Level-of-Effort Paradata and Nonresponse Adjustment Models for a National Face-to-Face Survey

James Wagner, Richard Valliant, Frost Hubbard, Charley Jiang, University of Michigan

August 2013

Introduction

Survey samples are designed to produce unbiased estimates. Unfortunately, nonresponse may lead to bias if the responders and nonresponders are different with respect to the survey variables. One common approach to addressing nonresponse after data collection has been completed is to differentially weight responding cases such that the respondents match the full sample on the selected characteristics. The selection of the characteristics is a modeling step that assumes that conditional upon the selected characteristics, responders and nonresponders are equivalent. This method is known as nonresponse weighting. The method relies upon having data available for the entire sample that predicts both response and the survey variables themselves. These data can come from either the sampling frame or from paradata (Couper, 1998; Couper and Lyberg, 2005), that is, from process data created during data collection. If the available data are only useful for predicting response and not for predicting the survey variables, then adjustments based upon these data can only add noise to estimates. This is true even when the true probability of responding is known. In practice, the true probability is never known and estimates of it have associated sampling error and, possibly, misspecification error which may also add noise to estimates.

Unfortunately, many surveys in the US have only very weak predictors of both response and the survey variables available on the sampling frame (Kreuter et al., 2010; Biemer et al., 2013). Paradata, on the other hand, include measures of effort which are frequently strongly predictive of response. These measures may or may not be predictive of the survey variables depending upon the survey content.

In this paper, we evaluate the utility of data available from the sampling frame and paradata for the creation of nonresponse adjustments for the Health and Retirement Study (HRS, http://hrsonline.isr.umich.edu/). We find that some paradata elements are useful predictors of both nonresponse and the key survey variables while other paradata (particularly those related to amount of field effort) are strongly related to response, but are not related to key survey variables collected by the HRS. Including these level-of-effort variables in nonresponse adjustment models may not reduce bias due to nonresponse and may needlessly add variability to both weights and survey estimates. Future waves of the HRS will seek to find paradata elements, including observations made by interviewers, which are related to key survey variables.

Background

Survey samples produce unbiased estimates of population quantities when every sampled unit responds. Unfortunately, in most surveys, complete response is never achieved. The pattern of nonresponse, to the extent that it is related to the variables measured by a survey, can lead to biased estimates. Little and Rubin (2002) describe three different patterns of missing data. The first type is missing completely at random (MCAR). In this pattern, the missingness is unrelated to any observed or unobserved data. The missingness is completely random and can be seen as just another stage of sampling. While the reduced sample size may lead to larger sampling errors, no adjustments to the observed data are needed in order to make unbiased estimates of some quantities like means. For population totals, weights do need to be adjusted even under MCAR when there is nonresponse; otherwise, estimated totals will be too small. The second pattern of missing data depends upon observed values. This pattern is known as missing at random (MAR). Under this pattern, if we condition our analyses upon the observed data, then our inference should be unbiased. As an example, imagine the only auxiliary variable we have on our sampling frame is Census Region. We note that response rates are different across the Regions. However, within each region, the responders and nonresponders would say the same thing on average in response to our survey questions. If we account for the different response rates between the regions, perhaps by differentially weighting the responders from each of the regions, then we can produce unbiased estimates. In the third pattern, the missingness depends upon unobserved values. In this case, the nonresponders are different than the responders in terms of the survey variables themselves. This is true even after we account for known information on the sampling frame. This pattern of missingness is known as not missing at random (NMAR). If the data are NMAR, then no adjustment strategy based on the observed data will be available to produce unbiased results. Strong modeling assumptions will be required for this situation (see, for example, Little, 1993).

We focus on methods for data that are MAR. If nonresponse is MAR, then it is important to use statistical adjustments to the data in order to produce unbiased estimates. The most common method for making these adjustments is known as nonresponse adjustment weighting (Kalton and Kasprzyk, 1986; Little 1986). The method assumes that the missingness is MAR and creates adjustment weights for each case that will account for the pattern of missing data. These weights can be formed in a variety of ways. One method is the weighting class approach (Holt and Elliott, 1992). Variables available for all cases are used to stratify the sample into classes. Within each class, the inverse of the response rate is used as an adjustment weight. Assuming that the responders and nonresponders within each class are equivalent with respect to the survey variables, these weights will produce unbiased estimates. A generalization of the approach uses response propensity models to estimate response probabilities. These response propensity models allow more flexibility than the weighting class approach. For example, these models allow for the inclusion of continuous predictors where the cell approach requires categorical variables. In addition, the response propensity modeling approach allows for the exclusion of interaction effects. The cell approach implicitly requires that all interactions between the variables used to form the cells be included. Little (1986) describes the propensity approach and notes that since the propensities are only estimates, it may be more robust to use the estimated propensities to create cells (e.g. deciles of the estimated propensities) that serve as the basis of a standard weighting class adjustment. He calls this approach "response propensity stratification."

The focus of these adjustment strategies is on response rates or, more generally, response propensities. However, nonresponse bias is the product of two components. The first is the nonresponse rate. The second is the differences between responders and nonresponders. In order to address this bias, weighting adjustments need to relate to both of these components. That is, the response propensities need to differ across the cells in a weighting class adjustment and the survey estimates need to differ across the cells as well. Kalton and Maligalig (1991) used a quasi-randomization approach in which every unit has a probability of responding to show that

$$Bias\left(\bar{y}\right) \approx \sum\left(y_i - \bar{Y}\right)\left(\phi_i - \bar{\phi}\right)/\left(N\bar{\phi}\right)$$

where y_i is the value of some variable for unit *i*, $\bar{y}$ is the survey-weighted mean for respondents, ϕ_i is the probability that unit *i* responds, *N* is the population size, $\bar{Y}$ is the population mean of y, $\bar{\phi}$ is the mean population response probability, and the sum is over the whole population. If respondents are put into cells, the bias formula above applies to each cell. Thus, the quasi-randomization bias can be removed either by putting units into cells so that the response probabilities, ϕ_i, are all the same, or the respondent cell means of y equal the population cell mean.

Using a model-based approach, Little and Vartivarian (2005) showed that if respondents are classified into $c = 1,\ldots,C$ cells and follow a model where the mean differs by cell, the model bias of the mean of the respondents is

$$b(\bar{y}) = \sum_{c=1}^{C} \pi_c \left(\mu_{Rc} - \mu_c\right)$$

where π_c is the population proportion in cell *c*, μ_{Rc} is the model-mean for respondents in cell *c,* and μ_c is the model-mean in cell *c.* Thus, from a model-based point-of-view (which conditions on the selected sample), the preferable approach is to create cells where the respondent mean equals the population mean. Little and Vartivarian (2005) emphasize this point in their evaluation of nonresponse adjustments. Their simulations show that if the variables used to define the cells predict response but do not relate to the survey measures, then the result of using these adjustments will be no reduction in bias and increases in variance. We use the following example to demonstrate this key point.

Suppose that an equal probability sample is selected and two nonresponse adjustment cells are formed based on a variable like gender (with levels denoted by *M* and *F*). Assume that the response probabilities for units in the two cells are π_M and π_F. The mean of the respondents is $\bar{y}_R = \left(\sum_{s_{RM}} y_k + \sum_{s_{RF}} y_k \right) \Big/ \left(n_{RM} + n_{RF} \right)$ where s_{RM} and s_{RF} are the sets of responding sample males and females and n_{RM} and n_{RF} are the numbers of respondents in each set. A nonresponse-adjusted estimator of the mean is

$$\bar{y}_R^* = \left(\sum_{s_{RM}} y_k / \pi_M + \sum_{s_{RF}} y_k / \pi_F \right) \Big/ \left(\sum_{s_{RM}} 1/\pi_M + \sum_{s_{RF}} 1/\pi_F \right).$$

The choice $\bar{y}_R^*$ is approximately unbiased with respect to the response distribution while $\bar{y}_R$ is not. However, if all units obey the same superpopulation model with $E_M \left(y_k \right) = \mu$, then both the unadjusted mean, $\bar{y}_R$, and the adjusted mean, $\bar{y}_R^*$, are model-unbiased in the sense that $E_M \left(\bar{y}_R \right) = E_M \left(\bar{y}_R^* \right) = \mu$. Thus, making the nonresponse adjustment in this simple case is unnecessary to create a model-unbiased estimate. Letting E_R denote expectation with respect to the response distribution, both estimators are unbiased in the sense that $E_M E_R \left(\bar{y}_R \right) = E_M E_R \left(\bar{y}_R^* \right) = \mu$. Thus, the variable weights in $\bar{y}_R^*$ would serve only to increase the variance of the estimated mean without decreasing either the model- or model-response bias. More generally, this reasoning leads to the conclusion that including variables in the nonresponse adjustment that are only related to response—not to the analysis variables—is inefficient.

This logic extends to response propensity models as well. Weights based on propensities that are uncorrelated with the survey variables can only increase variance. This is the case for known probabilities of response. However, the “noise” added is likely to be greater when the probabilities are estimated. The situation may be even worse if the model used to define the cells or for estimating the propensities is misspecified.

On the other hand, if the available predictors are related to the survey measures of interest, then we have the possibility to control bias and may also be able to control the estimated variance. As a result, if we had to choose, we would prefer to have predictors of the survey variables. Kreuter and Olson (2011) note that the problem is further complicated in multivariate modeling since it is possible that predictors in these models can have countervailing effects. For example, a predictor that appears to be related to both the survey variables and response propensities may be less effective for adjustments when combined with other predictors in a multivariate model. In general, population means are unknown, and the best we can do is to create cells where all respondents appear to follow a common mean model.

In practice, it is often difficult to find predictors that are strongly related to either response or survey measures. In their simulation study, Little and Vartivarian (2005) defined a strong correlation between the predictor and the survey variable as 0.8 and a weak correlation as 0.2. Kreuter and colleagues (2010) examined several studies and found empirically that the highest correlations between predictors drawn from paradata and survey measures were less than 0.2 and most of the correlations between such predictors and survey measures were less than 0.1.

There are two key sources of data available for nonresponse adjustment purposes – sampling frames, including commercially-available data, and paradata. In the case of large, area probability samples, the sampling frames are constructed from Census data. These data provide very general information about sampled neighborhoods and not about specific households. Since they are at the neighborhood-level and not the housing unit, many of these relationships with survey variables are likely to be attenuated (Biemer and Peytchev, 2012). The commercially-available data are merged to the selected sample. In the U.S., these data include information about the persons in the sampled housing units – age, sex, race, and ethnicity. However, this information is incomplete and sometimes incorrect.

The other source of data is paradata (Couper, 1998; Couper and Lyberg, 2005). These data are derived from the process of collecting survey data. They include, for example, call record data and interviewer observations. The variables related to effort (number of calls, ever refused; see Table 2, "Level-of-Effort Paradata") are often highly predictive of response (Drew and Fuller, 1980; Ahlo, 1990; Potthof et al., 1993; Groves and Couper, 1998; Beaumont, 2005; Wood et al. 2006; Durrant et al., 2009). For some surveys, they may also relate to the survey measures. For example, a study of time use may be biased if busier persons, who may be harder to contact, are included at lower rates. In such a study, a measure of contactiblity (the number of calls) may be related to both survey measures and response propensity.

A potential problem with paradata is that they can be measured with error. West (2013), for example, shows that interviewer observations can be measured with error and that these errors reduce the utility of these variables for nonresponse adjustment purposes. Biemer, Chen, and Wang (2013) found that interviewers in field studies do make errors in call records. They show through simulation that these errors can lead to biased estimates when variables derived from these data (e.g. number of calls) are used to make nonresponse adjustments. Thus, it appears to be difficult in practice to find variables that are useful for nonresponse adjustment.

In this paper, we explore the use of level-of-effort paradata as part of a nonresponse adjustment strategy for a large, face-to-face survey. This is an empirical question which depends upon the relationship of these data to both response probabilities and key survey variables. In the next section, we describe the survey, the data available on the sampling frame and paradata, and the modeling approach. We then examine the utility of level-of-effort variables and determine whether using them as part of nonresponse adjustment models will improve those adjustments. We conclude with some discussion about plans for future waves of the survey.

Methods

The Health and Retirement Study (HRS) is a national panel survey of persons over the age of 50 in the United States. Participants are interviewed every two years. The primary focus of the study is on the relationship between health and economic status in the years leading up to and following retirement. A new cohort is added every six years. These new cohorts are selected using a multi-stage area probability sample that screens for households with age-eligible persons. In households with age-eligible persons, interviews are conducted with up to two persons.

In 2004, the HRS recruited a new cohort of persons born between 1948 and 1953. This cohort, known as "Early Baby Boomers" (EBB), was interviewed in 2004 and then every two years following that including in 2010. During the 2004 recruitment, the HRS also "pre-recruited" persons for the next cohort – those born between 1954 and 1959, known as "Middle Baby Boomers" (MBB). This cohort would be added in 2010. However, there was additional funding made available in 2010 to increase the size of the sample of persons (especially persons from minority race and ethnicity groups) born between 1948 and 1959. The sample for this supplement was a multi-stage area probability sample. Since the 2010 sample was a supplement meant to increase the number of minorities in the panel, the sample was selected from areas with at least 10% black population or at least 10% Hispanic population. When combined with the earlier sample, the new sample of persons born between 1948 and 1959 is a fully representative national sample. Interviews were attempted with these expanded cohorts in 2010 and 2011.

We created a comprehensive set of adjustments for nonresponse for the persons recruited in 2004 and 2010 to these two new cohorts. As noted earlier, a nonresponse adjustment to weights is necessary, even if missingness is MCAR, in order for the weights to be properly scaled for estimating population totals. We made an adjustment in each of the several steps that sample had to pass through in order to be interviewed in 2010-2011. Figure 1 provides an overview of the different components of the sample and steps that each had to go through in order to be interviewed. We created a logistic regression model for each box in the figure. In other words, we modeled the probability that a case would be successfully screened in 2004 (the box in the upper left). We then modeled, conditional upon having been successfully screened as an eligible EBB, whether an eligible EBB would complete the main interview in 2004 (the next box to the right).

The one exception was for EBB cases that were interviewed in 2004. Rather than model the probability that they were interviewed in 2006, 2008, and 2010, we simply modeled the probability of whether they were interviewed in 2010 (i.e. we ignored the distinction between cases that dropped out in 2006, 2008, or 2010). This is symbolized by the broken line in Figure 1.

The HRS measures some characteristics of the person and some of the household. Therefore, it was necessary to have adjustments for both types of variables. As a result, we also modeled separately the probability that the household would respond and that particular persons would respond. Since we would interview up to two persons per household, it could happen that one of two eligible persons in a household would be interviewed. Table 1 lists all of the models estimated in the process of creating nonresponse adjustments for the EBB and MBB cohorts.

Figure 1. Overview of the Response Process: EBB and MBB Cohorts

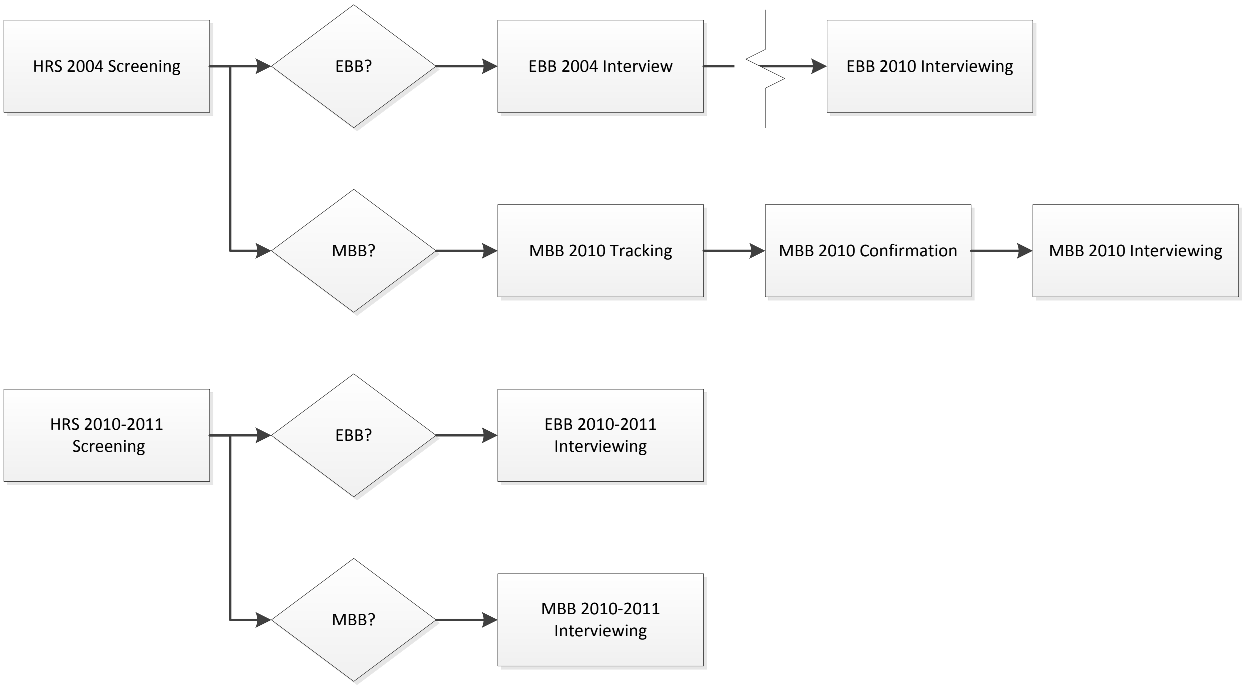


Table 1. Sequential Models of Response Process

Model 1	HRS 2004 Screening
Model 2	HRS EBB 2004 Interview – Person
Model 3	HRS EBB 2004 Interview – Household
Model 4	HRS EBB 2010 Interview – Person
Model 5	HRS EBB 2010 Interview – Household
Model 6	HRS MBB 2010 Confirmation[1]
Model 7	HRS MBB 2010 Interview – Person
Model 8	HRS MBB 2010 Interview – Household
Model 10	HRS 2010-2011 Screening
Model 11	HRS 2010-2011 Interview - Person
Model 12	HRS 2010-2011 Interview - Household

[1] Confirmation is the process of confirming that a HH screened as containing an MBB person in 2004 still contains that person in 2010.

The models were fit using the following procedures. First, all available data were fed into a stepwise regression model to determine a subset of predictors. Once an appropriate subset had been found, particular interactions were tested. Once a final model had been selected, cases were split into deciles based on the estimated response propensities. The means of several key statistics were then calculated for each decile.

We wanted to include data that would be predictive of measures of income and health since these are the key variables for the HRS. As is typical in household surveys, the sampling frame does not have very specific information. The variables used in our modeling are listed in Table 2. Since the sampling was done using Census data to create the frame, we had much of the data available from the Decennial Census 2000 and the American Community Survey (ACS). These measures are for the neighborhood (Census Block, Block Group, or Tract) of the selected housing unit. Some of these measures may be related to income (for example, Tract-level median income). Others may be indirectly related to health (for example, race and ethnic composition of the neighborhood). The utility of the data from the 2000 Decennial Census may have been reduced as they were used for data collected in 2010 and 2011. The most recent available versions of the ACS data were used. At the housing unit level, we had some commercially available data that can be merged to the addresses on the sampling frame. This information is incomplete (about 50% of housing units have some information) and can also be inaccurate (for example, 7.7% of successfully screened cases expected to be age eligible based on the commercial data were not). These issues may have also reduced the utility of these data.

We also have several paradata elements, including the number of call attempts on a case, whether there was ever resistance, and whether the housing unit was in a locked building or gated community. These paradata are generated from records of every call. These records include information about the time, date, and outcome of each call attempt. For instance, one interim outcome indicated that the case is in a locked building or gated community and could not be reached. If this code was ever assigned after an attempt on a case, then it was coded as being in locked building. Another interim outcome code specifies that the case was resistant to completing the interview. These cases may then be "converted" to an interview by an expert interviewer. If this interim code of resistance was ever assigned to a case, then it was coded as having been "ever resistant." Since a record is generated for each attempt, this information can be summarized up to a case level. We tried several transformations of the number of calls to test for nonlinear relationships, including creating categories and the natural logarithm of the number of calls. Variables generated using these data are described in Table 2.

Table 2: Variables Included in Stepwise Regression Procedure

Variable Origin	**Variable Description**
Commercial Database (Data Matched from Commercial Sources at the Address Level)	Surname Matched to Address (Yes, No)
	Expected Age Eligibility for HRS 2010 and 2011 Addresses - HH Contains a Person 50-62 Years Old (Age Eligible, Age Ineligible, No Age Data Matched to Address)
	Expected Age Eligibility for HRS 2011 Addresses Only - HH Contains a Person 50-62 Years Old (Age Eligible, Age Ineligible, No Age Data Matched to Address)
	Estimated head of household (HoH) Race/Ethnicity (Black, non-Hispanic; Hispanic; Other Race/Ethnicity, No Race/Ethnicity Data Matched to Address)
	Expected HH Level HRS Age Cohort and Hispanicity Status (MBB, Hispanic; MBB, non-Hispanic; EBB, Hispanic; EBB, non-Hispanic; Age Ineligible; No Age and Hispanicity Data Matched to Address)
	Expected Head of Household (HoH) Gender (Male, Female, No Gender Data Matched to Address)
	Expected Number of Children
	Expected HoH Marital Status (Single, Married, No Marital Status Data Matched to Address)
	Expected HoH Education Level
	Expected HH Ownership Status (Own, Rent, No HH Ownership

	Status Data Matched to Address)
	Expected HH Income Category (Less than $40K, $40 – 75K, $75K+)
Paradata (Data All at the Address Level)	Number of Face-to-Face Contact Attempts Made Category (0-1, 2-3, 4-7, 8+)
	Number of Telephone Contact Attempts Made Category (0, 1-2, 3+)
	HH Residents Ever Resistant to Answer Screening Questions (Yes, No)
	Address in a Locked Building (Yes, No)
	The Year the Address was Listed (2004 ,2010, or 2011)
	Address Part of a Multiple Unit Structure (e.g. Apartment Building)
ACS 2005-09 Census Tract and Block Group Level Data	Block Group level - Number of Occupied Housing Units (HUs)
	Tract level: Median Year Residents Moved into the Tract
	Tract level: HH Median Income
	Tract level: HH Median Income Quintiles
	Tract level: % of population that are College Graduates
	Tract level: % of population that are High School Graduates
	Tract level: % of population age 45-49
	Tract level: % of population age 50-54
	Tract level: % of population age 55-59
	Tract level: % of population age 60-64
	Tract level: % of persons age 16+ that are civilian and employed
	Tract level: % of persons age 45-54 who moved into tract over the past year
	Tract level: % of persons age 55-64 who moved into tract over the past year
	Tract level: % of persons age 45-64 who are married
	Tract level: % of population Black
	Tract level: % of population that are Black and age 45-64
	Tract level: % of persons age 16+ that are Black, age 16-64, civilian and employed
	Tract level: % of persons age 25+ that are Black and have a BA or higher
	Tract level: % of population that are Black and moved into tract over past year

	Tract level: % of population that are Hispanic
	Tract level: % of population that are Hispanic and age 45-64
	Tract level: % of persons age 16+ that are Hispanic, age 16-64, civilian and employed
	Tract level: % of persons age 25+ that are Hispanic and have a BA or higher
Census 2000 Tract and Block Group Level Data	Block Group Level: Race/Ethnicity Population Distribution (2: <10% Hispanic HHs and 10%+ Black, non-Hispanic HHs; 3: 10%+ Hispanic HHs and < 10% Black, non-Hispanic HHs; 4: 10%+ Hispanic HHs and 10%+ Black, non-Hispanic HHs)
	Tract level: % vacant HUs
	Tract level: "Hard to Count" score from Census 2000 Planning Database which indicates the level of difficulty the Census Bureau had in enumerating the tract.[1]
	Tract level: % single unit structures
	Tract level: % multi-unit structures with 10+ people
	Tract level: % mobile home
	Tract level: % renter occupied HUs
	Tract level: % unemployed
	Tract level: % primary HH language is Spanish
	Tract level: % occupied HUs moved into in past year
	Block level: Census Region
	Block level: Area total in square miles

Results

The impact of several variables was consistent across the estimated response propensity models. For brevity sake, we present the results from one of the eleven models. Table 3 lists the estimated odds ratios for the predictors in the model for one of the ten models listed in Table 1 (Model 10). This model estimates the probability of completing screening interview in the 2010-2011 sample. Variables related to income, wealth, race, ethnicity, and household size were

[1] More information on the Census Hard to Count score can be found here: http://www.census.gov/2010census/partners/pdf/TractLevelCensus2000Apr_2_09.pdf

valuable predictors in these models. For example, the quintiles of median household income at the Census Block Group level from the ACS and commercially-purchased estimates of household income were both useful predictors. These variables are important as they are also related to the key statistics measured by the HRS.

Table 3. Probability of Response for Screening Interviews Conducted in 2010-2011 (Model 10), Estimated Odds Ratios and 95% Confidence Limits (CI). Variables whose CI does not cover 1 are marked with an asterisk.

Variable Origin	**Predictor**	**Odds Ratio Estimate**	**95% Wald Confidence Limits**	
Commercial Database	No Age Data Matched to the 2010 or 2011 Address (reference category)			
	Expected Age Ineligible for the 2010 or 2011 address	1.073	0.916	1.257
	Expected Age Eligible for the 2011 Address (reference category)			
	Not Expected Age Eligible for the 2011 Address	1.740*	1.308	2.314
	No Age Data Matched to the 2011 Address	1.788*	1.378	2.320
	Expected HoH Other Race/Ethnicity (reference category)			
	Expected HoH Black, non-Hispanic	0.875	0.690	1.111
	Expected HoH Hispanic	1.073	0.876	1.314
	No Race/Ethnicity Data Matched to Address	0.765*	0.642	0.912
	Expected Single (reference category)			
	Expected Married	1.370*	1.095	1.714
	No Marital Status Data Match to Address	1.122	0.888	1.418
	No HoH Income Data Matched to Address (reference category)			
	Expected HoH Income Less Than $40K	1.095	0.861	1.393
	Expected HoH Income $40K to $75K	0.776*	0.605	0.996
	Expected HoH Household Income $75K+	0.647*	0.520	0.804
Paradata	Face-to-Face Contact Attempts Made: 8+ (reference category)			
	Face-to-Face Contact Attempts Made: 0-1	4.348*	3.693	5.120

	Face-to-Face Contact Attempts Made: 2-3	3.001*	2.629	3.426
	Face-to-Face Contact Attempts Made: 4-7	1.797*	1.599	2.020
	Telephone Contact Attempts Made: 3+ (reference category)			
	Telephone Contact Attempts Made: 0	20.101*	17.687	22.843
	Telephone Contact Attempts Made: 1-2	2.840*	2.493	3.236
	HH Residents Ever Refused to Answer Screening Questions	4.378*	3.956	4.846
	Address Not in a Locked Building	1.616*	1.416	1.845
	Segment Level: Address in Segment Listed in 2011 (Reference Category)			
	Segment Level: Address in Segment Listed in 2004	3.801*	3.068	4.709
	Segment Level: Address in Segment Listed in 2010	4.411*	3.616	5.380
	Address Level: Multiple Unit Structure (reference category)			
	Address Level: Not Multiple Unit Structure	0.844*	0.727	0.980
ACS 2005-2009	Tract Level: Median Income (Continuous)	1.000*	1.000	1.000
	Tract Level: Median Income Quintile 5 (Highest – reference category)			
	Tract Level: Median Income Quintile 1 (Lowest)	1.796*	1.225	2.632
	Tract Level: Median Income Quintile 2	1.400*	1.015	1.930
	Tract Level: Median Income Quintile 3	1.486*	1.133	1.948
	Tract Level: Median Income Quintile 4	1.611*	1.286	2.018
	Tract Level: % of Population that are Black, non-Hispanic and Ages 45-64	1.025*	1.006	1.045
	Tract Level: % of Persons Age16+ that are Civilian and Employed	0.970*	0.956	0.984
	Tract Level: % of Persons Age 16+ that are Hispanic, Ages 16-64,Civilian and Employed	0.984*	0.974	0.994
	Tract Level - % of Persons Age 25+ that are Black, non-Hispanic	0.982*	0.966	0.999
	Tract Level - % of Population that have at Least a High School Diploma or GED	1.017*	1.010	1.025

	Tract Level: % of Population that are Ages 50-54	1.030*	1.001	1.060
	Tract Level: % of Population that are Ages 60-64	0.922*	0.895	0.950
	Block Group Level: Number of Occupied HUs	1.000*	1.000	1.000
	Block Group Level: % of Population that are Hispanic	4.205*	2.490	7.100
	Block Group Level: % of Population that are Black, non-Hispanic	1.954*	1.287	2.967
	Block Group Level: % of Population that are Black, non-Hispanic	1.954*	1.287	2.967
Census 2000	Block Group Level: Race/Ethnicity Sampling Domain 4 (10%+ Black, non-Hispanic Population and 10%+ Hispanic Population) (reference category)			
	Block Group Level: Race/Ethnicity Sampling Domain 2 (10%+ Black, non-Hispanic Population)	1.168	0.999	1.365
	Block Group Level: Race/Ethnicity Sampling Domain 3 (10%+ Hispanic Population)	0.777*	0.680	0.887
	Trace Level: % Vacant HUs	0.990*	0.980	0.999
	Trace Level: % Single Unit Structures	0.996*	0.994	0.999
	Trace Level: % Mobile Homes	1.008*	1.003	1.013
	Trace Level: % Unemployed	0.924*	0.902	0.947
	Trace Level: % Primary HH Language is Spanish	1.023*	1.010	1.037

Across all the models estimated, a key finding was that the level of effort data from the call records (in particular, the number of calls and whether the case had ever been resistant were highly predictive of response. The model results in Table 3 demonstrate this. The call record data are used to create predictors regarding the number face-to-face calls made, the number of telephone calls made, and whether someone at the housing unit was ever resistant to completing the screening interview (some of these resistant cases are later “converted”). A case with resistance had a much lower probability of ever completing a screening interview. Cases without resistance relative to those that did had an odds ratio of about 4.4, indicating that cases without resistance had a much higher probability of completing the screening interview. The model had good fit with the area under the curve (AUC) at 0.884.

In contrast, the estimated propensities from the models including level-of-effort paradata were not associated with the key statistics. Figure 3 shows, for example, the estimates of Mean Wealth A (HRS 2010 Total HH Wealth including secondary residence with missing values imputed)) and B (same as Wealth A but excluding secondary residences) by deciles of the propensities estimated from the model for responding to the screener in 2010-11 in Table 3. These are unweighted estimates of the mean, which is appropriate for the purposes of creating nonresponse adjustments (Little and Vartivarian, 2003). The correlation between these propensities and Wealth A is -0.013 (p=0.51).

Figure 2. Mean Wealth A and B by Decile of Estimated Propensity (Model Includes Level-of-Effort Paradata)

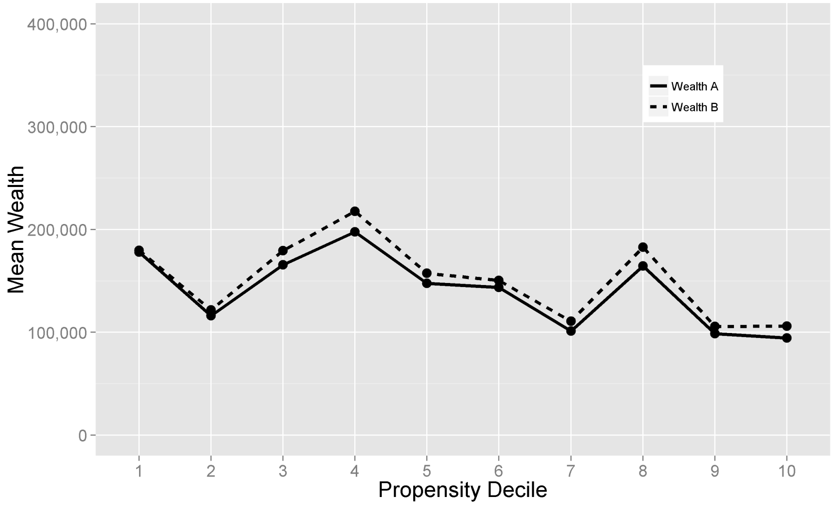
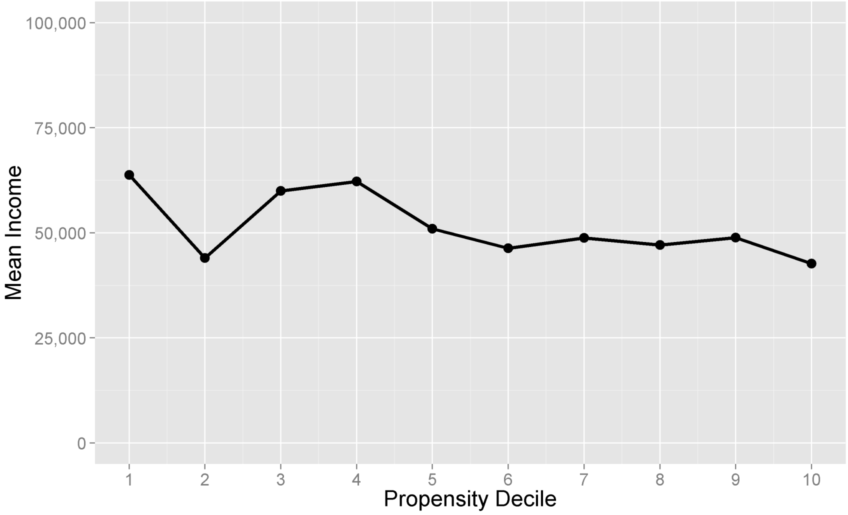


Figure 3 shows a similar pattern. The propensities do not appear to be related to Mean Household Income (total household income with missing values imputed as reported by HRS households during 2010 data collection). The correlation is -0.035 (p=0.08).

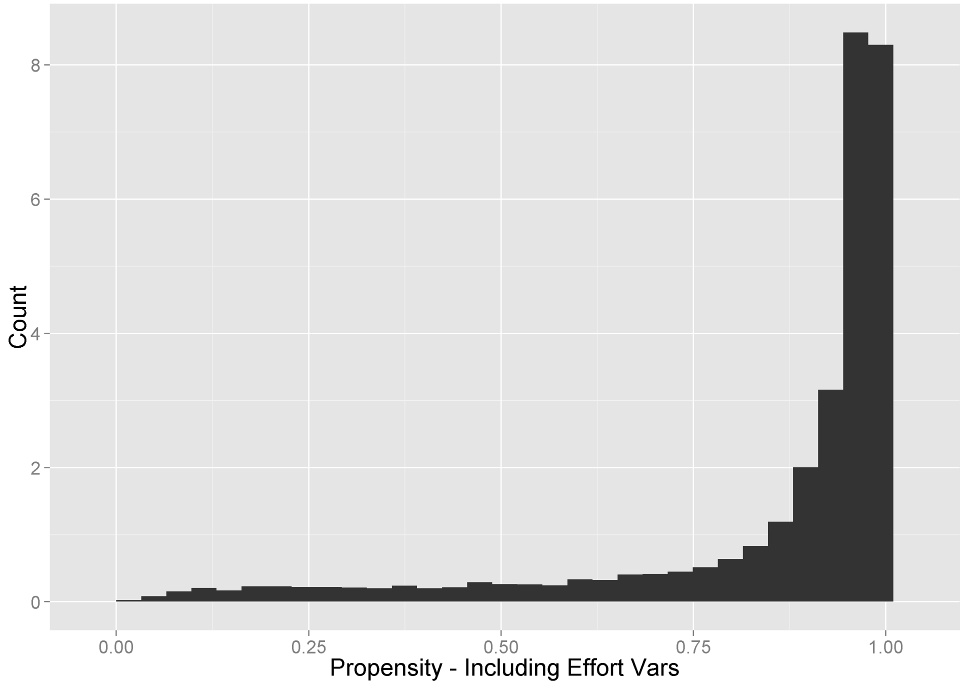
Figure 3. Mean Household Income by Decile of Estimated Propensity (Model Includes Level-of-Effort Paradata)



There are several reasons that may explain why estimated contact and cooperation probabilities are not related to key statistics in this survey. First, being difficult to contact may not be associated with higher or lower income. Second, given that the fieldwork is under the control of interviewers, the choices they make may add noise to these effort variables. For instance, as an extreme example, one case may be called repeatedly on weekday days and receive the same number of calls as another case that is called repeatedly in the evening. These two treatments are clearly not the same, but the model does not distinguish them. We tried using the natural logarithm of the number of calls to remedy this problem, as well as indicator variables for various levels of calling (e.g. 1-3, 4-7, 8+). Third, there is evidence that the number of calls can be systematically underreported (Biemer, Chen, and Wang, 2013). This underreporting can lead to biased estimates of coefficients related to the number of calls since the underreported calls are more likely to be noncontacts.

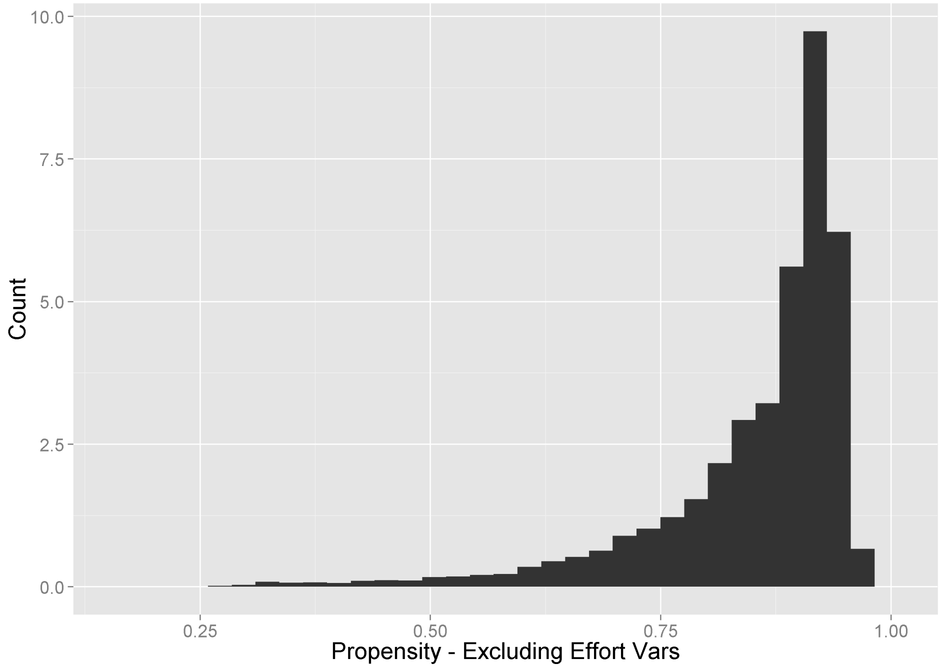
Figure 4 shows the distribution of the estimated response propensities from the model in Table 3. Although many of these propensities are greater than 0.9, the range is quite large. There were cases with estimated propensities as low as 0.019. The 5th and 95th quantiles were 0.293 and 0.989 respectively. If these propensities were used to form nonresponse weighting adjustments, they would lead to highly variable weights. These highly variable weights would lead to increases in estimated variances (Kish, 1992; Little and Vartivarian, 2005). Since cases with different weights do not have different average means of key survey variables, these variable weighting factors would not lead to changes in estimates nor to reduction in model-bias.

Figure 4. Distribution of Estimated Response Propensities from Model in Table 3



Since the weights derived from propensity models including level-of-effort paradata could not lead to changes in estimates but could increase estimates of variance, the call number and ever-resistant status variables were removed and the propensity models were re-estimated. The fit of the resulting models predicting response was not as good (AUC=0.706). However, the variability of the estimated propensities was reduced relative to those from the models that include level-of-effort paradata. The minimum of the estimated propensities from the models that excluded level-of-effort paradata was 0.199 ; the range was also reduced (see Figure 5). The 5th and 95th quantiles were 0.624 and 0.941 respectively.

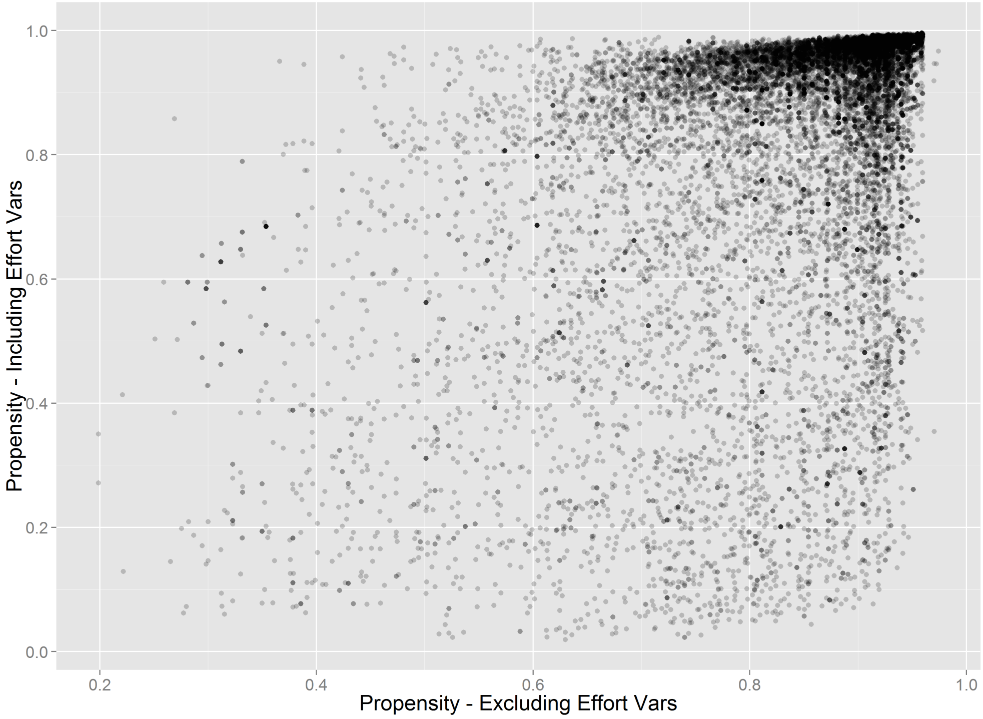
Figure 5. Distribution of Estimated Propensities from Model Excluding Level-of-Effort Variables



One commonly used method for judging the potential design effect due to weighting is the "1+L" statistic described by Kish (1988). This statistic uses the relvariance (plus one) of the weights to determine the inflation of the variance that the weights could potentially have on the analysis. This method assumes that the weights are unrelated to the survey variables. As we have seen here, the weights are somewhat related to several variables from the survey. Still, the "1+L" can be thought of as the maximal inflation of variance estimates due to having weights that are not all equal. The weights that were based on the models including the level-of-effort variables had a "1+L" of 2.75. The weights based on the models which excluded these variables had a "1+L" of 2.03.

Figure 6 shows a scatterplot of the two estimates of the propensity – those from the model with the effort variables included plotted against the estimates from the model that excludes these effort variables. As the figure illustrates, the weights for individual cases can be substantially different using the two models. Although full population estimates may be similar using the two models for nonresponse, domain estimates could be quite different with the two approaches.

Figure 6. Propensities Estimated With and Without Level-of-Effort Paradata



The key statistics were somewhat more associated with the propensities estimated from models that excluded the level-of-effort variables. Figure 7 shows the estimates of the same wealth statistics as presented in Figure 4 across the propensity deciles estimated from the model that excluded the level-of-effort variables. In this case, there does seem to be an association between the propensities and wealth. The correlation between these propensities and Wealth A is -0.102 (p<0.0001). The correlation between these propensities and mean household income in Figure 8 is -0.148 (p<0.0001).

Figure 7. Mean Wealth A and B by Decile of Estimated Propensity (Model Excludes Level-of-Effort Paradata)

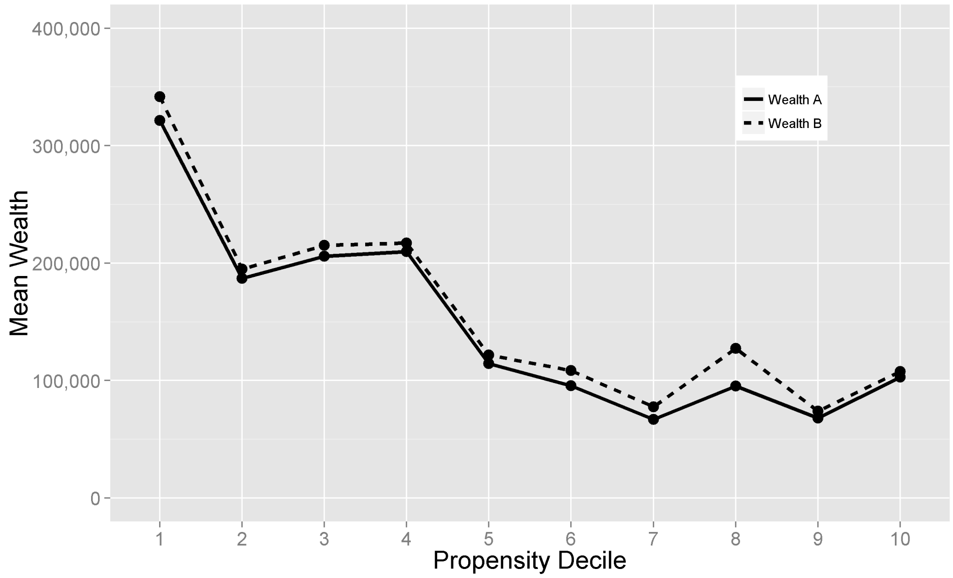
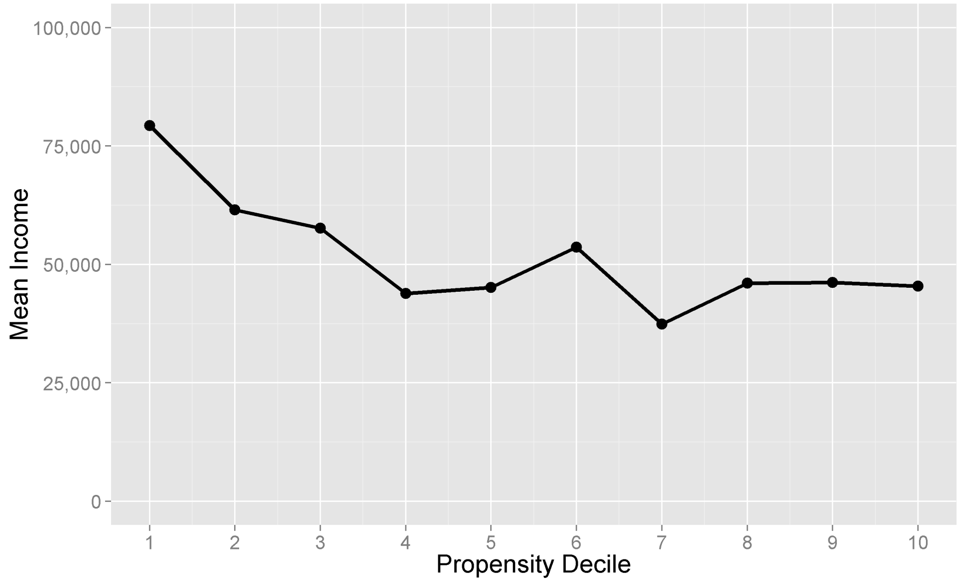


Figure 8. Mean Household Income by Decile of Estimated Propensity (Model Excludes Level-of-Effort Paradata



The estimated propensities from the model that excludes the level-of-effort variables meets both criteria for a good adjustment model. There is evidence that nonresponse bias will be reduced since there is variation in the propensities (albeit less than the initial model with the level-of-effort variables), and the propensities are correlated with the key survey variables. Therefore, the models without the effort variables were selected to be the final models.

As a final check, we developed nonresponse adjustments based on propensities estimated from each model. The approach was the same for each adjustment model – estimate the propensities, create deciles of those propensities, use the inverse of the response rate in each decile as an adjustment weight. Table 4 shows estimates of several key variables and their standard errors estimated using both of these sets of weights. None of the estimates are significantly different based on jackknife replication estimates of the variance of the difference between the two estimates. The standard errors are generally similar when the weight based on the model excluding level-of-effort variables is used.

Table 4: HRS EBB and MBB Cohorts Household Level Estimates of Key Statistics when Effort Variables are Excluded and Included in the Nonresponse Adjustment Propensity Models

Key Statistic	N	Mean - No Effort Variables	Standard Error - No Effort Variables		Mean - Effort Variables	Standard Error - Effort Variables
% of HHs where at least 1 member of the family unit is currently employed	6099	67.9%	1.2%		67.4%	1.1%
% of HHs where HoH rates Health as “Fair” or “Poor”	6199	25.3%	1.3%		25.8%	1.2%
% of HHs with Any Other Debts Outside of Mortgage, Car Loans, or Money Owed on Other Assets	5853	48.2%	0.9%		46.1%	1.1%
% of HHs that Own a Second Home	5964	15.5%	0.7%		14.2%	0.9%
% of HHs that Own Vehicle for Transportation	5906	87.6%	0.8%		87.7%	0.8%
% of HHs that Donate to Charity	5870	47.7%	1.6%		46.3%	1.6%
Mean HH Income (Imputed	6084	$96,894	$4,705		$87,825	$4,141

where missing)						
Total HH Wealth Excluding 2nd Residence (Imputed where missing)	6084	$355,046	$22,256		$323,300	$22,256
Total HH Wealth Including 2nd Residence (Imputed where missing)	6084	$373,219	$23,448		$342,418	$24,372

Conclusion

As Little and Vartivarian (2005) demonstrate, effective nonresponse adjustments require a model that predicts well both nonresponse and the quantity to be estimated from the survey. As a result, the search for a nonresponse adjustment needs to be conducted along both dimensions more or less simultaneously. In a multivariate setting, this may require building models that predict response, testing those models against the variables from the survey, and then iteratively refitting the model until the model converges to something that is effective along both dimensions.

The problem is more complicated for multi-purpose surveys. Our approach is to consider a range of “key” statistics that may stand as a sample of all the statistics that could be produced by a survey. The selected model should predict this range of statistics well in order to be robust across the many statistics that can be estimated from the survey. Other solutions to this problem may be possible. This is a problem (multipurpose design) that is also faced by sample design. Using ‘weighted’ combinations of key statistics is another useful approach (Kish, 1988; Valliant and Gentle, 1997).

We found that predictors from the sampling frame related to income, wealth, race, ethnicity, and household size were useful in predicting both nonresponse and the key survey statistics. This is logical since the content of the survey is about health and income for those approaching retirement. We also found several elements of the available paradata such as indicators for whether the housing unit was in a locked building or multi-unit structure were useful predictors.

On the other hand, we found predictors of response that were seemingly unrelated to the survey data. These predictors were drawn from the paradata and represented levels of effort. It may be that these predictors are only weak proxies for contact and cooperation. This could be due to measurement problems in the call records or due to variability in strategies that interviewers use to contact and interview persons in households. Some variables collected in conjunction with field work, related to difficulty of contact, resistance by sample cases, and other paradata items, are associated with the particular field personnel and the way in which they behave. Some cases may be ignored by a field interviewer; others may be attempted repeatedly over a short period of time. Response probabilities estimated with such fieldwork variables are not stable, repeatable values that would be found in any other edition of a survey. Although these variables are powerful predictors of response, measured by model fit statistics like pseudo-R^2 or AUC, these statistics are subject to sampling error and other issues, such as overfitting, and models with higher values on these statistics may not be closer to the truth than models with lower values on these statistics.

Given the nature of these fieldwork variables, it is credible that contact and cooperation, given the levels of response achieved by the survey, are not related to the outcomes. In order to definitively answer this question, we would need the survey data for the nonresponders. Consequently, the level-of-effort predictors were dropped from our nonresponse models. Leaving them in the models led to adjustments that did not change the estimates (relative to adjustments based on models that dropped them) but did inflate variances. This is in concordance with the simulation results of Little and Vartivarian (2005).

Although we determined to exclude the level of effort variables from our models, this is an empirical question for each study. Other studies have found some types of paradata variables are useful. As such, it is not a general principle to exclude them. Rather, each study needs to determine whether these predictors might be useful. We also found other paradata elements that were more useful for adjustment purposes – for example, whether the sampled unit was in a locked building. Further, level-of-effort paradata are useful for other purposes, including monitoring field work.

Finally, paradata are largely under the control of the data collector. It would be useful to tailor the collection of paradata to the content of the survey. This might mean collecting interviewer observations about sampled units. These observations can be designed to be related to the survey variables. For example, the National Survey of Family Growth has interviewers guess whether the selected person is in a sexually active relationship with a person of the opposite sex. These observations have been shown to be correlated with key variables collected by that survey (Kreuter, et al. 2010). Collecting these observations can be difficult. Interviewers can make errors, which reduce their effectiveness (West, 2013). Reducing these errors in paradata may require careful thought about their design and additional training effort. These additional costs will need to be justified. If the reduction in nonresponse bias from adjustments using such data is small, then the budget may be better spent elsewhere. Future waves of this study will seek to expand the paradata relevant for nonresponse adjustments.

References

Alho, J. M. (1990). "Adjusting for Nonresponse Bias Using Logistic Regression." Biometrika **77**(3): 617-624.

Beaumont, J. (2005). "On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment." Survey Methodology **31**(2): 227.

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

Biemer, P. P. and A. Peytchev (2012). "Census Geocoding for Nonresponse Bias Evaluation in Telephone Surveys: An Assessment of the Error Properties." Public Opinion Quarterly.

Couper, M. and L. Lyberg (2005). The Use of Paradata in Survey Research. Proceedings of the International Statistical Institute Meetings.

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

Drew, J. H. and W. A. Fuller (1980). Modeling nonresponse in surveys with callbacks. Proceedings of the Section on Survey Research Methods of the American Statistical Association.

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

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

Holt, D. and D. Elliot (1991). "Methods of Weighting for Unit Non-Response." The Statistician **40**(3): 333-342.

Kalton, G. and D. Kasprzyk (1986). "Treatment of missing survey data." Survey Methodology **12**: 1-16.

Kalton, G. and Maligalig, D. (1991). "A comparison of methods of weighting adjustment for nonresponse." Census Bureau Annual Research Conference, 409-428.

Kish, L. (1988). "Multipurpose Sample Designs." Survey Methodology **14**(1): 19-32.

Kish, L. (1992). "Weighting for unequal P i." Journal of Official Statistics **8**(2): 183-200.

Kreuter, F. and K. Olson (2011). "Multiple auxiliary variables in nonresponse adjustment." Sociological Methods & Research **40**(2): 311-332.

Kreuter, F., K. Olson, J. Wagner, T. Yan, T. M. Ezzati-Rice, C. Casas-Cordero, M. Lemay, A. Peytchev, R. M. Groves and T. E. Raghunathan (2010). "Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys." Journal of the Royal Statistical Society: Series A (Statistics in Society) **173**(2): 389-407.

Little, R. J. A. (1986). "Survey Nonresponse Adjustments for Estimates of Means." International Statistical Review / Revue Internationale de Statistique **54**(2): 139-157.

Little, R. J. A. (1993). "Pattern-Mixture Models for Multivariate Incomplete Data." Journal of the American Statistical Association **88(**421): 125-134.

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

Little, R. J. and S. Vartivarian (2003). "On weighting the rates in non-response weights." Statistics in Medicine **22**(9): 1589-1599.

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

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

Schenker, N., and Gentleman, J. (2001). "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals." The American Statistician, 55, 182-186.

Valliant, R. and J. E. Gentle (1997). "An application of mathematical programming to sample allocation." Computational Statistics & Data Analysis **25**(3): 337-360.

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 (Statistics in Society) **176**(1): 211-225.

Wood, A. M., I. R. White and M. Hotopf (2006). "Using number of failed contact attempts to adjust for non-ignorable non-response." Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(3): 525-542.