *historical* method, we simply applied the coefficients estimated from the 2015 data to the weekly records in 2017. Here the variables and coefficients did not change, but the response propensities could change weekly as time-varying covariates (e.g., paradata) accumulated.

Lastly, we generated predictions of the final response propensity using the *Bayesian* method. Using the model coefficients that were estimated from the 2015 data, as well as their standard errors, and assuming approximate normality of the maximum likelihood estimates for regression parameters (Rao et al. 2008), we generated normal priors for each covariate in the predictive model, $N\sim(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance of the coefficient estimate. We ignored correlation in these parameters, treating them as independent. This is a limitation of this work and we return to this point in the Discussion. Using PROC MCMC in the SAS 9.4 programming language, we then generated posterior estimates for each model coefficient by taking 4,000 Monte Carlo simulations (20,000 thinned by 5) of the posterior distributions of the coefficients based on the priors and likelihood which had information up through week t. We then estimated the response propensities from each of those draws, and averaged them in order to arrive at a predicted response propensity for each case in each week.

In order to evaluate the three models against each other, we calculated the difference between the weekly case-level estimate for the final response propensity generated by each method, $\hat{\rho}_{it}^{m}$, and the target response propensity, ρ_i^T. We then used these estimates of prediction error to look at the weekly distribution of prediction error for all nonresponding cases during that week, as well as the weekly mean prediction bias (MPB) and RMSE. MPB averages the individual case-level errors described above over the number of open cases, n:

$$MPB_t^m = Bias(\hat{\bar{\rho}}_t^m) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\rho}_{it}^m - \rho_i^T) \quad ,$$

and the daily RMSE for the mth method is defined as:

$$RMSE_t^m = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\hat{\rho}_{it}^m - \rho_i^T)^2} \quad .$$

Figure 25 through Figure 27 compare boxplots of the error in predictions of final response propensity on a weekly basis for each of the three methods. Figure 25 displays the earliest part of data collection, weeks 1-6. For the first four weeks of data collection, the *historical* method results in central tendencies of the error that are closest to zero, while the *Bayesian* method outperforms the *current* method. However, starting in week 5, the *Bayesian* method has a median error near zero, and an intraquartile range that is smaller than the *current* method. This suggests that even as early as week 5, the *Bayesian* method can improve predictions of final response propensity by leveraging both historical and current survey data.

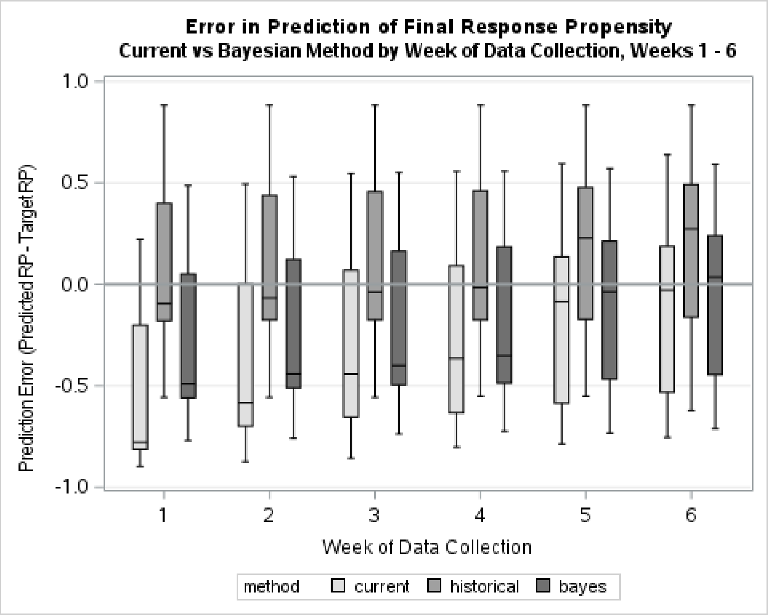


Figure 25. Error in Predictions of Final Response Propensity for Open Cases, Weeks 1-6

In fact, in historical rounds of the NSCG (Coffey, Reist and Miller 2019), week 6 is generally the earliest point in data collection when cases are identified for adaptive interventions, so these early improvements are valuable, in a practical sense.

Figure 26 depicts the middle portion of data collection, from week 7 through week 16. In the NSCG, this is when the bulk of contact attempts occur, and when the paper questionnaire and outbound telephone operations (CATI) are introduced.

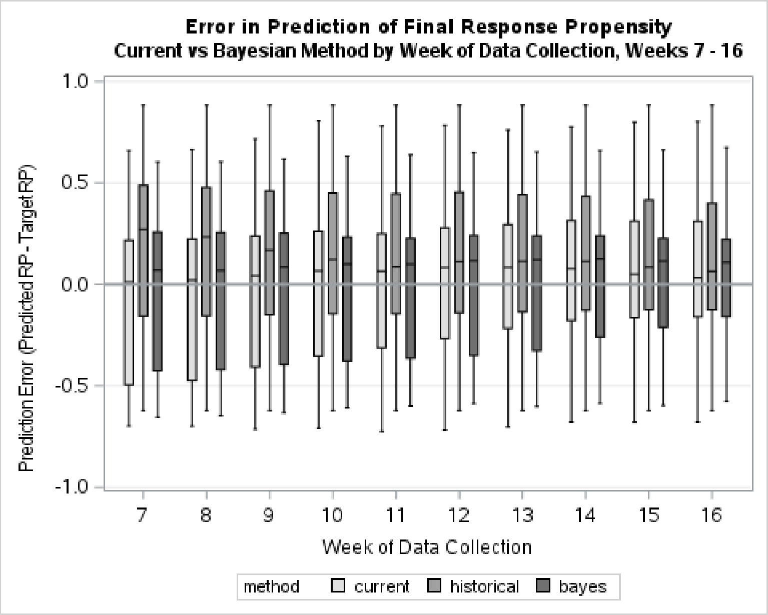


Figure 26. Error in Predictions of Final Response Propensity for Open Cases, Weeks 7-16

Up through week 12, the *Bayesian* method continues to have median and intraquartile ranges of prediction error more centered around zero than the *historical* method, as well as a smaller total range than the *current* method. After week 12, the *current* method begins to outperform the other methods when examining the median error, though the *Bayesian* method still produces a smaller total range of error.

Figure 27 displays the end of data collection, from week 17 through week 26. Towards the end of data collection, the all three methods converge toward an error of zero, as more paradata is accumulated, and fewer cases remain open. By the end of data collection the prediction error in the *current* method is zero, as our target prediction is the final response propensity under the current method. However, the error in the other two methods also shrinks toward zero.

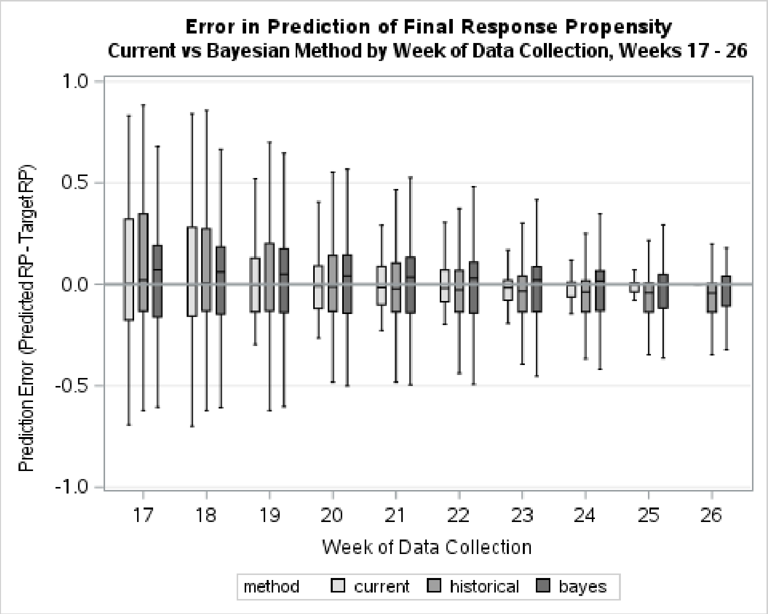


Figure 27. Error in Predictions of Final Response Propensity for Open Cases, Weeks 17-26

In addition to the distributions of prediction error for each of the methods, we can also look at the mean prediction error generated by the three different methods. Figure 28 is a plot of the weekly mean bias generated by each of the three prediction methods. This plot provides evidence that, on average, the *Bayesian* method outperforms the other methods from approximately week 5 through week 20, and consistently outperforms the *current* method until the end of data collection.

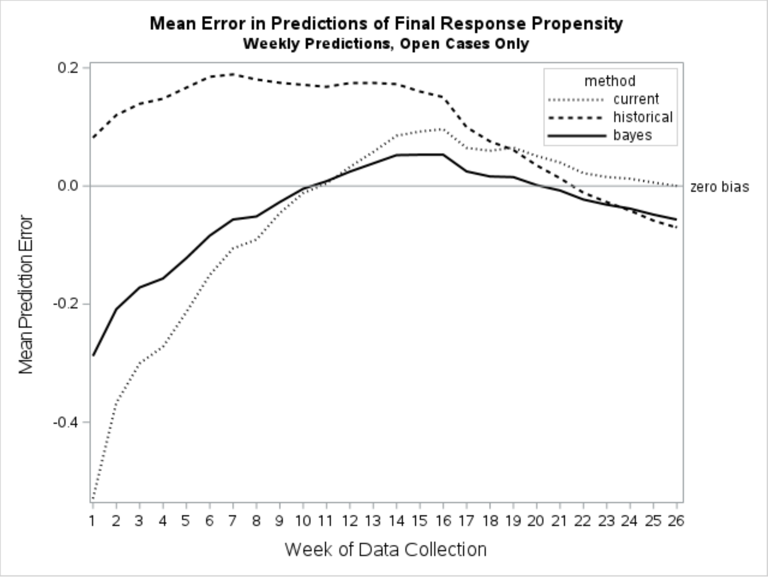


Figure 28. Mean Bias in Predictions of Final Response Propensity for Open Cases, Weeks 1-26

Figure 29 shows a similar pattern in the estimated RMSE of predictions of final response propensity. Right at the start of data collection, the historical method performs best, but starting at week 5, and continuing throughout most of data collection, the *Bayesian* method results in consistent (albeit small) improvements in the RMSE of predicted response propensity.

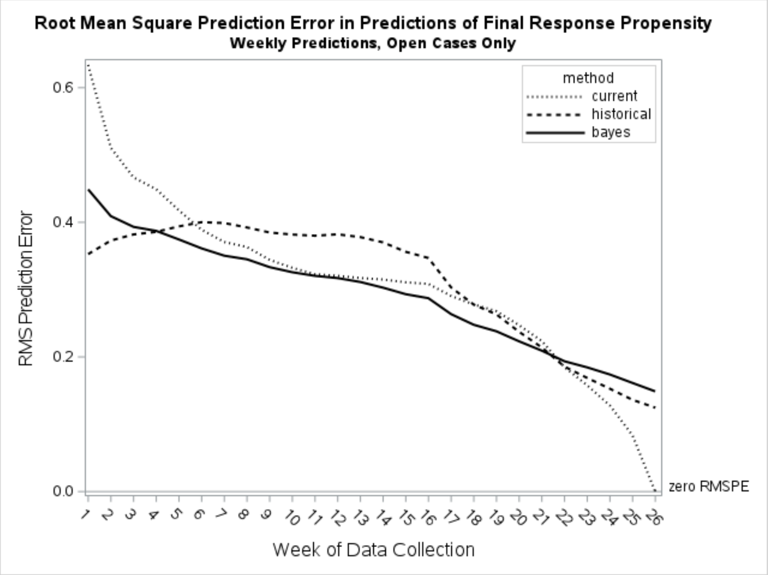


Figure 29. RMSE in Predictions of Final Response Propensity for Open Cases, Weeks 1-26

Most intervention opportunities in the NSCG occur between weeks 6 and 16, and so the Bayesian framework is a reasonable way to combine current and external data to improve predictions of response propensity without reducing the quality of those predictions.

4.3.2 Survey Response, Self-Reported Salary

In order to determine how much information a sample unit would contribute to the quality of estimates generated from survey response data, we also need to understand how a sample unit would actually respond in the NSCG. We selected one key survey item that is important to the NSCG, self-reported salary. The NSCG is often used by education and employment researchers estimating mean salaries by different types of occupations, or for different sampling domains, so any interventions should be applied in a way to avoid increasing error in the estimate of mean of salary. We therefore need to ensure that we have high quality predictions of respondent salary so that our interventions are based on reasonable estimates for both individual case-level predicted salary, and predicted mean salary. While we are focusing on one response item here, surveys generally have many key estimates. We return to this point in the Discussion.

The NSCG requests sample members report their salary, a continuous variable. Because salary often has a right-skewed distribution, we examined several transformations of the variable, in order to determine how best to obtain a linear model. Figure 30 below shows the distribution of the actual response variable, followed by three transformations: the log, square root and cube root. Taking the cube root of salary results in a distribution most similar to the normal distribution, based on visual examination. As a result, we will take the cube root of salary for the 2015 NADC group, and estimate a linear model of the form:

$$(y_i^L)^{1/3} = X_i^T \hat{\beta} + \epsilon_i$$

where y_i^L is the estimated salary for the i^{th} case using the L^{th} method, X_i is the vector of predictive covariates for the i^{th} case, and β is the vector of predictive model coefficients.

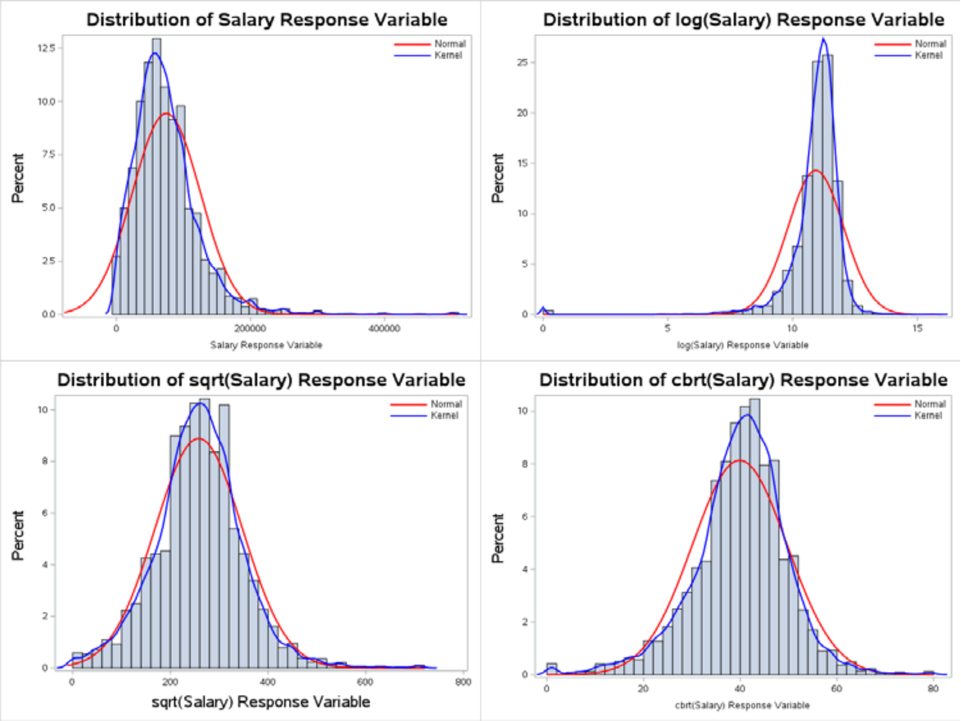


Figure 30. Distributions of Transformations of Self-Reported Salary vs Normal Distribution in 2015 NADC

Table 11 below lists all variables with a p-value of 0.25 or less. The full list of binary indicator variables, coefficients and standard errors is in Appendix E.

Table 11. Retained Predictors for Reported Salary Model, 2015 NADC

Variable Name
ACS Response Mode
Major Field of Degree
Demographic Group
Disability Indicator
Census Geographic Division
Working Status/Financial Quintile Variable
Highest Degree Earned
Broad Occupation Category
Science and Engineering Occupation Indicator
Sex
Central City, MSA, Outside of MSA
Health Insurance Coverage Indicator
Absent from Work Indicator
Looking for Work Indicator
Poverty Status Indicator
Public Healthcare Coverage Indicator
Science and Engineering Field of Degree Indicator
Home Rental/Ownership Status
Number of Vehicles at Housing Unit
Weeks Worked Past 12 Months
Access Internet by DSL Service Indicator
No Internet Access on Handheld Device Indicator
Number of Bedrooms in Housing Unit
Hours of Work in a Week
Cube Root of Personal Income
Cube Root of Wages
Cube Root of Retirement Income

Figure 31 below illustrates some diagnostics for the in-sample performance of the final model. While the model only has an R-squared of 0.56, it is notable that when looking in the bottom quadrant of the diagnostic plot (Residuals), over 60% have a residual very close to zero. That percentage of cases is higher than expected, based on the kernel of the normal distribution included in that plot. However, looking at the remainder of the diagnostics, it is clear that while the center of the distribution of self-reported salary is well explained by this model, the residuals along the tails can grow large – this model overestimates the true value of salary for cases at the low end of the distribution, and underestimates the true value for those at the high end.

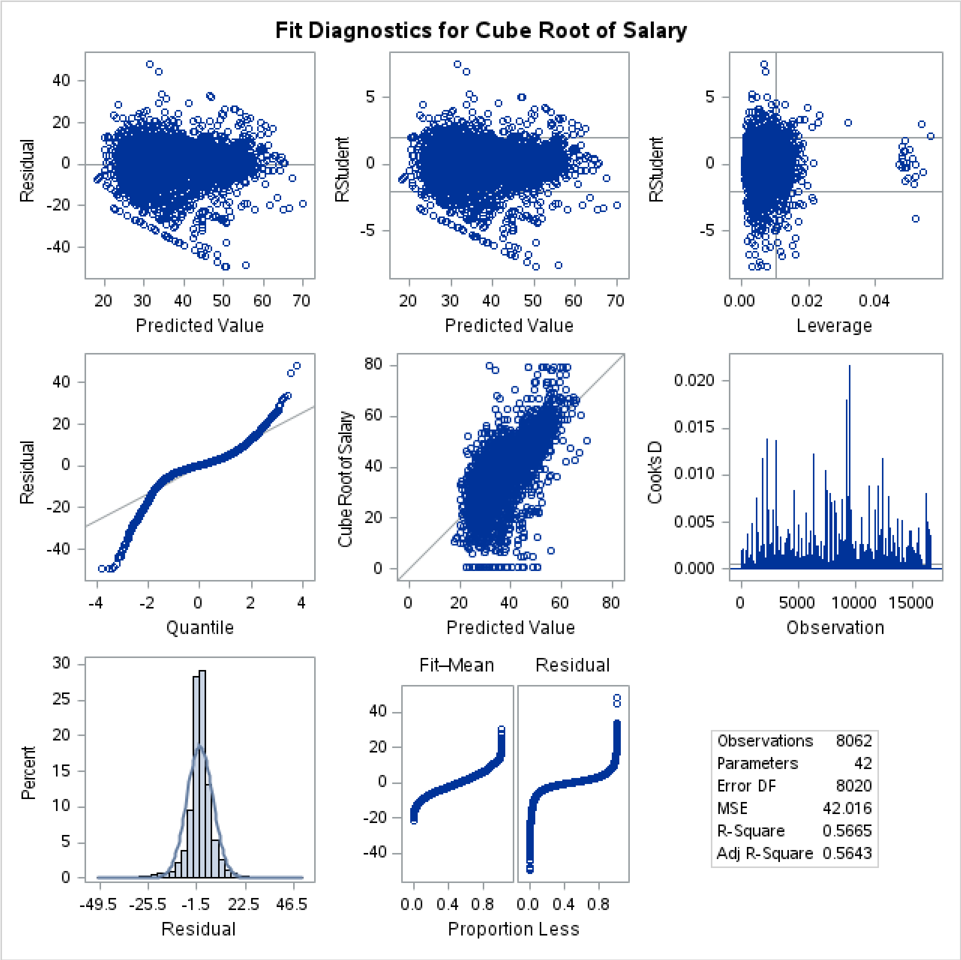


Figure 31. Model Fit Diagnostics for In-Sample Prediction of Self Reported Salary

This does not mean that the predictive model is not useful – it is more important to how the three potential methods, *current*, *historical*, and *Bayesian* methods perform relative to each other. Similar to the final response propensity models, we retained the list of predictors found significant in the 2015 NADC for use in estimating a predictive model for self-reported salary. We refit the model weekly to generate predictions based on coefficients estimated from frame information and the accumulating 2017 response data. We used the actual survey responses of salary as the “target" self-reported salary. Additionally, the parameters and standard errors from Appendix E were retained for use in the *historical* and *Bayesian* methods for predicting self-reported salary in the 2017 NSCG. Unlike the predictions of response propensity, the quality of predictions of survey responses can only be evaluated for cases that ultimately responded in the NSCG.

With the model coefficients that were estimated from the 2015 data, their standard errors, and borrowing from the approximate normality of maximum likelihood estimates of regression coefficients, we generated Normal priors for each covariate in the predictive model, $N{\sim}(\mu, \sigma^2)$, , where μ is the mean and σ^2 is the variance of the coefficient estimate. Similar to the model construction for response propensity, here we ignored correlation in these parameters, treating them as independent. Using PROC MCMC in the SAS 9.4 programming language, we then generated posterior estimates for each model coefficient by taking 2,000 Monte Carlo simulations (10,000 thinned by 5) of the posterior distributions of the coefficients based on the priors and the likelihood, which included information up through week w. We then estimated the value of salary from each of those draws, and averaged them in order to arrive at a predicted response salary for each case in each week.

Final nonrespondents would not have a response value to compare to the predicted values; therefore, while we generate predictions for all open cases, we drop final nonrespondents from the evaluation in this section. For this evaluation, we do not use design-adjusted variance estimates. We are concerned primarily with prediction of outcomes within a single survey sample, and therefore focused on internal validity of the predictions, rather than attempting to create predictions or estimates for the full target population. The NSCG does have very variable weights, however, and so we return to this point in the Discussion.

In order to evaluate the three models against each other, we calculated the difference between the weekly case-level prediction of salary generated by each method, $\hat{y}_{it}^{m}$, and the target value of salary, y_i^T, based on actual survey responses. We then used these estimates of prediction error to look at the weekly distribution of prediction error for all open cases during that week, as well as the weekly mean prediction bias and RMSE. MPB averages the individual case-level errors described above over n, the number of open cases that ultimately responded to the survey question in the 2015 NADC group:

$$MPB_t^m = Bias(\hat{\bar{y}}_t^m) = \frac{1}{n}\sum\nolimits_{i=1}^{n} (\hat{y}_{it}^{m} - y_i^T) \quad ,$$

and the weekly RMSE for the mth method is defined as:

$$RMSE_t^m = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} (\hat{y}_{it}^{m} - y_i^T)^2} \quad .$$

Figure 32 through Figure 34 compare boxplots of the error in predictions of self-reported salary on a weekly basis for each of the three methods. In general, there is little difference between the three methods when looking at the boxplots. Figure 32 displays the earliest part of data collection, weeks 1-6. Both the *historical* method and the *Bayesian* method provide small benefits by reducing the full range and intraquartile range of errors slightly. Additionally, the median error in the *Bayesian* method is squarely at zero, a very small improvement over the other methods.

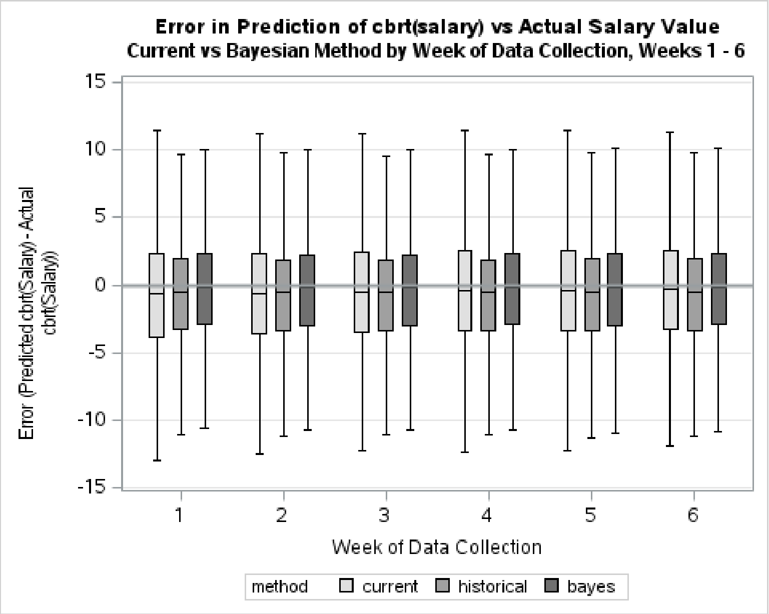


Figure 32. Error in Predictions of Final Response Propensity for Open Cases, Weeks 1-6

The fact that even the *current* method performs similarly to both the *historic* method and the *Bayesian* methods suggests that even early in data collection, when response rates are low, it is possible to use the accumulating response data to generate a model nearly as good as when a full data collection period worth of historical data is used.

Figure 33 and Figure 34 illustrate the middle and later parts of data collection. All three models perform similarly to each other, with the Bayesian method producing slight improvements in the median and overall range of prediction error. Starting around week 9, the median error of the *historical* and *Bayesian* methods moves slightly closer to zero, suggesting that the current data collection period is slightly different from the historical period.

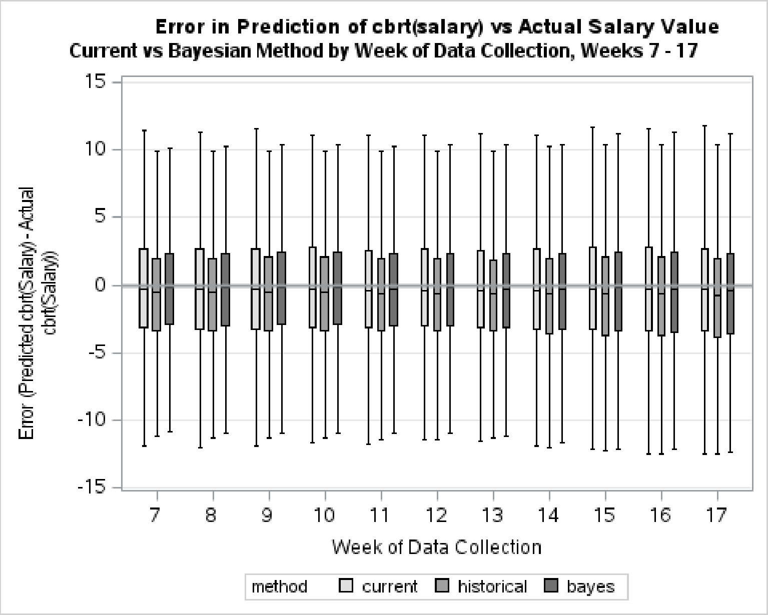


Figure 33. Error in Predictions of Final Response Propensity for Open Cases, Weeks 7-17

One interesting observation in Figure 34 below is for the last week of data collection. The *current* method does a worse job of predicting the value of salary for open cases at the end of data collection, while both the *historical* and *Bayesian* method have median errors closer to zero and reduced overall ranges of error. This suggests that individuals responding very late may actually be different (at least in their self-reported salary) than individuals responding early, because the model coefficients estimated from current accumulating data generates worse predictions of salary than either of the other two methods.

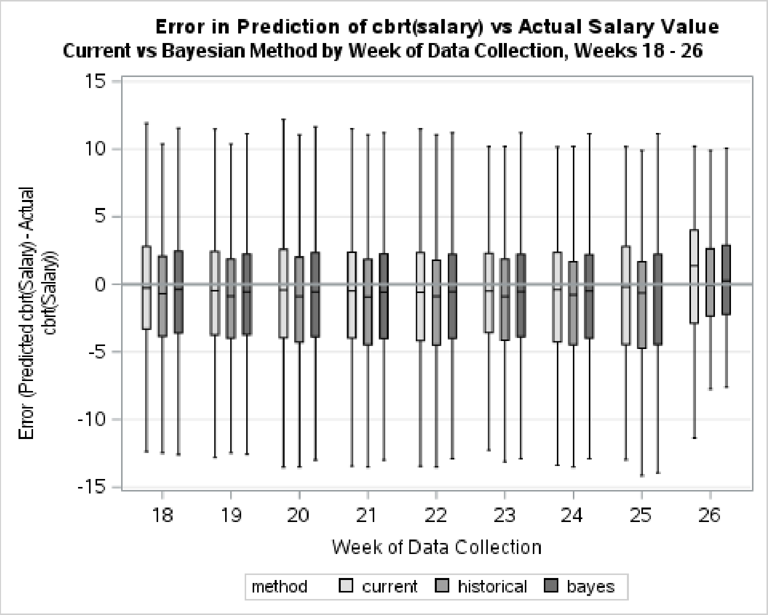


Figure 34. Error in Predictions of Final Response Propensity for Open Cases, Weeks 18-26

Figure 35 plots the mean bias in predictions of salary for open cases on a weekly basis. Early in data collection the *Bayesian* method generates predictions for open cases with the least bias, on average, when compared to the actual survey responses from those cases. This is a similar observation as in Figure 28, which depicts the mean bias in estimates of response propensity for open cases. Throughout the remainder of data collection, the *Bayesian* and *current* methods perform similarly, and both outperform the *historical* method on this metric.

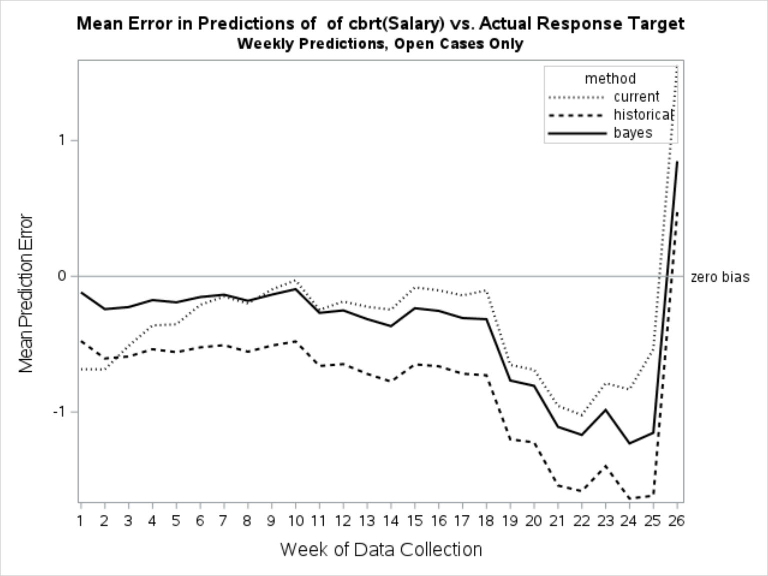


Figure 35. Mean Bias in Predictions of Cube Root of Salary for Open Cases, Weeks 1-26

While the improvements in prediction are not as clear here, in the prediction of self-reported salary, as they were in the predictions of response propensity, here we see that the *Bayesian* method performs at least as well as the *current* or *historical* methods.

While the improvements in prediction are small, using the Bayesian method early in data collection helps to hedge against the lower performance of either the historical or current methods early in data collection, This is evidenced in Figures 28 and 35 by smaller mean prediction error of the Bayesian method than either of the other two methods. Taken together, these Figures suggest that external data and currently accumulating data can be combined in a Bayesian framework without jeopardizing the quality of predictions of SDCPs during the data collection period.

4.3.3 Data Collection Costs

Data collection costs are generally driven by two major factors: what mode a case responds in, and when during the data collection period a sample member responds. This is because modes of contact and response have very different costs, and because the longer a sample member remains a nonrespondent, the more data collection resources are typically expended on trying to obtain cooperation response. These generalities are true in the NSCG, as well. Sample members can choose to respond using one of three modes in the NSCG: mail, paper, or CATI. In order to obtain the average accumulated costs for each potential response mode in each of the four phases described in Table 8, we constructed case-level cost estimates for the 2017 NADC group. Available historical operational data includes detailed information on what data collection features were applied to a particular case, and cost documentation, like what is summarized in Table 9, provided estimated costs per feature, including mailing costs, costs of an unsuccessful contact attempt in CATI, etc.

As shown in Figure 23, web is offered early and exclusively, and so nearly all response in Phase 1 occurs by web. However, as data collection continues, new modes are introduced, allowing sample members to respond by paper questionnaire or telephone interview, and potentially increasing data collection costs. Table 12 summarizes the total cumulative costs associated with being a respondent by phase and mode, or to be a final nonrespondent, assuming the standard data collection strategy.

Table 12. Costs by Response Phase and Mode, 2017 NADC

Response Mode	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**
Web	$17.05	$34.10	$43.22	$69.96
Mail	$41.77	$58.82	$67.94	$94.68
CATI	$67.28	$84.33	$93.45	$120.19
Nonresponse				$69.96

It is clear that, as data collection continues, obtaining a response becomes more expensive, regardless of the mode of response, and a web response is always the cheapest mode of response. However, web respondents' costs still increase during data collection because a case that does not respond until late in data collection will still receive all of the mailings and telephone calls that applied to nonrespondents.

To illustrate how interventions could reduce data collection costs, Table 13 summarizes the costs of response by phase and mode if, for example, a case was not sent to CATI in phase 3, and did not receive any phase 4 mailings or invitations. While the costs in phase 1 and phase 2 are the same as in Table 12, we see the costs of phase 3 and phase 4 reduced in two ways. First, no telephone calls would be made in phases 3 or 4, so a case could not respond in CATI. However, the cost of unsuccessful CATI calls would not be added to either the costs of late phase response costs *or* the final nonresponse cost. Similarly, by reducing mailings, the cost of ignored mail packages is removed from all phase 4 costs. In this instance, a case *could* still respond by mail, using the questionnaire that was sent in Phase 2. However, the reduction in additional mailings and unsuccessful CATI calls still reduces the cost of a mail response occurring in phase 4.

Table 13. Costs by Response Phase and Mode, 2017 NADC, Fewer Contacts

Response Mode	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**
Web	$17.05	$34.10	$34.81	$34.81
Mail	$41.77	$58.82	$59.53	$59.53
CATI	$67.28	$84.33	N/A	N/A
Nonresponse				$34.81

It is important to note that the costs in each of the table cells include the costs of nonresponse in prior phases. For example, Table 13 shows that a response via paper questionnaire in phase 2 costs \$58.82. That cost all costs incurred in phase 1 and phase 2, as well as the cost of a mail response in phase 2. This structure is useful operationally when trying to predict costs by response phase and mode.

Most simply, the estimated cost for a case can be expressed as an expected value:

$$E(C_i) = \hat{p}_i\left(\hat{C}_i^R\right) + (1 - \hat{p}_i)\left(\hat{C}_i^N\right) \ ,$$

where $\hat{p}_i$ is the response propensity for the i^{th} case, C_i^R is the estimated cost of a response, and C_i^N is the estimated cost for a nonresponse. The estimate for $\hat{p}_i$ is generated by the predictions discussed in Section 4.3.1 above. In order to generate estimates for $\hat{C}_i^R$ and $\hat{C}_i^N$, we need to consider not only the costs associated with response in a particular phase and mode (or final nonresponse), but also how likely response is to occur in a particular phase and mode.

As Table 12 and Table 13 illustrate, there are 13 potential response classes a sample member can fall into: response in one of three modes, during any of the four data collection phases, or a final nonrespondent. Table 14 shows the proportion of 2017 sample members that fell into each of the thirteen categories.

Table 14. End of Phase Response Propensity, Conditional on Prior Phase Nonresponse, 2017 NADC

Response Mode	**Phase 1**	**Phase 2**	**Phase 3**	**Phase 4**
Web	39.33%	5.49%	4.83%	4.40%
Mail	0.01%	3.10%	0.60%	1.30%
CATI	1.00%	0.83%	1.79%	2.36%
Nonresponse				34.98%

This means we can estimate $\hat{C}_i^N$ using the predicted response propensity from 4.3.1, and the estimated cost of being a final nonrespondent from Table 12. Estimating $\hat{C}_i^R$, however requires the estimated probability of responding by each mode in remaining phases, *conditional* on prior phase nonresponse and response by the end of data collection. Table 15 shows the conditional response probabilities for future phases, assuming the case will respond, estimated from the 2017 NADC,

Table 15. Response Proportions by Phase Assuming Final Response

Response	**P1**	**P2**	**P3**	**P4**
	Assuming Nonresponse After Phase 1			
Web	N/A	22.23%	19.54%	17.82%
Mail	N/A	12.56%	2.43%	5.27%
CATI	N/A	3.34%	7.24%	9.57%
	Assuming Nonresponse After Phase 2			
Web	N/A	N/A	31.59%	28.81%
Mail	N/A	N/A	3.93%	8.51%
CATI	N/A	N/A	11.70%	15.47%
	Assuming Nonresponse After Phase 3			
Web	N/A	N/A	N/A	54.57%
Mail	N/A	N/A	N/A	16.12%
CATI	N/A	N/A	N/A	29.30%

To illustrate the process of constructing an estimated cost, assume that at the end of phase 2, an open case has an estimated final response propensity (estimated using the method described in Section 4.3.1) of $\hat{p}_i = 0.42$. The estimated cost can be first written as:

$$E(C_i) = 0.42\left(\hat{C}_i^R\right) + 0.58\left(\hat{C}_i^N\right) \; .$$

The overall nonresponse cost is \$69.96, taken directly from Table 12. The estimated cost of a response, $\hat{C}_i^R$, can be calculated as follows:

$$\hat{C}_i^R = \sum_{p\geq 3}\sum_{m} \hat{r}_i^m c_i^m \quad ,$$

where $\hat{r}_i^m$ is the likelihood of the i^{th} case to respond in the m^{th}mode in the p^{th} phase, and c_i^m is the estimate of the data collection costs for the i^{th}case to respond in the m^{th}mode in the p^{th} phase. Using the conditional distribution of response phase and mode, conditional on nonresponse through phase 2, and the case being a final respondent, we can substitute the appropriate values and obtain:

$$\hat{C}_i^R = 0.3159 * \$43.22 \; + \; 0.0393 * \$67.94 \; + \; 0.1170 * \$93.45 \; + \; 0.2881 * \$69.96 \\ + \; 0.0851 * \$94.68 \; + \; 0.1547 * \$120.19 = \$74.06$$

The overall estimated cost, $E(C_i)$, then is equal to $0.42(\$74.06) + 0.58(\$69.96) = \$71.68$. We can construct these cost tables assuming different adaptive strategies, where certain features are applied or withheld.

We attempted to build a multinomial logistic regression model to predict the response phases and modes displayed in Table 14, using covariates similar to those used to predict response propensity and salary. However, the available covariates showed little ability to discriminate any mode of response from nonresponse, likely because of the rarity of mail and CATI respondents. The resulting models had very high specificity, as nearly all open cases were simply predicted to be nonrespondents, but sensitivity was extremely low. As a result, we made the decision to assign prior round conditional probabilities to the current round in order to generate likelihoods to respond via each mode during each phase. This means that, for example, in the 2019 NSCG, all nonresponding cases after Phase 1 would be assigned an 9.2% chance of responding in the web in Phase 2, based on the proportion of cases that responded by web in Phase 2 in 2017 $(5.49\%/(100\% - 40.34\%)$. Similarly, we use the average costs for response by phase and mode to assign estimated costs to cases. This means that our inputs to the 2019 experiment (which come from 2017 data) were treated as though they have no error, leading to a deterministic prediction for expected costs, rather than being drawn from a predictive distribution. While we use these estimates of mean costs for the purposes of cost prediction in our experiment, not all cases responding by a particular mode in a particular phase have *exactly* the same costs – there is some variability. We return to this issue in the Discussion.

4.4 2019 NSCG Adaptive Design Experiment

Methodologically, adaptive design experiments in past implementations of the NSCG have focused on improving representativeness, a measure of data quality. While we have intervened on cases in order to meet this goal, we have not done this under any formal decision rule framework, nor have we estimated the expected change in response or data collection costs by implementing an intervention on a particular subset of cases. In other words, the cost savings and relatively stable response rates have been fortunate outcomes from making sensible interventions during data collection. However, it is possible that we could have obtained better results by maintaining a neutral cost per case, and intervening on more cases (or reducing contact effort on fewer cases), which could have further improved response rates. In order to evaluate the effect of adaptive design on a key survey estimate in the NSCG while applying cost-efficient decision rules, we implemented an adaptive design experiment under a Bayesian framework (Elliott 2017).

4.4.1 Experimental Design

The goal of the 2019 experiment was to use dynamic adaptive design to reduce costs without causing a large increase in the RMSE of a key survey estimate, the mean of self-reported salary. Additionally, we wanted to leverage both historical and current accumulating information in a Bayesian framework in order to maximize the information used to generate interim predictions of response propensity and salary.

The experiment compares one group of cases that were managed by responsive design decisions (Treatment), and one group managed using the standard production methodology (Control). The experiment included 16,000 cases selected from the new cohort (i.e., their first NSCG interview) split evenly across the Treatment and Control groups. The cases were drawn using a systematic random sample with a cluster of two cases. NSCG sampling variables, including ACS response information and demographic information, as well as other operational variables like having a valid mailing address and phone number on the frame were used in the systematic sampling to ensure the cases in the Treatment and Control were representative of the NSCG sample in general, and had comparable characteristics to each other. Once the sample was drawn, the first case in the cluster was assigned to the Treatment and the second to the Control.

In order to reduce data collection costs, we needed to apply interventions at times when data collection costs would be increasing for a typical nonrespondent. Referring back to Figure 23 and Table 8, there are three main points where data collection costs increase more than linearly for nonresponding cases. At week eight, a paper questionnaire is introduced; at week 12, outgoing telephone data collection (CATI) begins, and at week 18, a late mail strategy combines several mailings to encourage late non-responders to complete the survey. In particular, the week 8 and 12 interventions appear to increase data collection costs significantly. A paper questionnaire in more expensive to mail, and requires significantly more processing after return (e.g., scanning, keying, etc.) than a response in the web instrument. CATI introduces interviewers, further increasing data collection costs. As a result, we identified these three points for the interventions described in Table 16.

Table 16. Intervention Strategies for 2019 Experiment

Data Collection Week	Standard Strategy	Alternate Strategy
8	Send paper questionnaire to nonrespondents	Withhold questionnaire; Send additional web invite letter
12	Begin CATI operation for nonrespondents	Exclude case from CATI; Send additional web invite letter
18	Conduct late mail operations for nonrespondents	Exclude case from late mail operations

Using these intervention points, we constructed all possible sets of data collection features that could be applied to a case, shown below in Figure 36. The top (yellow) pathway is the standard production data collection pathway, where at each decision point, we keep the case in the standard strategy. All cases in the control group followed this pathway during data collection. Additionally, in the treatment group, cases that were not a part of any of the interventions also followed this pathway. At the other extreme, the bottom (orange pathway) represents the minimum data collection strategy, where at each decision point, we reduce the effort, and therefore the cost, of features applied to a case. Here, at Week 8, we replace the mail questionnaire with a web invitation; at Week 12, we do not send the case to CATI, and send another web invitation; and at Week 18, we do not send the final set of contact mailings. In the middle, we see all of the available combinations.

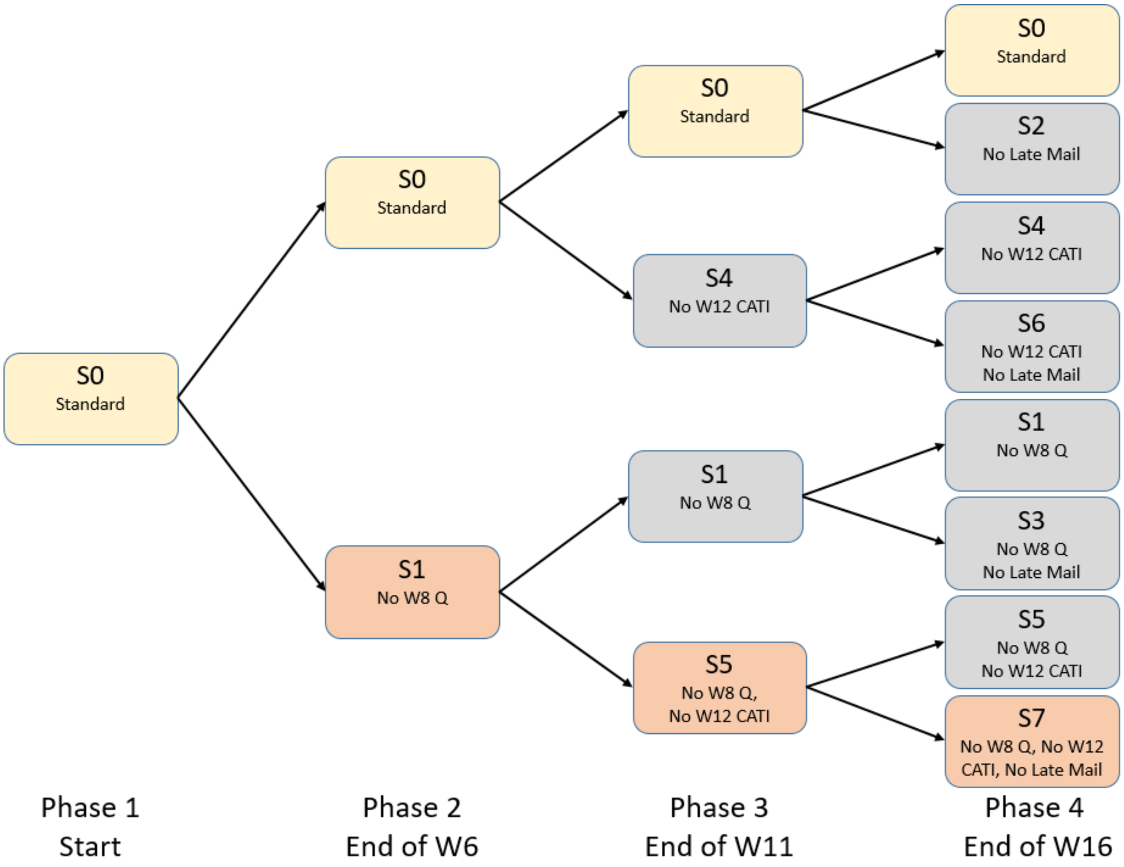


Figure 36. All Possible Case-Level Data Collection Pathways Based on Intervention Decision Points

At each decision point, there are only two options for each case in the next phase – the case either remains in the current strategy, receiving all standard data collection features going forward, or is switched to the alternate strategy for the next feature, which reduces effort and cost. This binary choice allows us to compare the predicted effect on response propensity, cost, and quality of allocating sets of cases to a reduced data collection strategy from the one they are in at any given point in time. While we were able to generate predictions for response propensity, cost and data quality using the models discussed in Section 4.3, estimating the expected effect of switching a case to an alternate strategy was not a parameter we could estimate easily. These particular data collection interventions had not been carried out as part of a randomized experiment in the past, and so we did not have an obvious choice of predicted effect size. We used the results presented in Coffey, Reist and Miller (2019) to generate an estimate of the expected change in response rate due to adaptive interventions. Cases who were assigned to reduced effort interventions saw their response rates drop by approximately 5% when compared to similar cases that were not assigned to reduced effort interventions. Cases were not assigned to those interventions in a randomized way, but the interventions presented in that publication were the most similar to those utilized here. As a result, we assumed a 5% reduction in the predicted response rate for a case each time a case is assigned to a lower effort strategy in this experiment as well.

Figure 36 illustrates that, at each intervention point, a decision must be made about which cases to assign to the alternate data collection feature. This decision could be made in a number of ways. Cases could be randomly assigned to each of the two features; cases with the lowest predicted response propensities could be allocated to the lower (or higher) effort and cost feature; or as we elected to do, cases with predicted salary values close to the predicted mean salary value would be allocated to the lower effort and cost data collection feature. Similarly, the decision about *how many* cases to switch to the alternate data collection strategy could be carried out in a number of ways. An arbitrary case count could be used (e.g., move 10% of nonresponding cases); an external threshold could be used (e.g., move all cases with a predicted response propensity under 35%); or as we chose, cases should be moved until we find the point where the tradeoffs in cost savings and loss of data quality (increase in RMSE) are optimized.

4.4.2 Optimization Steps

We wanted to identify sample units that would have the smallest impact on the mean of a key survey estimate, and assign those units to a less expensive set of data collection features, in order to reduce data collection costs, while still giving them an opportunity to respond in the less expensive mode. A given sample unit may not contribute much information to the mean of a key survey estimate when the sample unit-level response is very similar to the overall mean. By simulating a change in strategy over increasing sets of nonrespondents based on how similar the predicted case-level value for salary would be to the predicted mean value for salary, we were able to balance the reduction in cost and increase in RMSE in an optimal way. At each of the three intervention time points, t, we carried out a set of optimization steps in order to determine which cases would receive the alternate, lower cost, strategy. In addition to the textual description below, a visual step-by-step graphic is included in Appendix F. At each intervention point:

1) Impute estimate of salary for all non-responding cases, using the methodology discussed in Section 4.3.2.

2) Generate an estimate for mean salary, $\hat{\bar{y}}^T$, using the survey response for respondents and imputed estimate for nonrespondents. This will be considered the unbiased “target” parameter of interest. This is calculated as:

$$\hat{\bar{y}}_t^T = \left(\sum_{i \in S} \delta_{it} y_i + (1 - \delta_{it}) \hat{y}_i\right) \ ,$$

where y_i is the actual value of self-reported salary for the i^{th} case, provided by each of the respondents, and $\hat{y}_i$ is the imputed value of salary for the i^{th} case, provided by each of the nonresponding, cases at intervention time t. $\delta_{it} = 1$ for respondents, and $\delta_{it} = 0$ for nonrespondents.

3) Use the estimate of mean salary to generate a distance, $\hat{d}_{it} = (\hat{y}_{it} - \hat{\bar{y}}_t^T)$ for all nonresponding cases at intervention time t.

4) Sort cases by increasing values of $\hat{d}_{it}$. At each intervention point, cases will be identified for interventions by increasing values of $\hat{d}_{it}$.

5) At each of the three intervention points, t, generate two expected data collection cost estimates for each case, one assuming the case remains in its current strategy, $\hat{c}_{it}^{S_0}$, and one assuming the case is switched to the alternate strategy, $\hat{c}_{it}^{S_A}$. These costs are generated using the methodology explained in Section 4.3.3.

6) At each of the three intervention points, obtain two posterior estimates of response propensity for all unresolved cases, one assuming the case remains in its current strategy, $\hat{p}_{it}^{S_0}$, and one assuming the case is switched to the alternate strategy, $\hat{p}_{it}^{S_A}$. The response propensity for the current strategy, $\hat{p}_{it}^{S_0}$, is generated

using the methodology explained in Section 4.3.1. The response propensity for switching a case to the alternate strategy, $\hat{p}_{it}^{S_A}$, is calculated by penalizing $\hat{p}_{it}^{S_0}$ by five percentage points, so $\hat{p}_{it}^{S_A} = \left(\hat{p}_{it}^{S_0} - 0.05\right)$.

7) Using a Bernoulli random trial, with $(p, n) \sim (\hat{p}_{it}^{S_j}, 1)$, where $j = \{0, A\}$, determine stochastically whether the case will be a respondent under each strategy. A nonresponding case at intervention point t that is predicted to be a respondent has a predicted response indicator value $\hat{\delta}_{it} = 1$.

8) Generate a new estimate of mean salary, $\hat{\bar{y}}_t^{S_0}$, assuming all cases stay in the current strategy, S_0, dropping all cases who are considered nonrespondents in step 7, when $j = 0$. Using this new estimate of mean salary, which accounts for predicted nonresponse, and the unbiased target mean from Step (2), generate a measure of RMSE of the mean value of salary. RMSE for strategy S_0 is defined:

$$RMSE\left(\hat{\bar{y}}_t^{S_0}\right) = \left(\hat{\bar{y}}_t^{S_0} - \hat{\bar{y}}_t^{T}\right)^2 + \mathrm{Var}\left(\hat{\bar{y}}_t^{S_0}\right) \quad ,$$

where the first term represents the squared bias of the predicted mean, $\hat{\bar{y}}_t^{S_0}$, versus the target mean, $\hat{\bar{y}}_t^{T}$, and the second term is the variance of the predicted mean,$\hat{\bar{y}}_t^{S_0}$. The variance term is defined:

$$Var\left(\hat{\bar{y}}_t^{S_0}\right) = \frac{1}{\sum_{i \in S}\left(\delta_{it} + (1 - \delta_{it})\hat{\delta}_{it}\right)} \sum_{i \in S} \Big(\delta_{it}\left(y_{it} - \hat{\bar{y}}_t^{S_0}\right)^2$$

$$+ (1 - \delta_{it})\hat{\delta}_{it}\left(\hat{y}_{it} - \hat{\bar{y}}_t^{S_0}\right)^2\Big) \quad ,$$

where the squared differences of observed respondents' ($\delta_{it} = 1$) reported values of salary, y_{it} versus the predicted mean $\hat{\bar{y}}_t^{S_0}$, and the squared differences of the case-level predictions of salary, $\hat{y}_{it}$, versus the predicted mean for nonresponding cases predicted to be respondents ($\delta_{it} = 0, \hat{\delta}_{it} = 1$) are summed and averaged over the total number of observed respondents and predicted respondents.

9) Using the ranking from Step (3), simulate switching open cases into the alternate strategy, starting with the case with the minimum $\hat{d}_{it}$. Switch over open cases in two-percentile increments, e.g., 2%, 4%, 6%, etc. Estimate $RMSE\left(\hat{\bar{y}}_t^{S_A}\right)$, where $A = \{2, 4, 6, \ldots, 96, 98, 100\}$ based on the percentage of open cases being switched to the alternate strategy using the formula from Step (8), substituting $\delta_{it}^{S_A}$ and $\hat{\bar{y}}_t^{S_A}$ for each intervention group. Each of these groups is increasing in size, likely leading to larger cost savings, and larger increases in RMSE, and represents a potential intervention group.

10) Using the same process as Step (5), substitute the expected data collection costs in the alternate strategy for cases that are switched. Generate a total expected data collection cost for each simulated switch (e.g., 2%, 4%, 6%, etc. of open cases).

11) For each potential intervention group, calculate the ratio of the RMSE, $\left(\frac{RMSE\left(\hat{\bar{y}}_t^{S_A}\right)}{RMSE\left(\hat{\bar{y}}_t^{S_0}\right)}\right)$, and cost, $\left(\frac{\hat{C}_t^{S_A}}{\hat{C}_t^{S_0}}\right)$ for the intervention vs. baseline. If the strategy results in a cost savings, the cost ratio will be below one. If the strategy results in an increase in RMSE, the RMSE ratio will be above one.

12) For each potential intervention group, S_A, find the product of the RMSE and cost ratios, $O^{S_A} = \left(\frac{RMSE\left(\hat{\bar{y}}_t^{S_A}\right)}{RMSE\left(\hat{\bar{y}}_t^{S_0}\right)}\right)\left(\frac{\hat{C}_t^{S_A}}{\hat{C}_t^{S_0}}\right)$. The optimal set of cases to switch to the alternate strategy satisfies $\min_A\{O^{S_2}, O^{S_4}, O^{S_6}, \dots, O^{S_A}, \dots, O^{S_{100}},\}$.

4.5 Summary of Interventions

For each intervention time period, we provide the following information:

1) The unweighted response rate. Interventions are limited to nonrespondents at the time of intervention;

2) A plot of the ratio of cost vs. the ratio of RMSE when different percentages of open cases are switched to the alternate strategy.

3) Information about the identified intervention, including the number of cases switched to the alternate strategy, the expected effect on data collection costs, and the expected effect on the RMSE of salary.

4.5.1 Week 8: Replace Paper Questionnaire with Web Invite Letter

The first cost-saving intervention occurred at week 8, when open cases would be sent a paper questionnaire. For the majority of the sample, this is the first time a paper questionnaire is available to the respondent. While offering new modes of response can encourage additional response, sending a paper questionnaire and processing a paper response are both more expensive operations than sending web invites and processing internet data. It may not be worth it to incur those additional costs for cases that are not contributing new information to the survey estimate. Cases must be identified for intervention at the end of week 6 for this intervention in order to allow the mail centers to prepare the correct packages for individual cases.

Table 17 shows the response rate at the end of week 6 in the treatment and control groups, and the mean of predicted salary. When calculating the mean we used the actual survey response variable if the sample member responded; otherwise, we used the predicted value.

Table 17. Information Prior to First Intervention

	Treatment (NADT)	Control (NADC)
Unit Response Rate	31.33%	30.67%
Mean Predicted Salary	$62,997.28	$62,287.51

We carried out the steps in Section 4.4.2, simulating the effects on RMSE of the mean of salary and data collection costs of moving up to 16 sets of open cases (2%, 4%, …, 32%). The "most optimal" intervention would reduce our estimates of cost and RMSE as close to zero as possible. Realistically though, interventions that reduce effort likely reduce the response rate and increase the RMSE. Therefore, we are looking for the best tradeoff between cost reduction and RMSE increase. We can evaluate intervening on different subsets of the open cases by comparing ratios of cost and RMSE of a given iteration versus the baseline.

Our goal was to find the $O^{S_A} = \left(\frac{RMSE\left(\hat{\bar{y}}_t^{S_A}\right)}{RMSE\left(\hat{\bar{y}}_t^{S_0}\right)}\right)\left(\frac{\hat{C}_t^{S_A}}{\hat{C}_t^{S_0}}\right)$ from Section 4.4, Step (12). Figure 36 shows the effect on data collection costs plotted along the y-axis versus the effect on RMSE of mean salary, plotted along the x-axis.

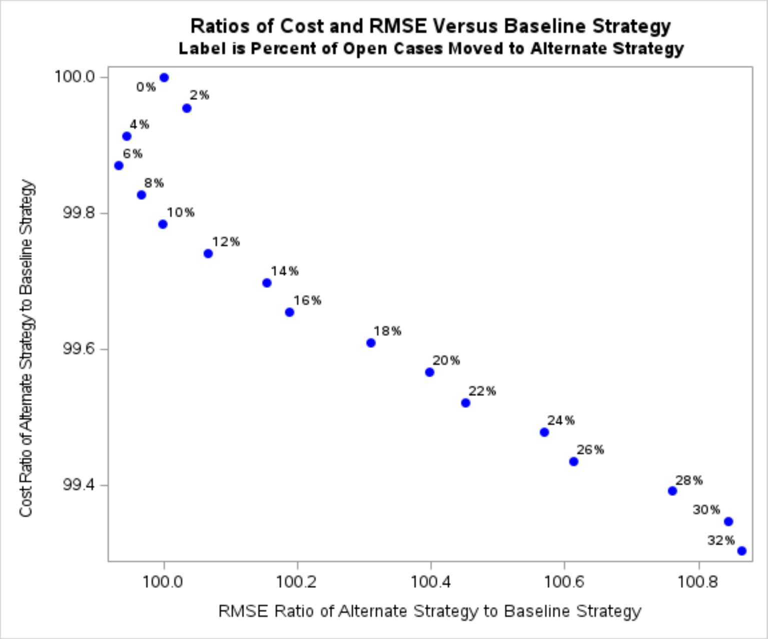


Figure 37. Ratios of Cost and RMSE to Minimize Optimization Function

The general trend is as expected; as more cases have their mail questionnaire replaced with a web invite letter, the cost decreases, but due to the falling response rate from those who would respond by paper but not web, the RMSE increases. In this instance, both the differences are small. Even when 30% of open cases are moved to the alternate data collection strategy, overall data collection costs are only predicted to be reduced by approximately 0.6%. Similarly, the RMSE is only expected to increase 0.8%.

The small effects on both cost and RMSE may be for a few reasons. First, this intervention simply removes a mailed questionnaire from the Week 8 mailing. This does not preclude cases from receiving a later mailed questionnaire (in week 18), and so there is still a chance (though a small one) that cases could respond by paper, thus incurring the cost of processing paper questionnaires. Further, this intervention does not preclude a case from being included in CATI in week 12 or in the late mail strategy in week 18. Lastly, this intervention is made in week 6, which is relatively early in the NSCG data collection period. As a result, cases may still have relatively high response propensities, so the Bernouilli check that determines whether a case is a respondent or not may still classify most cases as respondents, thus reducing the effect on RMSE as well.

In Figure 37, we plot the optimization function, O^{S_A} against the percent of cases switched to the alternate data strategy. The plot clearly shows that when 10% of open cases are switched to the alternate strategy of receiving a web invite instead of a questionnaire, the function is minimized.

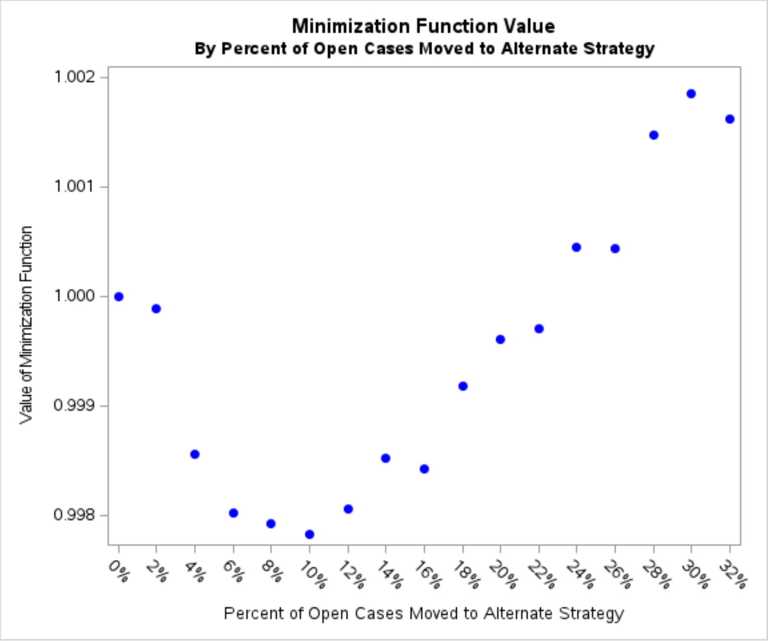


Figure 38. Minimization Function by Percent of Cases Switched to Alternate Strategy

Therefore, we will send approximately 550 cases in the treatment to the alternate strategy. All other open cases continue to follow the standard production methodology. Figure 36 shows that we should expect to a see a reduction of 0.2% in data collection costs, and an increase in RMSE of 0.00%. Therefore, we do not expect this particular intervention to have a large effect on data collection outcomes.

4.5.2 Week 12: Replace Telephone Nonresponse Follow-Up with Web Invite Letter

Our second cost-saving intervention came in week 12, when we can choose to hold cases out of CATI nonresponse follow-up. As shown in Table 12, CATI is a large cost driver in the NSCG, and so sending fewer cases to the operation could result in significant data collection cost savings.

Table 18 shows the counts of respondents and nonrespondents at the end of week 11 in the treatment and control, and the mean of predicted salary. When calculating the mean and standard deviation, we used the actual survey response variable if the sample member responded; otherwise, we used the predicted value.

Table 18. Information Prior to Second Intervention

	Treatment (NADT)	**Control (NADC)**
Unit Response Rate	40.91%	42.08%
Mean Predicted Salary	$71,731.16	$71,752.54

We again carried out the steps in Section 4.4.2, simulating the effects on RMSE of salary and data collection costs of moving up to 50 sets of open cases (2%, 4%, …, 100%) to the alternate data collection strategy. Figure 38 shows the effect on cost plotted along the y-axis versus the effect on RMSE, plotted along the x-axis. Here, we plotted all 50 possible scenarios, from where none of the cases are held out of CATI, to where 100% of cases are held out of CATI. We were interested in the effect of CATI on cost and RMSE, as it is such a large cost driver.

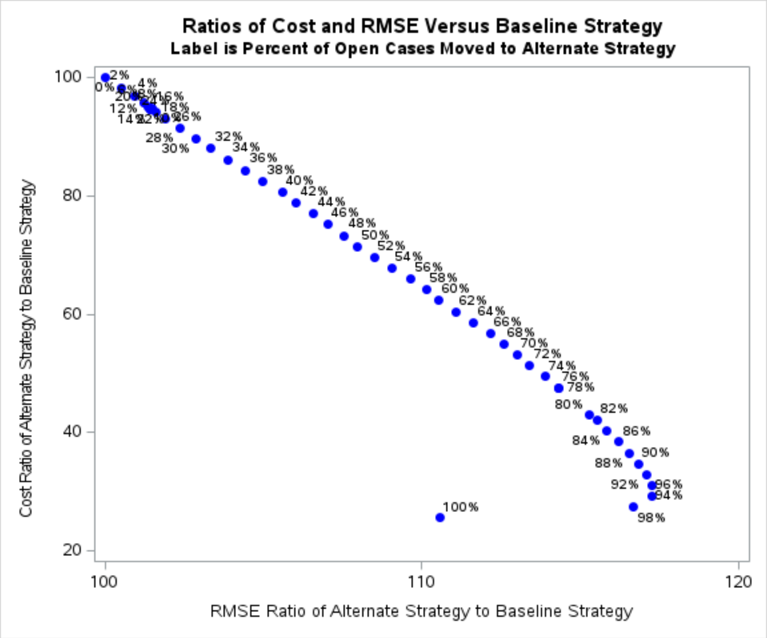


Figure 39. Ratios of Cost and RMSE to Minimize Optimization Function

Again, the general trend is as expected; as more cases are not sent to CATI, the cost decreases, but due to the falling response rate, the RMSE increases. However, in this figure, there seem to be varying effects on RMSE of moving small (< 30%) of cases to the alternate strategy. Then there is a nearly linear relationship between cost decreases and increases in RMSE up through 80% of cases being moved to the alternate strategy. After 80%, further shifts seem to have smaller effects on cost, and the RMSE starts to decrease. This may be evidence that cases with predictions of salary that are very different from the target estimate becoming nonrespondents. This would reduce the RMSE by shrinking the tails of the distribution of salary, which would not be beneficial to analysts who rely on this data to understand the distribution of salary among this population, or for those carrying out subgroup analyses.

Figure 39 is the same plot, but restricted to switching less than 30% of open cases to the alternate strategy. When fewer than 6% or more than 24% of cases are switched, we again see a linear relationship, but in between those two percentages, there does seem to be a tradeoff between cost and RMSE.

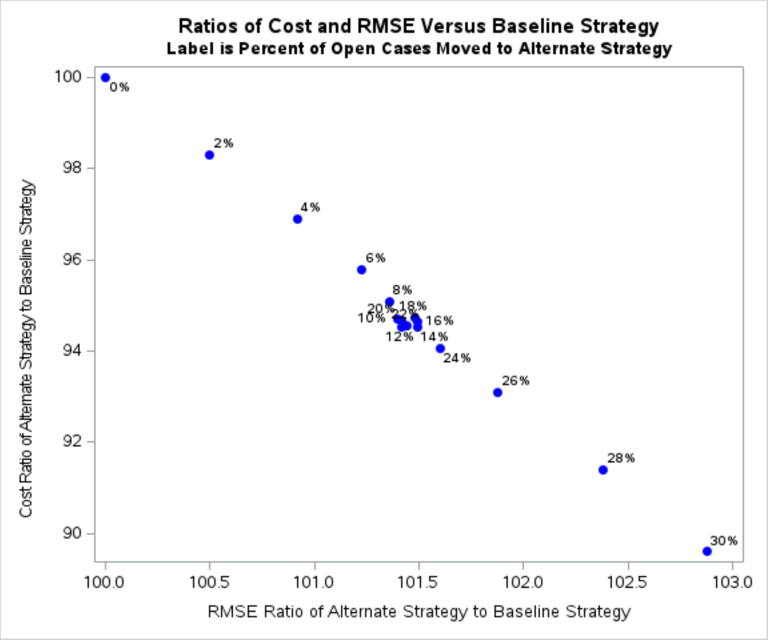


Figure 40. Ratios of Cost and RMSE to Minimize Optimization Function, Less than 30% of Cases Switched

When we plot the minimization function in Figure 40, we see that there is, in fact, a local minimum of the optimization function at around 12% of cases.

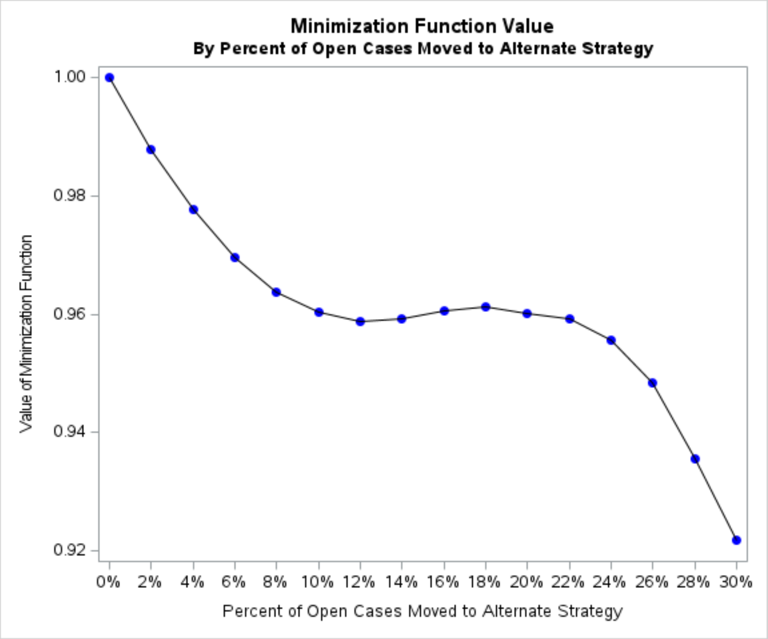


Figure 41. Minimization Function by Percent of Cases Switched to Alternate Strategy

Therefore, we will send approximately 565 cases in the treatment to the alternate strategy. Looking again at Figure 39, we expect to see a reduction of about 5.5% of data collection costs, and an increase of about 1.5% in the RMSE of salary.

Again, open cases not part of this group of 565 continue to follow the standard production methodology. This means that, potentially, a case could have been selected for the Week 8 intervention, but *not* selected for the Week 12 intervention. In practice, this would mean that case did not receive a paper questionnaire, but did receive CATI nonresponse follow-up contacts starting in Week 12.

4.5.3 Week 18: Withhold Late Contact Strategies

The final cost saving intervention came in week 18, when sample members receive a set of late mailings, including an additional questionnaire, a web invite, and a final reminder. We can choose to not send any mailings to open cases; however, this is a binary choice. Either the case receives all three mailings, or receives none of them. This decision had to be made by the end of week 16 to allow the mailings centers time to prepare the correct packages.

Table 19 shows the counts of respondents and nonrespondents at the end of week 16 in the treatment and control, and the mean of predicted salary. When calculating the mean, we used the actual survey response variable if the sample member responded; otherwise, we used the predicted value.

Table 19. Information Prior to Third Intervention

	Treatment (NADT)	**Control (NADC)**
Unit Response Rate	48.92%	50.17%
Mean Predicted Salary	$74,253.20	$74,213.25

We carried out the optimization steps in Section 4.4.2 one final time, simulating the effects on RMSE of salary and data collection costs of moving up to 50 sets of open cases (2%, 4%, …, 100%) to the alternate data collection strategy. Figure 41 shows the effect on cost plotted along the y-axis versus the effect on RMSE, plotted along the x-axis. Here, we plotted all 50 possible scenarios, from all of the open cases receiving the final contact strategy, to no cases receiving the final contact strategy. While the final contact mailings have a smaller effect on cost than the CATI operations in phase 3, we see a pattern similar that seen in Figure 38. Between 12% and 24%, we see a potential spot where there are tradeoffs between cost reductions and RMSE increases. Switching between 24% and 60% of cases results in a similarly linear relationship that we saw in Figure 38.

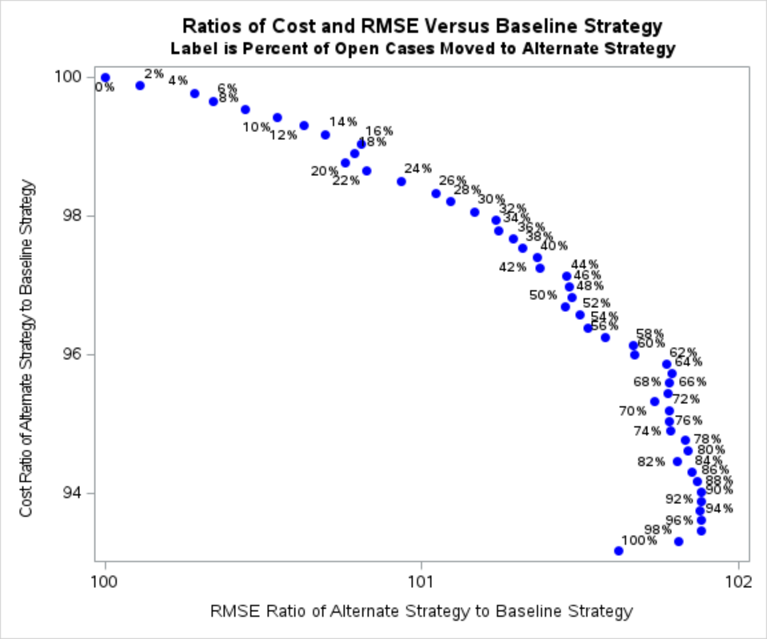


Figure 42. Ratios of Cost and RMSE to Minimize Optimization Function

Figure 42 is the same plot, again restricted to switching less than 30% of open cases to the alternate strategy. When fewer than 14% or more than 22% of cases are switched, we again see a linear relationship, but in between those two percentages, there does seem to be a tradeoff between cost and RMSE. All open cases not selected for this intervention received the Week 18 late contact strategies, following the standard production methodology.

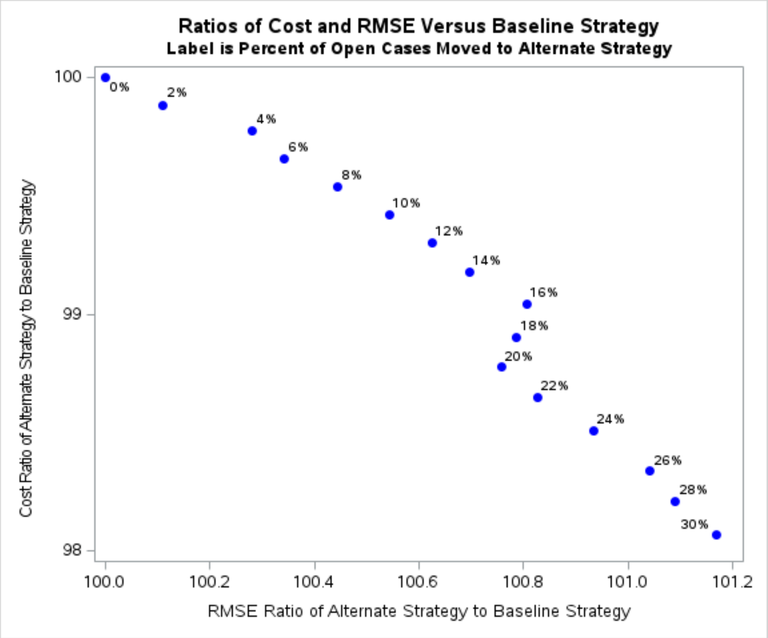


Figure 43. Ratios of Cost and RMSE to Minimize Optimization Function, Less than 30% of Cases Switched

When we plot the minimization function in Figure 43, we do not see a local minimum in the optimization function like we saw in Figure 40. However, there does seem to be an inflection point around 20%. After that point, the reductions in cost are smaller and increases in RMSE are larger.

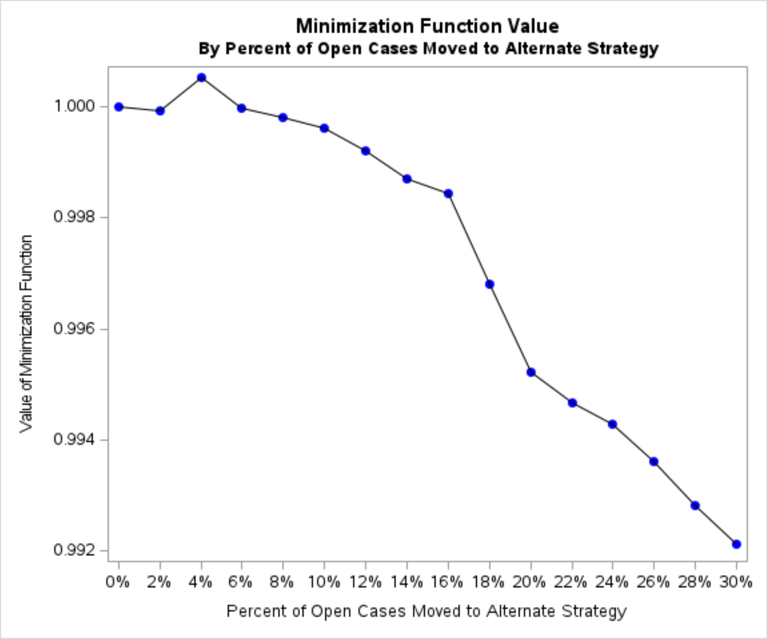


Figure 44. Minimization Function by Percent of Cases Switched to Alternate Strategy

Therefore, we will send approximately 800 cases in the treatment to the alternate strategy. Looking again at Figure 42, we expect to see a reduction of about 1.2% of data collection costs, and an increase of about 0.8% in the RMSE of salary.

4.6 Results

Our main interest was to see if we could intervene on cases in a strategic way so that we would be able to save data collection resources without jeopardizing the quality of a key survey estimate, the mean of self-reported salary. Successfully reducing expended data collection resources while maintaining quality in an optimization framework would allow survey managers to make decisions about how to allocate resources during data collection without harming the quality of the information produced by the survey.

4.6.1 Estimated Data Collection Costs (Treatment vs. Control)

During the experiment, we used estimated costs for particular data collection features (like a letter mailing or an unsuccessful telephone contact attempt) based on 2017 production cost data. Cumulatively, we expected our three interventions to reduce costs by approximately 6.9% (0.2% for the first intervention, 5.5% for the second intervention, and 1.2% for the third intervention).

After this experiment ended, we were able to use the actual costs for 2019 data collection features provided by the data collection production staff. By adding up the costs for all applied data collection features in the treatment versus control group, we were able to generate summary statistics about the data collection costs. Table 20 displays the median, mean, and total cost per case in the treatment versus control.

Table 20. Cost Comparison between Treatment and Control

	Sample Size	**Median Cost-per-Case**	**Mean Cost-per-Case**	**Total Data Collection Costs**
Treatment	8,000	$20.22	$26.81	$214,569.06
Control	8,000	$27.81	$29.57	$236,620.15

The median cost-per-case is approximately 25% lower in the treatment than in the control, while the average cost per case is approximately 9.3% lower in the treatment. Similarly, the total data collection costs for all cases was approximately 9.3% lower in the treatment than in the control. The 9.3% reduction in data collection costs is similar to the 6.9% reduction we expected based on our data collection interventions.

Using a two-tailed, two-sample z-test to estimate the difference in the mean costs of the treatment and control with $\alpha = 0.05$ results in a test statistic of -6.68, an so we reject the null hypothesis.

$$z = \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\left(\left(\frac{\sigma_T^2}{n_T}\right) + \left(\frac{\sigma_C^2}{n_C}\right)\right)}} = \frac{(26.81 - 29.57)}{\sqrt{\left(\left(\frac{25.34^2}{8000}\right) + \left(\frac{26.89^2}{8000}\right)\right)}} = -6.68$$

The mean costs per case are significantly different, with the treatment group having a lower average cost-per-case. Additionally, a 10% reduction in overall and average costs and a 25% reduction in median costs is a practically significant result that is important for evaluating the effect of this optimization on operational realities.

4.6.2 Self-Reported Salary (Treatment vs. Control)

During the experiment, we stochastically determined which cases would respond under each data collection strategy using a Bernoulli test centered around the predicted final response propensity for each case. Determining who responded allowed us to then obtain estimates of mean salary based on a combination of survey responses and our predictions for nonrespondents. After data collection we have the complete set of response data with which to compare actual survey responses. In order to determine the impact our data collection interventions had on the quality of this statistic, we compare both the RMSE and estimates of the mean salary in the treatment and control to determine whether our intervention resulted in large shifts in this key survey estimate.

Table 21 displays information about which sample members are included in the estimation of unweighted mean salary, as well as the median, mean and standard deviation within each treatment group. Estimating mean salary in the NSCG requires logic that excludes a number of cases, including those that respond but are not working (and so do not report a salary), and those who have responded, and may be working, but skip that item in the survey.

Table 21. Summary Statistics about Survey Response

	Treatment	**Control**
Sample Size	8,000	8,000
	Proportion of Sample	
Nonrespondents[1]	42.95%	41.80%
Respondents, Not Working[1]	7.68%	7.28%
Respondents, Working, Item Skipped[1]	5.35%	5.71%
Respondents Reporting Salary[2]	44.03%	45.21%
	Mean Unweighted Estimates	
Median Salary	$72,000.00	$73,000.00
Mean Salary	$84,082.10	$85,336.98
Standard Deviation of Salary	$62,776.24	$77,278.77

[1] *Excluded from Estimates of Salary*
[2] *Included in Estimates of Salary*

After taking these exclusions into consideration, we estimated the unweighted mean and median values for salary for the respondents. The median salary in the treatment group is $1,000 less than in the control group. The mean salary is approximately $1,255 lower in the treatment than in the control, representing a 1.5% difference in the point estimate of the mean. We used a two-tailed z-test for means with an $\alpha = 0.05$ to evaluate whether the mean salary in the two groups were significantly different. The sample sizes were based upon the study group sample size and the respondents that reported salary, as those were the only individuals included in the estimation of mean salary.

$$z = \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\left(\left(\frac{\sigma_T^2}{n_T}\right) + \left(\frac{\sigma_C^2}{n_C}\right)\right)}} = \frac{(84{,}082.10 - 85{,}336.98)}{\sqrt{\left(\left(\frac{62{,}776.24^2}{3{,}522.4}\right) + \left(\frac{77{,}278.77^2}{3616.8}\right)\right)}} = -0.75$$

The test, however, leads to a test z-score of -0.75, and so we do no not reject the null hypothesis. The mean salary estimates in the treatment and control are not statistically significantly different.

Based on our interventions, we expected the RMSE in the treatment group to be approximately 2.3% larger than in the control group (0.0% increase for the first intervention, 1.5% for the second intervention, and 0.8% for the third intervention). In order to estimate the RMSE for the treatment and control groups, we assume the mean of the control group is the true estimate, and use the formula:

$$RMSE(\hat{\hat{y}}_m) = (\hat{\hat{y}}_m - \bar{y})^2 + Var(\hat{\hat{y}}_m) \ ,$$

where $\hat{\hat{y}}_m$ is the estimated mean of salary for method m. The RMSE of salary in the treatment and control groups, respectively, were 62,788.78 and 77,278.77. This represents an 18.75% reduction in the RMSE in the treatment group versus the control group. In order to explain why this result was so different from our expectations, Table 22 displays the bias and RMSE for salary in the treatment and control when salary was restricted to less than $1,000,000 versus unrestricted for eligible respondents.

Table 22. Comparison of Bias and RMSE for Different Salary Cutoffs

Salary Cutoff for Estimation	**$1,000,000**		**No Limit**	
Treatment Group	Treatment	Control	Treatment	Control
% Respondents Included	100.00%	99.94%	100.00%	100.00%
Mean Salary ($)	84,082.10	84,250.02	84,082.10	85,336.98
RMSE Salary	62,776.47	61,940.82	62,788.79	77,278.77
Bias in Mean Salary ($)	-167.92		-1,254.88	
% Difference RMSE	1.35%		-18.75%	

When salary is restricted to all responses under $1,000,000, all of the respondents in the treatment group are included in the estimate. However, a small proportion (< 0.06%) are excluded from the estimate because the reported salaries are above the $1,000,000 threshold. When this restriction is placed on the cases included in the estimate of mean salary, we see that the bias in mean salary in the treatment group versus the control group is less than $200, and the RMSE of the treatment group is 1.35% larger than the control group. This is much closer to what the predicted effect of our interventions would be.

When we remove the restriction on salaries that are included in the estimate, a few more cases with extreme salaries are included, increasing the variance of salary in the control group to the point where the treatment group shows a large reduction in RMSE. Due to the very small number of influential responses in the treatment group that not only cause a relatively large shift in the mean (>$1,000), these cases are likely outliers.

When they are excluded, we see a slight increase in the RMSE to balance out the reduction in cost discussed in 4.6.1, as expected based on our experimental predictions. This is an important finding, as it demonstrates that we can meaningfully reduce costs, as shown in Section 4.6.1 without causing a sizeable increase in the RMSE of mean salary.

4.6.3 Unweighted Response Rate (Treatment vs. Control)

As a secondary measure, we wanted to evaluate the effect of our interventions on unweighted response rate. While this experiment focuses on RMSE of salary as the measure of data quality, survey managers commonly utilize response rate as another measure of data quality. Table 23 displays the response rates and percentages of response coming from different modes in the treatment and control groups.

Table 23. Summary Statistics about Unweighted Response Rates

	Treatment	**Control**
Sample Size	8,000	8,000
Unweighted Response Rate	57.08%	58.23%
Percent of Response from Web	85.92%	83.50%
Percent of Response from Mail	8.59%	10.32%

Percent of Response from CATI	5.50%	6.18%

The treatment group has a slightly lower unweighted response rate. Using a two-tailed z-test for two proportions with $\alpha = 0.05$, however, leads to a test z-score of -1.47.

$$z = \frac{(\hat{p}_T - \hat{p}_C)}{\sqrt{\left(\hat{p}(1-\hat{p})\left(\frac{1}{n_T} + \frac{1}{n_c}\right)\right)}} = \frac{(0.5708 - 0.5823)}{\sqrt{\left((0.5766)(0.4234)\left(\frac{2}{8{,}000}\right)\right)}} = -1.47$$

As a result of the test, we do no not reject the null hypothesis. The final response rates between treatment and control are not significantly different. We also observed small shifts in response behavior due to our data collection interventions. The proportion of respondents that responded via web was 2.4% higher in the treatment group versus the control group (p <0.05), and fewer respondents responded by mail or telephone, though those differences were not significant. These shifts make sense – cases identified for our interventions had a reduced number of modes by which they could respond. Cases could always respond by web, but we removed both paper questionnaires and telephone in order to save costs. So, if cases that were identified for interventions responded to the NSCG, they were more likely to respond by web, accounting for the small increase in the proportion of response coming from web in the treatment versus the control. Again, these results show that our intervention methodology was able to meaningfully reduce data collection costs without causing a negative effect on response rates.

4.7 Discussion and Future Work

The goal of this research was to implement a responsive design experiment using Bayesian methods in order to leverage external data to minimize a function of cost and

RMSE of a key survey estimate, salary. This research was largely successful. Results from our experiment indicate that it is possible to implement Bayesian prediction methods during data collection to inform adaptive interventions. Further, this methodology resulted in reduced costs without reductions in data quality. We were able to significantly reduce the mean and median estimates of data collection costs-per-case by 9% and 25%, respectively, without causing significant changes to estimates of the mean salary, the RMSE of mean salary, or unweighted response rates.

Building upon recent work (West et al. 2019; Wagner et al. 2020; Coffey et al. 2020), the first part of this chapter illustrated improvements in predictive models of both final response propensity and salary provided by the use of Bayesian methods. In particular, we found that Bayesian methods lead to smaller estimates of bias in predictions of final response propensity and salary, especially early in data collection when the accumulating data from the current survey implementation may be sparse.

From there, we illustrated how we used the predictive models for final response propensity and salary, as well as deterministic models for estimated costs, in order to intervene in data collection at points when the cost of new data collection features in the NSCG is increasing. We identified cases that had predicted values of salary closest to our estimated average salary in order, as they would have the least effect on the overall distribution of salary. We simulated the effect on estimated data collection cost and estimated RMSE of salary when different percentages of open cases were moved to an alternate, less expensive, data collection strategy. By minimizing an optimization function, we were able to find the percentage of cases that led to the optimal tradeoff between cost and RMSE, based on our predictive models. After data collection we found that data collection costs were 10% lower in the treatment group than in the control, which was significantly different, and we found no significant differences in the unweighted average estimate of salary or in the unweighted response rate.

This research does have limitations, suggest potential areas for future work. Most significantly, this experiment focused on the RMSE of one key survey item. However most demographic surveys collect data for many items. It is possible that the cases we identified as low impact with respect to salary will be highly impactful with respect to another survey item. A multivariate approach to identifying cases for intervention would help ensure that data quality was maintained across many key survey items as data collection costs were reduced. The multivariate approach might also help with determining the optimal cutoff when local or global minima in the optimization function are not obvious, as we saw in Section 4.5.3. Interventions could be restricted to cases considered to be low impact across all key survey items (i.e., the intersection), or perhaps a proportion (e.g., low impact for 50% of key survey items).

Second, while Bayesian versions of predictive models were incorporated into the optimization framework, there are areas of improvement in the model construction steps. First, when estimating model coefficients for use in our Bayesian priors, we ignored correlation between the covariates, instead treating them as independent. Coefficients would be more accurate, and Bayesian predictions might show further improvement if the correlation was accounted for. Further, we did not successfully develop Bayesian models, or any predictive models at all, for the estimated data collection costs. While the deterministic models performed reasonably well based on our experimental results, large shifts in response behavior or in costs related to data collection operations could render the deterministic models useless. One of the benefits of Bayesian methods is the continuous statistical learning that protects against poor assumptions. In other words, if our Bayesian priors, based on historical cost information, were wildly incorrect, at least the current cost information would be incorporated, hopefully improving the posterior predictions of data collection costs. Therefore, more work on the models predicting costs and response by phase and mode is important. Classification algorithms other than logistic regression may improve the predictions of response phase and mode. Wagner et al. (2020) evaluates different methods for predicting data collection costs, such as Bayesian Adaptive Regression Trees (BART), that could be applied in this setting. Alternatively, models like piecewise exponential models explored by Li and colleagues (2012; 2015) could help overcome the difficulties we found during model specification.

Third, this work should be extended to incorporate survey weights. In this experiment, our measure of impact was the distance a case’s value of predicted salary was from the mean predicted value of salary. However, when weights are highly variable (as they are in the NSCG), the weighted mean may be significantly different from the unweighted mean. Further, highly variable weights may have an impact on measures of the RMSE when they are incorporated. Base weights could be incorporated into most of the statistical models here to evaluate their effect on which cases might be selected for intervention. Additionally, weighting procedures like nonresponse adjustments could be incorporated during the steps that generate point estimates and the RMSE of key survey items. Weight variability can increase the variance of survey estimates once nonresponse is taken into account, which could have an impact on the RMSE of different data collection strategies and allocations.

Finally, with respect to experimentation, this work would be enhanced by experiments that include a “random” intervention group in addition to the control group. A random intervention group would demonstrate the benefit of the optimization methods over selecting a random subset of cases to move to an alternate data collection strategy. While this work can be partially completed through simulation, it is not possible to completely isolate the effect on cost and response of switching sample members into alternative data collection strategies without a randomized experiment. Additionally, a sequential, multiple assignment, randomized trial (SMART) design could be used to isolate the effects of including or excluding specific data collection features. In a SMART design, cases that were nonrespondents at decision points in data collection would be randomized among different available treatment options (Lei et al. 2012). This would allow for post-experimental analysis to estimate the effect of particular data collection interventions (e.g., replacing a paper questionnaire with a web invite), rather than simply penalizing the predicted response propensity by an arbitrary amount to obtain an estimate of the response propensity for a case under an alternate data collection strategy.

Despite these limitations, the results of this experiment clearly demonstrate that both Bayesian methods and optimization methods can be implemented in a production setting to intervene in data collection and reduce data collection costs without hurting data quality. Carrying out research identified above to overcome the limitations of this study would improve specific outcomes, and also improve the generalizability of the findings. These benefits could lead to increased adoption of these techniques for balancing cost-quality tradeoffs in large survey data collections.

Appendix A: Significant Predictors of Final Screener Response Propensity

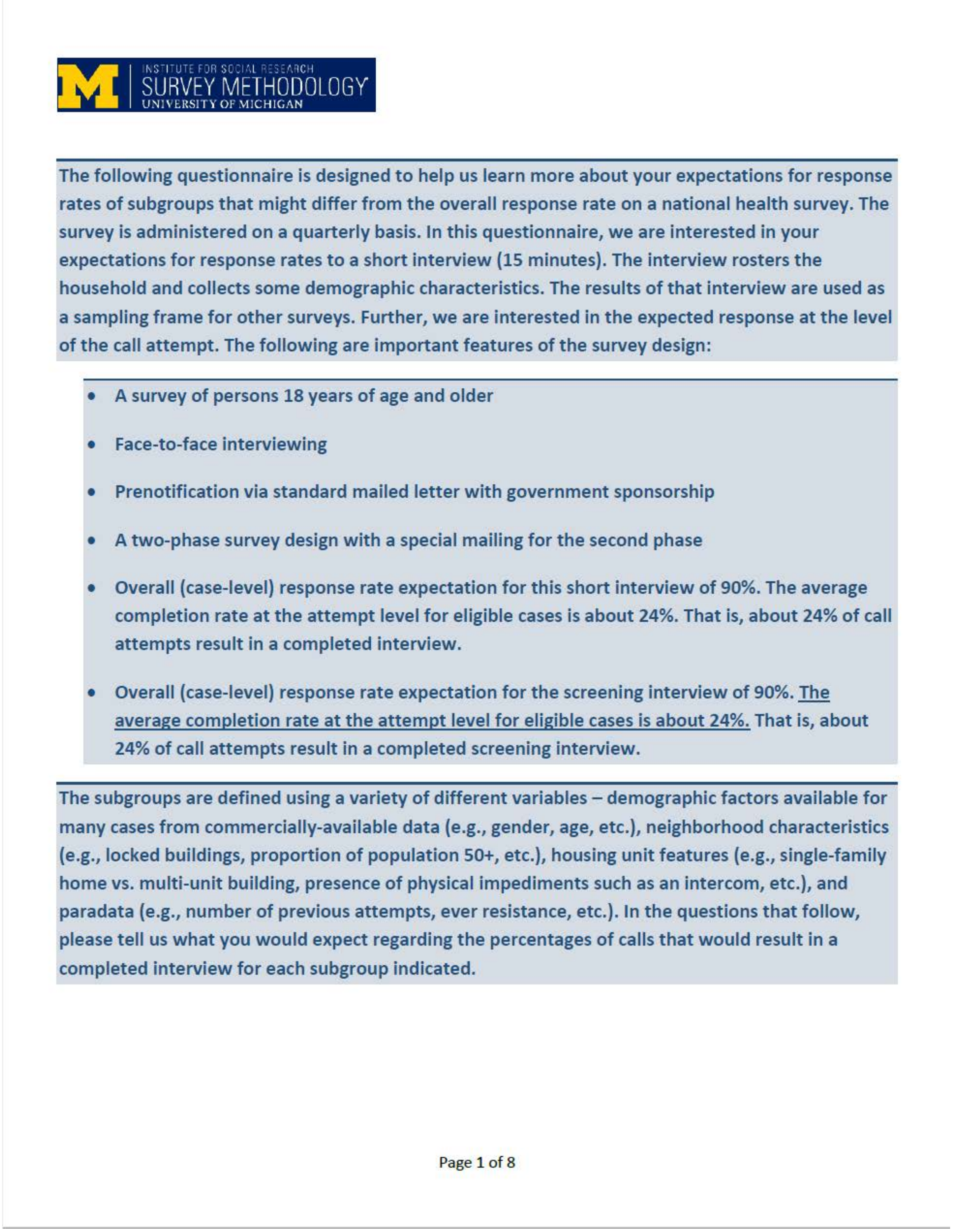
Appendix A displays the coefficients and standard errors for all retained predictors of screener response propensity in the final discrete time logit model for call-level data from the eight most recent quarters, after applying backward selection (*n* = 119,981 calls; Nagelkerke pseudo R-squared = 0.09; AUC = 0.66).

Predictor	**Coefficient**	**StdErr**
Intercept	-2.56	0.32
Mail Delivery Point Type: Missing	0.08	0.03
Mail Delivery Point Type: A	0.03	0.02
Mail Delivery Point Type: B	-0.04	0.03
Mail Delivery Point Type: C	-0.09	0.03
Interviewer-Judged Eligibility: Missing	2.46	0.10
Interviewer-Judged Eligibility: No	0.63	0.07
Segment Listed: Car Alone	0.03	0.02
PSU Type: Non Self-Representing	0.06	0.03
PSU Type: Self-Representing (Not Largest 3 MSAs)	0.03	0.03
Previous Call: Contact	3.97	0.28
Previous Call: Different Window	-0.12	0.02
Previous Call: Building Ever Locked	0.32	0.05
Previous Call: Building Locked	2.16	0.14
Previous Call: Strong Concerns Expressed	0.26	0.04
Previous Call: No Contact	2.26	0.13
Previous Call: Other Contact, No Concerns Expressed	-1.35	0.25
Previous Call: Concerns Expressed	-1.58	0.26
Previous Call: Soft Appointment	-1.03	0.30
Previous Call: Call Window Sun.-Thurs. 6pm-10pm	0.07	0.03
Previous Call: Call Window Fri.-Sat. 6pm-10pm	0.08	0.02
No Access Problems in Segment	-0.05	0.02
Evidence of Other Languages (not Spanish)	-0.09	0.03
Census Division: G	-0.14	0.03
Census Division: B	-0.32	0.03
Census Division: D	-0.22	0.03
Census Division: H	-0.24	0.03
Census Division: C	-0.20	0.03
Census Division: F	-0.27	0.04
Census Division: E	-0.20	0.03
Census Division: A	-0.19	0.04
Contacts: None	-0.68	0.24
Contacts: 1	-0.54	0.22

Contacts: 2 to 4	-0.42	0.19
Segment Domain: <10% Black, <10% Hispanic	-0.04	0.02
Segment Domain: >10% Black, <10% Hispanic	-0.04	0.02
Segment Domain: <10% Black, >10% Hispanic	0.01	0.03
Percentage of Segment Non-Eligible (Census Data)	-0.01	<0.01
Interviewer-Estimated Segment Eligibility Rate	-0.55	0.12
Interviewer-Estimated Household Eligible	-0.09	0.02
Segment Type: All Residential	0.04	0.02
Log(Number of Calls Made)	-0.60	0.03
Log(Number of Calls Made) x No. Prev. Contacts	-0.04	0.01
CML* HoH Age: 35-64	-0.12	0.02
CML Adult Count: Missing	-0.13	0.04
CML Adult Count: 1	-0.09	0.03
CML Adult Count: 2	0.01	0.03
CML Asian in HH: Missing	0.21	0.04
CML Asian in HH: No	0.20	0.05
CML HoH Gender: Missing	-0.03	0.02
CML HoH Gender: Female	-0.01	0.02
CML HoH Income: $35k-$70k	0.12	0.02
CML HoH Income: less than $35k	0.14	0.02
CML HH Own/Rent: Missing	-0.06	0.03
CML HH Own/Rent: Owned	-0.02	0.02
CML Age of 2nd Person: Missing	-0.13	0.03
CML Age of 2nd Person: 18-44	-0.15	0.03
No Respondent Comments	0.08	0.04
Non-Contacts: None	-0.51	0.08
Non-Contacts: 1	-0.25	0.05
Non-Contacts: 2-4	-0.03	0.03
Occupancy Rate of PSU	-0.26	0.10
Respondent Other Concerns	0.18	0.06
Physical Impediment to Housing Unit: Locked	-0.35	0.03
Day of Quarter	0.01	<0.01
Respondent Concerns Expressed: None	-1.25	0.15
Respondent Concerns Expressed: Once	0.15	0.09
Single Family Home / Townhome	-0.22	0.03
Structure with 2-9 Units	-0.29	0.04
Structure with 10+ Units	-0.21	0.04
Respondent Concern: Survey Voluntary?	-0.46	0.15
Respondent Concern: Too Old	0.60	0.15

** CML denotes that the variable came from a commercial data source.*

Appendix B: Questionnaire for Expert Elicitation



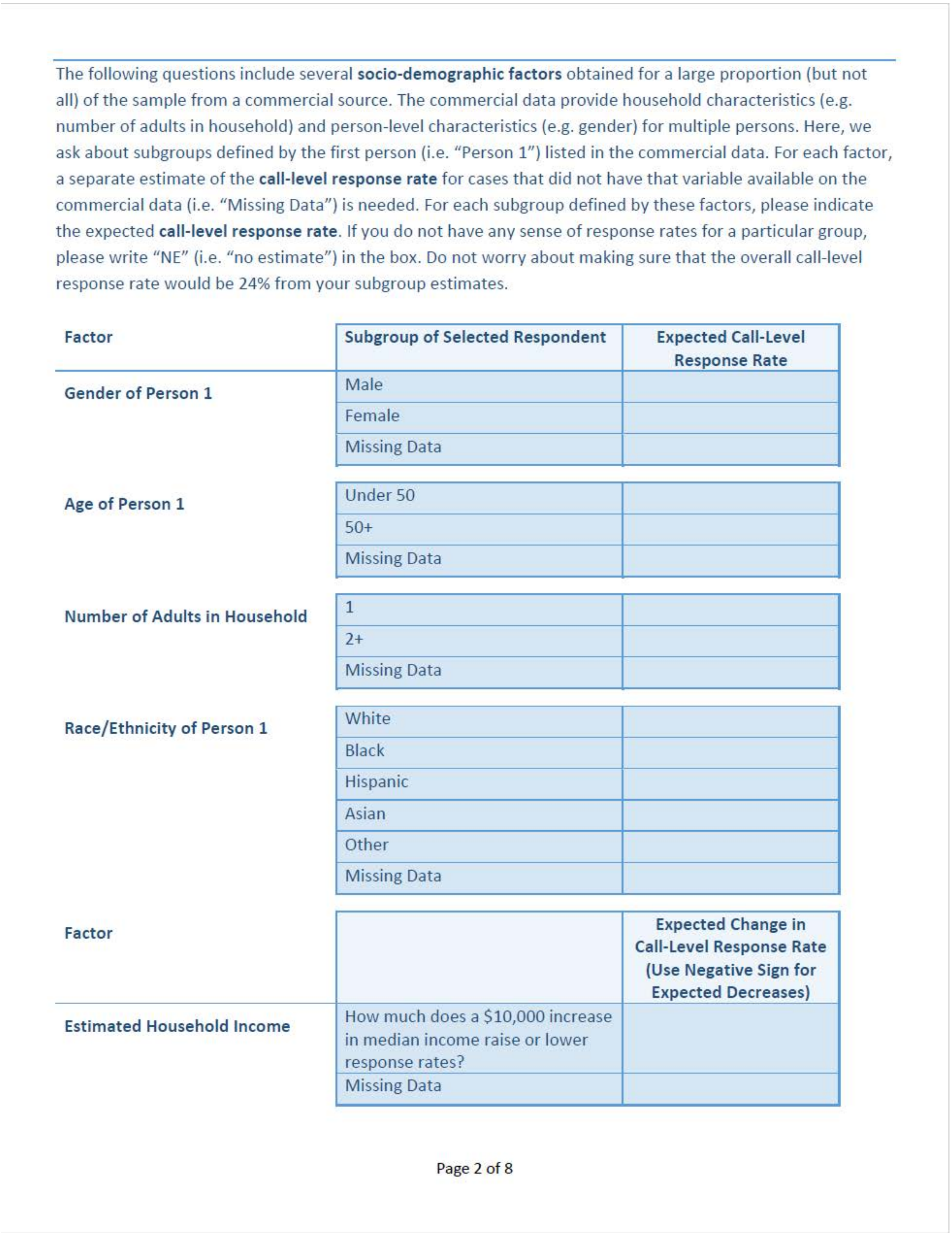
INSTITUTE FOR SOCIAL RESEARCH
SURVEY METHODOLOGY
UNIVERSITY OF MICHIGAN

The following questionnaire is designed to help us learn more about your expectations for response rates of subgroups that might differ from the overall response rate on a national health survey. The survey is administered on a quarterly basis. In this questionnaire, we are interested in your expectations for response rates to a short interview (15 minutes). The interview rosters the household and collects some demographic characteristics. The results of that interview are used as a sampling frame for other surveys. Further, we are interested in the expected response at the level of the call attempt. The following are important features of the survey design:

- A survey of persons 18 years of age and older
- Face-to-face interviewing
- Prenotification via standard mailed letter with government sponsorship
- A two-phase survey design with a special mailing for the second phase
- Overall (case-level) response rate expectation for this short interview of 90%. The average completion rate at the attempt level for eligible cases is about 24%. That is, about 24% of call attempts result in a completed interview.
- Overall (case-level) response rate expectation for the screening interview of 90%. The average completion rate at the attempt level for eligible cases is about 24%. That is, about 24% of call attempts result in a completed screening interview.

The subgroups are defined using a variety of different variables – demographic factors available for many cases from commercially-available data (e.g., gender, age, etc.), neighborhood characteristics (e.g., locked buildings, proportion of population 50+, etc.), housing unit features (e.g., single-family home vs. multi-unit building, presence of physical impediments such as an intercom, etc.), and paradata (e.g., number of previous attempts, ever resistance, etc.). In the questions that follow, please tell us what you would expect regarding the percentages of calls that would result in a completed interview for each subgroup indicated.

Page 1 of 8

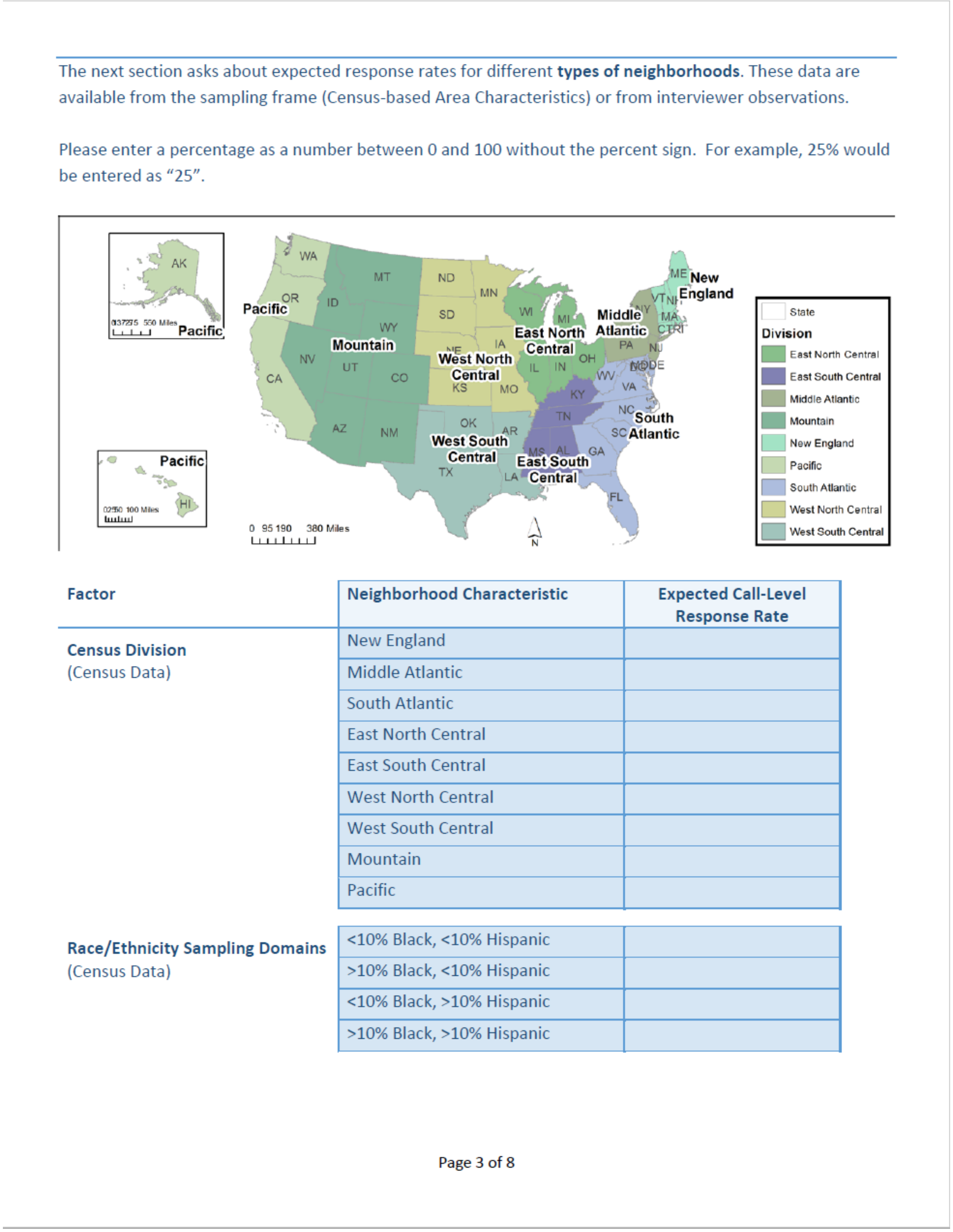


The following questions include several **socio-demographic factors** obtained for a large proportion (but not all) of the sample from a commercial source. The commercial data provide household characteristics (e.g. number of adults in household) and person-level characteristics (e.g. gender) for multiple persons. Here, we ask about subgroups defined by the first person (i.e. "Person 1") listed in the commercial data. For each factor, a separate estimate of the **call-level response rate** for cases that did not have that variable available on the commercial data (i.e. "Missing Data") is needed. For each subgroup defined by these factors, please indicate the expected **call-level response rate**. If you do not have any sense of response rates for a particular group, please write "NE" (i.e. "no estimate") in the box. Do not worry about making sure that the overall call-level response rate would be 24% from your subgroup estimates.

Factor	**Subgroup of Selected Respondent**	**Expected Call-Level Response Rate**
Gender of Person 1	Male	
	Female	
	Missing Data	
Age of Person 1	Under 50	
	50+	
	Missing Data	
Number of Adults in Household	1	
	2+	
	Missing Data	
Race/Ethnicity of Person 1	White	
	Black	
	Hispanic	
	Asian	
	Other	
	Missing Data	

Factor		**Expected Change in Call-Level Response Rate (Use Negative Sign for Expected Decreases)**
Estimated Household Income	How much does a $10,000 increase in median income raise or lower response rates?	
	Missing Data	

Page 2 of 8



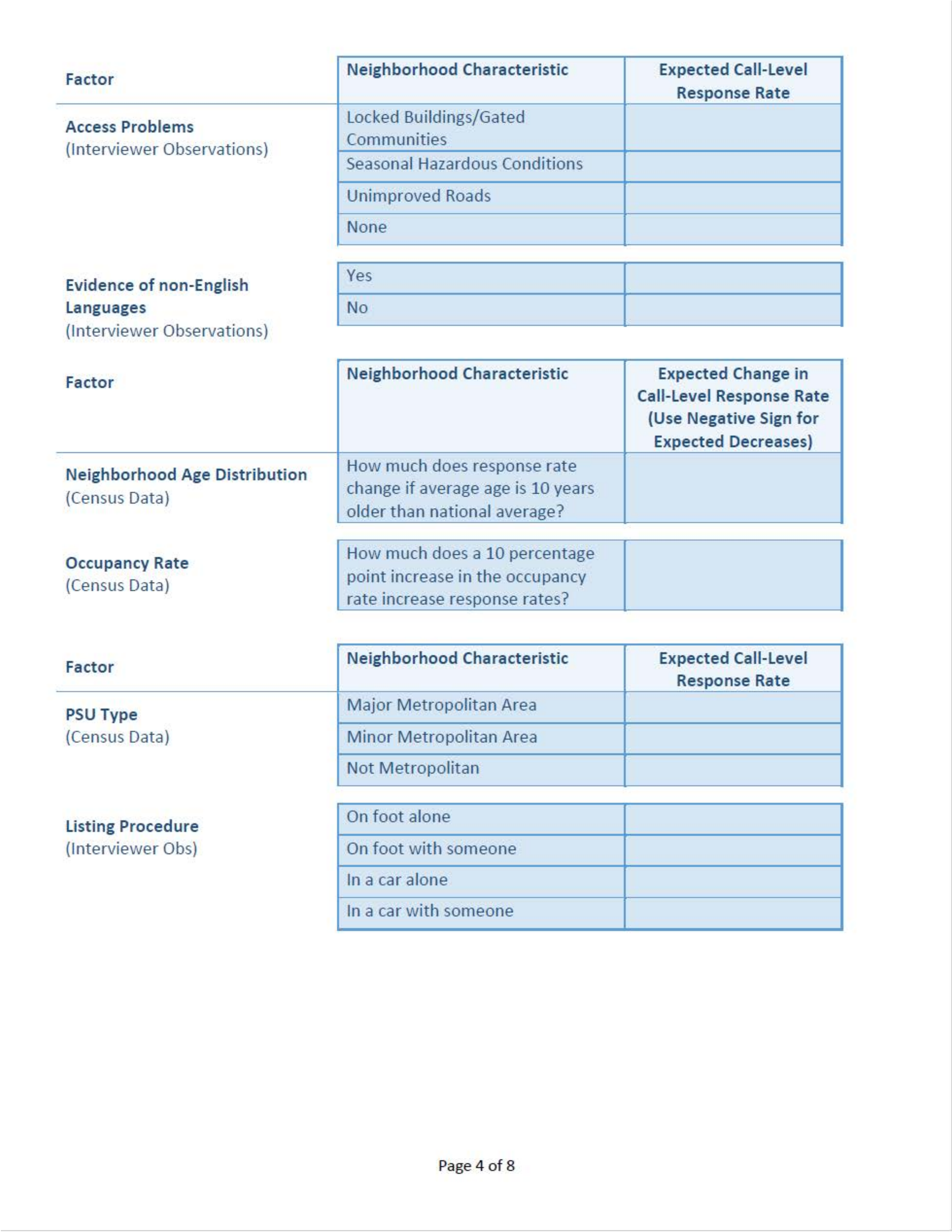
The next section asks about expected response rates for different **types of neighborhoods**. These data are available from the sampling frame (Census-based Area Characteristics) or from interviewer observations.

Please enter a percentage as a number between 0 and 100 without the percent sign. For example, 25% would be entered as "25".

AK
0 137 275 550 Miles
Pacific
WA
OR
Pacific
CA
ID
MT
WY
Mountain
NV
UT
CO
AZ
NM
ND
SD
NE
West North Central
KS
MN
IA
MO
OK
AR
West South Central
TX
LA
WI
MI
East North Central
IL
IN
OH
KY
TN
MS
AL
East South Central
Middle Atlantic
NY
PA
NJ
New England
ME
VT
NH
MA
CT
RI
MD
DE
WV
VA
NC
SC
GA
FL
South Atlantic
Pacific
HI
0 25 50 100 Miles
0 95 190 380 Miles
N
State
Division
East North Central
East South Central
Middle Atlantic
Mountain
New England
Pacific
South Atlantic
West North Central
West South Central

Factor	Neighborhood Characteristic	Expected Call-Level Response Rate
Census Division (Census Data)	New England	
	Middle Atlantic	
	South Atlantic	
	East North Central	
	East South Central	
	West North Central	
	West South Central	
	Mountain	
	Pacific	
Race/Ethnicity Sampling Domains (Census Data)	<10% Black, <10% Hispanic	
	>10% Black, <10% Hispanic	
	<10% Black, >10% Hispanic	
	>10% Black, >10% Hispanic	

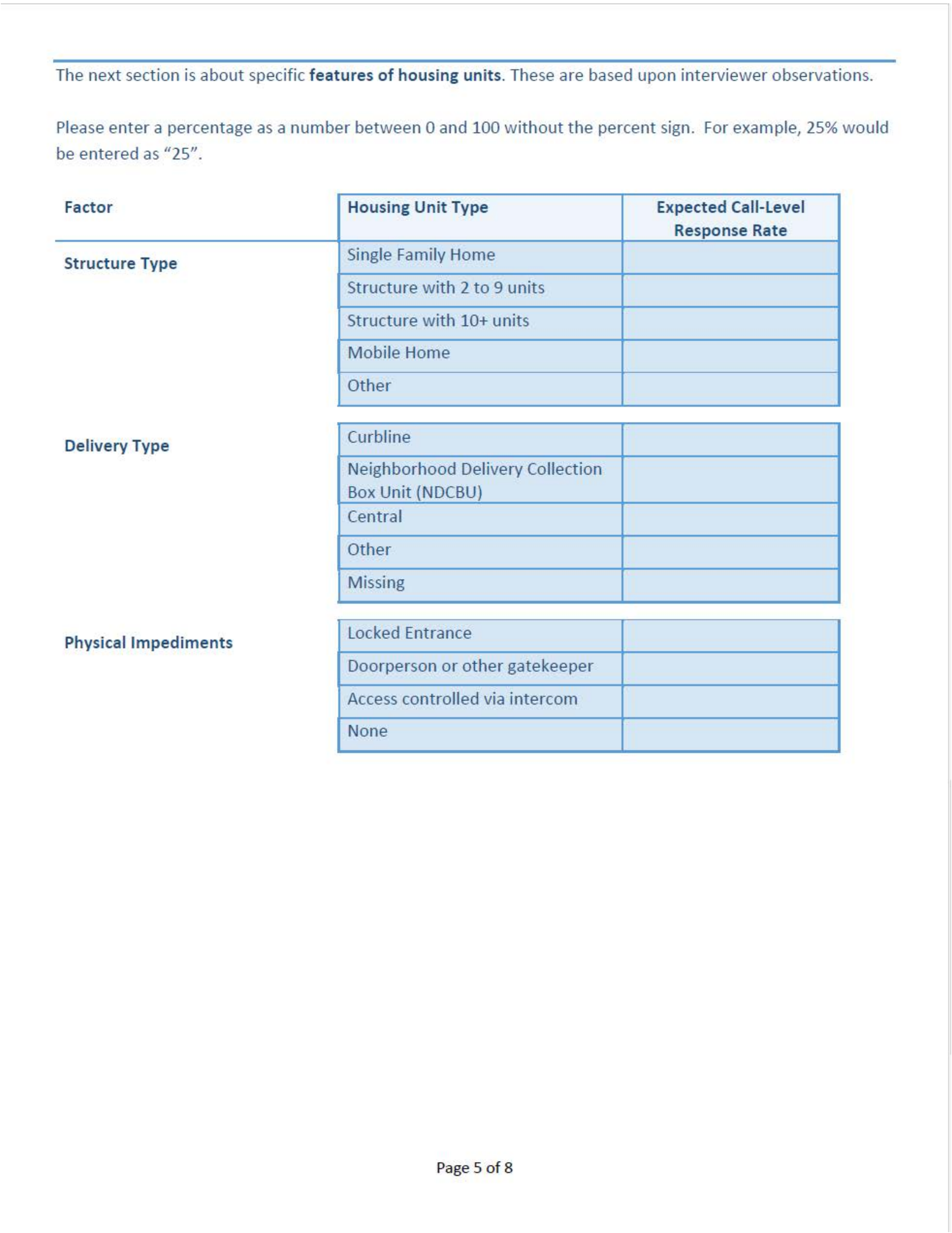
Page 3 of 8



Factor	Neighborhood Characteristic	Expected Call-Level Response Rate
Access Problems (Interviewer Observations)	Locked Buildings/Gated Communities	
	Seasonal Hazardous Conditions	
	Unimproved Roads	
	None	
Evidence of non-English Languages (Interviewer Observations)	Yes	
	No	

Factor	Neighborhood Characteristic	Expected Change in Call-Level Response Rate (Use Negative Sign for Expected Decreases)
Neighborhood Age Distribution (Census Data)	How much does response rate change if average age is 10 years older than national average?	
Occupancy Rate (Census Data)	How much does a 10 percentage point increase in the occupancy rate increase response rates?	

Factor	Neighborhood Characteristic	Expected Call-Level Response Rate
PSU Type (Census Data)	Major Metropolitan Area	
	Minor Metropolitan Area	
	Not Metropolitan	
Listing Procedure (Interviewer Obs)	On foot alone	
	On foot with someone	
	In a car alone	
	In a car with someone	

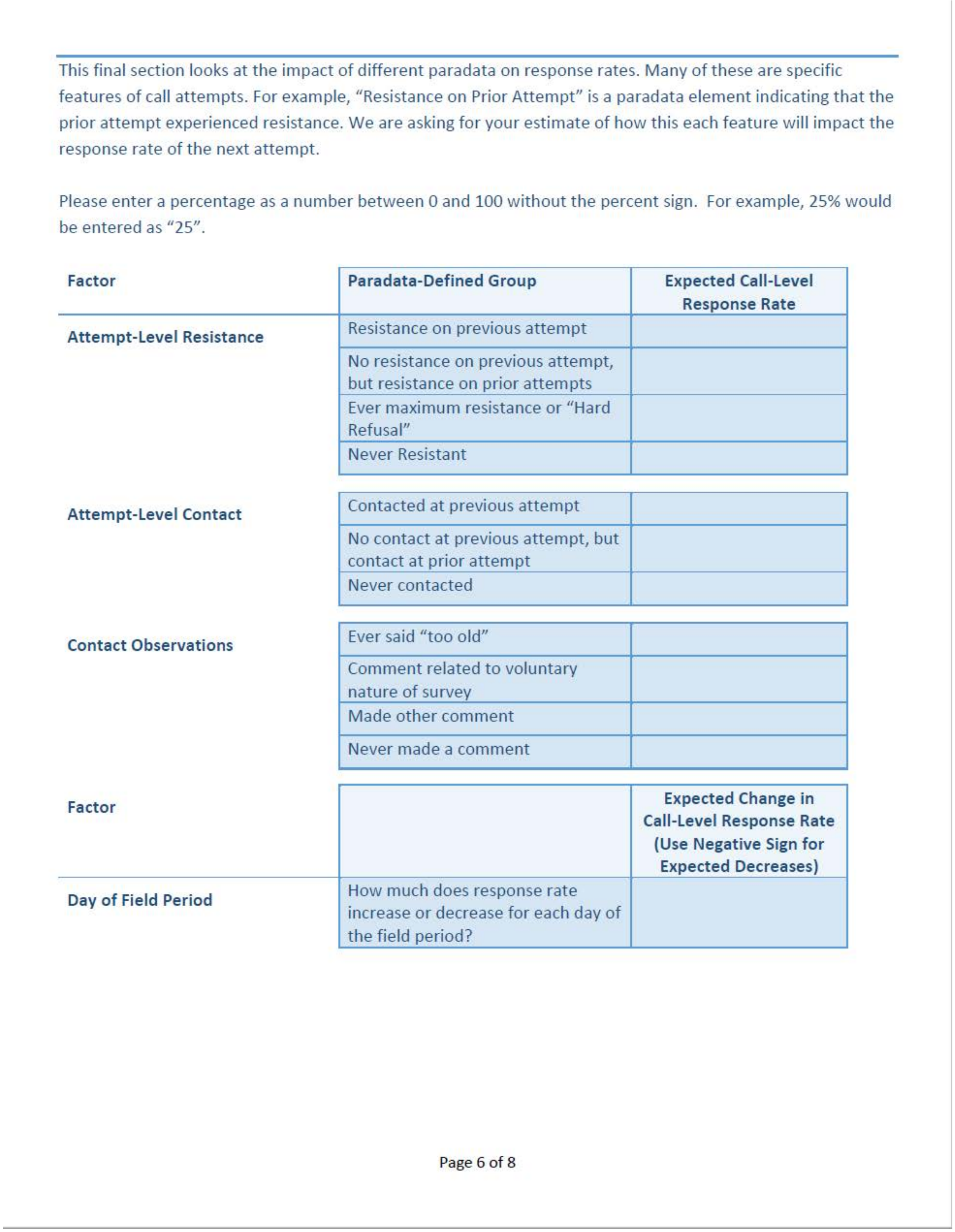


The next section is about specific **features of housing units**. These are based upon interviewer observations.

Please enter a percentage as a number between 0 and 100 without the percent sign. For example, 25% would be entered as "25".

Factor	Housing Unit Type	Expected Call-Level Response Rate
Structure Type	Single Family Home	
	Structure with 2 to 9 units	
	Structure with 10+ units	
	Mobile Home	
	Other	
Delivery Type	Curbline	
	Neighborhood Delivery Collection Box Unit (NDCBU)	
	Central	
	Other	
	Missing	
Physical Impediments	Locked Entrance	
	Doorperson or other gatekeeper	
	Access controlled via intercom	
	None	

Page 5 of 8

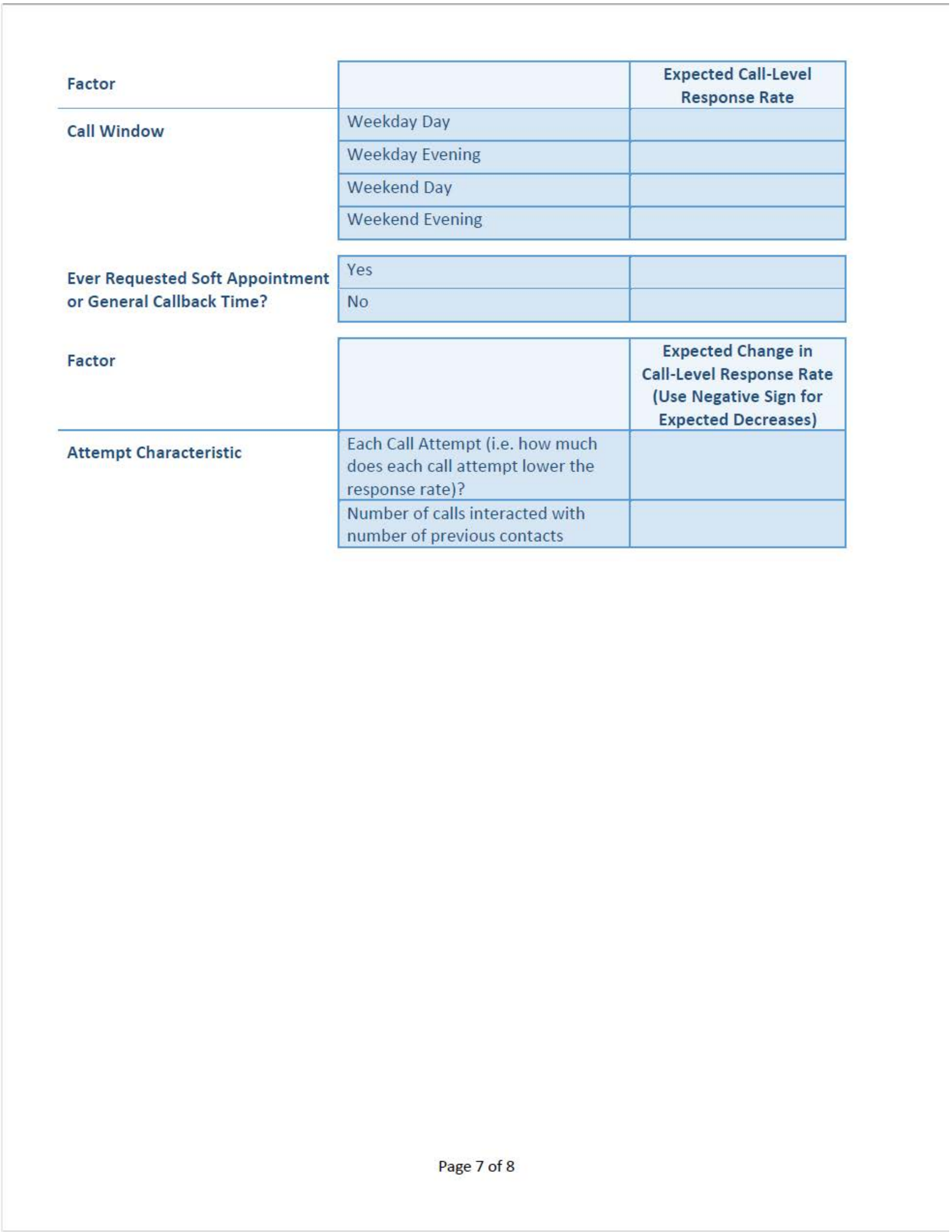


This final section looks at the impact of different paradata on response rates. Many of these are specific features of call attempts. For example, "Resistance on Prior Attempt" is a paradata element indicating that the prior attempt experienced resistance. We are asking for your estimate of how this each feature will impact the response rate of the next attempt.

Please enter a percentage as a number between 0 and 100 without the percent sign. For example, 25% would be entered as "25".

Factor	**Paradata-Defined Group**	**Expected Call-Level Response Rate**
Attempt-Level Resistance	Resistance on previous attempt	
	No resistance on previous attempt, but resistance on prior attempts	
	Ever maximum resistance or "Hard Refusal"	
	Never Resistant	
Attempt-Level Contact	Contacted at previous attempt	
	No contact at previous attempt, but contact at prior attempt	
	Never contacted	
Contact Observations	Ever said "too old"	
	Comment related to voluntary nature of survey	
	Made other comment	
	Never made a comment	
Factor		**Expected Change in Call-Level Response Rate (Use Negative Sign for Expected Decreases)**
Day of Field Period	How much does response rate increase or decrease for each day of the field period?	

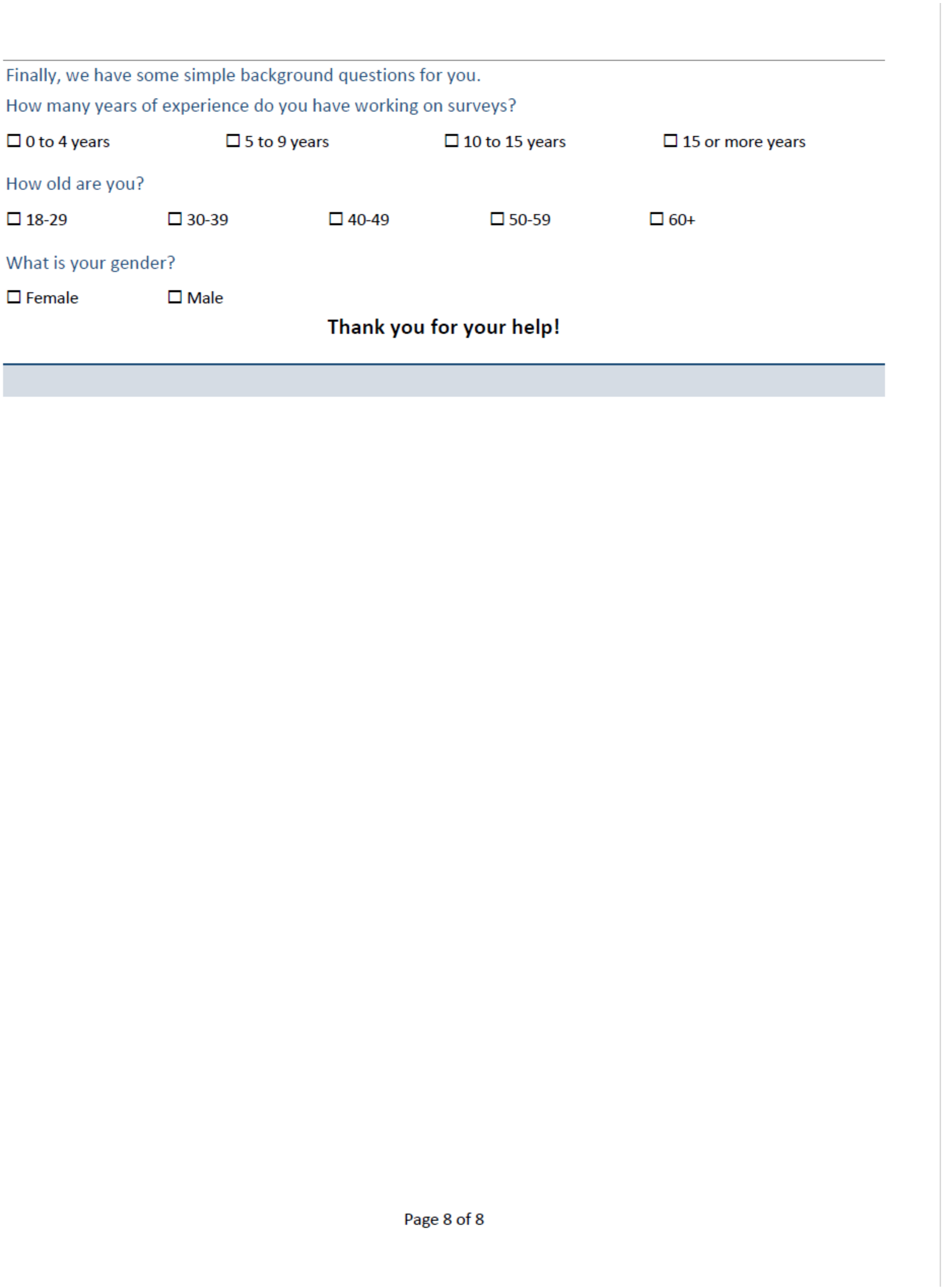
Page 6 of 8



Factor		Expected Call-Level Response Rate
Call Window	Weekday Day	
	Weekday Evening	
	Weekend Day	
	Weekend Evening	

Ever Requested Soft Appointment or General Callback Time?	Yes	
	No	

Factor		Expected Change in Call-Level Response Rate (Use Negative Sign for Expected Decreases)
Attempt Characteristic	Each Call Attempt (i.e. how much does each call attempt lower the response rate)?	
	Number of calls interacted with number of previous contacts	



Finally, we have some simple background questions for you.

How many years of experience do you have working on surveys?

□ 0 to 4 years □ 5 to 9 years □ 10 to 15 years □ 15 or more years

How old are you?

□ 18-29 □ 30-39 □ 40-49 □ 50-59 □ 60+

What is your gender?

□ Female □ Male

Thank you for your help!

Page 8 of 8

Appendix C: Coefficients and Standard Errors of Priors Based on Expert Elicitation

Appendix C displays the standard normal prior definitions, $\left(\hat{\hat{\beta}}_{jk}, SE\left(\hat{\hat{\beta}}_{jk}\right)\right)$, for the same predictors included in the NSFG response propensity model described in Appendix A. The table notes which categories served as reference categories in the prior generation process, and also notes how many responses (out of a maximum of 20) that we received for each category.

	All Respondents (max n = 20)		
Questions and Categories	**Count of Responses**	**Mean Beta**	**StdErr Beta**
Gender of Primary Householder (vs. Male)			
Female	20	0.336	0.063
Missing	14	-0.465	0.257
Age of Primary Householder (vs. 50 or Over)			
< 50	20	-0.370	0.108
Missing	15	-0.831	0.293
Number of Adults in HH (vs. 2 or More)			
1	20	0.066	0.198
Missing	12	-0.732	0.219
Race/Ethnicity of Primary Householder (vs. Asian)			
White	18	0.532	0.121
Black	18	-0.031	0.173
Hispanic	18	-0.118	0.112
Other	13	-0.348	0.233
Missing	12	-0.326	0.292
Household Income Effect			
+$10,000	17	0.466	0.235

Questions and Categories	All Respondents (max n = 20)		
	Count of Responses	Mean Beta	StdErr Beta
Masked Census Division (vs. Region I)			
G	14	0.020	0.129
B	14	-0.205	0.138
D	14	0.041	0.141
H	14	0.060	0.161
C	14	0.133	0.170
F	15	0.294	0.150
E	15	0.057	0.145
A	16	-0.050	0.192
Race/Ethnicity Sampling Domain (vs. > 10% Black, > 10% Hispanic)			
< 10% Black, < 10% Hispanic	16	0.696	0.202
> 10% Black, < 10% Hispanic	16	0.535	0.132
< 10% Black, > 10% Hispanic	16	0.364	0.143
Access Problems (vs. Other)			
Locked Buildings/Gated Communities	19	-0.687	0.190
Seasonal Hazardous Conditions	18	-0.418	0.153
Unimproved Roads	17	0.267	0.164
None	10	1.091	0.189
Evidence of Non-English Languages (vs. No)			
Yes	15	-0.725	0.163
Neighborhood Age Effect			
10 years older than national average	17	0.520	0.099
Occupancy Rate Effect			
10% increase in occupancy rates	16	0.187	0.170
PSU Type (vs. Major Metropolitan Area)			
Minor Metropolitan Area	18	0.155	0.155
Not Metropolitan	17	0.398	0.158
Listing Procedure (vs. On Foot Alone)			
On Foot With Someone	11	0.787	0.607
In a Car Alone	11	-0.066	0.135
In a Car With Someone	11	0.795	0.614
Structure Type (vs. Other)			
Single Family Home	5	1.172	0.567
Structure with 2-9 Units	5	0.788	0.602
Structure with 10+ Units	5	0.600	0.617
Mobile Home	5	0.728	0.462

	All Respondents (max n = 20)		
Questions and Categories	**Count of Responses**	**Mean Beta**	**StdErr Beta**
Delivery Type (vs. Other)			
Curbline	3	0.917	0.590
Neighborhood Delivery Collection Box	3	0.199	0.289
Central	3	0.069	0.384
Missing	3	0.000	0.000
Physical Impediments (vs. Other)			
Locked Entrance	19	-0.096	0.206
Doorperson or Gatekeeper	19	-0.627	0.117
Access controlled via Intercom	19	-0.371	0.106
None	14	1.076	0.155
Attempt-Level Concerns Expressed (vs. No Concerns)			
Concerns Expressed on Previous Attempt	17	-1.347	0.434
Concerns Expressed Not on Previous but Prior Attempt	17	-1.451	0.244
Strong Concerns Ever Expressed	15	-2.228	0.593
Attempt-Level Contact (vs. Never Contacted)			
Contacted at Previous Attempt	15	1.367	0.329
Not Previous but Prior Contact	15	1.009	0.298
Contact Observations (vs. Other)			
Ever Said "Too Old"	14	-0.532	0.336
Comment re: Voluntary Nature of Survey	17	0.335	0.489
Any Other Comments	14	0.118	0.182
Never Made Comment	13	0.325	0.205
Day of Field Period Effect			
Change in RR for Each Day of Field Period	12	0.213	0.078
Call Window (vs. Weekday Day)			
Weekday Evening	19	1.203	0.193
Weekend Day	19	1.052	0.166
Weekend Evening	19	0.426	0.220
Ever Requested Call-Back/Soft Appointment (vs. No)			
Yes	18	0.564	0.339
Contact Attempt Effect			
Change in RR for Each Additional Contact	17	-0.058	0.109
*Contact*Contact Interaction Effect*			
Change in RR for Each Add'l Call*Contact	13	0.177	0.228

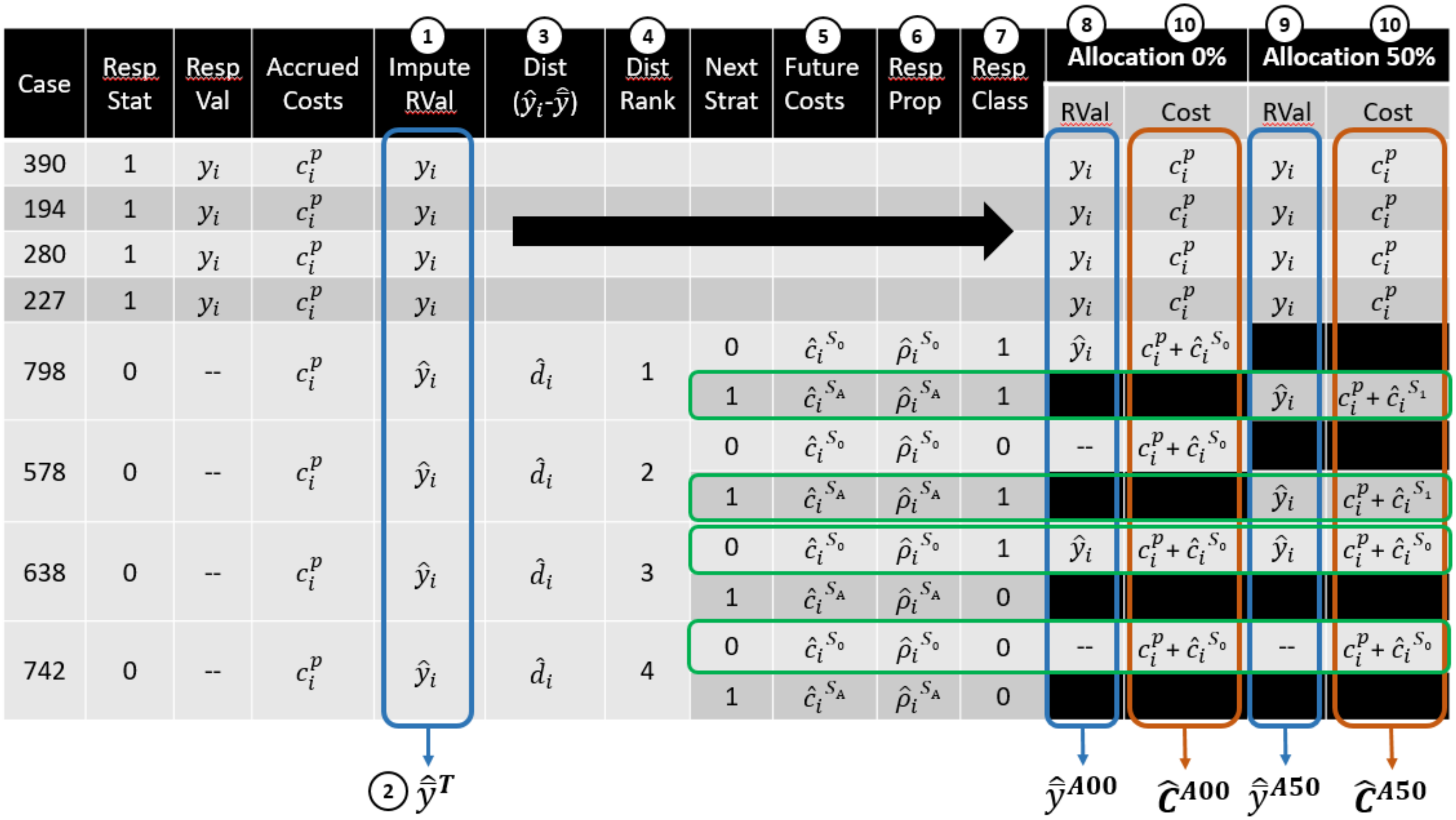
Appendix D: Coefficients and Standard Errors from Historical (2017) Data for Predicting Response Propensity in the NSCG

Variable Name	Level	DF	Estimate	SE	p-value
Intercept		1	-3.3751	0.4065	<.0001
Age Group: 35-39	1	1	0.0834	0.0489	0.0880
Age Group: 40-44	1	1	0.2031	0.0505	<.0001
Age Group: 45-49	1	1	0.0824	0.0519	0.1120
Age Group: 50-54	1	1	0.1609	0.053	0.0024
Age Group: 55-59	1	1	0.2724	0.0554	<.0001
Age Group: 60-64	1	1	0.2419	0.0589	<.0001
Age Group: 65-69	1	1	0.4631	0.0754	<.0001
Age Group: 70-75	1	1	0.5735	0.1064	<.0001
Demographic Group: Hispanic	1	1	-0.1747	0.0466	0.0002
Demographic Group: Black	1	1	-0.2615	0.0451	<.0001
Demographic Group: Asian	1	1	-0.198	0.0425	<.0001
Demographic Group: non-USCAB, Hispanic	1	1	-0.2948	0.0869	0.0007
Demographic Group: non-USCAB, Asian	1	1	-0.1745	0.0473	0.0002
Demographic Group: non-USCAB, Other	1	1	-0.244	0.0503	<.0001
Census Division: Other	1	1	0.1992	0.1431	0.1641
Census Division: West	1	1	0.2373	0.0593	<.0001
Census Division: East South Central	1	1	0.1483	0.0718	0.0389
Census Division: Mountain	1	1	0.0824	0.0556	0.1382
Masters - Highest Degree Held	1	1	0.1758	0.0289	<.0001
Doctorate – Highest Degree Held	1	1	0.3743	0.0718	<.0001
Non-S&E Degree	1	1	-0.0886	0.0348	0.0110
Works 30 Hours or Less per Week	1	1	-0.0588	0.0377	0.1187
Works 50 Hours or More per Week	1	1	-0.1204	0.0467	0.010
Indicator for Veteran’s Service	1	1	0.1379	0.0672	0.040
Indicator for No Internet Access at Home	1	1	-0.1449	0.0743	0.0512
Missing Information about Internet Access	1	1	-0.7603	0.1771	<.0001
Responded to the ACS in Paper	1	1	-0.2297	0.0338	<.0001
Responded to ACS in CATI	1	1	-0.431	0.0664	<.0001
Responded to ACS by Personal Visit	1	1	-0.4938	0.0441	<.0001
Indicator for Incentive Sent at Week 0	1	1	0.0972	0.04	0.0152
One Trip to Research for New Contact Info	1	1	-1.1251	0.0626	<.0001
Two or More Trips to Research for New Info	1	1	-1.5734	0.142	<.0001
Indicator for a Refusal During Operations	1	1	-1.5515	0.0741	<.0001

Appendix E: Coefficients and Standard Errors from Historical (2017) Data for Predicting Cube Root of Salary in the NSCG

Variable Name	Level	DF	Estimate	SE	p-value
Intercept		1	24.868	0.2276	<.0001
Responded to ACS in CATI	1	1	-1.022	0.2198	0.0294
Responded to ACS by Personal Visit	1	1	-0.552	0.0739	0.0422
Responded to ACS by Group Qtr Operation	1	1	-4.631	2.2038	0.0018
Field of Degree: Computer/Math Sciences	1	1	0.864	0.0825	0.0026
Field of Degree: Engineering	1	1	1.29	0.0405	<.0001
Demographic Group: Asian	1	1	0.828	0.0668	0.0014
Indicator for Disability	1	1	-1.019	0.1361	0.0057
Census Division: Pacific	1	1	0.478	0.0421	0.0197
Income in 3rd Quintile of ACS Respondents	1	1	1.48	0.0638	<.0001
Income in 4th Quintile of ACS Respondents	1	1	3.7	0.0824	<.0001
Income in 4th Quintile of ACS Respondents	1	1	6.735	0.1449	<.0001
Masters - Highest Degree Held	1	1	0.424	0.0279	0.0112
Doctorate – Highest Degree Held	1	1	0.897	0.1067	0.006
Occupation Code: Computer Scientist	1	1	0.793	0.1119	0.0178
Occupation Code: Non S&E High Interest	1	1	1.108	0.0693	<.0001
Non-S&E Occupation	1	1	-0.743	0.0449	0.0005
Female	1	1	-0.807	0.0292	<.0001
Living Outside Central City and MSA	1	1	-1.823	0.082	<.0001
Not Working & Not Looking for Work	1	1	-0.317	0.0989	0.3131
Poverty Indicator	1	1	3.388	0.2007	<.0001
Private Healthcare Coverage	1	1	-1.837	0.1242	<.0001
Public Healthcare Coverage	1	1	-1.228	0.146	0.0013
Home is Owned with No Mortgage	1	1	-0.876	0.0558	0.0002
No Vehicles at the Housing Unit	1	1	0.885	0.1422	0.0189
Worked 40-47 Hours per Week in Last Year	1	1	-0.786	0.1114	0.0186
Accesses Internet with DSL	1	1	-0.509	0.0354	0.0068
No Handheld Computer in Housing Unit	1	1	-0.746	0.079	0.0079
Four Bedrooms in Housing Unit	1	1	0.666	0.0342	0.0003
Five or More Bedrooms in Housing Unit	1	1	1.056	0.0922	0.0005
Works 30 Hours or Less per Week	1	1	-2.617	0.0715	<.0001
Cube Root of Personal Income	1	1	0.274	0.0003	<.0001
Cube Root of Wages	1	1	0.082	0.0002	<.0001
Cube Root of Retirement Income	1	1	-0.174	0.0002	<.0001

Appendix F: Illustration of Optimization Steps at Intervention Points



Case	Resp Stat	Resp Val	Accrued Costs	(1) Impute RVal	(3) Dist $(\hat{y}_i\text{-}\hat{\bar{y}})$	(4) Dist Rank	Next Strat	(5) Future Costs	(6) Resp Prop	(7) Resp Class	(8) Allocation 0%	(10) Allocation 0%	(9) Allocation 50%	(10) Allocation 50%
											RVal	Cost	RVal	Cost
390	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
194	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
280	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
227	1	y_i	c_i^p	y_i							y_i	c_i^p	y_i	c_i^p
798	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	1	0	$\hat{c}_i{}^{S_0}$	$\hat{\rho}_i{}^{S_0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S_0}$		
							1	$\hat{c}_i{}^{S_A}$	$\hat{\rho}_i{}^{S_A}$	1			$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S_1}$
578	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	2	0	$\hat{c}_i{}^{S_0}$	$\hat{\rho}_i{}^{S_0}$	0	--	$c_i^p + \hat{c}_i{}^{S_0}$		
							1	$\hat{c}_i{}^{S_A}$	$\hat{\rho}_i{}^{S_A}$	1			$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S_1}$
638	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	3	0	$\hat{c}_i{}^{S_0}$	$\hat{\rho}_i{}^{S_0}$	1	$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S_0}$	$\hat{y}_i$	$c_i^p + \hat{c}_i{}^{S_0}$
							1	$\hat{c}_i{}^{S_A}$	$\hat{\rho}_i{}^{S_A}$	0				
742	0	--	c_i^p	$\hat{y}_i$	$\hat{d}_i$	4	0	$\hat{c}_i{}^{S_0}$	$\hat{\rho}_i{}^{S_0}$	0	--	$c_i^p + \hat{c}_i{}^{S_0}$	--	$c_i^p + \hat{c}_i{}^{S_0}$
							1	$\hat{c}_i{}^{S_A}$	$\hat{\rho}_i{}^{S_A}$	0				

Appendix G: Code for Generating Predictions of Expected Lag in Chapter 2

```
###############################################################################
# Predictions of Expected Lag Based Using Four Methodologies 2:34 PM 10/26/2020
###############################################################################

lagfunc1 <- function(b, d, eligday, b0033, b0050, b0100, b0200, b0300, w0033,
       w0050, w0100, w0200, w0300, outstuff)
{
  #########################################################################
  #Create 3 month evaluaton dataset and 1 month prediction dataset
  #3-month evaluation dataset comes from summary dataset
  #1-month prediction dataset comes from the contact-level dataset

  s <- b
  f <- s+2
  p <- s+3

  evalfile <- subset(summary, t >= s & t <= f)
  predbase <- subset(contact, t == p)

  #Create calday shifts for graphing boxplots
  predbase$shift_act <- predbase$calday + 0
  predbase$shift_m1 <- predbase$calday + 0.15
  predbase$shift_m2 <- predbase$calday + 0.30
  predbase$shift_m3 <- predbase$calday + 0.45
  predbase$shift_m4_033 <- predbase$calday + 0.60
  predbase$shift_m4_050 <- predbase$calday + 0.75
  predbase$shift_m4_100 <- predbase$calday + 0.90
  predbase$shift_m4_200 <- predbase$calday + 1.05
  predbase$shift_m4_300 <- predbase$calday + 1.20
  head(predbase)

  #Assign priors
  pr_bin_0033 <- b0033
  pr_bin_0050 <- b0050
  pr_bin_0100 <- b0100
  pr_bin_0200 <- b0200
  pr_bin_0300 <- b0300
  pr_wei_0033 <- w0033
  pr_wei_0050 <- w0050
  pr_wei_0100 <- w0100
  pr_wei_0200 <- w0200
  pr_wei_0300 <- w0300

  #########################################################################
  #Prediction Method #1: Prior 3 Month Mean (Single Estimate)
  mean_3mo <- mean(evalfile$lag_att_con)
  predbase$m1_estlag <- mean_3mo
  predbase$m1_res <- predbase$m1_estlag - predbase$lag_att_con
  predbase$m1_resabs <- abs(predbase$m1_estlag - predbase$lag_att_con)

  #########################################################################
  #Create datasets for Day d File
  predday <- subset(predbase, eval(as.name(eligday)) == 1)
  predday$daylag <- ifelse(predday$conday > d, (d - predday$calday + 1),
       predday$lag_att_con)
  predday$daycens <- ifelse(predday$conday > d, 1, 0)
```

```
#########################################################################
#Prediction Method #2: Prior 3 Month Regression Parameters (Single Set of
     Parameters, Apply to Each Day)

#Logistic Regression Predicting Positive Lag from Historical Parameters
p_poslag_m2 <- glm(poslag ~ pct_College_ACS_10_14 +
                pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                pct_Mobile_Homes_ACS_10_14 +
                expbas0_1 + calday +
                pct_Vacant_Units_ACS_10_14*calday,
                data = evalfile, family = "binomial")
summary(p_poslag_m2)

#Weibull Regression Predicting Lag Length from Historical Parameters
p_laglngth_m2 <- survreg(Surv(lag_att_con,status) ~ wkldro_22 + wkldro_23 +
                  wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
                  BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
                  data=subset(evalfile, lag_att_con > 0),
                  na.action=na.omit, dist="weibull", model=FALSE, x=FALSE,
                  y=TRUE, score=TRUE)
summary(p_laglngth_m2)
1/p_laglngth_m2$scale

#########################################################################
#Method 2: Score attempted cases with historical model parameters
predday$m2_poslag <- predict(p_poslag_m2, predday, type="response")
predday$m2_laglngth <- predict(p_laglngth_m2, predday, type="response")
#Calculate Expected lag (Pr(poslag)*E(lag)) and residuals
predday$m2_estlag <- predday$m2_poslag*predday$m2_laglngth
predday$m2_res <- predday$m2_estlag - predday$lag_att_con
predday$m2_resabs <- abs(predday$m2_estlag - predday$lag_att_con)
head(predday)

#########################################################################
#Prediction Methods #3 & #4: Applied Daily
# #3: Current Month Regression Parameters (Estimate per Day of Interest)
# #4: Fully Bayesian Estimates of Regression Parameters (Estimate per Day of
     Interest)
#########################################################################
#Method 3: Predict lag using only current data
p_poslag_m3 <- glm(poslag ~ pct_College_ACS_10_14 +
                pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                pct_Vacant_Units_ACS_10_14*calday, data = predday,
                family = "binomial")
summary(p_poslag_m3)
predday$m3_poslag <- predict(p_poslag_m3, predday, type="response")
p_weilag_m3 <- survreg(Surv(daylag,daycens) ~ wkldro_22 + wkldro_23 +
                wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
                BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
                data=subset(predday, daylag > 0), na.action=na.omit,
                dist="weibull", model=FALSE, x=FALSE, y=TRUE, score=TRUE)
summary(p_weilag_m3)
predday$m3_laglngth <- predict(p_weilag_m3, predday, type="response")
#Calculate Expected Lag and Residuals
predday$m3_estlag <- predday$m3_poslag*predday$m3_laglngth
```

```
predday$m3_res <- predday$m3_estlag - predday$lag_att_con
predday$m3_resabs <- abs(predday$m3_estlag - predday$lag_att_con)
head(predday)
###########################################################################
#Method 4: Predict lag using Bayesian posteriors combining historical and
     current data
#Positive Lag Prediction -- 5 values of "c"
p_poslag_0033 <- brm(poslag ~ pct_College_ACS_10_14 +
                  pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                  pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                  pct_Vacant_Units_ACS_10_14*calday , data = predday,
                  family = bernoulli(), prior <- pr_bin_0033, chains = 3,
                  iter = 4000, warmup = 2000)
summary(p_poslag_0033)
assign("p_poslag_0033", p_poslag_0033, envir = .GlobalEnv)
p_poslag_0050 <- brm(poslag ~ pct_College_ACS_10_14 +
                  pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                  pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                  pct_Vacant_Units_ACS_10_14*calday , data = predday,
                  family = bernoulli(),prior <- pr_bin_0050,
                  chains = 3, iter = 4000, warmup = 2000)
summary(p_poslag_0050)
assign("p_poslag_0050", p_poslag_0050, envir = .GlobalEnv)

p_poslag_0100 <- brm(poslag ~ pct_College_ACS_10_14 +
                  pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                  pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                  pct_Vacant_Units_ACS_10_14*calday , data = predday,
                  family = bernoulli(),prior <- pr_bin_0100,
                  chains = 3, iter = 4000, warmup = 2000)
summary(p_poslag_0100)
assign("p_poslag_0100", p_poslag_0100, envir = .GlobalEnv)
p_poslag_0200 <- brm(poslag ~ pct_College_ACS_10_14 +
                  pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                  pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                  pct_Vacant_Units_ACS_10_14*calday , data = predday,
                  family = bernoulli(),prior <- pr_bin_0200,
                  chains = 3, iter = 4000, warmup = 2000)
summary(p_poslag_0200)
assign("p_poslag_0200", p_poslag_0200, envir = .GlobalEnv)
p_poslag_0300 <- brm(poslag ~ pct_College_ACS_10_14 +
                  pct_Vacant_Units_ACS_10_14 + pct_No_Health_Ins_ACS_10_14 +
                pct_Not_HS_Grad_ACS_10_14 + pct_URBANIZED_AREA_POP_CEN_2010 +
                  pct_Mobile_Homes_ACS_10_14 + expbas0_1 + calday +
                  pct_Vacant_Units_ACS_10_14*calday , data = predday,
                  family = bernoulli(),prior <- pr_bin_0300,
                  chains = 3, iter = 4000, warmup = 2000)
summary(p_poslag_0300)
assign("p_poslag_0300", p_poslag_0300, envir = .GlobalEnv)

#Predict Estimated Response Propensities based on c value
predday$m4_poslag_0033 <- predict(p_poslag_0033, predday,
     type="response")[,1]
```

```
predday$m4_poslag_0050 <- predict(p_poslag_0050, predday,
     type="response")[,1]
predday$m4_poslag_0100 <- predict(p_poslag_0100, predday,
     type="response")[,1]
predday$m4_poslag_0200 <- predict(p_poslag_0200, predday,
     type="response")[,1]
predday$m4_poslag_0300 <- predict(p_poslag_0300, predday,
     type="response")[,1]
head(predday)

#Positive Lag Prediction -- 5 values of "c"
p_weilag_0033 <- brm(daylag|cens(daycens) ~ wkldro_22 + wkldro_23 +
     wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
     BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
     data = subset(predday, daylag > 0),
     family = weibull(link = "log", link_shape = "log"),
     prior <- pr_wei_0033, chains=3, iter=4000, warmup = 2000)
summary(p_weilag_0033)
assign("p_weilag_0033", p_weilag_0033, envir = .GlobalEnv)
p_weilag_0050 <- brm(daylag|cens(daycens) ~  wkldro_22 + wkldro_23 +
     wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
     BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
     data = subset(predday, daylag > 0),
     family = weibull(link = "log", link_shape = "log"),
     prior <- pr_wei_0050, chains=3, iter=4000, warmup = 2000)
summary(p_weilag_0050)
assign("p_weilag_0050", p_weilag_0050, envir = .GlobalEnv)
p_weilag_0100 <- brm(daylag|cens(daycens) ~ wkldro_22 + wkldro_23 +
     wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
     BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
     data = subset(predday, daylag > 0),
     family = weibull(link = "log", link_shape = "log"),
     prior <- pr_wei_0100, chains=3, iter=4000, warmup = 2000)
summary(p_weilag_0100)
assign("p_weilag_0100", p_weilag_0100, envir = .GlobalEnv)
p_weilag_0200 <- brm(daylag|cens(daycens) ~ wkldro_22 + wkldro_23 +
     wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
     BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
     data = subset(predday, daylag > 0),
     family = weibull(link = "log", link_shape = "log"),
     prior <- pr_wei_0200, chains=3, iter=4000, warmup = 2000)
summary(p_weilag_0200)
assign("p_weilag_0200", p_weilag_0200, envir = .GlobalEnv)
p_weilag_0300 <- brm(daylag|cens(daycens) ~ wkldro_22 + wkldro_23 +
     wkldro_25 + wkldro_29 + wkldro_31 + calday + prior_FR_1p +
     BARS_1 + WHEELCHAIR_1 + HHAGE_2 + ADDR_COND_2 + ACCESS_1,
     data = subset(predday, daylag > 0),
     family = weibull(link = "log", link_shape = "log"),
     prior <- pr_wei_0300, chains=3, iter=4000, warmup = 2000)
summary(p_weilag_0300)
assign("p_weilag_0300", p_weilag_0300, envir = .GlobalEnv)


#Predict Estimated Positive Lag Lengths based on c value
predday$m4_laglngth_0033 <- predict(p_weilag_0033, predday,
     type="response")[,1]
predday$m4_laglngth_0050 <- predict(p_weilag_0050, predday,
     type="response")[,1]
```

```
predday$m4_laglngth_0100 <- predict(p_weilag_0100, predday,
     type="response")[,1]
predday$m4_laglngth_0200 <- predict(p_weilag_0200, predday,
     type="response")[,1]
predday$m4_laglngth_0300 <- predict(p_weilag_0300, predday,
     type="response")[,1]
head(predday)

#Calculate estimated lag from propensity and predicted lag
predday$m4_estlag_0033 <- predday$m4_poslag_0033*predday$m4_laglngth_0033
predday$m4_estlag_0050 <- predday$m4_poslag_0050*predday$m4_laglngth_0050
predday$m4_estlag_0100 <- predday$m4_poslag_0100*predday$m4_laglngth_0100
predday$m4_estlag_0200 <- predday$m4_poslag_0200*predday$m4_laglngth_0200
predday$m4_estlag_0300 <- predday$m4_poslag_0300*predday$m4_laglngth_0300
#Calculate residuals
predday$m4_res_0033 = predday$m4_estlag_0033 - predday$lag_att_con
predday$m4_res_0050 = predday$m4_estlag_0050 - predday$lag_att_con
predday$m4_res_0100 = predday$m4_estlag_0100 - predday$lag_att_con
predday$m4_res_0200 = predday$m4_estlag_0200 - predday$lag_att_con
predday$m4_res_0300 = predday$m4_estlag_0300 - predday$lag_att_con
#Calculate absolute value of residuals
predday$m4_resabs_0033 = abs(predday$m4_estlag_0033 - predday$lag_att_con)
predday$m4_resabs_0050 = abs(predday$m4_estlag_0050 - predday$lag_att_con)
predday$m4_resabs_0100 = abs(predday$m4_estlag_0100 - predday$lag_att_con)
predday$m4_resabs_0200 = abs(predday$m4_estlag_0200 - predday$lag_att_con)
predday$m4_resabs_0300 = abs(predday$m4_estlag_0300 - predday$lag_att_con)
head(predday)

#Create MSE vector for cases that are have been attempted but not contacted
     by selected day
msefile <- data.table(subset(predday, predday$conday >= d))
summ1 <- msefile[, list(time_id = b, day = d, method="m1", num= .N,
     actual=mean(lag_att_con), pred=mean(m1_estlag),
     bias=(mean(m1_estlag)-mean(lag_att_con)),
     var=var(m1_estlag), diffsq=sum(m1_res^2)), by=calday]
summ2 <- msefile[, list(time_id = b, day = d, method="m2", num= .N,
     actual=mean(lag_att_con), pred=mean(m2_estlag),
     bias=(mean(m2_estlag)-mean(lag_att_con)),
     var=var(m2_estlag), diffsq=sum(m2_res^2)), by=calday]
summ3 <- msefile[, list(time_id = b, day = d, method="m3", num= .N,
     actual=mean(lag_att_con), pred=mean(m3_estlag),
     bias=(mean(m3_estlag)-mean(lag_att_con)),
     var=var(m3_estlag),  diffsq=sum(m3_res^2)), by=calday]
summ4_0033 <- msefile[, list(time_id = b, day = d, method="m4_0033", num= .N,
     actual=mean(lag_att_con),pred=mean(m4_estlag_0033),
     bias=(mean(m4_estlag_0033)-mean(lag_att_con)),
     var=var(m4_estlag_0033),  diffsq=sum(m4_res_0033^2)), by=calday]
summ4_0050 <- msefile[, list(time_id = b, day = d, method="m4_0050", num= .N,
     actual=mean(lag_att_con), pred=mean(m4_estlag_0050),
     bias=(mean(m4_estlag_0050)-mean(lag_att_con)),
     var=var(m4_estlag_0050),  diffsq=sum(m4_res_0050^2)), by=calday]
summ4_0100 <- msefile[, list(time_id = b, day = d, method="m4_0100", num= .N,
     actual=mean(lag_att_con), pred=mean(m4_estlag_0100),
     bias=(mean(m4_estlag_0100)-mean(lag_att_con)),
     var=var(m4_estlag_0100),  diffsq=sum(m4_res_0100^2)), by=calday]
summ4_0200 <- msefile[, list(time_id = b, day = d, method="m4_0200", num= .N,
     actual=mean(lag_att_con), pred=mean(m4_estlag_0200),
     bias=(mean(m4_estlag_0200)-mean(lag_att_con)),
       var=var(m4_estlag_0200),  diffsq=sum(m4_res_0200^2)), by=calday]
  summ4_0300 <- msefile[, list(time_id = b, day = d, method="m4_0300", num= .N,
       actual=mean(lag_att_con), pred=mean(m4_estlag_0300),
       bias=(mean(m4_estlag_0300)-mean(lag_att_con)),
       var=var(m4_estlag_0300),  diffsq=sum(m4_res_0300^2)), by=calday]

  mse_summ <- data.frame(rbind(summ1, summ2, summ3, summ4_0033, summ4_0050,
                               summ4_0100, summ4_0200, summ4_0300))
  mse_summ$mse <- mse_summ$diffsq/mse_summ$num
  head(mse_summ)
  mse_summ <- mse_summ[order(mse_summ$calday, mse_summ$method),]


  #Return List for Function
  outobjs = list()
  outobjs[[1]] = mse_summ
  outobjs[[2]] = predday
  outobjs[[3]] = p_poslag_m2
  outobjs[[4]] = p_laglngth_m2
  outobjs[[5]] = p_poslag_m3
  outobjs[[6]] = p_weilag_m3
  outobjs[[7]] = p_poslag_0033
  outobjs[[8]] = p_poslag_0050
  outobjs[[9]] = p_poslag_0100
  outobjs[[10]] = p_poslag_0200
  outobjs[[11]] = p_poslag_0300
  outobjs[[12]] = p_weilag_0033
  outobjs[[13]] = p_weilag_0050
  outobjs[[14]] = p_weilag_0100
  outobjs[[15]] = p_weilag_0200
  outobjs[[16]] = p_weilag_0300

  objnames <- c("mse_summ", "predday", "p_poslag_m2", "p_laglngth_m2",
       "p_poslag_m3", "p_weilag_m3",
                "p_poslag_0033", "p_poslag_0050", "p_poslag_0100",
       "p_poslag_0200", "p_poslag_0300",
                "p_weilag_0033", "p_weilag_0050", "p_weilag_0100",
       "p_weilag_0200", "p_weilag_0300")
  names(outobjs) <- objnames

  saveRDS(outobjs, file=paste0("/data/san_caesapp6/nscg/lagattcon/", outstuff,
       ".rds"))

  return(outobjs)
}
```

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests: Sensitivity and specificity. BMJ (Clinical research ed.), 308(6943), 1552.

Axinn, W.G., Link, C.F., and Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, 48, 1127-1149.

Bates, N., J. Dahlhamer, P. Phipps, A. Safir and L. Tan (2010). Assessing Contact History Paradata Quality Across Several Federal Surveys. Proceedings of the American Statistical Association 2010 Joint Statistical Meeting, Vancouver, Canada.

Biemer, P., Chen, P., Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A*, 176 (1). pp. 147-168

Boulet, S., Ursino, M., Thall, P., Landi, B., Lepère, C., Pernot, S., Burgun, A., Taieb, J., Zaanan, A., Zohar, S., Jannot, A.-S. (2019). Integration of elicited expert information via a power prior in Bayesian variable selection: Application to colon cancer data. *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280219841082.

Census Bureau (2004), "American Community Survey Design and Methodology (January 2014), Chapter 4: Sample Design and Selection. Design and Methodology Report. https://www2.census.gov/programssurveys/acs/methodology/design_and_methodology/acs_design_methodology_ch04_2014.pdf?#

Census Bureau (2008). A Compass for Understanding and Using American Community Survey Data. https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf [October, 2017].

Census Bureau (2016). Planning Database Documentation. https://www.census.gov/research/data/planning_database/2016/docs/PDB_Block_Group_2016-07-28a.pdf [December, 2017].

Census Bureau (Census). (2019). American Community Survey Technical Documentation – Data Dictionary. Retrieved from: https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2013-2017.pdf?#

Chapman, C. (2014). "National Center for Education Statistics Adaptive Design Overview", Federal Committee on Statistical Methodology Conference, Washington, DC, December 16th.

Coffey, S., Reist, B. (2015). “Implementing Static Adaptive Design in the National Survey of College Graduates: Results of an Incentive Timing Experiment”. *Joint Statistical Meetings*, Seattle, WA, August 11th.

Coffey, S., Reist, B., Miller, P.V. (2019). Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates, *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smz026

Coffey, S., West, B.T., Wagner, J., Elliott, M.R. (2020). What Do You Think? Using Expert Opinion to Improve Predictions of Response Propensity Under a Bayesian Framework. *methods, data, analysis*. In press.

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

Couper, M. (2017). Birth and Diffusion of the Concept of Paradata (in Japanese – translated by W. Matsumoto). *Advances in Social Research*, 18, 14-26. Retrieved from: http://jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf [October, 2019].

Dallow, N, Best, N, Montague, TH. (2018). Better decision making in drug development through adoption of formal prior elicitation. *Pharmaceutical Statistics*, 17: 301-316. https://doi.org/10.1002/pst.1854

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

Elliott, M.R. (2017). *Preliminary Ideas for Responsive Design Implementation*. Internal Memo at the Census Bureau. Unpublished.

Fecso R.S., Frase M.J., and Kannankutty N. (2012), Using the American Community Survey as the Sampling Frame for the National Survey of College Graduates. Working Paper NCSES 12-201. Arlington, VA: National Science Foundation, National Center for Science and Engineering Statistics. Available at http://www.nsf.gov/statistics/ncses12201/.

Finamore, J., and Dillman, D. (2013), “An Experimental Evaluation of How Mode Sequence for Offering Internet, Mail and Telephone Options Affects Responses to NSCG”, *European Survey Research Association Conference*, Ljubljana. Available at http://www.websm.org/uploadi/editor/1392686326Finamore_Dillman_2013_An_Experimental_Evaluation_Of_How_Mode_Sequence.pdf

Gelman, A.; Carlin, J.B.; Stern, H.; Dunson, D.; Vehtari, A.; Rubin, D. *Bayesian Data Analysis*. Boca Raton: Chapman Hall. 2013. Print.

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

Groves, R. M., Heeringa, S. G. (2006), “Responsive designing for household surveys: tools for actively controlling survey errors and costs”, Journal of the Royal Statistical Society A, 169, 439-457.

Hall, D., Cohen, S., Finamore, J., and Lan, F. (2011), "NSCG sampling issues when using an ACS-based sampling frame", *Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings 2011*; 3954–3963. Available: https://www.amstat.org/sections/srms/proceedings/y2011/Files/302332_68268.pdf.

Hampson, L. V., Whitehead, J., Eleftheriou, D. and Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33, 4186- 4201. https://doi.org/10.1002/sim.6225

Hosack, G. R., Hayes, K. R., & Barry, S. C. (2017). Prior elicitation for Bayesian generalised linear models with application to risk control option assessment. *Reliability Engineering & System Safety*, 167, 351-361.

Kim, D., Han, S., Youngblood M. (2018). Sequential patient recruitment monitoring in multi-center clinical trials. *Communications for Statistical Applications and Methods*, 25(5), 501-512. https://doi.org/10.29220/CSAM.2018.25.5.501

Klein, J.P., Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer Science+Business Media. 2003. Print.

Laflamme, F., and Karaganis, M. (2010). “Development and implementation of responsive design for CATI surveys at Statistics Canada”, European Quality Conference, Helsinki. Retrieved from: https://www.researchgate.net/publication/228583181_Implementation_of_Responsive_Collection_Design_for_CATI_Surveys_at_Statistics_Canada

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. A. (2012). A “SMART” design for building individualized treatment sequences. *Annual Review of Clinical Psychology, 8,* 14.1 – 14.28.

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J., and Gu, H., (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth. Vital and Health Statistics, 2(158).

Li, Y., Gail, M.H., Preston, D.L., Graubard, B.I. and Lubin, J.H. (2012), Piecewise exponential survival times and analysis of case-cohort data. *Statistics in Medicine*, 31: 1361-1368. https://doi.org/10.1002/sim.4441

Li, Y., Panagiotou, O.A., Black, A., Liao, D. and Wacholder, S. (2016), Multivariate piecewise exponential survival modeling. Biometrics, 72: 546-553. https://doi.org/10.1111/biom.12435

Ma, L.; Yan, M.; Weng, J. Modeling traffic crash rates of road segments through a lognormal hurdle framework with flexible scale parameter. *Journal of Advanced Transportation*, 2015, 49: 928-940.

Maitland, A., Hubbard, R., Edwards, B. (2016). The use of CARI and Feedback to Improve Field Interviewer Performance. Presentation at the 2016 American Association for Public Opinion Research (AAPOR) National Conference. May 13, 2016. Austin, TX.

Mason, A. J., Gomes, M., Grieve, R., Ulug, P., Powell, J. T., & Carpenter, J. (2017). Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE trial. *Clinical Trials*, 14(4), 357–367. https://doi.org/10.1177/1740774517711442

Mullahy, J. Specification and testing of some modified count data models. *Journal of Econometrics*, 1986, 33: 341-365.

National Center for Health Statistics (NCHS) (2018). “National Health Interview Survey: Survey Description”. Retrieved from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2017/srvydesc.pdf [October, 2019].

O’Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73, 69-81.

Opsomer, J. D., Breidt, F. J., White, M., and Li, Y. (2016), "Successive Difference Replication Variance Estimation in Two-Phase Sampling," *Journal of Survey Statistics and Methodology*, 4: 43–70, https://doi-org.proxy-um.researchport.umd.edu/10.1093/jssam/smv033

Peytchev, A., Rosen, J., Riley, S., Murphy, J., and Lindblad, M. (2010). “Reduction of Nonresponse Bias through Case Prioritization”, *Survey Research Methods*, 4, 21-29.

Rao, C., Toutenburg, H., Shahlab, S., Heumann, C. (2018). *Linear Models and Generalizations: Least Squares and Alternatives*. Heidelberg, Germany: Springer.

Roberts, C., Vandenplas, C. and Stahli, M.E. (2014). “Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs”, *Survey Research Methods*, 8, 67-80.

Rosen, J.A., Murphy, J., Peytchev, A., Holder, T., Dever, J.A., Herget, D.R., and Pratt, D.J. (2014). Prioritizing low-propensity sample members in a survey: Implications for nonresponse bias. *Survey Practice*, 7(1).

Schouten, B., Mushkudiani, N., Shlomo, N., Durrant, G., Lundquist, P., Wagner, J. (2018). A Bayesian Analysis of Design Parameters in Survey Data Collection. *Journal of Survey Statistics and Methodology*, 6, 431-464.

Schouten, B., Peytchev, A., Wagner, J. (2017). *Adaptive Survey Design*. Boca Raton, Florida: CRC Press.

Schouten, B., Calinescu, M., and Luiten, A. (2013). “Optimizing quality of response through adaptive survey design”, *Survey Methodology*, 39, 29-58.

Singer, J.D., Willett, J.B. (2003). Fitting Basic Discrete Time Hazard Models. In *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (pp.357-406). New York: Oxford University Press. Print.

Smithson, M., Merkle, E. (2013). Chapter 5: Count Variables. In *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. New York: Chapman and Hall/CRC. Print.

Spiegelhalter, D. J., K. R. Abrams and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester ; Hoboken, NJ, John Wiley & Sons.

Thompson, J., and Kaputa, S. (2017). Investigating adaptive non-response follow-up strategies for small businesses through embedded experiments. *Journal of Official Statistics*, 33(3), 835-856.

U.S. Bureau of Labor Statistics and U.S. Census Bureau. (2006), “Design and Methodology; Current Population Survey”, Technical Report 66. Available at: https://www.census.gov/prod/2006pubs/tp-66.pdf

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G., and Kruger-Ndiaye, S. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28, 477-499.

Wagner, J. and Hubbard, F. (2014). Producing unbiased estimates of propensity models during data collection. *Journal of Survey Statistics and Methodology*, 2, 323-342.

Wagner, J., B. T. West, H. Guyer, P. Burton, J. Kelley, M. P. Couper and W. D. Mosher (2017). The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth. *Total Survey Error in Practice*. P. P. Biemer, E. de Leeuw, S. Eckman et al. New York, Wiley.

United States Postal Service(USPS). (2009). Intelligent Mail Barcode – Technical Resources Guide. Retrieved from: https://postalpro.usps.com/node/221

Wagner, J., West, B.T., Elliott, M.R. (2020). Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a Responsive Survey Design Context. *Journal of Official Statistics.* In Press.

West, B.T. and Groves, R.M. (2013). The PAIP Score: A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly*, 77, 352-374.

West, B.T. (2013). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society, Series A*, 176, 211-225.

West, B.T., Kreuter, F. (2013). Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth (NSFG). *Public Opinion Quarterly*, 77(2), 522-548.

West, B.T., Wagner, J., Gu, H. and Hubbard, F. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, 3, 240-264.

Williams, D., Brick, J.M. (2018). Trends in U.S Face-to-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology*, 6(2),186-211.

West, B.T., Wagner, J., Coffey, S., and Elliott, M.R. (2020). The Elicitation of Prior Distributions for Bayesian Responsive Survey Design: Historical Data Analysis vs. Literature Review. Retrieved from https://arxiv.org/ftp/arxiv/papers/1907/1907.06560.pdf.

White, M. and Opsomer, J. (2011), "Variance Estimation Issues in the 2010 NSCG Two-Phase Sample Design," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 3987–4000.

White, M. and Opsomer, J. (2012), "Replicate Variance Estimation in a Two-Phase Sample Design Setting—Simulation Study with 2010 National Survey of College Graduates Data," Proceedings of the Section on Survey Research Methods of the American Statistical Association, pp. 3911– 3923.

Willimack, D.K., Dalzell, J.L., (2006) An Examination of Non-Contact as a Source of Nonresponse in a Business Survey. Retrieved from http://www.asasrms.org/Proceedings/y2006/Files/JSM2006-000027.pdf [October, 2019].

Zhang, X., Long, Q. (2012). "Modeling and prediction of subject accrual and event times in clinical trials: a systematic review." Clinical Trials, 9, 681-688.

Zou, K., O'Malley, J., Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Test and Predictive Models. *Circulation*, 115:654-657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929