ABSTRACT

Title of Dissertation: ADJUSTMENTS FOR NONRESPONSE, SAMPLE QUALITY INDICATORS, AND NONRESPONSE ERROR IN A TOTAL SURVEY ERROR CONTEXT

Cong Ye
Doctor of Philosophy, 2012

Directed By: Dr. Roger Tourangeau
Joint Program in Survey Methodology

The decline in response rates in surveys of the general population is regarded by many researchers as one of the greatest threats to contemporary surveys. Much research has focused on the consequences of nonresponse. However, because the true values for the non-respondents are rarely known, it is difficult to estimate the magnitude of nonresponse bias or to develop effective methods for predicting and adjusting for nonresponse bias. This research uses two datasets that have records on each person in the frame to evaluate the effectiveness of adjustment methods aiming to correct nonresponse bias, to study indicators of sample quality, and to examine the relative magnitude of nonresponse bias under different modes.

The results suggest that both response propensity weighting and GREG weighting, are not effective in reducing nonresponse bias present in the study data. There are some reductions in error, but the reductions are limited. The comparison between response propensity weighting and GREG weighting shows that with the same set of auxiliary variables, the choice between response propensity weighting and GREG weighting makes little difference. The evaluation of the R-indicators and the penalized R-indicators using the study datasets and from a simulation study suggests that the penalized R-indicators perform better than the R-indicators in terms of assessing sample quality. The penalized R-indicator shows a pattern that has a better match to the pattern for the estimated biases than the R-indicator does. Finally, the comparison of nonresponse bias to other types of errors finds that nonresponse bias in these two data sets may be larger than sampling error and coverage bias, but measurement bias can be bigger in turn than nonresponse bias, at least for sensitive questions. And postsurvey adjustments do not result in substantial reduction in the total survey error.

We reach the conclusion that 1) efforts put into dealing with nonresponse bias are warranted; 2) the effectiveness of weighting adjustments for nonresponse depends on the availability and quality of the auxiliary variables, and 3) the penalized R-indicator may be more helpful in monitoring the quality of the survey than the R-indicator.

ADJUSTMENTS FOR NONRESPONSE, SAMPLE QUALITY INDICATORS, AND NONRESPONSE ERROR IN A TOTAL SURVEY ERROR CONTEXT

By

Cong Ye

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
2012

Advisory Committee:
Dr. Roger Tourangeau, Chair
Dr. Steven G. Heeringa
Professor Frauke Kreuter
Professor Partha Lahiri
Professor Michael S. Rendall

Dedication

This dissertation is dedicated to my grandparents, Minggao Ye and Meiying Li, and to my parents, Sunrong Ye and Aiyu Fang, for their continuous encouragement.

Acknowledgments

I owe thanks to many people for advice, guidance and help.

First and foremost among them is my dissertation advisor, Roger Tourangeau. I am thankful for the enormous amount of time he spent helping me stay focused, take one step at a time, and reach my goals. This dissertation would not be possible without his guidance and help. In addition to Roger, I would also like to thank the other members of my committee—Steve G. Heeringa, Frauke Kreuter, Partha Lahiri, and Michael S. Rendall. The inspiring questions and constructive comments I received from them on my dissertation stimulated my thinking and improved my work significantly. For the same reason, I also wanted to thank Richard Valliant who was on my committee but a last-minute emergency prevented him from attending my dissertation defense.

This Dissertation is based upon two datasets. One dataset was collected for a study supported by a grant from the National Science Foundation (SES: 550385) to Roger Tourangeau and Robert Groves. I think Roger for providing the data and Courtney Kennedy for providing technical assistance for the use of the data. The other dataset comes from the Joint Program in Survey Methodology 2005 Practicum survey. I would like to think Roger and Frauke for generously providing me with the data and think Mirta Galesic, Joseph W. Sakshaug, and Duane Gilbert for providing technical assistance for the use of the data.

Finally, I would like to think the faculty, my fellow graduate students, and the staff for the administrative, technical and moral support, and the expert reviews I received on my research.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The decline in response rates in surveys of the general population (de Leeuw and de Heer 2002; Groves and Couper 1998; Hox and de Leeuw 1994) has been a major concern in the survey industry. This challenge is regarded by some as the greatest threat to contemporary surveys (Tourangeau 2004). The increase in nonresponse rates creates concerns that nonresponse bias is also increasing. Nonresponse leads to a reduced sample size, which implies larger variances for the estimates derived from the resulting sample. In addition, nonresponse bias is a function of the nonresponse rate and the difference between respondents and nonrespondents, which means that increases in the nonresponse rate can lead to a larger nonresponse bias if such differences exist. Between the two effects, survey researchers are particularly concerned about the nonresponse bias. If the response process is viewed as a deterministic process, the bias in the respondent mean for respondents can be expressed as

$$B\left(\bar{Y}_r\right)=\frac{n-r}{n}\left(\bar{Y}_r-\bar{Y}_{nr}\right), \tag{1.1}$$

in which n is the sample size, r is the number of respondents, $\bar{Y}_r$ is the respondent mean, and $\bar{Y}_{nr}$ is the nonrespondent mean.

Responding to this challenge, researchers have put forward many proposals for boosting response rates; meanwhile, many researchers have examined the consequences of nonresponse and methods to counter the effects of nonresponse on survey statistics. This research follows the latter approach. It estimates the effectiveness of adjustment methods aiming to correct nonresponse bias, indicators for nonresponse bias, and the relative magnitude of nonresponse bias as compared to other types of survey errors under different modes. This dissertation tries to answer the following three questions: How effective are the adjustment methods in correcting nonresponse bias? How informative are sample quality indicators? And how important is nonresponse bias?

1.1 Dissertation Overview

The outline of this dissertation is as follows. Chapter One reviews the current literature on declining response rates, factors affecting the sample member's response behavior, the relation between response rates and nonresponse bias, adjustment methods, sample quality indicators, and comparisons of nonresponse bias to other types of survey errors. In this chapter, we also briefly discuss the data on which this research is based.

In Chapter Two, we examine the effectiveness of the two weighting methods – response propensity weighting and generalized regression (GREG) weighting. Both weighting methods are explicitly model-based, but response propensity weighting does not ensure that the marginal distributions conform to the population marginal distributions, whereas GREG weighting does. They have been compared to each other and to other weighting methods, but the comparison studies (except simulation studies) rarely have a "gold" standard for the evaluation; in contrast, our study has records available from the frame for every sample member.

In Chapter Three, we propose a modified R-indicator based on the existing R-indicator and evaluate the performance of our new indicator and the existing one in two settings. We first examine the performance of R-indicators and modified R-indicators at different call attempt levels, taking advantage of the records in the study datasets. In addition, we carry out a simulation study to further examine the performance of the original and modified R-indicators.

Chapter Four examines the magnitude of nonresponse bias relative to coverage bias, measurement bias, and sampling error. In addition, it assesses the amount of error reduction that postsurvey adjustments can achieve.

Chapter Five summarizes the findings in this dissertation research, presents general remarks, and points to future research.

1.2 Declining Response Rates and the Efforts to Combat the Trend

Survey researchers would like to get responses from every sample member. However, survey nonresponse has been with survey research since the first sample survey (Hansen and Hurwitz 1946). In the 1990s, more and more researchers have called attention to the declining response rates (e.g., Hox and de Leeuw 1994; Harris-Kojetin and Tucker 1999), and finally a study of nonresponse trends in 16 countries over a 20-year period ending in the 1990s found that both noncontact and refusal rates had been increasing over time (de Leeuw and de Heer 2002). An analysis of several major U.S. surveys (e.g., the Current Population Survey, the National Crime Victimization Survey, the Surveys of Consumers) ending in the 2000s also shows similar findings (Groves et al. 2009).

Groves et al. (2009) distinguish three major types of nonresponse: 1) the failure to deliver the survey request; 2) the refusal to participate; and 3) the inability to participate. Aiming to reduce different types of nonresponse, a variety of techniques has been developed to combat the declining response rates. A particular technique may remove the obstacles from one or more types of nonresponse.

To get the sample member to respond, the first step is to deliver the survey request to the sample member. However, this may not happen for several reasons. This failure may occur if the address, email address, or telephone number is wrong and the survey cannot be delivered. Therefore, efforts are often made to correct the contact information. This failure to make contact may also happen when no one answers or the interviewer cannot get access to a locked building or gated community. More calls/visits help increase the likelihood of contact (Purdon, Campanelli, and Sturgis 1999); it also helps to schedule the calls/visits at different days and time slots (Weeks, Kulka, and Pierson 1987). A longer data collection period also increases the likelihood of contact (Groves and Couper 1998).

After contact is made, there are a number of factors that can affect the sample member's decision to participate. These include 1) survey sponsor, 2) pre-notification, 3) follow-up efforts, 4) incentives, 5) topic interest, and 6) personalization of the request. The findings for each of these variables are clear. First, sample members are more likely to cooperate when the survey sponsor is governmental or academic than when it is commercial (e.g., Groves and Couper 1998). Second, pre-notification generally increases the likelihood of response (e.g., Traugott and Goldstein 1993; Traugott, Groves, and Lepkowski 1987; see, de Leeuw et al. 2007, for a review). Third, increasing the number of attempts brings better response rates (e.g., Heberlein and Baumgartner 1978). Fourth, offering various types of incentives has been tested in many studies and is effective in increasing cooperation (for a recent review, see Singer and Ye, forthcoming). Fifth, Heberlein and Baumgartner (1978) show that interest in the topic of the questionnaire strongly correlates with response rates (see also Groves, Presser, and Dipko 2004). This is not a surprise, as “not interested in the topic” has often been offered as a reason for survey refusals (Bates, Dahlhamer, and Singer 2008; Kulka et al. 1991). Finally, interviewers with high cooperation rates tailor their introductory behavior to individual respondents (Groves and Couper 1998). Similarly, personalizing mailings can increase response rates to mail surveys (Dillman 2007). Other factors, including interview length, privacy concerns, and survey question difficulty, also have an impact on cooperation.

Health problems and language problems are the main reasons for the inability to participate. In order to reduce this type of failure, many surveys offer additional language options for sample members who do not speak English.

1.3 Response Rates and Nonresponse Bias

Concerned about the lower response rates that federal surveys often get these days, the U.S. Office of Management and Budget (OMB) has established an 80 percent response rate standard, which states that agencies should conduct studies to examine the potential nonresponse bias in federal surveys with a response rate less than 80 percent (OMB 2006). And high response rates are seen as a gold standard by some textbook authors (Alreck and Settle 1995; Singleton and Straits 2005). Although high response rates are desirable for surveys, studies show that lower response rates do not necessarily imply higher nonresponse bias. For example, Keeter et al. (2000) compare a standard telephone survey to a more rigorous telephone survey, and find few differences on the survey estimates. A later replication study came to the same conclusion (Keeter et al.

2006). Curtin, Presser, and Singer (2000) show there is no effect of response rates on estimates from the Survey of Consumer Attitudes. Similarly, Merkle and Edelman (2002) find no relationship of nonresponse rates and nonresponse bias in exit polls. In a meta-analysis of 59 studies, Groves and Peytcheva (2008) conclude that large nonresponse biases often exist but there is no clear relationship between nonresponse rates and nonresponse bias.

However, Groves, Singer, and Corning (2000) reason that when a common factor correlates with both nonresponse and nonresponse bias, response rates can affect nonresponse bias. Therefore, instead of focusing on the response rate solely, researchers should focus on whether response propensities and the survey variables are correlated. Assuming every sample member has some nonzero probability of responding, Bethlehem (2002) expresses nonresponse bias in the following form:

$$B\left(\bar{Y}_r\right) \approx \frac{Cov\left(P,Y\right)}{\bar{P}}, \tag{1.2}$$

where P is the response propensity for a sample member, $\bar{P}$ is the population mean of the response propensities, and $Cov\left(P,Y\right)$ is the covariance between the response propensity and the outcome variable Y. Because nonresponse bias is a function of the covariance of Y and P, it should vary over different estimates in a survey.

1.4 Nonresponse Adjustments

Because of nonresponse and its potential to bias estimates, weighting adjustments are often carried out. Therefore, instead of using a simple sample mean,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{1.3}$$

where n is number of respondents and y_i is the value of survey variable y, a weighted mean is used to achieve a better estimate,

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \tag{1.4}$$

where w_i is the adjustment weight assigned for each respondent based on some criteria (of course, weights are often needed to correct for differential selection probabilities as well). Often, auxiliary information, especially demographic information, is used in weighting adjustments, but the adjustment can include any information available for the entire population.

Many weighting methods have been developed for postsurvey adjustments. Kalton and Flores-Cervantes (2003) offer a useful review on this topic. The most common weighting methods used in practice include poststratification, raking, response propensity weighting, and GREG weighting.

1.4.1 Poststratification

Poststratification is a commonly used weighting method (e.g., Oh and Scheuren 1983; Little 1986) that is also known as cell weighting or ratio adjustment. It uses one or more categorical variables to form “cells” (or strata), and assigns a weighting adjustment to all cases in the same stratum. The weights align the sample joint distribution on these variables to the population joint distribution. For example, to generate weights to compensate for nonresponse, the weight for case i in stratum h is computed as

$$w_i = \frac{N_h / N}{n_h / n}, \tag{1.5}$$

where N_h is the number of sample elements in stratum h, and n_h is the number of respondents in stratum h. If the nonresponding units are missing at random (Little and Rubin, 2002) within each adjustment stratum, poststratification will eliminate nonresponse bias. And if the strata are homogeneous with respect to the survey variable, the variance of the estimate for this particular survey variable will also be reduced (Holt and Smith 1979).

One problem with poststratification is that as more variables are used, more post-strata are formed. As a result, there is a risk of creating empty post-strata; this makes the weighting impossible. Another problem with poststratification is that it requires exact information on all cases in the frame, but often information about the nonrespondents is limited. We may only know the marginal totals on some variables. When only marginal totals are available, raking may be the best method available to compensate for nonresponse.

1.4.2 Raking

Raking (Deming and Stephan 1940; Oh and Scheuren 1983) aligns the respondent *marginal* distribution of auxiliary variable to the sample distribution. Therefore, unlike poststratification, raking only needs marginal totals for the population. Raking uses an iterative method to force the subsample row totals and column totals to conform to sample row totals $Z_{r\bullet}$ and column totals $Z_{\bullet c}$. The estimated size for a respondent cell rc (z_{rc}) at t iteration is

$$\hat{z}_{rc}^{(t)} = \begin{cases} \hat{z}_{rc}^{(t-1)} \dfrac{Z_{r\bullet}}{\hat{Z}_{r\bullet}^{(t-1)}} & \text{if } t \text{ is odd,} \\ \hat{z}_{rc}^{(t-1)} \dfrac{Z_{\bullet c}}{\hat{Z}_{\bullet c}^{(t-1)}} & \text{if } t \text{ is even.} \end{cases} \tag{1.6}$$

This procedure continues until convergence is achieved; after t iterations, if the raking procedure converges, we have

$$\hat{z}_{rc}^{(t)} = z_{rc}^{(0)} \prod_{\substack{i=1 \\ i\ odd}}^{t} \frac{Z_{r\bullet}}{\hat{Z}_{r\bullet}^{(i)}} \prod_{\substack{i=2 \\ i \text{ even}}}^{t} \frac{Z_{\bullet c}}{\hat{Z}_{\bullet c}^{(i)}}. \tag{1.7}$$

However, convergence is not guaranteed (Ireland and Kullback 1968).

The weight w_i is computed in the same way as in Equation 1.5. As with poststratification, raking can eliminate nonresponse bias if after controlling for the auxiliary variables, the nonresponse units are missing at random. In addition, there is no interaction effect between the row and column variables, a condition that is not required in poststratification.

1.4.3 GREG Weighting

GREG weighting (Särndal and Lundström 2005; Särndal, Swensson, and Wretman 1992) is another method of forcing subsample totals conform to the sample totals. GREG estimation is motivated by the linear model, which describes the linear relationship between the outcome variable and a vector of x variables. The GREG estimator of the population total takes the form of an adjusted total:

$$\hat{Y}_{GREG} = \hat{Y} + \left(X - \hat{X}\right)^{T} \hat{\beta}, \tag{1.8}$$

in which $\hat{Y}$ is the sample estimate of the total for the outcome variable, and $\hat{X}$ is the sample estimate of the totals for the x variables, and X is the known population total for the vector of x variables. The set of weights resulting from GREG calibration is

$$w_i = d_i \left[1 + \left(X - \hat{X} \right)^T \left(X_*^T D V^{-1} X_* \right)^{-1} x_i \Big/ v_i \right], \tag{1.9}$$

in which d_i is the base weight before the GREG adjustment, X_*^T is the $n \times p$ matrix of x variables for the respondent sample, $D = diag\left(d_i\right)$, v_i is the variance of residuals from the model $y_i = x_i^T \beta + \varepsilon_i$, and $V = diag(v_i)$.

As with poststratification and raking, GREG weighting eliminates the bias if after controlling for the vector of x variables, the nonresponse units are missing at random.

1.4.4 Response Propensity Weighting

As discussed in Section 1.3, if every sample member has some nonzero probability of responding, we can estimate the response propensities using a logistic regression model:

$$\log\left(\frac{p_i}{1 - p_i} \right) = x_i^T \beta, \tag{1.10}$$

where x_i^T is a vector of x variables, and β is a vector of logistic regression coefficients. The vector of x variables must be known for both the respondents and nonrespondents in order to fit the model. The response propensity is predicted as

$$\hat{p}_i = \frac{\exp\left(x_i^T \hat{\beta} \right)}{1 + \exp\left(x_i^T \hat{\beta} \right)}. \tag{1.11}$$

and the weight w_i is equal to the inverse of the response propensity $\hat{p}_i$ (Rosenbaum and Rubin 1983):

$$w_i = \frac{1}{\hat{p}_i}. \tag{1.12}$$

The bias goes to zero if the response propensities depend only on the *x* variables. In other words, response propensity weighting eliminates nonresponse bias if, after controlling for the *x* variables, the nonresponse units are missing at random.

1.4.5 Auxiliary Variables for Weighting

The weighting adjustment methods are effective only if the right set of variables is available to survey researchers. Therefore, searching for auxiliary variables is a critical effort in successful nonresponse weighting. Different types of auxiliary variables can be identified from difference sources.

National registers and other administrative data. Some European countries maintain population registers, which contain information on individuals and households. Similarly, some organizations maintain administrative databases that have rich auxiliary information. If the individuals and households can be linked to the elements of the survey sample, the auxiliary information can be used in the nonresponse weighting methods reviewed above (e.g., Bethlehem, Cobben, and Schouten 2011; Särndal and Lundström 2005) and more generally to improve survey estimation and inference.

Commercial databases. In the United States, some commercial companies (e.g., Survey Sampling International) offer commercial databases that incorporate other supplementary information to US addresses and telephone numbers. By linking the cases in the survey sample to the databases or sampling using the databases as the frame, the rich information on the frame can be used in nonresponse weighting (e.g., Link and Lai

2011). However, we should note that incomplete information is common in commercial databases.

Two other sources of information can be used in the adjustment process—paradata and aggregate data. Paradata are by-products of the survey itself. They can be used in nonresponse weighting (e.g., Kreuter et al. 2010; Olson, forthcoming). This type of data includes call records, disposition codes, interviewer characteristics, interviewer observation of sample members, and keystroke data. Benchmark surveys (e.g., the Current Population Survey) can provide aggregate data on some useful variables for surveys of general population. Many surveys calibrate the marginal distribution of some demographic variables to that of the Current Population Survey (CPS).

1.4.6 Comparing Response Propensity Weighting and GREG Weighting to Other Weighting Methods

The response propensity weighting method has been used in various surveys, and comparisons of this method with other weighing methods have been reported. For example, Lepkowski, Kalton, and Kasprzyk (1989) compared a propensity score weighting method to the traditional cell weighting method (in which the response propensity for a given respondent is estimated by the inverse of the response rate within his or her weighting cell) for the Survey of Income and Program Participation (SIPP). Data collected in the initial interview were used to predict the response status in successive waves of SIPP. Their analyses found that the two weighting methods did not differ much. Also using SIPP data, Folsom and Witt (1994) and Rizzo, Kalton, Brick, and Petroni (1994) compared response propensity weighting methods to cell weighting method and/or CHAID, but neither study found clearly different results in nonresponse adjustment. Carlson and Williams (2001) compared the propensity method to the weighting cell method, and found little difference between the two methods in their analysis of the Community Tracking Study (CTS) survey. Ekholm and Laaksonen (1991) compared response propensity weighting to poststratification weighting and found the results are similar. Smith et al. (2001) tried to use propensities of obtaining adequate provider data to adjust nonresponse bias in the National Immunization Survey (NIS). However, estimates were not that different between using this propensity method and the original poststratification method. Kreuter et al. (2010) used both demographic variables and paradata in the response propensity weighing for five different surveys, but found the weighted estimates did not change much from unweighted ones. Several other studies (e.g., Battaglia et al. 1995; Brick, Waksberg, and Keeter 1996; Duncan and Stasny 2001; Hoaglin and Battaglia 1996) evaluated weighting methods that use the propensities of being a nontelephone (or transient telephone) household to make adjustments to the weights in a telephone survey, but there is no way of knowing which weighting procedure was better in these comparisons, because there is no validating data available. In contrast, with court records available for bias calculation, Lin and Schaeffer (1995) tried to make adjustments for nonresponse using number of call attempts (methods 1) or call results (method 2) to classify sample members. However, they found that these methods did not reduce bias. In a simulation study, Biemer and Link (2008) suggest response propensity weighting adjustment based upon a callback model should be used, either in lieu of poststratification or in a combination with it. Also in a simulation study, Garren and Chang (2002) concluded that using estimated propensities of being nontelephone households to make adjustments can result in reduction in coverage bias. Still, it seems that more empirical studies are needed to access how effective the propensity method is and when it is effective.

Interest in the GREG weighting methods has been growing in recent years. A number of studies have been conducted to investigate the potential use of the method. Notable examples include the application of the GREG weighting methods on Canadian population censuses by Bankier and his colleagues (Bankier, Rathwell, and Majkowski 1992; Bankier, Houle, and Luc 1997; Bankier and Janes 2003) and on the American Community Survey (ACS) by Fay (Fay 2005; Fay 2006). A few studies have compared the GREG weighting method to other weighting methods in nonresponse adjustment. Fuller, Loughin, and Baker (1994) applied regression calibration estimation to adjust for nonresponse for the National Food Consumption Survey. Using the 1999 National Household Survey on Drug Abuse data, Folsom and Singh (2000) compared raking and GREG weighting for nonresponse adjustments. They found that the estimates for use of cigarettes, alcohol, marijuana, and cocaine were close to each other under the different weighting schemes. Bethlehem and Schouten (2004) applied the GREG weighting method with different models to the 1998 Dutch Integrated Survey on Household Living Conditions (POLS). They found that biases could be reduced but still remained after the weighting. Comparison of raking, response propensity weighting and GREG weighting methods has been done on the Education Longitudinal Study data by Siegel, Chromy, and Copello (2005). They also found similar results in nonresponse adjustments. These studies do not have validation data to check the effectiveness of each weighting method in comparison with the true values for the parameter estimates.

1.5 Quality Indicators

As suggested by the literature reviewed in Section 1.3, response rates are not good indicators of sample quality. Lower nonresponse rates do not necessarily imply lower nonresponse bias in the survey estimates. Although more and more researchers have questioned the value of response rates as indicators of nonresponse error or overall survey quality, no alternatives have been accepted in the field (Groves et al. 2008). The weighting methods reviewed in the previous section aim to correct potential nonresponse bias in the survey estimates, but do not provide a measure indicating how good the survey sample is either before or after weighting.

Two sets of alternative indicators have gained some attention in recent years (see Wagner 2012, for a review of alternative indicators). Schouten, Cobben and Bethlehem (2009) proposed the R-indicator, which is a measure based on the variation of the response probabilities estimated from a model with a set of auxiliary variables. Särndal (2011) proposed three balance indicators (BI_1-BI_3), which measures differences of the response means and sample means of auxiliary variables. As Särndal points out, BI_1 is similar and “sometimes identical” to the R-indicator (Särndal 2011, p12).

Rather than relying on remedies for correcting nonresponse error after the data are collected, the quality of the survey can be monitored continuously and the researchers can actively intervene into the recruitment protocol to achieve a sample that is close to the targeted one. The approach to modify the design based on process data during the data collection period has been labeled “responsive design” (Groves and Heeringa 2006). This idea has invited various applications (e.g., Laflamme and Karaganis 2010; Mohl and

Laflamme 2007; Peytchev et al. 2010). Because the quality indicators can be computed at any stage of the data collection period, it is natural to use them to monitor and guide the field work. Furthermore, partial R-indicators are developed particularly to guide data collection decisions in adaptive and responsive survey designs (Schouten, Shlomo, and Skinner 2011).

Although the R-indicator has gained some attention from survey researchers, few studies examining the R-indicator have been published except by the authors who proposed it originally (Schouten, Cobben, and Bethlehem 2009; Schouten et al. 2012; Schouten, Shlomo, and Skinner 2011). In addition, no study has evaluated its performance using frame records.

1.6 Nonresponse Error in a Total Survey Error Context

From a survey quality perspective, nonresponse error is only one component of the total error in the survey estimate. Other major sources of errors include coverage error, sampling error, measurement error, and even adjustment error (Groves et al. 2009). Because every survey only has limited resources that can be allocated, survey researchers face the problem of minimizing the total survey error given a fixed cost. To optimize the allocation of the resources, the relative magnitude of each source of errors has to be evaluated.

Coverage error usually results from an imperfect frame from which some of the units have been missed, resulting in undercoverage. If the frame contains units that do not belong to the target population, overcoverage errors can also occur. *Sampling error* occurs because only a fraction of the units in the population of interest is included in the sample. Often, some of the sample members fail to respond to the survey and this introduces *nonresponse error*. *Measurement error* emerges when the responses obtained from the sample member do not agree with the true values. The instrument and mode of data collection often play important roles in this type of error. *Processing error* arises after data collection and before estimation. Some editing rules aiming to resolve illogical answers alter responses and cause missing values. Coding for open questions also inevitably introduces error. Although there are both bias and variance components in each type of error except sampling error, we only intend to investigate the variance produced by sampling error and the bias produced by other types of error. In practice, postsurvey adjustments are used to reduce the effects of coverage and nonresponse biases on the estimates, but they can introduce errors of their own. Groves et al. (2009) present a figure (see Figure 1.1) that depicts these errors in the context of the main survey stages.

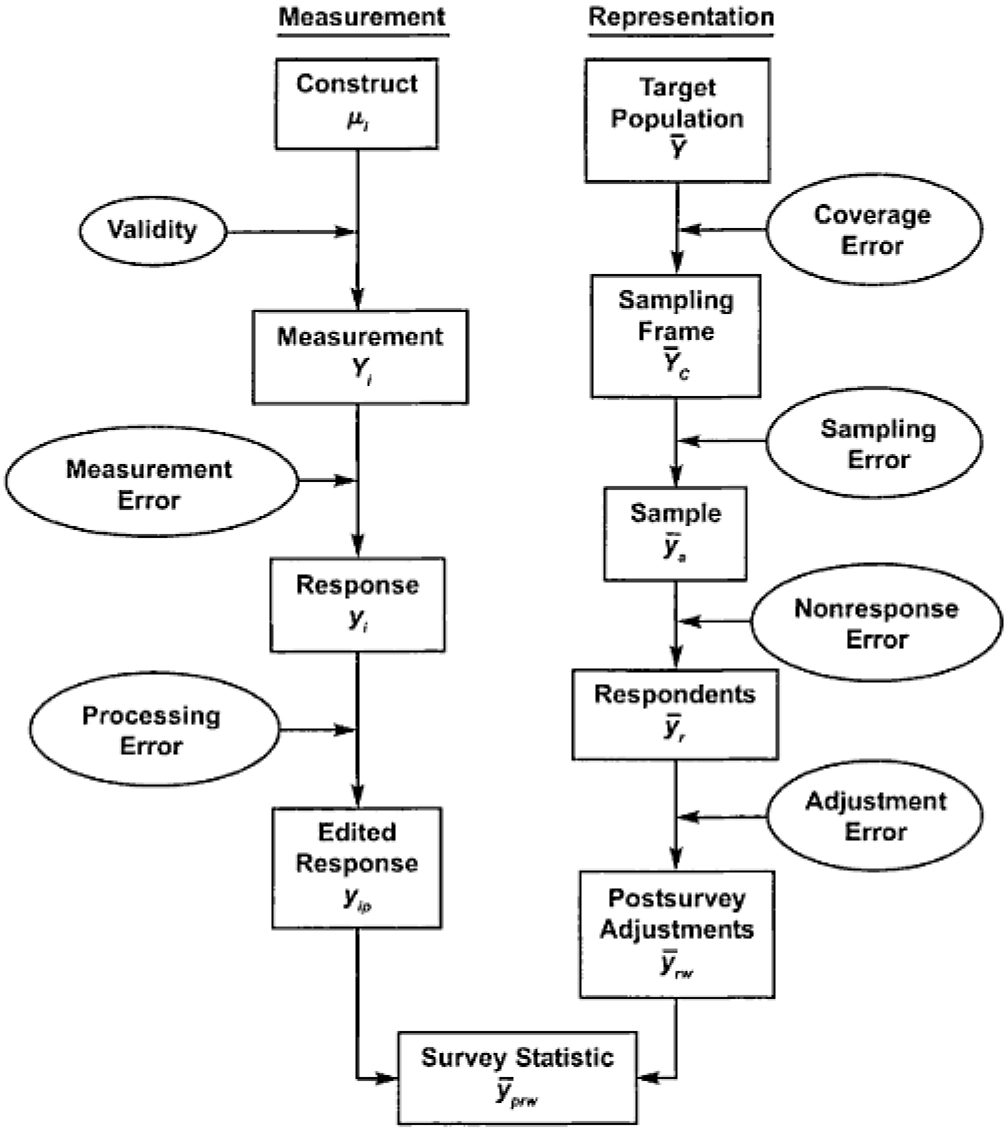

Figure 1.1. Survey lifecycle from a quality perspective (Source: Groves et al. 2009: 48)

Individual errors have been the focus of many investigations, and comparison of multiple errors rarely goes beyond two types of errors. Peytchev, Carley-Baxter, and Black (2011) investigate both coverage bias and nonresponse bias in their study which compared a landline telephone survey with a cell phone survey and a follow-up survey on nonrespondents. They found that coverage biases and nonresponse biases were in opposite directions and that coverage biases were larger.

Using court records as benchmarks, Schaeffer, Seltzer, and Klawitter (1991) and Olson (2006) compared the magnitudes of nonresponse biases and measurement biases, but found inconsistent results. Schaeffer, Seltzer, and Klawitter (1991) showed that measurement bias was higher than nonresponse bias for amounts of support owed and paid; however, for proportion of cases with any support owed and paid, nonresponse bias was greater than measurement bias. Olson (2006) found that nonresponse bias was greater than measurement bias for the mean length of marriage and the mean number of marriages, but for mean time elapsed since divorce, the reverse was true. Mixed results were also found by Biemer (2001), who used a reinterview survey design. Nonresponse bias was higher than measurement bias for some items (e.g., whether the sample member ever stopped smoking for at least one day during the past 12 months), while measurement bias was larger for some items (e.g., whether the sample member would like to quit smoking completely). For the CATI survey, measurement bias was higher than nonresponse bias for whether the sample member had smoked at least 100 cigarettes during the lifetime, but the magnitudes of the two sources of error reversed for the face-to-face survey. For the item whether there is firearm in or around the sample member’s home, nonresponse bias is greater than measurement bias for the CATI survey, while the reverse is observed for the face-to-face survey. Using the report of abortions in the audio computer-assisted self-interviewing (ACASI) mode as a “gold standard” to assess the report of abortions in the computer-assisted personal interviewing (CAPI) mode, Peytchev, Peytcheva, and Groves (2010) found that CAPI respondents in the lowest response propensity quintile tended to be more likely to underreport their abortions. Tourangeau, Groves, and Redline (2010) find that measurement bias was about twice as large as nonresponse bias for two voting behaviors, using voter registration records as true values. In an investigation using the same alumni dataset, Sakshaug, Yan, and Tourangeau (2010) suggest that measurement bias tends to be the larger than nonresponse bias for estimates of socially undesirable characteristics, but not for estimates of socially desirable or neutral characteristics, where nonresponse biases were larger.

1.7. Study Datasets

This dissertation re-examines these issues using two datasets in which accurate data are available from the frame for both respondent and nonrespondent members of the two samples.

1.7.1 The Maryland Registered Voters Dataset

The first dataset (see Tourangeau, Groves, and Redline 2008 for more information about the data) consists of a list of 50,000 Maryland residents who were registered to vote. These data were purchased from Aristotle (http://www.aristotle.com). The Aristotle database contained fields for voting history, various demographic variables, and contact information. Two strata, voters and nonvoters, were created for sample selection. Registered residents who voted in either the 2004 or 2006 general election were classified as voters. A pretest was carried out using a random sample of 500 voters and 500 nonvoters. The 49,000 remaining records were sorted by Congressional district, party registration, and predicted quality of matching the telephone number. A systematic sample of 2,689 records was drawn for the main study, with selections done separately for voters and nonvoters. The final sample had 1,346 voters and 1,343 nonvoters. Because this study focused on the nonresponse bias, comparing sample values to respondent values, there were no initial weights used in the analysis.

The study included a mode experiment. 1,669 cases were assigned to receive a mail survey; 1,020, a telephone survey. The response rate for the telephone survey was 34.3 percent (AAPOR RR1); for the mail survey, the response rate was 33.2 percent. There were also an incentive experiment and a framing experiment. The cases in both modes were randomly assigned to either receive $5.00 cash incentives or no incentive, and under both modes, the survey topic was identified as "Health & Lifestyles" for a random half of the sample and as "Politics, Elections, and Voting" for the other half. The distribution of cases by experimental condition and stratum is shown in Table 1.1.

Table 1.1. Distribution of cases by experimental condition and stratum, The Maryland registered voters dataset

Condition	Sample Members	Completes	%
Overall	2689	904	33.6
Telephone	1020	350	34.3
Mail	1669	554	33.2
$5 incentive	1349	591	43.8
No incentive	1340	313	23.4
Nonvoter	1343	348	25.9
Voter	1346	556	41.3
Politics, elections, and voting	1346	441	32.8
Health and lifestyles	1343	463	34.5

The Aristotle database contained information on voting history in the 2004 and

2006 general elections. It also had a variety of auxiliary variables. Fourteen frame variables were identified. Nine of them were dichotomous variables (whether the person ever donated to various organizations, whether he or she was computer owner, whether he or she had a home business, party identification, whether the person was on the Federal Do Not Call list, whether the person was a head of household, whether the person reported on religion, sex, and whether the person reported on ethnicity), and the other five variables were continuous variables (number of persons in the household, age, and income) or treated as continuous variables (education level and home ownership level—renter, probable renter, probable homeowner, and homeowner).

1.7.2 The University of Maryland Alumni Dataset

The second dataset (see Kreuter, Presser, and Tourangeau 2008 for more information about the data) comes from a survey conducted by the 2005 Practicum class at the Joint Program in Survey Methodology (JPSM) at the University of Maryland. The target population of survey was University of Maryland alumni who received undergraduate degrees from 1989 to 2002. A random sample of 20,000 graduates was drawn from 55,320 individuals listed as graduates during this period in the Registrar's records. Telephone numbers were matched to only 10,325 of the 20,000 sampled. After excluding those who were listed as residents of Puerto Rico, the Virgin Islands, or on military bases (14), those having the same telephone number as another graduate (151; only one of whom was randomly picked), and those used in the pretest sample (1975), 7,591 of them were selected and received at least one call for the main study (Table A1 lists final dispositions for all cases). The sample alumni were contacted by telephone for a brief screener survey and the screener completes were randomly assigned to one of the three modes of data collection (telephone, Web, and IVR). A total of 1,501 cases completed the telephone screener (AAPOR Response Rate 1: 31.9%) and were randomly assigned to one of the three modes of data collection (telephone, Web, and IVR). Cases assigned to the telephone mode continued with the main survey immediately after the screener; those assigned to the Web option were told to follow the instructions in a letter they were sent; and those assigned to IVR were switched to IVR after they completed the screener. The response rates for this final stage were 94.7%, 56.8%, and 61.1% for telephone, Web, and IVR, respectively.

The University of Maryland alumni dataset contained variables on the graduates' academic performance and on their relationship with the University. These variables could be checked against records available from the Registrar's Office or the Alumni Association. These variables included the sampled member's GPA, whether he or she dropped a class, received an unsatisfactory grade, received an academic warning or was being placed on probation, received academic honors, was a member of the Alumni Association, and donated money to the University of Maryland after graduation or in last year (2004). Some additional auxiliary variables were also available, including state of residence, age, sex, and time of getting the degree.

Chapter 2: Assessing Effectiveness of Nonresponse Adjustment Methods: Response Propensity Weighting and Generalized Regression Calibration Estimation

2.1 Introduction

Nonresponse can pose serious problems for researchers, because sample members with some characteristic of interest may be more or less likely to provide data than those without it. If that is the case, the estimate from the sample will be a biased estimate of the true population value (Lessler and Kalsbeek 1992). The bias in the respondent mean can be expressed as

$$B\left(\bar{Y}_r\right)=\frac{n-r}{n}\left(\bar{Y}_r-\bar{Y}_{nr}\right), \tag{2.1}$$

in which n is the original sample size, r is the number of respondents, $\bar{Y}_r$ is the respondent mean, and $\bar{Y}_{nr}$ is the nonrespondent mean. This expression reflects the deterministic view of the response process. Alternatively, the stochastic view assumes every sample member has some nonzero probability of responding, and the bias arises when the response propensity and outcome variable covary. Under this framework, the bias in the respondent mean as an estimate of the population mean is given by (Bethlehem 2002):

$$B\left(\bar{Y}_r\right)\approx\frac{Cov\left(P,Y\right)}{\bar{P}}, \tag{2.2}$$

where P is the response propensity for a sample member, $\bar{P}$ is the population mean of the response propensities, and $Cov\left(P,Y\right)$ is the covariance between the response propensity and the outcome variable Y. As Rosenbaum and Rubin (1983) suggest, if we know the response propensity for each sample member, we can make a nonresponse adjustment by assigning a weight W_i that is equal to the inverse of the response propensity P_i:

$$W_i = \frac{1}{P_i}. \tag{2.3}$$

However, in practice, the true response propensities are unknown; typically, they are estimated using a logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta , \tag{2.4}$$

where x_i^T is a vector of x variables, and β is a vector of logistic regression coefficients. The fitted response propensity is

$$\hat{p}_i = \frac{\exp\left(x_i^T \hat{\beta}\right)}{1+\exp\left(x_i^T \hat{\beta}\right)}. \tag{2.5}$$

The bias goes to zero if the data are missing at random (MAR), in the terminology introduced by Little and Rubin (2002). MAR states that

$$P\left(R \mid X^T, (Y_r, Y_{nr})\right) = P\left(R \mid X^T, Y_r\right). \tag{2.6}$$

That is, the response propensities depend only on the x variables. After controlling for the x variables, the nonresponse units are missing at random (MAR).

The response propensity weighting method is widely used in survey practice. Lee and Valliant (2007) review the use of propensity adjustments in telephone surveys (see Valliant and Dever [2011] for the use of propensity adjustments in web surveys). Little (1986) suggests creating weighting classes based on propensity scores to avoid large variation in the weights, which might offset the gain in bias reduction. And when large variation in the weights is observed, the conventional choice in practice is to create five subclasses (following Cochran 1968; Little and Rubin 2002) based on the quintiles of the propensity scores.

The response propensity weighting method has been used in various surveys, and comparisons of this method with other weighing methods have been reported. For example, Lepkowski, Kalton, and Kasprzyk (1989) compared propensity score weighting to the traditional cell weighting method (in which the response propensity for a given respondent is estimated by the inverse of the response rate within his or her weighting cell) for the Survey of Income and Program Participation (SIPP). Data collected in the initial interview were used to predict the case’s response status in successive waves of SIPP. Their analyses found that the two weighting methods did not differ much. Also using SIPP data, Folsom and Witt (1994) and Rizzo, Kalton, Brick, and Petroni (1994) compared response propensity weighting methods to cell weighting method and/or CHAID, but neither study found clearly different results in nonresponse adjustment. Carlson and Williams (2001) compared the propensity method to the weighting cell method, and found little difference between the two methods in their analysis of the Community Tracking Study (CTS) survey. Ekholm and Laaksonen (1991) compared response propensity weighting to poststratification weighting and found the results were similar under the two weighting schemes. Smith et al. (2001) tried to use the propensities of obtaining adequate provider data to adjust for nonresponse bias in the National Immunization Survey (NIS). However, the estimates were not that different using this propensity method from those based on the original poststratification method. Kreuter et al. (2010) used both demographic variables and paradata in the response propensity weighing for five different surveys, but found the weighted estimates did not change much from unweighted ones. Several other studies (e.g., Battaglia et al. 1995; Brick, Waksberg, and Keeter 1996; Duncan and Stasny 2001; Hoaglin and Battaglia 1996) evaluated weighting methods that use the predicted propensities of being a nontelephone (or transient telephone) household to make adjustments to the weights in a telephone survey, but there is no way of knowing which weighting procedure was better in these comparisons, because there were no validating data available. In contrast, with court records available for bias calculation, Lin and Schaeffer (1995) tried to make adjustments for nonresponse using number of call attempts (methods 1) or call results (method 2) to classify sample members. However, they found that these methods did not reduce bias. In a simulation study, Biemer and Link (2008) argued that response propensity weighting adjustment based upon a callback model should be used, either in lieu of poststratification or in a combination with it. Also in a simulation study, Garren and Chang (2002) concluded that using estimated propensities of being nontelephone households to make adjustments can result in reduction in coverage bias. Overall, these studies do not provide clear-cut evidence regarding the effectiveness of nonresponse weighting adjustments on the advantages of any specific weighting method. More empirical studies are needed to access how effective the propensity method is and when it is effective.

There are a number of weighting methods used in practice, and response propensity weighting is just one of them. Response propensity weighting does not ensure that the marginal distributions conform to known population marginal distributions, as calibration methods such as poststratification or raking does. Like propensity weighting, calibration techniques use the auxiliary variables to create postsurvey weighting adjustments. Calibration finds a new set of weights that have minimal distance from the original weights, but the new set of weights reproduce population totals on the auxiliary variables exactly. Two common postsurvey adjustments, poststratification and raking, are special cases of calibration estimation, as discussed in Deville and Särndal (1992). However, we can use various other linear calibration models, such as linear, bounded linear, raking, bounded raking, and logit calibration functions. These weighting methods do not explicitly appeal to an underlying model. In contrast, GREG weighting (Särndal and Lundström 2005; Särndal, Swensson, and Wretman 1992) is an explicitly model-based weighting procedure that incorporates auxiliary variables in linear regression, and thus allows specifying main effects and interaction effects among the auxiliary variables. This explicit formulation of a model is an advantage shared by response propensity weighting. GREG estimation is motivated by the linear model, which describes the linear relationship between the outcome variable and a vector of x variables. The GREG estimator of the population total takes the form of an adjusted total:

$$\hat{Y}_{GREG} = \hat{Y} + \left(X - \hat{X} \right)^T \hat{\beta}, \tag{2.7}$$

in which $\hat{Y}$ is the standard estimated total for the outcome variable, $\hat{X}$ represents the sample estimates of the totals for the x variables, and X represents the known population totals for a vector of x variables. The set of weights results from calibration is

$$w_i = d_i \left[1 + \left(X - \hat{X} \right)^T \left(X_*^T D V^{-1} X_* \right)^{-1} x_i \big/ v_i \right], \tag{2.8}$$

in which d_i is the base weight before calibration, X_*^T is the $n \times p$ matrix of x variables for the respondent sample, $D = diag\left(d_i\right)$, v_i is the variance of residuals from the model in which y is regressed on the x variables, and $V = diag(v_i)$. One of the main goals of weight calibration is to adjust for nonresponse. When there is potential coverage bias, calibration is often used as a second stage of postsurvey adjustment following some other form of weighting. Chang and Kott (2008) propose a different calibration method for nonresponse adjustments. They first use a response propensity model and estimate the totals with these estimated propensities; they then calibrate the weights to the estimated totals. This method is not considered in this research.

Interest in the GREG weighting methods has been growing in recent years. A number of studies have investigated the potential use of the method. Notable examples include the application of the GREG weighting methods on Canadian population censuses by Bankier and his colleagues (Bankier, Rathwell, and Majkowski 1992; Bankier, Houle, and Luc 1997; Bankier and Janes 2003) and on the American Community Survey (ACS) by Fay (Fay 2005; Fay 2006). A few studies have compared the GREG weighting method to other weighting methods in nonresponse adjustment. Fuller, Loughin, and Baker (1994) applied regression calibration estimation to adjust for nonresponse for the National Food Consumption Survey. Using the 1999 National Household Survey on Drug Abuse data, Folsom and Singh (2000) compared raking and GREG weighting for nonresponse adjustments. They found that the estimates for use of cigarettes, alcohol, marijuana, and cocaine were close to each other under the different weighting schemes. Bethlehem and Schouten (2004) applied the GREG weighting method with different models to the 1998 Dutch Integrated Survey on Household Living Conditions (POLS). They found that biases could be reduced but still remained after the weighting. Comparisons of raking, response propensity weighting and GREG weighting methods have been made on the Education Longitudinal Study data by Siegel, Chromy, and Copello (2005). They also found similar results for the different nonresponse adjustments. None of these comparisons have validation data to check the effectiveness of each weighting method. They compare the final estimates under different weighting schemes, but not the final biases.

This study examines two common and explicitly model-based postsurvey weighting strategies—response propensity weighting and GREG weighting. The reasons to focus on these two weighting methods are that 1) both methods are explicitly model-based; 2) both methods can incorporate continuous variables as well as interaction terms in the models; 3) other calibration methods are equivalent to the GREG in large samples; and 4) there is good reason to expect a difference between the two methods because response propensity weighting does not ensure that the marginal distributions conform to the population marginal distributions, whereas GREG weighting does.

2.2 Study 1

2.2.1 Study Dataset

The Maryland registered voters dataset. We will refer to this data set as the voters data. A list of 50,000 Maryland residents who were registered to vote was purchased from Aristotle (http://www.aristotle.com). The Aristotle database contained fields for voting history, various demographic variables, and contact information. Two strata, voters and nonvoters, were created for subsample selection. Registered residents who voted in either the 2004 or 2006 general election were classified as voters. A pretest was carried out using a random of 500 voters and 500 nonvoters. The 49,000 remaining records were sorted by Congressional district, party registration, and predicted quality of matching the telephone number. A systematic sample of 2,689 records was drawn for the main study, with selections done separately for voters and nonvoters. The final sample had 1,346 voters and 1,343 nonvoters. This study focuses on the nonresponse bias by comparing sample values to respondent values. It does not use the base weights in the analysis.

The original study included a mode experiment. A total of 1,669 cases were assigned to receive a mail survey; 1,020, a telephone survey. The response rate for the telephone survey was 34.3 percent (AAPOR RR1); for the mail survey, the response rate was 33.2 percent. There were also an incentive experiment and an experiment on the "framing" or description of the survey to prospective respondents. Cases in both modes were randomly assigned to either receive a $5.00 cash incentive or no incentive; similarly, the survey topic was identified as "Health & Lifestyles" for a random half of the sample and as "Politics, Elections, and Voting" for the other half (see Tourangeau, Groves, and Redline 2008 for more information about the data).

2.2.2 Variables

Variables of interest. Table 2.1 lists the survey variables of interest and auxiliary variables. The Aristotle database contained information on voting history in the 2004 and 2006 general elections. The surveys also asked questions about whether the respondent had voted in the 2004 and 2006 general elections. These variables are dichotomous variables. Although we have reported values from the respondents on these two variables of interest, to avoid potential measurement errors, this study only uses the frame values in all analyses.

Table 2.1. Variables from the voters dataset

Variable	**Short Name**
Variable of interest	
Voted in 2004 vs. other	Vote04
Voted in 2006 vs. other	Vote06
Auxiliary variable	
Ever donate to various organizations vs. other	Donate
Computer owner vs. other	Comp_own
Home business vs. other	Home_biz
Party identification (Dem. vs. Rep.)	Party_code
On Federal Do Not Call list vs. other	Do Not Call List
Head of household vs. other	HOH
Religious vs. other	Religious
Sex (male vs. female)	Sex
Known Ethnicity vs. other	Ethnicity
Home ownership (renter, probable renter, probable homeowner, and homeowner)	Home_own
Number of persons in the household	Persons_HH
Age in years	Age
Education level (5 levels)	Edu_level
Income level (12 levels)	Income
Experimental variable	
Incentive vs. no incentive	Incentive
Description of the survey topic	Topic
Paradata	
Contact or not	Contact
Number of call attempts	#_calls

Auxiliary variables. Fourteen frame variables were used as auxiliary variables. Nine of them were dichotomous variables (whether the person ever donated to various organizations, whether he or she was computer owner, whether he or she had a home business, party identification, whether the person was on the Federal Do Not Call list, whether the person was a head of household, whether the person reported on religion, sex, and whether the person reported on ethnicity), and the other five variables were continuous variables (number of persons in the household, age, and income) or treated as continuous variables (education level, and home ownership level—renter, probable renter, probable homeowner, and homeowner). The two experimental variables were the incentive and the description of the survey topic, both of which have two categories. Two variables were created from call records (number of call attempts, whether contact was made or not). These two variables are relevant only to the telephone sample.

2.2.3 Imputation for Missing Values

Although rich information is available for the sample, some of the variables do have missing values. Missing rates were less than 2%. Discarding the missing cases would affect all the variables in the model, including the dependent variable. Therefore, we imputed values so the data form a rectangular data table without missing data. Frame variables with missing values were imputed using the multivariate sequential regression imputation method (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001). This procedure imputes one variable at a time, using complete and imputed values for all other variables. The two variables of interest (*vote04* and *vote06*) were excluded from the imputation models. Only one imputed dataset was requested. This single imputation treatment will underestimate the variance in the imputed variables, but it is ignored here because the main concern in this study is bias and how bias is affected by different weighting methods.

2.2.4 Correlation between Response/Voting Status and Auxiliary Variables

Figure 2.1 shows the correlations between response to the surveys and voting status and the auxiliary variables for the telephone sample. As shown in the figure, the correlations between whether the respondent voted in 2004 and the auxiliary variables are similar to those between whether the respondent voted 2006 and the auxiliary variables. In fact, these two sets of correlations are almost identical. The correlations between response status and auxiliary variables are not that different from the previous two sets of correlations; however, apparent differences are observed for incentive group, whether the sample person was contacted or not, the number of calls, the sample person's sex, which level of home ownership he or she was on, the value of the home, and the income level variable. Response status hardly shows any correlations with these variables, except for the incentive group. The incentive has little correlation with the voting behaviors. This is in line with prior research which suggests that incentives hardly correlate with survey variables (for a review, see Singer and Ye, forthcoming). Including the variable incentive is likely to inflate the variance of estimates while leaving bias unaffected. Therefore, we excluded incentive as a candidate variable.

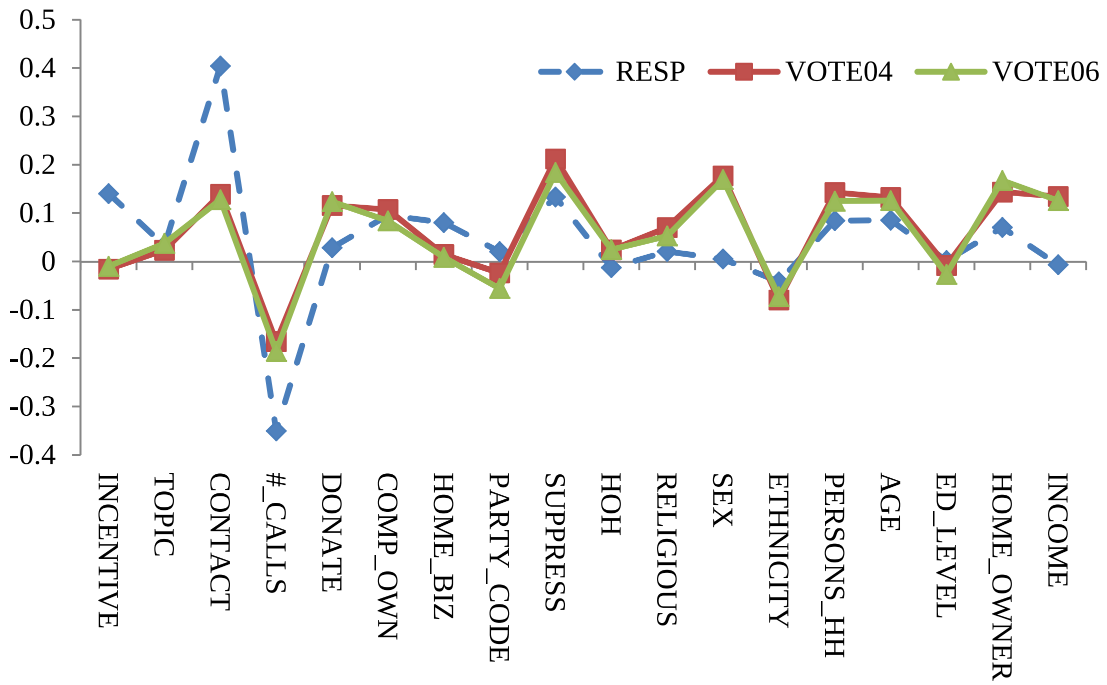

Figure 2.1. Correlations between frame variables and whether the case is a respondent and whether the case voted in 2004/2006, telephone sample

A similar pattern of relationships is also observed for the mail sample (Figure 2.2). The two sets of correlations between the voting variables and the auxiliary variables are almost identical. Differences between these two sets of correlations and those for response status are observed for incentive group, home business, Do Not Call List, sex, education level, number of persons in the household, education level, home ownership, and income variables. In general, the correlations with the auxiliary variables are lower for the response status than for voting status, with one exception; the incentive is more strongly correlated with responding than with voting. As in the telephone sample, we excluded incentive as a candidate variable.

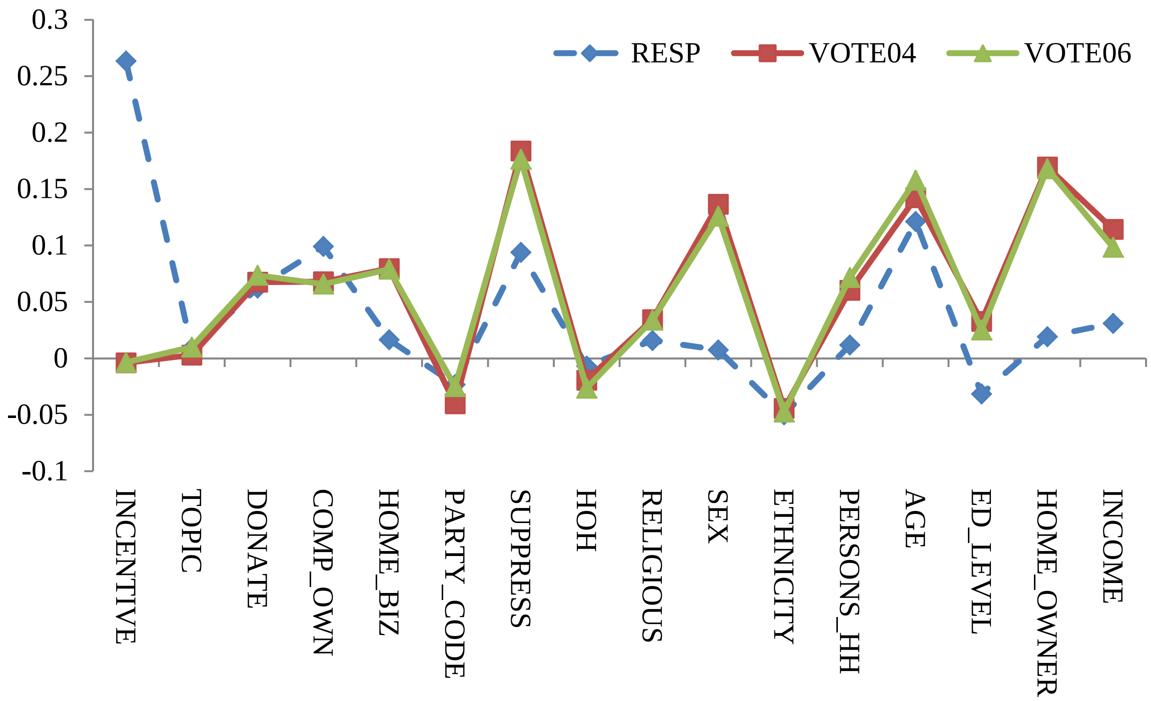

Figure 2.2. Correlations between auxiliary variables and whether the case is a respondent and whether the case voted in 2004/2006, mail sample

2.2.5 Model Selection, Final Models and Weighting Methods

A logistic procedure was used for propensity model selection, with stepwise selection of the variables included in the model. This method starts an intercept-only model, and uses a forward selection method to add one new variable at a time and a backward selection method to eliminate variables. The significance level for entering specified is 0.05 and the significance level for staying in the model is 0.05, which means that only variables that are significant at the 0.05 level will be included in the final model. However, because we also included second order interactions and specified that all effects contained in the interaction term must be present in the model, there are main effects in the model that are not significant at the 0.05 level.

Models were fitted separately for the telephone and mail samples. For the telephone sample, two response logistic models were fitted: one with all 17 variables as candidate variables; the other excluding the experimental (description of the topic) and paradata variables (contact and number of call attempts). The 17-variable model was intended for response propensity weighting. The 14-variable model was intended for GREG weighting because the excluded four variables do not have meaningful frame totals. The models for mail sample were fitted in the same way, dropping the two paradata variables. It is common to use paradata in response propensity weighting; however, because the models for response propensity weighting and GREG weighting are different in the current setting, it may not look like a fair comparison. Therefore, we also weighted the data using the 14-variable model for response propensity weighting, and refer to the 17-variable model (including the paradata and experimental variables) as RP0 weighting, and the 14-variable model as RP weighting.

As suggested by Little and Vartivarian (2005), the propensities need to predict the variables of interest to make the response propensity weighting effective. Therefore, it was desirable to include independent variables that were predictive to the variables of interest in the models. Because we have two variables of interest that were closely correlated, just for the purpose of comparison, we included them in the models to generate weights to estimate the counterpart variable. In other words, we used the weights from the model with the frame variable on whether the sample person voted in 2006 as a predictor to having voted in 2004, and vice versa.

Altogether, we selected six final models based on different sets of candidate variables for each survey mode, or, 12 models overall. We looked at the Hosmer and Lemeshow Goodness-of-Fit test as an indication of model fit. We cannot reject the models at the 0.05 level of significance except the one with all variables plus vote04 as candidate variables. A further investigation found that when number of call attempts entered the model, the validity of the model fit became questionable. A test for collinearity was conducted and number of call attempts showed much higher variance inflation (VIF) values than other variables. We excluded number of call attempts from the model, and the Hosmer and Lemeshow Goodness-of-Fit test suggested the model improved significantly.

The models are summarized in Table 2.2 and Table 2.3 for the telephone and mail samples, respectively. Cox and Snell's pseudo-R^2 is also shown for each model. As shown in Table 2.2, when we used all 17 variables as candidate variables, the final model included five main effects (contact, computer owner, home business, Do Not Call List, and number of persons in the household) and one interaction terms. Adding the variable whether the sample person voted in 2004 to the base of candidate variables resulted in a similar final model, with the main effect for this variable as an additional main effect and an interaction between this variable and home business. Similarly, adding the variable whether the sample person voted in 2006 to the base of candidate variables resulted in adding one main effect for whether the sample person voted in 2006.

When we excluded the experimental variable (description of the topic) and the two paradata variables (contact and number of call attempts), using the remaining 14 frame variables as candidate variables, the final model only included four main effects (computer owner, home business, on Do Not Call List, and age). Adding the variable whether the sample person voted in 2004 to the base of candidate variables resulted in a final model with five main effects and one interaction. The additional main effect is the variable whether the sample person voted in 2004 and the interaction is between this variable and age. Similarly, adding the variable whether the sample person voted in 2006 to the base of candidate variables resulted in a final model with additional main effect being the variable whether the sample person voted in 2006 and the interaction between this variable and age.

Table 2.3 summarizes the final models for the mail sample. The six models look very similar to each other. With all 14 frame variables and the experimental variable (description of the topic) as candidate variables, the final model for mail sample includes three main effects (computer owner, Do Not Call List, and age) and one interaction term (between age and computer ownership). When we excluded the experimental variable from the base of candidate variables, the final model is exactly the same because the experimental variable was not included through the selection process. Adding the variables voted in 2004 or 2006 to the base of candidate variables resulted in almost the same models with the variables voted in 2004 or 2006 as an additional main effect.

Several things are worth noting in comparing the models for the telephone and mail sample. First, computer owner, home business, and Do Not Call List appear in all six models for telephone sample, while for mail sample, they are computer owner, age, and Do Not Call List. Second, the two sets of models for the mail sample are identical, while there are variations among the six models for telephone sample, either in terms of number of terms or the specific variables included in the models. Third, the paradata are very predictive of response status. In the telephone sample, 21% of the variance in response is explained by the model from all 17 variables, but only 4% is explained when the paradata and the experimental variable are dropped.

For response propensity weighting, we used the predicted response propensities from the final models and assigned weights that were equal to the inverse propensities to respondents. For GREG weighting, we used the R® survey package to get the calibration weights for the respondents. Calibration weights were bounded within (1, infinity) so that each case at least represents itself. If this led to calibration failure, calibration weights were bounded within (0, infinity) to avoid negative values.

Table 2.2. Final response models selected from candidate variables with two-way interactions, telephone sample

All variables[†]	All + vote04	All + vote06	Frame variables[§]	Frame + vote04	Frame + vote06
contact	vote04	vote06	comp_own	vote04	vote06
ncall	contact	contact	home_biz	comp_own	comp_own
comp_own	comp_own	comp_own	suppress	home_biz	home_biz
home_biz	home_biz	home_biz	age	suppress	suppress
suppress	suppress	suppress		age	age
contact*suppress	contact*suppress	contact*suppress		vote04*age	vote06*age
	home_biz*vote04				
RSQ: 0.21	0.22	0.22	0.04	0.06	0.07

[†] 17 variables in total.

[§] Excluding three variables: topic, contact, and number of call attempts.

Table 2.3. Final response models selected from candidate variables with two-way interactions, mail sample

All variables[†]	All + vote04	All + vote06	Frame variables§	Frame + vote04	Frame + vote06
comp_own	vote04	vote06	comp_own	vote04	vote06
suppress	comp_own	comp_own	suppress	comp_own	comp_own
age	suppress	suppress	age	suppress	suppress
age*comp_own	age	age	age*comp_own	age	age
	age*comp_own	age*comp_own		age*comp_own	age*comp_own
RSQ: 0.04	0.05	0.06	0.04	0.05	0.06

[†] 15 variables in total.

[§] Excluding one variable: topic.

Table 2.4. Distribution of nonresponse adjusted weights and design effects due to weighting, telephone sample

	RP0 Weights			RP & GREG Weights					
Weight	**All variables[†]**	**all + vote04**	**all + vote06**	**frame variables[§]**		**frame + vote04**		**frame + vote06**	
				RP	**GREG**	**RP**	**GREG**	**RP**	**GREG**
0-1	–	–	–	–	–	–	0.9		0.9
1-5	68.3	69.0	68.3	92.8	98.9	85.0	99.1	85.3	95.7
5-10	1.6	0.6	3.7	7.3	1.1	15.0	–	14.7	3.4
10+	30.1	30.4	27.9	–	–	–	–	–	–
total	100.0	100.0	100.0	100	100.0	100	100.0	100	100.0
1+relvar	2.23	2.39	2.45	1.09	1.08	1.12	1.11	1.14	1.13

[†] 17 variables in total.

[§] Excluding four variables: incentive, topic, contact, and number of call attempts.

Table 2.5. Distribution of nonresponse adjusted weights and design effects due to weighting, mail sample

	RP Weights			GREG Weights		
Weight	**All variables[†]**	**all + vote04**	**all + vote06**	**frame variables[§]**	**frame + vote04**	**frame + vote06**
0-1	–	–	–	0.4	1.1	1.4
1-5	96.2	92.7	89.9	99.3	97.7	95.5
5-10	3.8	7.3	10.1	0.4	1.3	3.1
10+	–	–	–	–	–	–
total	100.0	100.0	100.0	100.0	100.0	100.0
1+relvar	1.07	1.09	1.12	1.07	1.09	1.11

[†] 15 variables in total.

[§] Excluding two variables: incentive and topic.

2.2.6 Comparison of Weights

The distributions of the sets of weights from the twelve final models are shown in Table 2.4 and Table 2.5 for telephone sample and mail sample, respectively. As we can see in Table 2.4, the RP0 weights (from the model with paradata included) are much more variable than the RP weights (under the model without paradata included), and under the same model, RP weights are a little more variable than GREG weights, but they are very close. Models with the variable whether the sample person voted in 2004 or the variable whether the sample person voted in 2006 as predictors did not change the weights that much. However, models with these two variables as predictors forced some of the calibration weights to take values less than 1 because there are sparse subsets of cases. The response propensity weights for mail sample also seem to be very similar to the calibration weights. Some of the calibration weights were forced to take the value less than 1 to make the calibration work.

The attempt to correct nonresponse bias by weighting does not come without a cost. Increasing the variance of the weights will increase the variance of the estimator. This effect is called “design effect due to weighting”, a concept introduced by Kish (1965). The design effect due to weighting is defined as

$$deff\_w = 1 + rel\,\mathrm{var}(w) = 1 + \mathrm{var}(w)/\bar{w}^2, \tag{2.9}$$

where $\bar{w}$ is the mean of the weights. The $deff\_w$ statistics associated with each set of weights are shown (labeled as “1+relvar”) in the bottom rows of tables 2.4 and 2.5. As we can see, for the telephone sample, variances would more than double if we used the response propensity weights under the model with paradata included to estimate the variables of interest as compared to the weights under the models without paradata included. The $deff\_w$ statistics associated with the RP0 weights are high, ranging from 2.23 to 2.45. This suggests we should create subclasses to reduce variation in the weights. We check the effect of this approach in next section. The weights for mail sample are shown in Table 2.5. As we can see, the six sets of weights do not have this problem. The $deff\_w$ statistic (labeled as "1+relvar") ranges from 1.07 to 1.12.

2.2.7 Bias Reductions through Weighting

For each of the two variables of interest (whether, according to the frame, the sample person voted in 2004 and 2006), we first computed the sample mean which was the target that we were interested in estimating. We then calculated the bias 1) when no weights were applied to the respondent cases, 2) when the RP0 weights were applied, 3) when the RP0 weights including *vote04*/*vote06* as predictors were applied, 4) when the RP weights were applied, 5) when the RP weights including *vote04*/*vote06* as predictors were applied were applied, 6) when GREG weights were applied, and 7) when GREG weights including *vote04*/*vote06* as predictors were applied. Tables 2.6 show the results for telephone sample and mail sample.

As Table 2.6 shows, when no weights are applied to the telephone sample, the proportion of voting is overestimated by 9 to 14 percentage points. The estimated percentage of voters in the 2004 election is 12.4 percentage points higher than the actual mean for the sample, and the estimated percentage of voters in the 2006 election is 14.2 percentage points higher than the sample mean. The relbiases are 26.0% and 32.8% for the 2004 and 2006 elections, respectively (note: the reductions shown in the analysis may be slightly different if calculated using the numbers in the tables, due to more precise values in the original numbers.). The RP0 weighting reduces the bias in the estimate for the 2004 election from 12.4 to 4.1 percentage points, the RP weighting reduces it to 8.0, and the GREG weighting reduces it to 8.4. However, including the informative variable *vote06* in the models changes the results. The RP0 weighting overcorrects; the positive bias with no weights becomes negative although the magnitude is smaller. The RP and GREG weighting basically removes the bias in the estimate for the 2004 election completely. Neither response propensity weighting nor GREG weighting effectively reduce the bias in the estimates for the 2006 election. The 14.2 percentage-point bias is reduced by 4.7, 3.4, and 3.1 percentage points by the RP0 weighting, RP weighting, and GREG weighting, respectively. Including the variable *vote04* (or *vote06*) in the models worked well for the three weighting methods; the bias is reduced by 69.1% with the RP0 weighting, by 73.7% with the RP weighting, and by 68.6% with GREG weighting. On average, the absolute bias in the unweighted estimates is 13.3 percentage points. RP0 weighting reduces the error to 6.8 percentage points, a 48.7% reduction. RP weighting reduces it to 9.4 percentage points, a 29.6% reduction. GREG weighting reduces it to 9.7 percentage points, a 26.9% reduction. Further reductions resulted when informative variables were included in the models.

Table 2.6. Bias in estimated percentage of voters in the 2004 and 2006 elections

Variable	Sample mean	Bias (unwtd)	RP0 weighting		RP weighting		GREG weighting	
			Bias1	Bias2†	Bias1	Bias2†	Bias1	Bias2†
Telephone Sample								
Vote04	47.8	12.4	4.1	-3.6	8.0	-0.5	8.4	-0.1
Vote06	43.2	14.2	9.5	4.4	10.8	3.7	11.1	4.5
Average (absolute)		13.3	6.8	4.0	9.4	2.1	9.7	2.3
Mail Sample								
Vote04	47.5	9.5	–	–	6.4	-2.2	6.5	-2.0
Vote06	43.9	12.8	–	–	9.8	3.9	10.0	4.4
Average (absolute)		11.1	–	–	8.1	3.1	8.3	3.2

† Bias with the other voting variable included as one of the predictors in the response propensity model.

As we can see in Table 2.6, the results for the mail sample are similar. The two key variables are overestimated when no weights were applied to the data for the mail respondents. The estimated percentage of voters in the 2004 election is 9.5 percentage points higher than the sample mean, a relbias of 20.0%; the estimated percentage of voters in the 2006 election is 12.8 percentage points higher than the sample mean, producing a relbias of 29.1%. Response propensity weighting reduces the bias in estimate of voters in the 2004 election to 6.4 percentage points, a 33.3% reduction; it reduces the bias in estimate of voters in the 2006 election to 9.8 percentage points, a 23.3% reduction. GREG weighting produces similar reductions in these errors. The bias after weighting are 6.5 (a 31.4% reduction), and 10.0 (a 21.4% reduction), respectively. On average, the absolute bias in the unweighted estimates is 11.1 percentage points. Response propensity weighting reduces it to 8.1 percentage points, a 27.6% reduction. GREG weighting reduces it to 8.3 percentage points, a 25.6% reduction. On average, including highly informative variables in the models reduces the bias to 3.1 for response propensity weighting, a 72.4% reduction; the reduction in the error is similar with GREG weighting.

There is a great deal of variation in the RP0 weights for the cases in the telephone sample; in practice, subclasses are usually created to reduce the variation. We created five weight classes (Little and Rubin 2002) based on these propensities and estimated the variables of interest using the new weights. For any case in a propensity weighting class, the new weight is the mean propensity for the class. Table 2.7 shows the comparison of biases using response propensity weights with classes and without. As we can see in the table, the average biases are similar, or even smaller with weight classes when no voting variable was included in the model. The design effects due to weighting are slightly reduced when we replace individual propensities with smoothed propensities based on propensity quintiles.

Table 2.7. Bias in estimated percentage of voters in the 2004 and 2006 elections using response propensity weighting, telephone sample

	Without subclasses		**With subclasses**	
Variable	**Bias1**	**Bias2**[†]	**Bias1**	**Bias2**[†]
Vote04	4.1	-3.6	1.6	-10.0
Vote06	9.5	4.4	8.2	0.6
Average (absolute)	6.8	4.0	4.9	5.3
1+relvar	2.23	2.39 (with vote04) 2.45 (with vote06)	2.06	2.20 (with vote04) 2.24 (with vote06)

[†] Bias with the other voting variable included as one of the predictors in the response propensity model.

2.2.8 Mean Squared Error Reductions through Weighting

Because there is a trade-off between bias reduction and variance inflation when it comes to weighting, we compared the mean squared error (MSE) which reflects both.

The mean squared error is defined as the expected squared difference between the population mean and the estimated mean, and can be decomposed as

$$MSE = Var\left(\hat{\bar{Y}}\right) + \left[Bias(\hat{\bar{Y}})\right]^2, \qquad (2.10)$$

where MSE is the sum of variance and squared bias in the estimated mean.

As we can see in Table 2.8, both response propensity weighting and GREG weighting produce large reductions in mean squared errors. For the telephone sample, RP0 weighting produces larger reductions in mean squared errors than the RP weighting and GREG weighting does when excluding *vote04*/*vote06* as predictors. The average reduction in mean square error is 60.4% under RP0 weighting, 47.2% under RP weighting, and 43.5% under GREG weighting. When including *vote04*/*vote06* as predictors, the RP weighting and GREG weighting produce larger reductions in mean squared errors than the RP0 weighting does. The average reduction in mean square error is 81.2% under RP0 weighting, 91.9% under RP weighting, and 90.4% under GREG weighting. Similarly, for the mail sample, including informative variables in the models results in larger reductions in mean squared errors. Under both RP weighting and GREG weighting, the reductions are double under the models including informative variables. This is true for both the mail and telephone sample.

Table 2.8. Mean square error of estimated proportion of voters in the 2004 and 2006 elections

		RP0 weighting		RP weighting		GREG weighting	
Variable	**Unweighted**	**Bias1**	**Bias2**[†]	**Bias1**	**Bias2**[†]	**Bias1**	**Bias2**[†]
Telephone Sample							
Vote04	0.016	0.004 (77.5%)[‡]	0.003 (81.1%)	0.007 (55.9%)	0.001 (94.9%)	0.008 (52.0%)	0.001 (95.1%)
Vote06	0.021	0.011 (47.1%)	0.004 (81.3%)	0.012 (40.6%)	0.002 (89.5%)	0.013 (37.0%)	0.003 (86.7%)
Average	0.019	0.007 (60.4%)	0.003 (81.2%)	0.010 (47.2%)	0.002 (91.9%)	0.010 (43.5%)	0.002 (90.4%)
Mail Sample							
Vote04	0.010	–	–	0.005 (52.4%)	0.001 (89.8%)	0.005 (50.1%)	0.001 (90.8%)
Vote06	0.017	–	–	0.010 (39.8%)	0.002 (87.8%)	0.011 (36.7%)	0.002 (85.3%)
Average	0.013	–	–	0.007 (44.4%)	0.002 (88.5%)	0.008 (41.6%)	0.002 (87.3%)

[†] Bias with the other voting variable included as one of the predictors in the response propensity model.

[‡] Numbers in parenthesis indicate reductions in mean square errors relative to unweighted estimates.

2.2.9 Summary of Study 1

Study 1 suggests that paradata (at least those examined in this study) are good predictors for response, but may or may not be good predictors of the variables of interest, in this case, voting in the general election. Including paradata in the response propensity model increases the variation in the resulting weights; therefore the gains in bias reduction may be offset by the loss in precision. "Informative" variables are powerful in bias reduction, but they are difficult to identify in practice.

The results shows that both response propensity weighting and GREG weighting can lead to bias reduction, but do not completely remove the bias. The reduction in bias is generally less than 50%. Therefore, we should consider the trade-off between bias and variance when creating weights for surveys. In this study, we see reductions in mean squared errors under both weighting schemes, which suggests that both weighting methods are effective in reducing overall error.

The two weighting methods have their advantages and disadvantages. GREG weighting guarantees that estimates of totals for the frame variables are unbiased, but response propensity weighting can employ other types of variables, such as the paradata used here. However, that may lead to more variable weights, which in turn leads to higher variances. If there are concerns about the magnitude of design effect due to weighting, weights based on propensity quantiles are effective in limiting variation in the weights, but can achieve similar bias reduction. GREG weighting may produce weights less than 1 or even negative weights, which are undesirable for many practitioners.

2.3 Study 2

This study assesses the response propensity weighting and GREG weighting methods using a dataset that has many variables of interest but only a few auxiliary variables, a situation more commonly encountered by survey researchers than that in Study 1. This study explores fewer models but follows the general findings from Study 1 to assess the two weighting methods.

2.3.1 Study Dataset

The University of Maryland alumni dataset. We will refer to this dataset as the alumni dataset. The data come from a survey conducted by the 2005 Practicum class at the Joint Program in Survey Methodology (JPSM) at the University of Maryland. The target population of the survey was University of Maryland alumni who received undergraduate degrees at Maryland from 1989 to 2002. A random sample of 20,000 graduates was drawn from the 55,320 graduates listed in the Registrar's records. Telephone numbers were matched to only 10,325 of the 20,000 sampled. After excluding those who were listed as residents of Puerto Rico, the Virgin Islands, or on military bases, those having the same number as another graduate, and those used in the pretest sample, 7,591[1] of them were selected and received at least one call for the main study. The sample alumni were contacted by telephone for a brief screener survey and the screener completes were randomly assigned to one of the three modes of data collection (telephone, Web, and IVR). A total of 1,501 cases completed the telephone screener (AAPOR Response Rate 1: 31.9%) and were randomly assigned to one of the three modes of data collection (telephone, Web, and IVR). Cases assigned to the telephone mode continued with the main survey immediately after the screener; those assigned to the Web option were told to follow the instructions in a letter they were sent; and those assigned to IVR were switched to IVR after they completed the screener. The response rates for this final stage were 94.7%, 56.8%, and 61.1% for telephone, Web, and IVR, respectively (see Kreuter, Presser, and Tourangeau 2008 for more information about the study). Although nonresponse bias can be analyzed both at the screener stage and the main survey stage, as shown in Sakshaug, Yan, and Tourangeau (2010), the nonresponse bias at the main survey stage is much smaller than the bias at the screener stage. Therefore, to make the analysis simple and clear, this study only focuses on the screener stage. As with Study 1, this study uses frame values in all analyses to avoid potential measurement errors. In this study, the frame variables mainly came from the Registrar's records.

[1] The original study (Kreuter, Presser, and Tourangeau 2008) reported that 7,591 cases were fielded, but a re-analysis of the data (Sakshaug, Yan, and Tourangeau 2010) reported the results on 7,535 cases based on some exclusion criteria. Here we used the 7,591 cases reported in the original study.

2.3.2 Variables

Variables of interest. Table 2.9 lists the survey variables of interest and the auxiliary variables in the dataset. These variables could be checked against records available from the Registrar's Office or the Alumni Association. The variables of interest are grouped into two categories: undesirable characteristics and desirable characteristics. Undesirable items include GPA less than 2.5, dropping a class, getting an unsatisfactory grade, and receiving an academic warning or being placed on probation; desirable items include GPA higher than 3.5, receiving academic honors, being a member of the Alumni Association, and donating money to the University of Maryland after graduation or in last year (2004).

Auxiliary variables. Five frame variables were used in the analysis. State of residence (Maryland versus other), sex, and time of getting the degree (summer versus winter) are dichotomous variables. Age and year of getting degree are treated as continuous variables.

Table 2.9: Variables from the alumni dataset

Variable	**Short Name**
Undesirable characteristics	
GPA <2.5 vs. other	GPA below 2.5
At least one D or F vs. other	F or D
Ever dropped a class vs. other	Withdraw
Getting warning or on probation vs. other	Probation
Desirable characteristics	
GPA > 3.5 vs. other	GPA above 3.5
Getting honors vs. other	Honors
Ever donated to the University vs. other	Donated
Donated to the University in last year vs. other	Donated in last year
Members of Alumni Association to the University	Member
Auxiliary variable	
Marylanders vs. other	State
Sex (male vs. female)	Sex
Age in years	Age
Year of getting degree	Degyear
Time of getting degree (summer vs. winter)	Degmon

2.3.3 Imputation for Missing Values

The missing rates for the five auxiliary variables are all less than 0.5%. As with Study 1, frame variables with missing values were imputed using the multivariate sequential regression imputation method. Only one imputed dataset was created.

2.3.4 Correlation between Response/Variables of Interest and Auxiliary Variables

Figure 2.3 shows the correlations between response status and auxiliary variables, and the correlations between survey variables of interest and auxiliary variables. The correlations between response status/variables of interest and the auxiliary variables do not show clear patterns. The correlations between variables of interest and the auxiliary variables have both negative and positive values, and the magnitude also varies. Therefore, it is likely difficult to find a set of weights that work for all variables of interest.

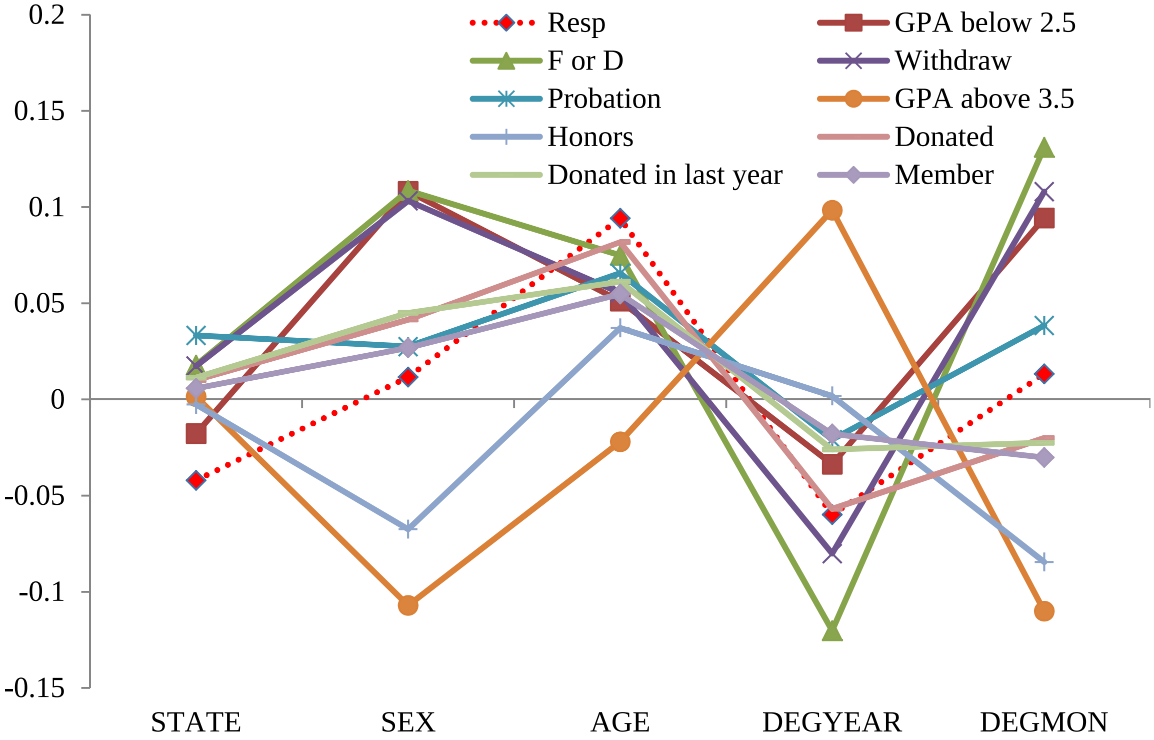

Figure 2.3. Correlations between auxiliary variables and whether the case is a respondent, and between auxiliary variables and variables of interest

2.3.5 Model Selection, Final Models and Weighting Methods

As with Study 1, logistic procedure was used for the model selection, with the stepwise automatic selection method. Again, the significance level for entering was specified as 0.05 and the significance level for staying was 0.05. Two-way interactions were candidates for inclusion and when an interaction was retained, all effects contained in the interaction term were included as well. The selected model was

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.45 - 0.21\mathit{state} + 0.03\mathit{Age},$$

with Cox and Snell's pseudo-R^2 equal to 0.010. The same model was selected when specifying the significance level for entering at 0.20 and the significance level for staying is 0.15, two commonly used values based on the suggestion by Hosmer and Lemeshow (2000: 118–119).

A likelihood ratio test was performed to assess the loss from the reduced model. This test was calculated as

$$LR = -2\log l(reduced) - 2\log l(full) = 7473.6\text{-}7451.0\text{=}22.6 \sim \chi^2_{15-2}\ ,$$

which is significant at 0.05 level. The test suggests the reduced model has lost some important contributing variables, compared to the full model with all five auxiliary variables and the associated two-way interactions. Also, the Hosmer and Lemeshow Goodness-of-Fit Test ($\chi^2_8 = 16.84,\ p < 0.05$) suggested the model did not fit the data well.

Because there are only five auxiliary variables available, and they are all important characteristics, it makes sense to at least include all five variables as main effects in the model. Redoing the model selection with this restriction, the final model is

$$\log\left(\frac{p_i}{1-p_i}\right) = -6.398 - 0.219state + 1.066sex + 0.048age + 0.002degyear + 0.034degmon - 0.030sex * age\,,$$

with Cox and Snell's pseudo-R^2 equal to 0.012. Likelihood ratio test

$$LR = 7461.1\text{-}7451.0\text{=}10.1 \sim \chi^2_{15-6}\ ,$$

is not significant. And Also, the Hosmer and Lemeshow Goodness-of-Fit Test ($\chi^2_8 = 15.20,\ n.s.$) suggested we cannot rejected the model at the 0.05 level of significance.

The same final model was used for response propensity weighting and GREG weighting. As in Study 1, for response propensity weighting, we used the predicted response propensities from the final models and assigned the inverse of the fitted propensities to respondents as weights. For GREG weighting, we used the R® survey package to get the calibration weights for the respondents. Calibration weights were bounded within (1, infinity). If this led to calibration failure, calibration weights were bounded within (0, infinity) to avoid negative values.

2.3.6 Comparison of Weights

The distributions of the two sets of weights are shown in Table 2.10. As we can see in the table, the two sets of weights are similar, although some of the calibration weights were forced to take values less than 1 to avoid calibration failure. The design effects due to weighting are both 1.04, which is small.

Table 2.10. Distribution of nonresponse adjusted weights and design effects due to weighting

Weight	RP weighting	GREG weighting
0-1	–	0.4
1-5	42.7	45.9
5-10	57.3	53.7
10+	–	–
Total	100.0	100.0
1+relvar	1.04	1.04

2.3.7 Error Reductions through Weighting

As we can see in Table 2.11, the sample proportions for the different survey variables show a great deal of variation, ranging from 2.6% to 70.8%. When estimating the sample proportions or means from the respondent cases without weighting, there are negative biases for the undesirable characteristics, except academic probation, a characteristic associated with only 2.6% of the sample. The biases in the estimates for these undesirable characteristics are small—all are generally less than 2 percentage points. The two weighting methods do not help in reducing the biases. Instead, they tend to make the biases a little worse for the undesirable characteristics. On average, there is a 1.2 percentage-point negative bias for the undesirable characteristics for the unweighted estimates. Both the response propensity weighting and GREG weighting resulted in 1.4

percentage-point negative bias, an increase of 16.7% in bias.

The unweighted estimates have positive biases for the desirable characteristics. The degree of overestimation ranges from 1.9 to 12.8 percentage points. The variable *donate* shows the biggest bias in terms of percentage points. The unweighted means overestimate the desirable characteristics by 12.5% to 106.5%, much more problematic than in the undesirable characteristics. The two weighting methods do not help much in bias reduction, although small decreases are seen in the estimates for every desirable variable. On average, the overestimation in the desirable variables is 4.0 percentage points; response propensity weighting reduces it by 0.1 percentage points, and GREG weighting reduces it by 0.2 percentage points.

With respect to mean squared error, because the design effects due to weighting are small for both weighting methods, the bias term plays a significant role in the mean squared error statistic. For undesirable characteristics, the magnitude of biases is small and the mean squared errors are relatively small compared to those for desirable characteristics. Overall, the weighting methods do not change mean squared errors so much. The weights slightly increase the mean squared errors for the estimated proportions with the undesirable characteristics, and decrease those for the estimated proportions with the desirable ones somewhat.

Table 2.11. Bias and mean squared error in estimated percentage of variables of interest

		Bias in percentage			Mean squared error		
Variable	Sample mean	Unweighted	RP	GREG	Unweighted	RP	GREG
Undesirable characteristics							
GPA below 2.5	15.4	-1.7	-2.1	-2.0	3.6	5.1	4.8
F or D	62.6	-1.7	-1.8	-1.7	4.4	5.1	4.6
Withdraw	70.8	-1.6	-1.8	-1.7	3.9	4.6	4.5
Probation	2.6	0.0	0.0	0.0	0.2	0.2	0.2
Average (absolute)		1.2	1.4	1.4	3.0	3.7	3.5
Desirable characteristics							
GPA above 3.5	18.6	2.3	2.0	1.9	6.5	5.3	4.9
Honors	9.4	1.9	1.5	1.4	4.2	2.8	2.7
Donated	25.1	12.8	12.1	12.1	166.4	148.5	148.6
Donated in last year	8.5	6.6	6.2	6.2	44.2	39.1	39.2
Member	7.0	7.5	7.3	7.3	56.9	53.9	54.2
Average (absolute)		6.2	5.8	5.8	55.6	49.9	49.9
Overall average (absolute)		4.0	3.9	3.8	32.2	29.4	29.3

2.3.8 Summary of Study 2

Study 2 suggests that not much can be done when we only have a limited set of auxiliary variables. Demographic variables are not good predictors of response status, nor are they good predictors for the variables of interest in this study. Both response propensity weighting and GREG weighting based on these variables are not effective in bias or mean squared error reduction.

2.4 General Discussion

Weighting as a method of nonresponse adjustment is appealing and commonly used in practice. Various weighting methods have been developed, providing a pool of choices to practitioners. Many of these weighting methods make an assumption of no “interactions” among the auxiliary variables. The two methods examined in this study do not make this assumption. Even so, weighting is only useful when we have variables that correlate with the outcome variables. This point is emphasized in a simulation study by Brick and Jones (2008). Brick and Jones showed that the choice of auxiliary variables to be included is far more important than the choice of the calibration method (e.g., raking or linear calibration). The comparison made in this research further showed that with the same set of auxiliary variables, the choice between response propensity weighting and GREG weighting did not make much difference.

We can also expect that nonresponse adjustment is effective when response status is a function of the outcome variables. However, Study 1 shows that some variables may be predictive of response status, but they are not necessarily good predictors for the outcome variables. For example, if the incentive variable was included in the adjustments, that would lead to huge variance inflation but no help with bias reduction, because the incentive variable is very predictive of response status, but not the voting variables. The same observation is made by Lepkowski, Kalton, and Kasprzyk (1989). It is important to collect a rich set of auxiliary variables and equally important is having the auxiliary variables correlate both with the response probability and the outcome variables.

The GREG weighting method in this study uses frame values as control totals. In practice, outside benchmarks are often used as control totals. We should notice that when the auxiliary variables are measured differently from the benchmark, bias may be introduced (Skinner 1999). Therefore, when searching for auxiliary variables, it is equally important to check to make sure that they are measured in the same way as the survey variables.

Both weighting methods inflated the variance in the estimates. The design effects due to weighting can be huge, and we should apply variance-reducing techniques to the weights, such as creating subclasses for the weights.

There are some limitations in this analysis. The target populations for both datasets examined here are not the general U.S. population, but only a fraction of it. Therefore, the results may not apply to the general population. The two weighting methods used explicit models. As with any models, these models have their limitations. The key survey variables examined here were potentially sensitive characteristics that were subject to social desirability effects. Therefore, the results should not be generalized to the non-sensitive questions without further investigation.

Chapter 3: Nonresponse Bias and Sample Quality Indicators

3.1 Introduction

Despite widespread concerns about declining response rates, it is often difficult to estimate the impact of nonresponse on survey estimates. There may be bias associated with the responding units, but one cannot know this for sure. In recent years, researchers have turned to auxiliary information in the hope of reducing potential bias. There are four major ways of using auxiliary information in surveys: 1) to draw an initial sample that is balanced on some important auxiliary variables; 2) to guide data collection efforts to achieve a balanced set of respondents; 3) to construct balance indicators for the achieved sample of respondents; and 4) to make postsurvey adjustments. This study focuses on the third use—investigating the measures of balance or representativeness in achieved samples.

Survey researchers have been using auxiliary information to draw stratified samples since the early days of surveys (e.g., Neyman 1934). Responsive survey designs (Groves and Heeringa 2006) aim to tailor the data collection strategy to make the most effective use of resources. The actions taken during the data collection process often are guided by the auxiliary information. For instance, during data collection, the response rate among older sample members may be low, and the data collection strategies may be adjusted to respond to this situation. Comparing the response rates across subgroups is a common tool for identifying lack of balance in the sample of respondents, although Peytcheva and Groves (2009) show in a meta-analysis of twenty three studies that the biases in the estimates of demographic variables do not seem to be related to the biases in the estimates of substantive variables.

Schouten, Cobben, and Bethlehem (2009) argue that more effective and easy-to-use indicators are needed. They propose the “R-indicator” (R stands for representativeness) as a tool for monitoring the effects of nonresponse. The R-indicator aims to measure the similarity between the sample selected initially and the responding units. It is based on the variance of the estimated response probabilities. The response probabilities are themselves defined as the conditional expectation of responding given a vector of auxiliary variables:

$$\hat{p}(z_i) = E(R_i = 1 \mid Z = z_i) = P(R_i = 1 \mid Z = z_i), \tag{3.1}$$

where the z’s are auxiliary variables with values known for all sample units. In practice, the response propensities are usually estimated using a logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta, \tag{3.2}$$

where x_i^T is a vector of auxiliary variables, and β is a vector of logistic regression coefficients. The response propensity is predicted as

$$\hat{p}_i = \frac{\exp\left(x_i^T \hat{\beta}\right)}{1+\exp\left(x_i^T \hat{\beta}\right)}. \tag{3.3}$$

Based on the estimated response probabilities, Schouten, Cobben, and Bethlehem (2009) define the R-indicator as:

$$R(\hat{p}) = 1 - 2S(\hat{p}) = 1 - 2\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{p}_i - \hat{\bar{p}})^2}\ . \tag{3.4}$$

$R(\hat{p})$ is bounded in the interval [0,1]. Särndal (2011) proposed three balance indicators (BI_1-BI_3) which measure differences of the response means and sample means of auxiliary variables. As Särndal points out, BI_1 is similar and “sometimes identical” to the R-indicator (Särndal 2011, p12). A related idea has been known to survey researchers for a long time. Kish (1965) introduces a concept called “design effect due to weighting” to measure the variation in the resulting weights. The design effect due to weighting is defined as

$$deff\_w = 1 + rel\,\mathrm{var}(w) = 1 + \mathrm{var}(w) / \bar{w}^2, \tag{3.5}$$

where $\bar{w}$ is the mean of the weights. Suppose that response propensity weights were generated from the estimated response propensities, the design effect due to weighting and the R-indicator are closely related in a sense that both of them measure the variation in the estimated response propensities in this case, but the R-indicator is a standardized measure which takes a value from 0 to 1.

Shlomo, Skinner, and Schouten (2012) argue that because there is sampling variation in the estimated response propensity, the estimate of the R-indicator is biased. They propose an approximation of the bias; the bias-adjusted R-indicator is given as:

$$R_B(\hat{p}) = 1 - 2\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{p}_i - \hat{\bar{p}})^2 - \frac{1}{n}\sum_{i=1}^{n} z_i^T (\sum_{j=1}^{n} z_j x_j^T)^{-1} z_i} \tag{3.6}$$

where $z_i = \nabla h\left(x_i^T \hat{\beta}\right) x_i$ and ∇h is the vector of first order derivatives with respect to β in Equation (3.2). Schouten, Shlomo, and Skinner (2011) also develop partial R-indicators, which take the value of square root of the between variance of the response propensity with respect to a particular auxiliary variable, to guide data collection decisions in adaptive and responsive survey designs. For example, for a categorical variable T with K categories,

$$Partial\ R(T,\hat{p}) = \sqrt{\frac{1}{n-1}\sum\nolimits_{k=1}^{K} n_K\,(\bar{\hat{p}}_k - \bar{\hat{p}})^2}\ . \tag{3.7}$$

There is a known problem with the R-indicators. All of the R-indicators are affected by both the level of the response rate and by the variation in response propensities (Groves et al. 2008). It is not uncommon that the values of the R-indicator take on a U-shape as the response rate goes up. This feature makes the R-indicator difficult to use and interpret. Here we propose a penalized R-indicator that reflects both the lack of balance in the response set and the overall level of response in one indicator. Our penalized R-indicator is defined as:

$$PR(\hat{p}) = e^{(-1+\sqrt{n_r/n_s})}\left(1-2S(\hat{p})\right) = e^{(-1+\sqrt{n_r/n_s})}\left(1-2\sqrt{\frac{1}{n-1}\sum\nolimits_{i=1}^{n}(\hat{p}_i - \bar{\hat{p}})^2}\right), \tag{3.8}$$

where n_r is the number of response units, and n_s is the number of sampled units. We call the multiplier $e^{(-1+\sqrt{n_r/n_s})}$ the penalizing factor. Bias correction can also be conducted in the same way as it is done for the original R-indicator given in Equation 3.3. The penalizing factor penalizes for low response rates, but keeps $PR(\hat{p})$ bounded in the interval [0,1]. Another way to see the penalizing factor is that it adjusts the representativeness "score" we give to the sample by the level of response we get. Figure 3.1 shows the graphs for the function $\sqrt{n_r/n_s}$ and function $e^{(-1+\sqrt{n_r/n_s})}$. Both functions are convex, which are desirable because like the variation of proportions, the deviation of response propensities is likely to be bigger when the mean response propensities are at the middle values than when they are at the two ends (0 and 1). Therefore, the R-indicator is likely to follow a concave pattern as the level of mean response propensities increases because it subtracts twice the deviation of response propensities from 1. The function

$e^{(-1+\sqrt{n_r/n_s})}$ has a smoother form and does not discount the representativeness completely as the value of n_r/n_s goes to 0.

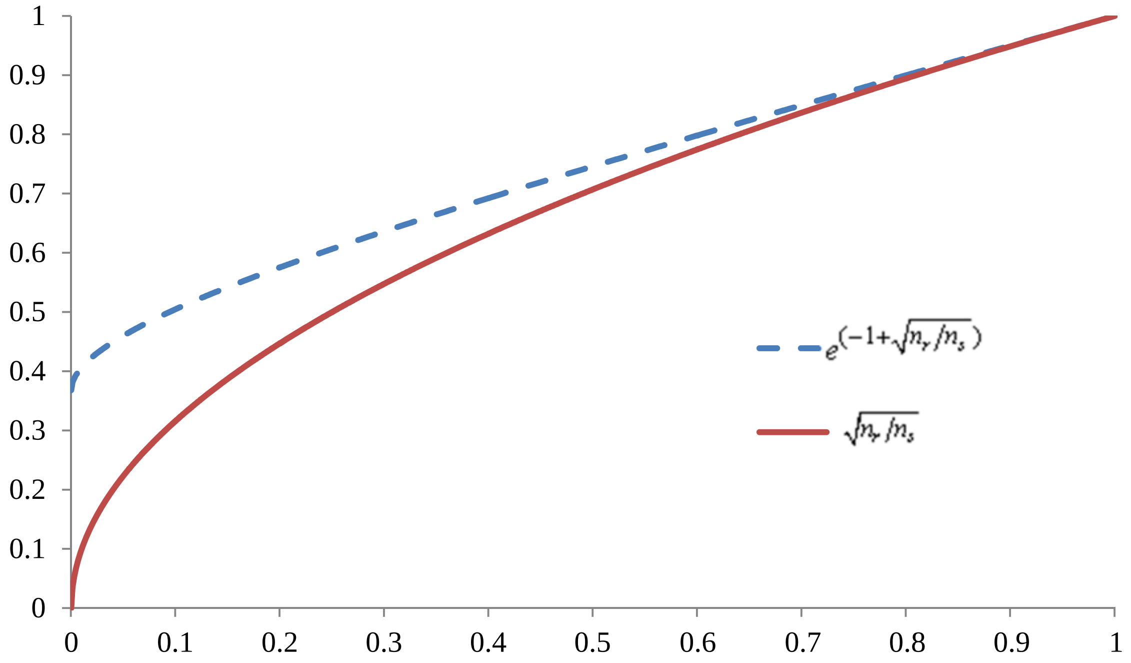

Figure 3.1. Values of the two functions at different response level

Although the R-indicator has gained some attention from survey researchers, few studies examining the R-indicator have been published except by the authors who proposed it (Schouten, Cobben, and Bethlehem 2009; Schouten et al. 2012; Schouten, Shlomo, and Skinner 2011). In addition, these studies do not have records available for the key survey items. Therefore, we do not know what a higher value of the R-indicator means for estimates involving these items. This study tries to address this gap by examining how well the R-indicator and the penalized R-indicator actually predict bias in sample estimates. It assesses the effectiveness of the R-indicators in predicting biases using two datasets with records available. It also conducts a simulation to compare the performance of the R-indicator and the penalized R-indicator.

3.2 Study 1

3.2.1 Study Dataset and Variables

The voters data. A description of the data set can be found in Section 1.7.1 of Chapter One. A list of 50,000 Maryland residents who were registered to vote was purchased from Aristotle (http://www.aristotle.com) for the study. The original study included a mode experiment. A total of 1,669 cases were assigned to receive a mail survey; 1,020, a telephone survey. The response rate for the telephone survey was 34.3 percent (AAPOR RR1); for the mail survey, the response rate was 33.2 percent. This study uses the telephone survey only because we are interested in how the indicators perform as the number of calls increases. The number of call attempts each of the cases received ranges from 1 to 9.

Variables of interest and auxiliary variables. The Aristotle database contained information on voting history in the 2004 and 2006 general elections. It also had a variety of auxiliary variables. This study uses only the frame values in all analyses. Fourteen auxiliary variables were identified. Nine of them were dichotomous variables (whether the person ever donated to various organizations, whether he or she was computer owner, whether he or she had a home business, party identification, whether the person was on the Federal Do Not Call list, whether the person was a head of household, whether the person reported on religion, sex, and whether the person reported on ethnicity), and the other five variables were continuous variables (number of persons in the household, age, and income) or treated as continuous variables (education level and home ownership level—renter, probable renter, probable homeowner, and homeowner). Auxiliary variables with missing values were imputed using the multivariate sequential regression imputation method (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001). Only one imputed dataset was created.

3.2.2 Response Propensity Model Selection

A logistic procedure was used for propensity model selection, with stepwise selection of the variables included in the model. This method starts with an intercept-only model, and uses forward selection method to add one new variable at a time and backward selection method to eliminate variables. The significance level for entering is 0.05 and the significance level for staying in the model is 0.05, which means only variables that are significant at 0.05 level will be included in the final model. However, because we also included first order interactions and specified that all effects contained in the interaction term must be present in the model, there are main effects in the model that are not significant at 0.05 level.

The response logistic models were fitted with the 14 auxiliary variables. The final model (shown in Table 3.1) included four main effects (computer owner, home business, Do Not Call List, and age). This model is the same as model 4 (labeled as “frame variables”) in Table 2.2 of Chapter Two.

Table 3.1. Logistic regression coefficients from the response propensity model

	Coefficient	**SE**
Intercept	-1.98	0.29
Computer owner	0.38	0.38
Home business	0.64	0.29
On Do Not Call List	0.63	0.16
Age	-0.01	0.00

Note: Model based on the 1,020 sample members of the telephone sample.

3.2.3 Response Rates in Subgroups

We first examined the variation in response rates for subgroups based on the selected independent variables. As Figure 3.2 shows, there was variation in the response rate across the subgroups. Computer owners were more likely to respond to the telephone survey than non-owners (41.2% vs. 31.3%, p<0.01). Home-business households had a higher response rate than the other households (51.0% vs. 33.4%, p<0.05). Oddly enough, sample members who were on the federal Do Not Call list were more likely to be respondents than those who were not (38.2% vs. 23.9%, p<0.001). Sample members older than 65 years old responded in a higher rate than those between 45 and 64 years old, who in turn responded at a higher rate than those between 18 and 44 years old (39.3% vs. 36.0% vs. 24.9%, p<0.05).

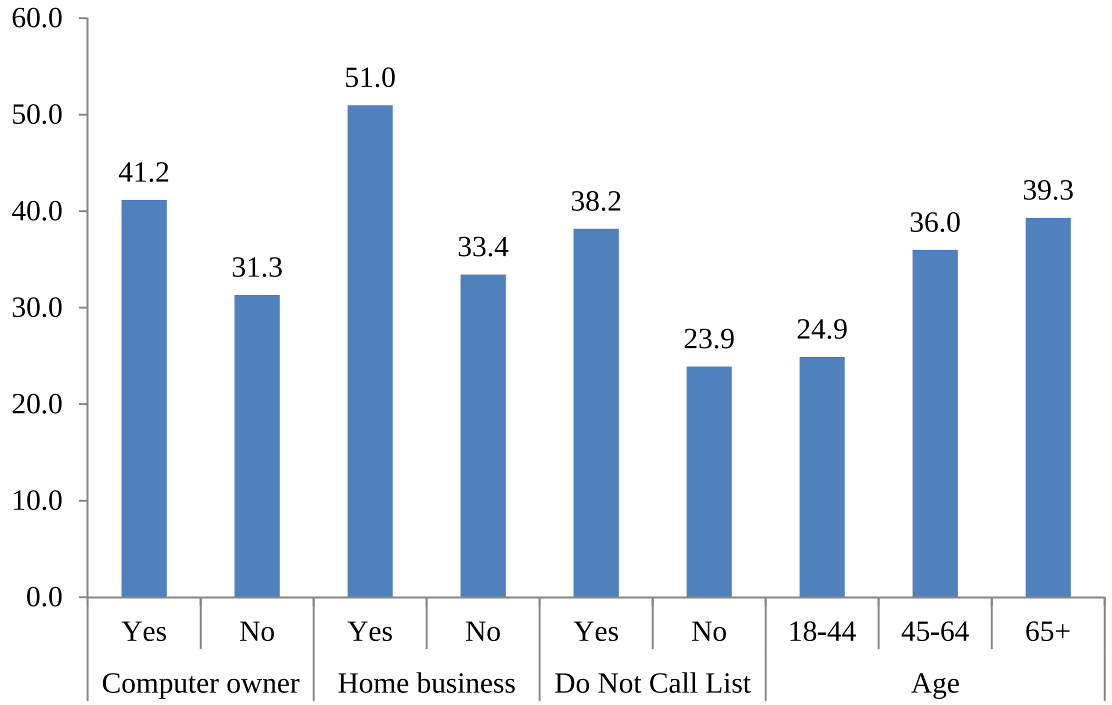

Figure 3.2. Response rates in percent for subgroups. Differences are significant for all four variables at the 0.05 level

3.2.4 Differences across Response Propensity Quintiles

We then examined the mean number of calls and the proportions voting in 2004 and 2006, grouping cases by estimated response propensity quintiles. The propensities were from the model in Table 3.1. In Figure 3.3, the predicted response propensities are increasing from left to right within each variable. The average numbers of call attempts are indicated by a diamond, and the confidence intervals for the mean are shown as a vertical line. As Figure 3.3 shows, the average number of call attempts is greater in the groups with lower response propensities. The average number of call attempts in the quintile with the highest response propensities is significantly lower than the average numbers in the two quintiles with the lowest propensities. And, generally speaking, the proportions voting in 2004 and 2006 are higher in the groups with higher response propensities, which means that those who were more likely to respond were more likely to vote.

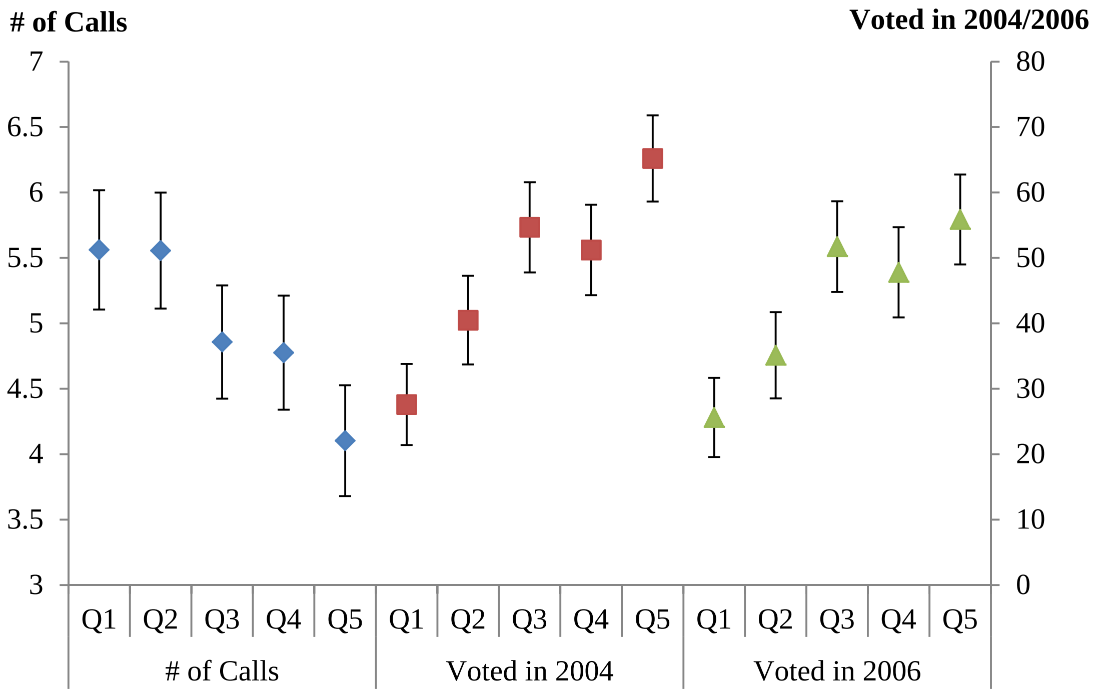

Figure 3.3. Average number of call attempts and proportion voting in 2004 and 2006, by response propensity quintile

3.2.5 Performance of the R-indicator and Penalized R-indicator

Figure 3.4 shows the values for the R-indicator and penalized R-indicator by the number of call attempts. It shows what would have happened in the survey had the call attempts been capped at a specific number. The relative biases in the estimated proportions who voted in 2004 and 2006 were also shown in the figure. As we can see in Figure 3.4, as call attempts increased (and more respondents were brought into the respondent pool), relative bias in the two estimates decreased. However, the R-indicator shows the opposite pattern; it decreased as the relative bias got smaller. In contrast, the penalized R-indicator increased as the relative bias went down. The correction for bias does not make much difference in the comparison.

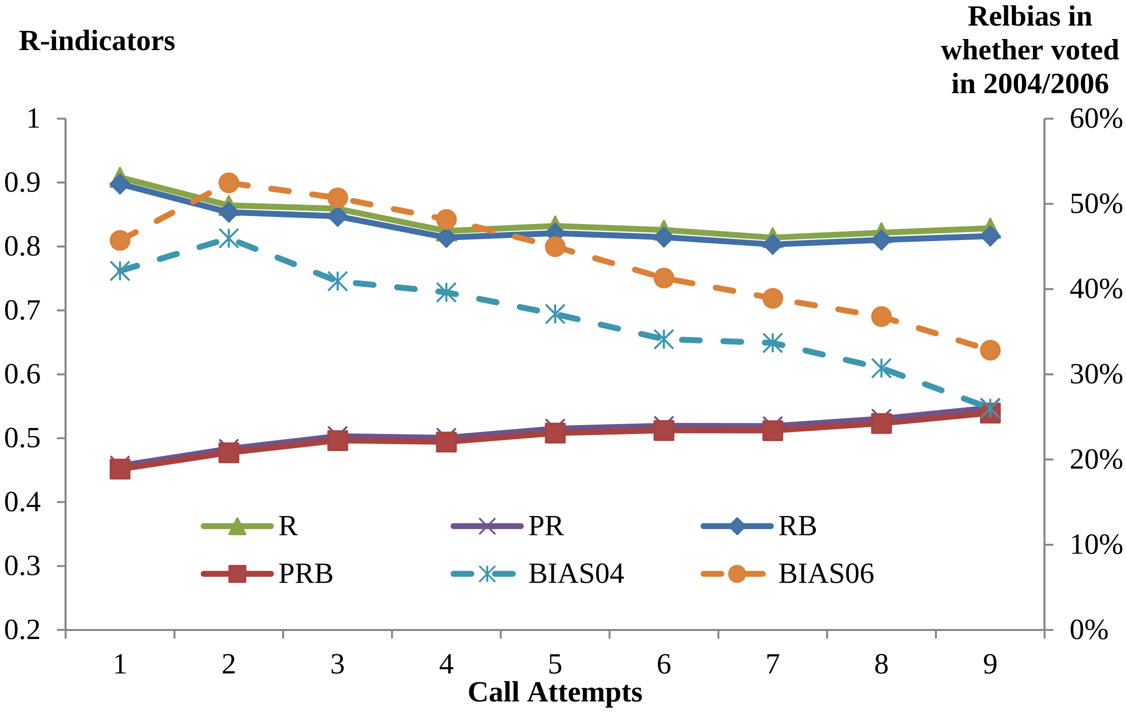

Figure 3.4. Relbiases in the estimated proportions voting in 2004 (BIAS04) and 2006 (BIAS06), the value of the R-indicator (R), the bias-adjusted R-indicator (RB), the penalized R-indicator (PR), and the penalized bias-adjusted R-indicator (PRB) at each level of call attempts

3.3 Study 2

3.3.1 Study Dataset and Variables

The alumni data. A description of these data can be found in Section 1.7.2 of Chapter One. A total of 7,591[2] sample members were selected and received at least one call for the main study. The sample alumni were contacted by telephone for a brief screener survey and the screener completes were randomly assigned to one of the three

[2] The original study (Kreuter, Presser, and Tourangeau 2008) reported that 7,591 cases were fielded, but a re-analysis of the data (Sakshaug, Yan, and Tourangeau 2010) reported the results on 7,535 cases based on some exclusion criteria. Here we used the 7,591 cases reported in the original study.

modes of data collection (telephone, Web, and IVR). A total of 1,501 cases completed the telephone screener (AAPOR Response Rate 1: 31.9%) and were randomly assigned to one of the three modes of data collection (telephone, Web, and IVR). We focused our analysis on the screener stage in this study. The number of call attempts each of the cases received ranges from 1 to 31. Because there were much fewer cases received more than 14 call attempts, we merged these cases with the cases receiving 14 call attempts.

Variables of interest and auxiliary variables. The alumni dataset contained variables on the graduates' academic performance and on their relationship with the University. These variables could be checked against records available from the Registrar's Office or the Alumni Association. These variables included the sampled member's GPA, whether he or she dropped a class, received an unsatisfactory grade, received an academic warning or was being placed on probation, received academic honors, was a member of the Alumni Association, and donated money to the University of Maryland after graduation or in last year (2004). Some additional auxiliary variables were also available, including state of residence, age, sex, and time of getting the degree. As with Study 1, frame variables with missing values were imputed using the multivariate sequential regression imputation method. Only one imputed dataset was created.

3.3.2 Response Propensity Model Selection

A logistic procedure was used for propensity model selection, with stepwise selection of the variables included in the model. This method starts with an intercept-only model, and uses a forward selection method to add one new variable at a time and a backward selection method to eliminate variables. Again, the significance level for entering was specified as 0.05 and the significance level for staying was 0.05. Two-way interactions were candidates for inclusion and when an interaction was retained, all effects contained in the interaction term were included as well. Because there are only five auxiliary variables available, and they are all important characteristics, it makes sense to at least include all five variables as main effects in the model. Redoing the model selection with this restriction, the final model is shown in Table 3.2. This model is the same as the final model for Study 2 in Chapter Two.

Table 3.2. Logistic regression coefficients from the response propensity model

	Coefficient	SE
Intercept	-6.398	18.880
State	-0.219	0.060
Sex	1.066	0.315
Age	0.034	0.007
Year of getting degree	0.048	0.009
Month of getting degree (summer vs. winter)	0.002	0.059
Age x Sex	-0.030	0.009

Note: Model based on the 7,591 fielded cases.

3.3.3 Response Rates in Subgroups

As before, we examined the variation in response rates for across subgroups based on the auxiliary variables. As Figure 3.5 shows, there was some variation across these subgroups. Alumni residing in states other than Maryland were less likely to respond to the screener than those living in Maryland (18.5% vs. 22.0%, p<0.001). Older alumni tend to be screener respondents more than younger alumni (22.9% vs. 16.9%, p<0.001). Alumni who graduated earlier were more likely to respond to the survey. The graduation year group that had highest response rate was 1990 and the group with lowest response rate was 2001.

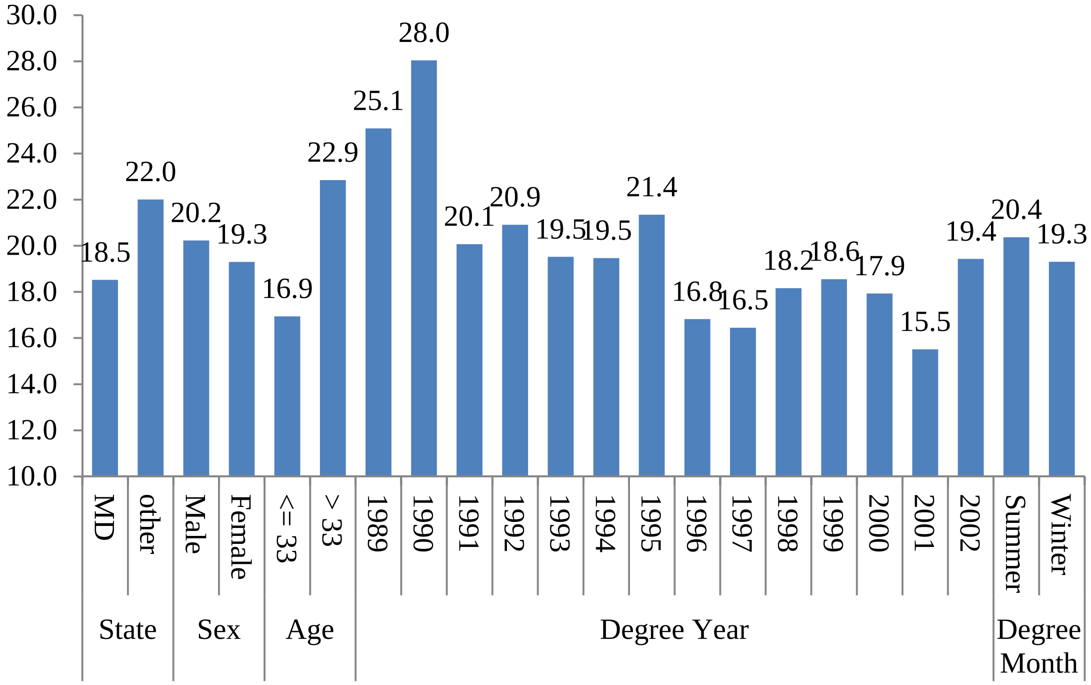

Figure 3.5. Response rates in percent by subgroup. Differences across the groups are significant for state, sex, age (treated as a continuous variable in the logistic model), and degree year (also a continuous variable in the model)

3.3.4 Differences across Response Propensity Quintiles

We then examined the mean number of calls and the proportion with a membership in the Alumni Association, grouping cases by predicted response propensity quintiles. We chose to show the proportion of cases who were Alumni Association members here because among the nine desirable and undesirable characters, the relative nonresponse bias for this estimate was the largest. The response propensities are increasing from left to right for within each variable in Figure 3.6. The mean number of call attempts is indicated by a diamond (and percent of membership is indicated by a square), and the confidence intervals for the mean are shown as a vertical line. Figure 3.6 shows that the average number of call attempts is not significantly different across the response propensity quintiles, while the proportion of Alumni Association members in the highest propensity quintile is quite different from those in the lowest three quintiles. The difference between the highest response propensity quintile and the third quintile in the proportions of cases who are Alumni Association members is significant.

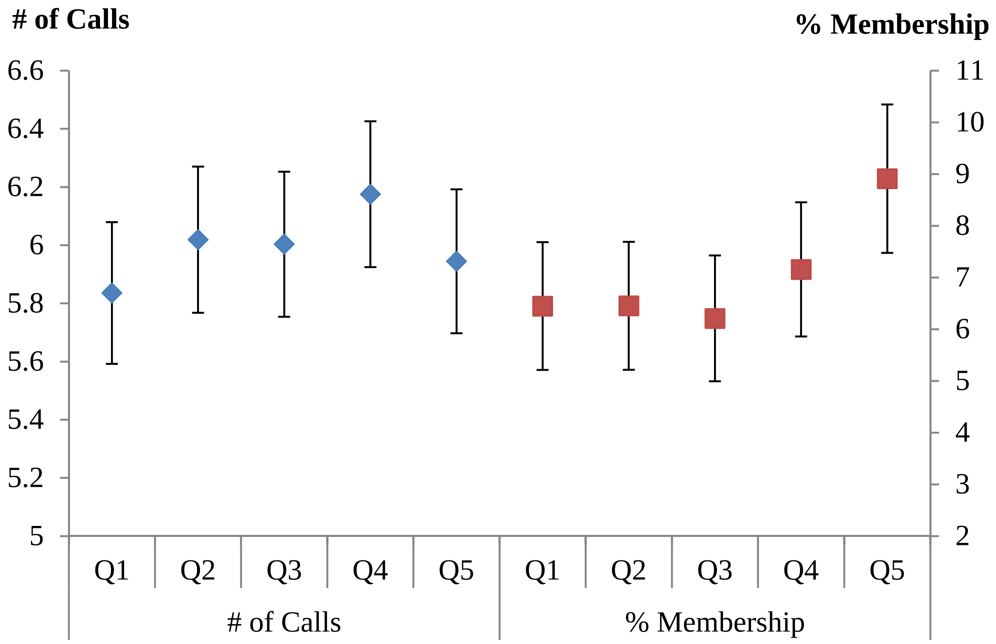

Figure 3.6. Average number of call attempts and proportion with alumni membership by response propensity quintile

3.3.5 Performance of the R-indicator and Penalized R-indicator

Figure 3.7 shows the values for the R-indicators and penalized R-indicators by the number of call attempts. The values were computed as if the call attempts had been capped at a specific number. The relative bias for the proportion of cases who were Alumni Association members is also shown in the figure because the relative nonresponse bias for this estimate was the largest among the estimates we examined (using other survey variables would reach the same conclusion). As we can see in Figure

3.7, the correction for bias has little impact on the values of the indicators. The bias-adjusted R-indicators do not differ much from the indicators without this adjustment, and we focus on the latter. As call attempts increased, the relative bias in the estimated proportion of Alumni Association membership generally decreased as more alumni became respondents. However, the R-indicators show the reverse pattern, decreasing as the relative bias goes down. In contrast, penalized R-indicators increase as the relative bias goes down. We noticed one exception that the relbias increased when the number of call attempts increased from one to two. This was true for Study 1 in Section 3.2.5, indicating that those who responded at the first attempt were different from those who responded later.

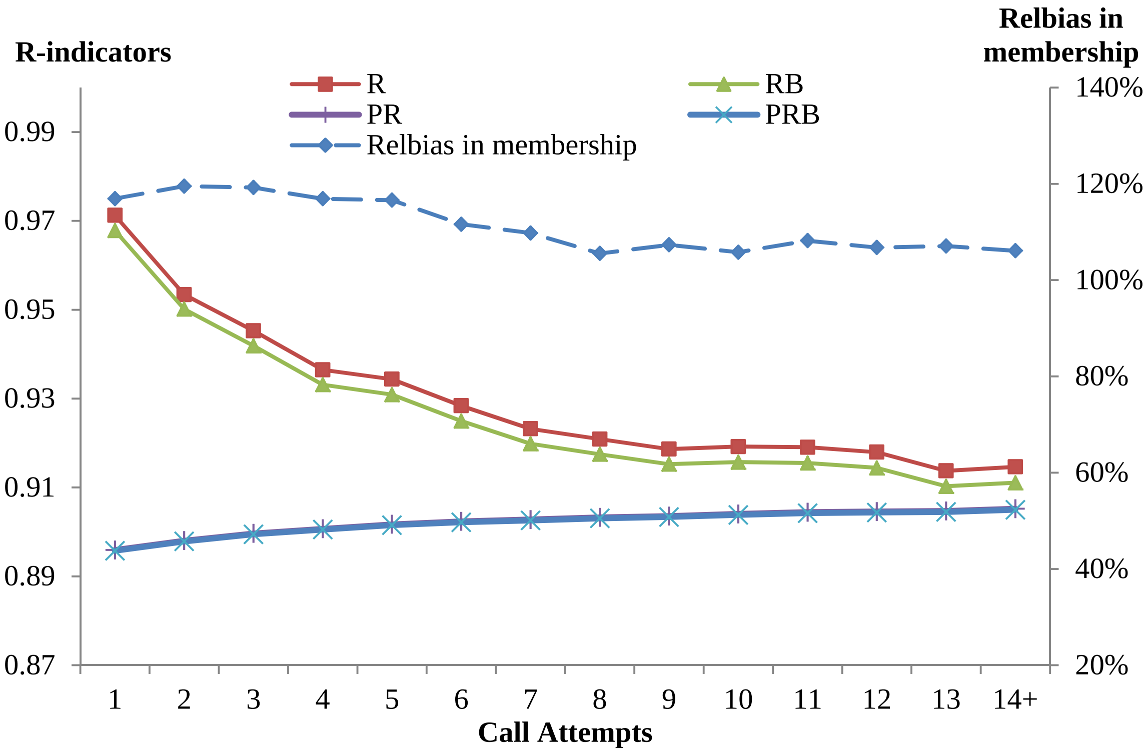

Figure 3.7. Relbias in estimated proportion with alumni association membership, R-indicator (R), bias-adjusted R-indicator (RB), penalized R-indicator (PR), and penalized bias-adjusted R-indicator (PRB) at each level of call attempts

3.4 Simulation Study

In order to compare the performance of the various indicators in different situations, we conducted a simulation study. In the simulation study, we varied the response probabilities and the associations of the variables of interest with the response probabilities. Because we know the true responses and the biases in the estimates, this will allow us to examine the performance of the indicators more in detail. Bias-corrected indicators do not make much difference in the analysis; therefore, I did not calculate them in the simulation.

3.4.1 Setup

We used the 7,591 cases with a telephone number available from the alumni data as our population and generated response probabilities using the following logistic model:

$$p_i = \frac{\exp\left(Intercept + \beta_1 * State + \beta_2 * Male + \beta_3 * Age\right)}{1 + \exp\left(Intercept + \beta_1 * State + \beta_2 * Male + \beta_3 * Age\right)}. \tag{3.9}$$

By varying the betas in the model, I generated eleven levels for the mean response probability (0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95). These cover nearly the full range of possible values.

I created five survey variables, y1 to y5; these were created by a model to achieve different levels of correlation with the response probabilities. More specifically, different levels of noise were added to the following model:

$$y_i = 100 * p_i + z * \varepsilon_i, \tag{3.10}$$

where ε_i is a random number from the normal distribution, and the values of *z* were varied to get the target correlation levels. The correlation between the response probabilities and these five survey variables for different mean response probability levels are shown in Table 3.3.

Table 3.3. Correlation between response probabilities and variables of interest at each response probability level

Mean Response Probability	y1	y2	y3	y4	y5
0.05	0.90	0.72	0.45	0.19	0.01
0.10	0.90	0.72	0.46	0.19	0.01
0.20	0.91	0.73	0.47	0.20	0.02
0.30	0.90	0.72	0.46	0.19	0.02
0.40	0.90	0.72	0.45	0.19	0.02
0.50	0.90	0.71	0.45	0.19	0.01
0.60	0.90	0.71	0.45	0.19	0.01
0.70	0.90	0.72	0.46	0.19	0.01
0.80	0.90	0.72	0.46	0.19	0.01
0.90	0.90	0.72	0.47	0.20	0.00
0.95	0.90	0.71	0.46	0.20	0.00

A simple random sample with replacement (SRSWR) was selected from the population, and 1,000 Monte Carlo replicates were generated. The response probabilities, which can be considered as the true response probabilities, were carried over to all of the resulting samples. A response indicator was then generated for each unit in each sample based on these response probabilities by making a random draw from a Bernoulli distribution with the “true” response probability. A logistic model was then fitted using the response indicator as the dependent variable. Because we generated the response propensities, we know the true model underlying the response propensities. This allows us to assess the impact of model misspecification. Two misspecified models were examined: a simpler one and a more complex one as defined below:

$$\text{Correct: } \log\left(\frac{p_i}{1-p_i}\right) = Intercept + \beta_1 * State + \beta_2 * Male + \beta_3 * Age, \qquad (3.11)$$

$$\text{Simpler}: \ \log\left(\frac{p_i}{1-p_i}\right) = Intercept + \beta_1 * Male + \beta_2 * Age, \tag{3.12}$$

$$\begin{aligned}\text{More complex: } \log\left(\frac{p_i}{1-p_i}\right) &= Intercept + \beta_1 * State + \beta_2 * Male + \beta_3 * Age \\ &+ \beta_4 * degree\_year + \beta_5 * state * degree\_year.\end{aligned} \tag{3.13}$$

We computed the biases in the estimates of interest (unweighted means) and the value of the R-indicators for each level of expected response rates.

3.4.2 Results

For each sample drawn from the population, we fitted a logistic model on the response indicators, and estimated the response propensities. The R-indicator and the penalized R-indicator were computed using these estimated response propensities. Biases in the estimates of the means of each y variable were computed using the means for the respondent cases and the means for all units in the sample.

Table 3.4 shows the correlation of the true response propensities and the estimated response propensities under the different models at each mean response propensity level. As we can see in the figure, the estimated response propensities from both the correct model and more complex model correlate with the true response propensities almost perfectly, but the estimated response propensities from the simpler model show poor correlations with the true response propensities, especially when the mean response propensity is at the range of 0.2 to 0.6. The results suggest that missing key variables in the model can have serious impact on the estimated response propensities, while adding redundant variables to the model does not have much of an effect.

Table 3.4. Correlation between estimated response propensities and true response propensities, by model and response probability level

Mean Response Probability	Correct Model	Simpler Model	More Complex Model
0.05	1.00	0.46	0.99
0.10	1.00	0.24	1.00
0.20	1.00	0.13	1.00
0.30	1.00	0.09	1.00
0.40	1.00	0.03	1.00
0.50	1.00	0.06	1.00
0.60	1.00	0.13	1.00
0.70	1.00	0.24	0.99
0.80	0.99	0.54	0.98
0.90	0.98	0.85	0.97
0.95	0.98	0.94	0.97

Figure 3.8 shows the average for each of the statistics from the correct model. The R-indicator shows the expected U-shape pattern. It starts at 0.91 when the mean response probability is 0.05, goes down to 0.53 when mean response probability is 0.30, and goes up gradually as mean response probability increases, reaching the highest value at 0.96 when mean response probability is 0.95. The U-shape pattern shown by the R-indicator is not necessarily bad because it corresponds to the reverse U-shape pattern shown by the biases in the estimated mean for the y variables, which it is what we want to see. The reverse U-shape pattern shown by the biases in the estimated mean for the y variables occurred because the y variables were defined as a linear function of response probability p_i. By definition of expectation, $E\left(\bar{Y}_R\right) = \frac{1}{\sum_{i \in s} p_i} * \sum_{i \in s} y_i p_i = \frac{100}{\sum_{i \in s} p_i} * \sum_{i \in s} p_i^2$. The

quadratic shape shown by the biases in the estimated mean for the y variables is due to the fact that the expected respondent mean is a quadratic function of p_i.

However, the values of R-indicator are similar when mean response probability is 0.10 and 0.70. It would be unwise to conclude that the two respondent samples were equally representative because that would give us wrong guidance for the biases in the estimates for the y variables. In contrast, the penalized R-indicator shows a pattern that has a better match to the pattern for the estimated biases. It also shows a U-shape pattern but with a short tail on the left and long tail on the right. Conclusions about representativeness of a respondent sample using the penalized R-indicator are more likely to be accurate than those based on the original R-indicator.

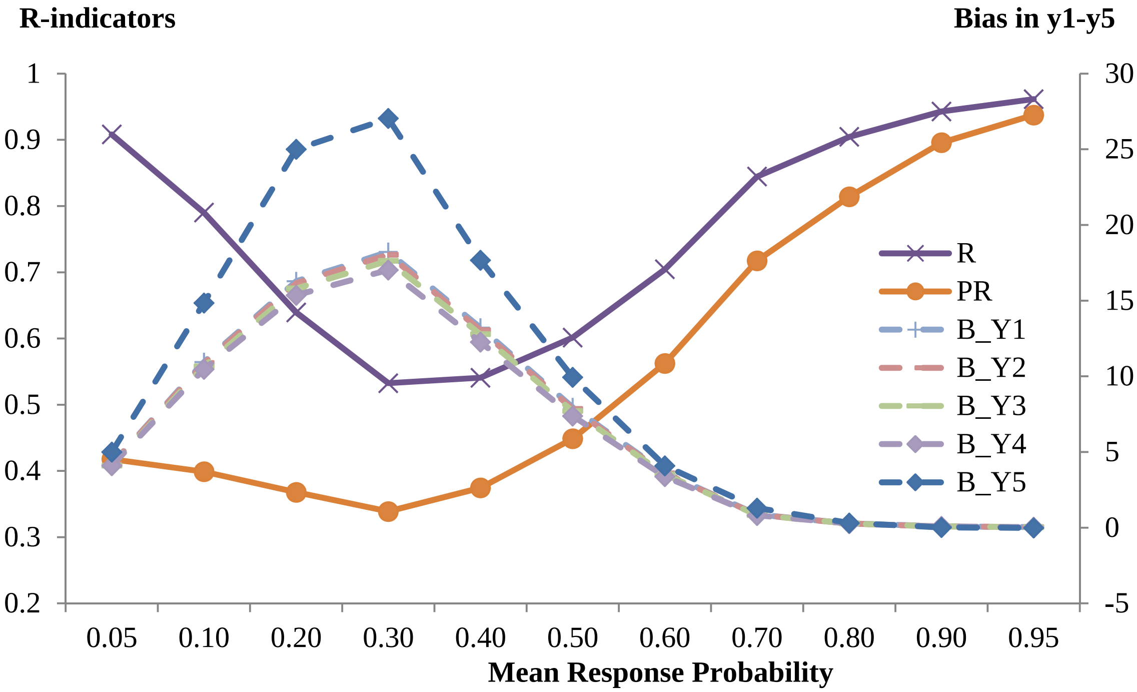

Figure 3.8. Bias in estimates of variables of interest y1-y5 (B_Y1-B_Y5), R-indicator (R), and penalized R-indicator (PR) at each level of mean response probabilities, correct model

We then examined the performance of the indicators under two misspecified models. The more complex model tells the same story as the correct model. Both the R-indicator and the penalized R-indicator show similar patterns for the complex model (Figure 3.10) as under the correct model. Under the simpler model, both R-indicator and penalized R-indicator perform very poorly (Figure 3.9). The R-indicator does not provide much information in this situation. It basically has the same value everywhere. The penalized R-indicator is dominated by the penalizing factor because of that. The penalized R-indicator increases over the whole spectrum. Judging from the biases in the estimated means for the y variables, the penalized R-indicator provides correct information about the sample representativeness when the mean response probability is greater than 0.30, but it provides incorrect information when mean response probability is less than 0.30.

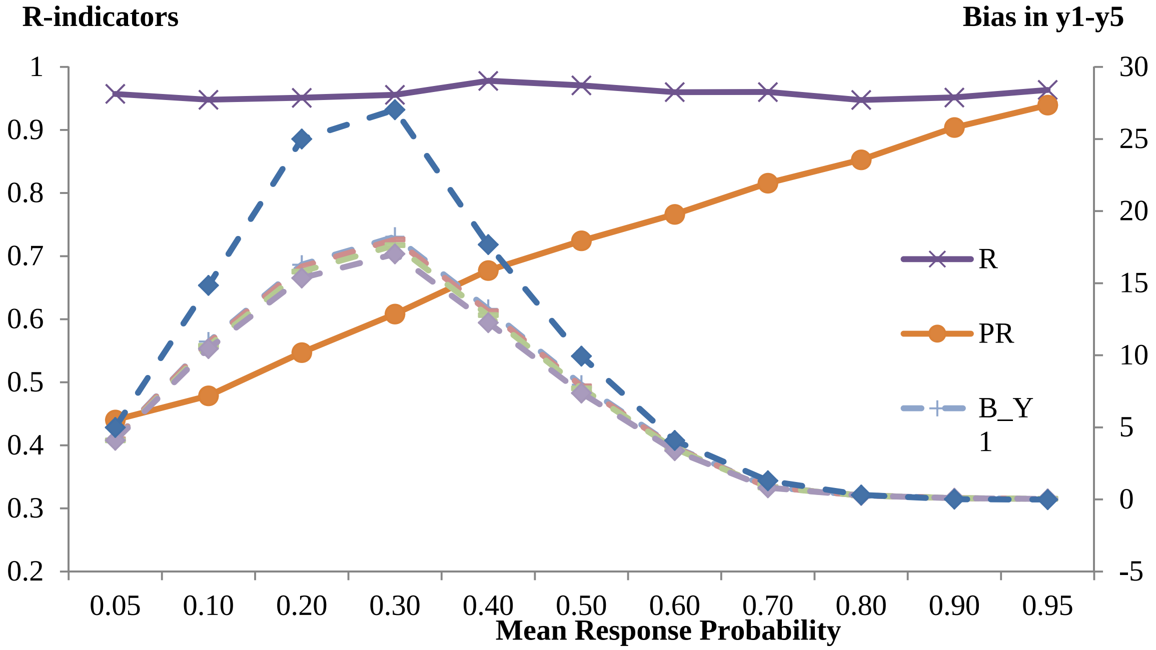

Figure 3.9. Bias in estimated means for variables y1-y5 (B_Y1-B_Y5), R-indicator (R), and penalized R-indicator (PR) at each level of mean response probabilities, simpler model

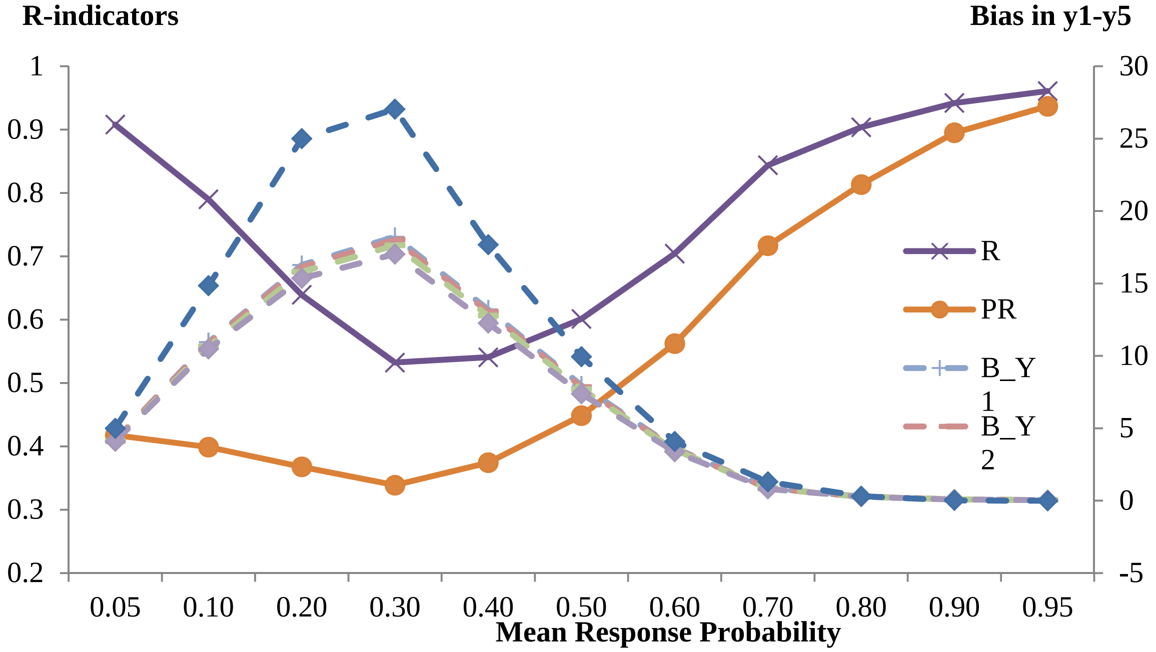

Figure 3.10. Bias in estimated means for variables of interest y1-y5 (B_Y1-B_Y5), R-indicator (R), and penalized R-indicator (PR) at each level of mean response probabilities, more complex model

In order to compare the performance of the R-indicator and penalized R-indicator more directly, we looked at the correlation between the bias in estimates for the y variables and the value of the R-indicators. The correlations were computed from the 1,000 replicates for each mean response probability level. Differences in the correlations were tested using Williams' (1959) t-test for comparing two correlations with one variable in common. As shown in Table 3.5, under the correct model, both the R-indicator and the penalized R-indicator show negative correlations with the biases in the estimated means for y1-y4, and near zero correlation with the biases in the estimated mean for y5. This suggests both indicators are in the right direction in indicating bias in estimates of variables of interest. However, the penalized R-indicator outperforms the R-indicator in most places. Take the first row in Table 3.5 for example; the correlation between the bias in the estimated mean for y1 and the R-indicator is -0.68. The corresponding number for the penalized R-indicator is -0.86. The difference in the two correlations is -0.18 (p<.001), which means the penalized R-indicator shows stronger negative correlations with the biases produced by nonresponse. For the 55 pairs of correlations shown in Table 3.5, 25 of them (45.5%) show the penalized R-indicator has stronger negative correlations with the biases than the R-indicator does; 5 of them (9.1%) indicate the opposite but the differences in correlations are close to zero (0.01 for all the five differences). The conclusion from Table 3.5 is that the penalized R-indicator generally shows stronger negative correlations with the biases produced by nonresponse, and when the correlations have higher negative values for the R-indicator, the differences are near zero. Exactly the same conclusion can be made under the more complex model as shown in Table 3.7. That is, the penalized R-indicator generally shows much stronger negative correlations with the variables of interest y1-y4, and both the R-indicator and penalized R-indicator show near zero and almost identical correlation with the biases in estimated mean for y5.

A stronger conclusion can be reached under the simpler model. As shown in Table 3.6, the penalized R-indicator universally shows stronger negative correlations with the biases in the estimated means for y1-y4, and both the R-indicator and penalized R-indicator show near zero correlation with the biases for y5. For the 55 pairs of correlations shown in Table 3.6, 27 of them (49.1%) show the penalized R-indicator has stronger negative correlations with the biases than the R-indicator does; none of them indicate the opposite.

Table 3.5. Correlation between the bias and the value of the R-indicators, under the correct model

Mean Response Probability	$Bias(\hat{\bar{y}}_1)$		Diff.	$Bias(\hat{\bar{y}}_2)$		Diff.	$Bias(\hat{\bar{y}}_3)$		Diff.	$Bias(\hat{\bar{y}}_4)$		Diff.	$Bias(\hat{\bar{y}}_5)$		Diff.
	R	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**	
0.05	-0.68	-0.86	**-0.18**	-0.57	-0.72	**-0.15**	-0.39	-0.50	**-0.11**	-0.21	-0.24	-0.03	-0.02	-0.01	0.01
0.1	-0.58	-0.80	**-0.22**	-0.43	-0.61	**-0.19**	-0.29	-0.40	**-0.11**	-0.14	-0.19	**-0.05**	-0.01	-0.01	0.00
0.2	-0.59	-0.76	**-0.18**	-0.43	-0.58	**-0.15**	-0.24	-0.33	**-0.09**	-0.16	-0.22	**-0.06**	-0.02	-0.02	0.00
0.3	-0.64	-0.77	**-0.14**	-0.49	-0.60	**-0.11**	-0.35	-0.42	**-0.07**	-0.13	-0.17	**-0.03**	-0.01	-0.02	0.00
0.4	-0.77	-0.86	**-0.09**	-0.60	-0.67	**-0.07**	-0.41	-0.46	**-0.05**	-0.13	-0.16	**-0.02**	-0.05	-0.04	0.00
0.5	-0.84	-0.89	**-0.05**	-0.67	-0.71	**-0.04**	-0.41	-0.44	**-0.03**	-0.13	-0.16	**-0.03**	0.01	0.01	0.00
0.6	-0.88	-0.90	**-0.02**	-0.71	-0.72	**-0.01**	-0.44	-0.45	-0.01	-0.13	-0.14	-0.01	0.01	0.01	-0.01
0.7	-0.89	-0.88	0.01	-0.70	-0.70	0.00	-0.43	-0.43	0.00	-0.16	-0.17	-0.01	-0.03	-0.02	0.01
0.8	-0.89	-0.87	**0.01**	-0.71	-0.70	**0.01**	-0.42	-0.42	0.00	-0.16	-0.17	0.00	-0.06	-0.06	0.00
0.9	-0.89	-0.88	**0.01**	-0.73	-0.72	0.01	-0.50	-0.49	0.01	-0.21	-0.21	0.00	-0.09	-0.09	0.00
0.95	-0.90	-0.88	**0.01**	-0.73	-0.72	**0.01**	-0.48	-0.48	0.01	-0.25	-0.25	0.00	-0.02	-0.01	0.01

Note: R means R-indicator; PR means penalized R-indicator; differences in bold are significant at .05 level.

Table 3.6. Correlation between the bias and the value of the R-indicators, under the simpler model

Mean Response Probability	$Bias(\hat{\bar{y}}_1)$		Diff.	$Bias(\hat{\bar{y}}_2)$		Diff.	$Bias(\hat{\bar{y}}_3)$		Diff.	$Bias(\hat{\bar{y}}_4)$		Diff.	$Bias(\hat{\bar{y}}_5)$		Diff.
	R	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**	
0.05	-0.46	-0.44	0.02	-0.40	-0.38	0.02	-0.27	-0.27	0.00	-0.17	-0.14	0.04	0.00	0.01	0.02
0.10	-0.37	-0.36	0.01	-0.30	-0.32	-0.02	-0.21	-0.21	0.00	-0.09	-0.09	0.00	-0.01	-0.01	0.00
0.20	-0.10	-0.20	**-0.10**	-0.12	-0.22	**-0.10**	-0.06	-0.13	**-0.07**	-0.04	-0.08	**-0.04**	0.00	0.01	0.01
0.30	-0.04	-0.18	**-0.13**	-0.08	-0.18	**-0.10**	-0.02	-0.08	**-0.07**	-0.01	-0.04	**-0.03**	-0.04	-0.04	0.00
0.40	-0.02	-0.15	**-0.12**	-0.06	-0.15	**-0.09**	-0.02	-0.09	**-0.06**	0.01	-0.03	**-0.04**	-0.09	-0.08	0.01
0.50	-0.10	-0.20	**-0.10**	-0.10	-0.18	**-0.08**	-0.05	-0.11	**-0.06**	-0.03	-0.07	**-0.04**	0.03	0.03	0.00
0.60	-0.14	-0.22	**-0.08**	-0.08	-0.15	**-0.07**	-0.12	-0.16	**-0.04**	-0.05	-0.07	**-0.02**	-0.09	-0.09	0.00
0.70	-0.24	-0.29	**-0.04**	-0.17	-0.21	**-0.04**	-0.10	-0.13	**-0.02**	-0.05	-0.08	**-0.02**	-0.01	0.00	0.01
0.80	-0.53	-0.56	**-0.02**	-0.43	-0.45	**-0.02**	-0.27	-0.29	**-0.02**	-0.09	-0.10	-0.01	-0.05	-0.05	0.00
0.90	-0.77	-0.78	-0.01	-0.64	-0.64	0.00	-0.47	-0.47	0.00	-0.17	-0.17	0.00	-0.10	-0.10	0.00
0.95	-0.86	-0.86	0.01	-0.70	-0.70	0.00	-0.49	-0.48	0.01	-0.24	-0.24	0.00	-0.01	0.00	0.01

Note: R means R-indicator; PR means penalized R-indicator; differences in bold are significant at .05 level.

Table 3.7. Correlation between the bias and the value of the R-indicators, under the complex model

Mean Response Probability	$Bias(\hat{\bar{y}}_1)$		**Diff.**	$Bias(\hat{\bar{y}}_2)$		**Diff.**	$Bias(\hat{\bar{y}}_3)$		**Diff.**	$Bias(\hat{\bar{y}}_4)$		**Diff.**	$Bias(\hat{\bar{y}}_5)$		**Diff.**
	R	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**		**R**	**PR**	
0.05	-0.72	-0.89	**-0.17**	-0.60	-0.75	**-0.14**	-0.41	-0.52	**-0.11**	-0.22	-0.24	-0.03	-0.01	0.00	0.01
0.10	-0.60	-0.81	**-0.22**	-0.44	-0.62	**-0.19**	-0.30	-0.41	**-0.11**	-0.14	-0.19	**-0.05**	0.00	0.00	0.00
0.20	-0.59	-0.77	**-0.17**	-0.43	-0.58	**-0.15**	-0.24	-0.33	**-0.09**	-0.16	-0.22	**-0.06**	-0.02	-0.02	0.00
0.30	-0.64	-0.77	**-0.14**	-0.49	-0.60	**-0.11**	-0.35	-0.42	**-0.07**	-0.13	-0.17	**-0.03**	-0.01	-0.02	0.00
0.40	-0.77	-0.86	**-0.09**	-0.60	-0.67	**-0.07**	-0.41	-0.46	**-0.05**	-0.13	-0.15	**-0.02**	-0.05	-0.05	0.00
0.50	-0.84	-0.89	**-0.05**	-0.67	-0.71	**-0.04**	-0.41	-0.44	**-0.03**	-0.13	-0.16	**-0.03**	0.01	0.01	0.00
0.60	-0.88	-0.89	**-0.02**	-0.71	-0.72	**-0.01**	-0.44	-0.45	-0.01	-0.13	-0.14	-0.01	0.02	0.01	-0.01
0.70	-0.88	-0.88	0.01	-0.70	-0.70	0.00	-0.43	-0.43	0.00	-0.16	-0.17	-0.01	-0.03	-0.02	0.01
0.80	-0.88	-0.87	**0.01**	-0.72	-0.70	**0.01**	-0.42	-0.42	0.00	-0.16	-0.17	0.00	-0.05	-0.05	0.00
0.90	-0.89	-0.88	**0.01**	-0.73	-0.72	0.01	-0.50	-0.49	0.01	-0.21	-0.21	0.00	-0.09	-0.09	0.00
0.95	-0.89	-0.88	**0.01**	-0.73	-0.72	0.01	-0.48	-0.47	0.01	-0.26	-0.26	0.00	-0.03	-0.02	0.01

Note: R means R-indicator; PR means penalized R-indicator; differences in bold are significant at .05 level.

3.5 General Discussion

In their applications of R-indicators, Schouten and his colleagues conclude that R-indicators can be “valuable tools in the comparison of different surveys and data collection strategies” (Schouten, Cobben, and Bethlehem 2009, p111). Furthermore, the authors suggest that “used together with R-indicators and response rates, survey managers can target data collection resources to specific subgroups contributing to the lack of representativity…” (Schouten, Shlomo, and Skinner 2011). However, as shown in this study, for the job of indicating the sample quality, the R-indicators are not as good as the penalized R-indictors, which correct (or penalize) for the observed response rates.

Using the two datasets, the penalized R-indictors show more compatible patterns with the bias in estimated means for the survey variables than the R-indicators. We should note that the target populations for both datasets examined here are not the general U.S. population, but only a fraction of it. We should also note because the final response rates for the studies were just above 30%, we do not know how the R-indicators might have performed had the response rate gone up (moved closer to 100%). However, the simulation study also shows that the penalized R-indicators have better correlation with the bias in estimates than the R-indicators. Note that in the simulation study, the bias in the estimated mean does not increase or decrease strictly with the response probability level.

However, as with the R-indicators, the penalized R-indicators are also dependent on the set of auxiliary variables. When the response propensities estimated from the model with the set of auxiliary variables have no correlation with the y variables, the penalized R-indictors are not informative about the potential bias in estimates, as shown with the variable y5 in the simulation. A second common weakness of the R-indicators and the penalized R-indicators is that they are sensitive to the response propensity models. As shown in the simulation, the choice of auxiliary variables to be included can have a big impact on the estimated response propensities. As a result, the estimated indicators may or may not be informative. A third common weakness lies in the fact that in order to compare the indicators across surveys, the response propensity models need to have a same set of variables. This may limit the model flexibility, and only a very limited set of variables can be included (e.g., sex and age).

Although there are some common weaknesses, the penalized R-indicators may be a better measure for survey practitioners, and more empirical research is needed to provide more evidence. Also, more theoretical research is needed to explore the properties of the penalized R-indictors.

Chapter 4: Nonresponse Error in a Total Survey Error Context

4.1 Introduction

Survey researchers advocate survey designs that minimize the total survey error for a given cost (see e.g., Biemer and Lyberg 2003; Groves 1989). However, studies are often forced to focus on one or two of the many sources of error, usually the sampling variance and perhaps one or two potential biases. Furthermore, the relative importance of the errors that are studied is rarely evaluated; this means researchers have no guidelines for resource allocation.

From the classical sampling perspective, the various types of survey errors are components of the deviation between the sample mean and the population mean, expressed as the concept of mean squared error (MSE). The MSE can be decomposed as

$$MSE = Var\left(\hat{\bar{Y}}\right) + \left[Bias(\hat{\bar{Y}})\right]^2, \qquad (4.1)$$

where MSE is the sum of variance and squared bias in the estimated mean. The bias refers to the difference between the expected value and the true value. The variance reflects the variation around the expected value over repeated trials. Each source of survey error can lead to bias and variance in the survey estimates.

Four of the major sources of error—coverage error, sampling error, nonresponse error, and measurement error—are regarded as "cornerstones of survey research" by Dillman (2007:10). *Coverage error* usually results from an imperfect frame from which some of the units have been missed, resulting in undercoverage. If the frame contains units that do not belong to the target population, overcoverage errors can also occur. *Sampling error* occurs because only a fraction of the units in the population of interest is included in the sample. Often, some of the sample members fail to respond to the survey and this introduces *nonresponse error*. *Measurement error* occurs when the responses obtained from the sample member do not agree with the true values. The instrument and mode of data collection often play important roles in this type of error. In practice, postsurvey adjustments are used to reduce the effects of coverage, sampling, and nonresponse errors on the estimates, but these adjustments can introduce errors of their own. Here we do not intend to study the processing error which is another important type of error. *Processing error* refers to the errors introduced by editing decisions, coding, or other operational errors.

To our knowledge, the four major types of errors have not been investigated together within one study; instead, individual errors have typically been the focus. Although there are both bias and variance components in each type of error except sampling error, we only intend to investigate the sampling error and bias components in other types of error. We reexamine data from two studies originally reported in Kreuter, Presser, and Tourangeau (2008) and Tourangeau, Groves, and Redline (2010) that together permit estimates of the magnitude of all five types of errors. Data from these two studies have frame records available for the entire sample. Previous analyses of these data (Kreuter, Presser, and Tourangeau 2008; Kreuter, Yan, and Tourangeau 2008; Sakshaug and Kreuter 2011; Sakshaug, Yan, and Tourangeau 2010; Tourangeau, Groves, and Redline 2010) focused only on nonresponse bias or/and measurement bias. Using these data, we try to study three major sources of error in the voters survey and four major sources of error in the alumni survey and evaluate their relative importance. In addition, we try to assess the amount of reduction in error that postsurvey adjustments can achieve.

Studies of multiple error sources. Peytchev, Carley-Baxter, and Black (2011) investigate both coverage bias and nonresponse bias in their study which compared a landline telephone survey with a follow-up survey on nonrespondents and an RDD cell phone survey. They found that coverage biases and nonresponse biases were in opposite directions and that the coverage biases were larger.

Using court records as benchmarks, Schaeffer, Seltzer, and Klawitter (1991) and Olson (2006) compare the magnitudes of nonresponse baises and measurement biases, but find inconsistent results across estimates. Schaeffer, Seltzer, and Klawitter (1991) show that measurement bias is higher than nonresponse bias for the *amounts* of support owed and paid; however, for the *proportion* of cases with any support owed and paid, nonresponse bias is greater than measurement bias. Olson (2006) finds that nonresponse bias is greater than measurement bias for mean length of marriage and mean number of marriages, while for mean time elapsed since divorce, the reverse was true. Mixed results are also found by Biemer (2001), using a reinterview survey design. Nonresponse bias was higher than measurement bias for some items (e.g., whether the sample member ever stopped smoking for at least one day during the past 12 months), while measurement bias was larger for some items (e.g., whether the sample member would like to quit smoking completely). For the CATI survey, measurement bias is higher than nonresponse bias for whether the sample members have smoked at least 100 cigarettes during the lifetime, but the magnitudes of the two reversed for the face-to-face survey.

For the item asking whether there was a firearm in or around the sample member's home, nonresponse bias was greater than measurement bias for the CATI survey, while the reverse was observed for the face-to-face survey. Using the report of abortions in the audio computer-assisted self-interviewing (ACASI) mode as a "gold standard" to assess the report of abortions in the computer-assisted personal interviewing (CAPI) mode, Peytchev, Peytcheva, and Groves (2010) found that CAPI respondents in the lowest response propensity quintile tended to be more likely to underreport their abortions. Tourangeau, Groves, and Redline (2010) find that measurement bias was about twice as large as nonresponse bias for two voting behaviors, using voter registration records as true values. The voting behavior items are thought to be prone to large measurement biases, especially social desirability biases. In a further investigation, Sakshaug, Yan, and Tourangeau (2010) suggest that measurement biases tended to be larger than nonresponse biases for estimates of socially undesirable characteristics, but not for estimates of socially desirable or neutral characteristics where nonresponse biases were larger.

Weighting adjustments to correct survey errors. In weighting a survey, a sampling weight is normally computed as the inverse of the sampling probability for each case:

$$W_{1i} = \frac{1}{\pi_i}. \tag{4.2}$$

where π_i is the probability that the case is selected. Such weights yield unbiased estimates but can increase their variance. However, adjustments to the base weights to compensate nonresponse or noncoverage are often not so simple. Most of the time, there is limited knowledge about the effects of nonresponse and noncoverage, and the weighting adjustments are based on some assumptions. Lin and Schaeffer (1995) created weights based on number of call attempts (methods 1) or call results (method 2), assuming these variables were predictive of response probabilities. However, they did not find this approach was effective in reducing bias. Bethlehem and Schouten (2004) employed more complicated models and used more types of variables in their study. Although they tried to fit weighting models separately for each of a set of target variables, biases still remained after the weighting.

A new weight is commonly computed by multiplying the base sampling weight with nonresponse adjustment weight (w_{2i}),

$$w_{3i} = w_{1i} w_{2i} \,. \tag{4.3}$$

Some survey researchers have tried to adjust the weights by identifying the households with an interruption in telephone service (Brick, Waksberg, and Keeter 1996; Frankel et al. 2003; Davern et al. 2004), assuming these households are similar to those without telephone service. However, there is no evidence for the effectiveness of this approach. Often times the new weight w_{3i} is calibrated to reproduce population totals, resulting in a final weight w_{4i}.

Taking advantage of the records available from the frame (the Aristotle database on Maryland residents who were registered to vote; the combined files from the Registrar and the Alumni Office at the University of Maryland), this study will examine nonresponse bias, measurement bias, sampling error, and adjustment error using the voters data, and coverage bias, nonresponse bias, measurement bias, sampling error, and adjustment error using the alumni data. To our knowledge, no study has done this before. The hypotheses for this study are: 1) nonsampling errors will generally be larger than sampling errors; 2) among the sources of nonsampling error, measurement biases will be larger than coverage or nonresponse biases, at least for the estimates based on responses to sensitive questions; and 4) adjustments can reduce the coverage and nonresponse biases, but sometimes may make the total error worse.

4.2 Study 1

4.2.1 Study Dataset

The voters data. A description of the data set can be found in Section 1.7.1 of Chapter One. The voters and nonvoters were drawn with different sampling rates to get roughly an equal number of sample cases. To keep the reported errors consistent with previous analysis (Tourangeau, Groves, and Redline 2010), we computed the expected population values using the sampling rates and treated the sample as equal probability sample.

The first two rows of Table 4.1 list the variables of interest and the wording of the survey questions on which they are based. The Aristotle (http://www.aristotle.com) database contained information on voting history in the 2004 and 2006 general elections. The surveys also asked questions about whether the respondent had voted in the 2004 and 2006 general elections. The Aristotle database also contained some demographic variables. To examine the errors in the estimates for these variables, we picked demographic characteristics—sex and age. These variables were used in the adjustment process.

Table 4.1. Variables from the voters dataset

Variable	Question asked in the survey
Voted in 2004	In the 2004 presidential election, did things come up that kept you from voting or did you happen to vote?
Voted in 2006	In the 2006 mid-term selection, did things come up that kept you from voting or did you happen to vote?
Sex (male)	Are you male or female?
Age	What is your date of birth?

4.2.2 Computing the Survey Errors

In sampling theory, the total error that combines the variance and the bias is a widely accepted measure. The total error, which is also often called the root mean square error, is defined as:

$$Error_{total} = \sqrt{Variance + Bias^2} = \sqrt{\sigma^2 + B^2}. \tag{4.4}$$

Kish (1987) suggests that this measure is useful in comparing designs. Kish (1987:221) also proposes a measure of relative size of bias, which is called “bias ratio” B / σ. The bias ratio measures the effect of a certain bias, although it will differ widely for different survey variables and domain statistics. This expression is related to *z*-statistic. For the estimated respondent mean, a *z*-statistic can be constructed as:

$$z = \frac{\bar{y}_{sR}^{O} - \bar{Y}_U}{\sqrt{\operatorname{var}\left(\bar{y}_{sR}^{0}\right)}}, \tag{4.5}$$

where $\bar{y}_{sR}^{O}$ is the mean for respondents *R* in the sample *s* based on reported values, $\bar{Y}_U$ is the population mean, and $\operatorname{var}\left(\bar{y}_{sR}^{0}\right)$ is the variance of $\bar{y}_{sR}^{O}$.

We further decomposed the bias component and estimated the total survey error as follows:

$$
\begin{aligned}
Error_{total} &= \sqrt{variance + \left(bias_{measurement} + bias_{nonresponse} + bias_{sample}\right)^2} \\
&= \sqrt{\frac{1}{n_{sR} * (n_{sR} - 1)} \sum_{i=1}^{n_{sR}} \left(y_{iR}^{O} - \overline{y}_{sR}^{O}\right)^2 + \left[(\overline{y}_{sR}^{O} - \overline{y}_{sR}^{A}) + (\overline{y}_{sR}^{A} - \overline{y}_{s}^{A}) + (\overline{y}_{s}^{A} - \overline{Y}_{U}) \right]^2}
\end{aligned}
$$

(4.6)

where n_R is the total number of respondents in the sample s, $\overline{y}_{sR}^{O}$ is the mean for respondents *R* in the sample *s* based on reported values, $\overline{y}_{sR}^{A}$ is the mean for respondents R in the sample s based on actual values, $\overline{y}_{s}^{A}$ is the mean for the full sample s (including respondents and nonrespondents) based on actual values, and $\overline{Y}_U$ is the population mean. The total error is also computed after all adjustments. The adjustment procedures are explained in the next section. Note that the last term $\overline{y}_{s}^{A} - \overline{Y}_U$ reflects the bias in this specific realized sample. For infinite random samples, the expected bias is zero. Also note that there is no coverage bias in the formula because contact information was available for all units in the frame which we considered as the target population. The coverage bias is addressed in the analysis in study 2 (Section 4.3). Further note that the biases as calculated in the formula do not consider the variance of the estimated biases.

After taking out the 1,000 cases used for pretest, the size for the frame is 49,000, with 1,020 cases picked for the telephone sample and 1,669 cases for the mail sample. The number of respondents was 350 for the telephone sample and 554 for the mail sample.

4.2.3 Adjustment Procedure

We developed weights and examined their impact on the total survey error. First, a set of weights compensating for nonresponse was created for the telephone and mail samples separately. The weights were estimated by the inverse of response propensities. The response propensity models were based on the available auxiliary variables (stepwise selection of the variables with the standard 0.05 significance level). For the telephone sample, the candidate independent variables include call records, experimental variables, and frame variables; only experimental variables and frame variables were included as candidate independent variables for the mail sample because there were no call records for the mail cases. Missing values of the frame variables were imputed using the multivariate sequential regression imputation method (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001). Only one imputed dataset was created. The final models are shown in Table 4.2. The final model for the telephone sample is the same as model 1 (labeled as "all variables") in Table 2.2 of Chapter Two, and the final model for the mail sample is the same as model 1 (labeled as "all variables") in Table 2.3 of Chapter Two.

Table 4.2. Logistic regression coefficients from response propensity models

	Telephone (n=1,020)		**Mail (n=1,669)**	
	Coefficient	**SE**	**Coefficient**	**SE**
Intercept	-2.03	0.46	-1.86	0.26
Contact	1.45	0.44	–	–
Number of calls	-0.07	0.03	–	–
Computer owner	0.48	0.16	-0.67	0.44
Home business	0.97	0.35	–	–
Do Not Call list	-0.49	0.48	0.47	0.13
Age	–	–	0.01	0.00
Contact* Do Not Call List	1.19	0.51	–	–
Age*computer owner	4.25	1.64	0.02	0.01

The weights were then calibrated to the population totals on sex and age, generating the final weights for the analyses. The total error after adjustments were computed in the way as defined in Equation 4.6, but $\operatorname{var}\left(\bar{y}_{sR}^{0}\right)$ and $\bar{y}_{sR}^{0}$ were estimated with the final weights.

Figure 4.1 shows the proportions voting in 2004 and 2006, the percent of the respondents who were male, and mean age by weight quintile for the telephone sample. The weights increase from left to right within each variable. The mean for each group of respondents who voted in 2004 is indicated by a diamond and the confidence intervals for each mean are shown as a vertical line. As Figure 4.1 shows, the proportions of male respondents in the highest three weight quintiles are higher than in the lowest two quintiles. This pattern indicates that the weights will change the estimate for the sex variable greatly—the weighted estimate will have a higher percent of male than the unweighted estimates. There are no such clear patterns for the other three variables—the proportions voting in 2004 and 2006, and mean age.

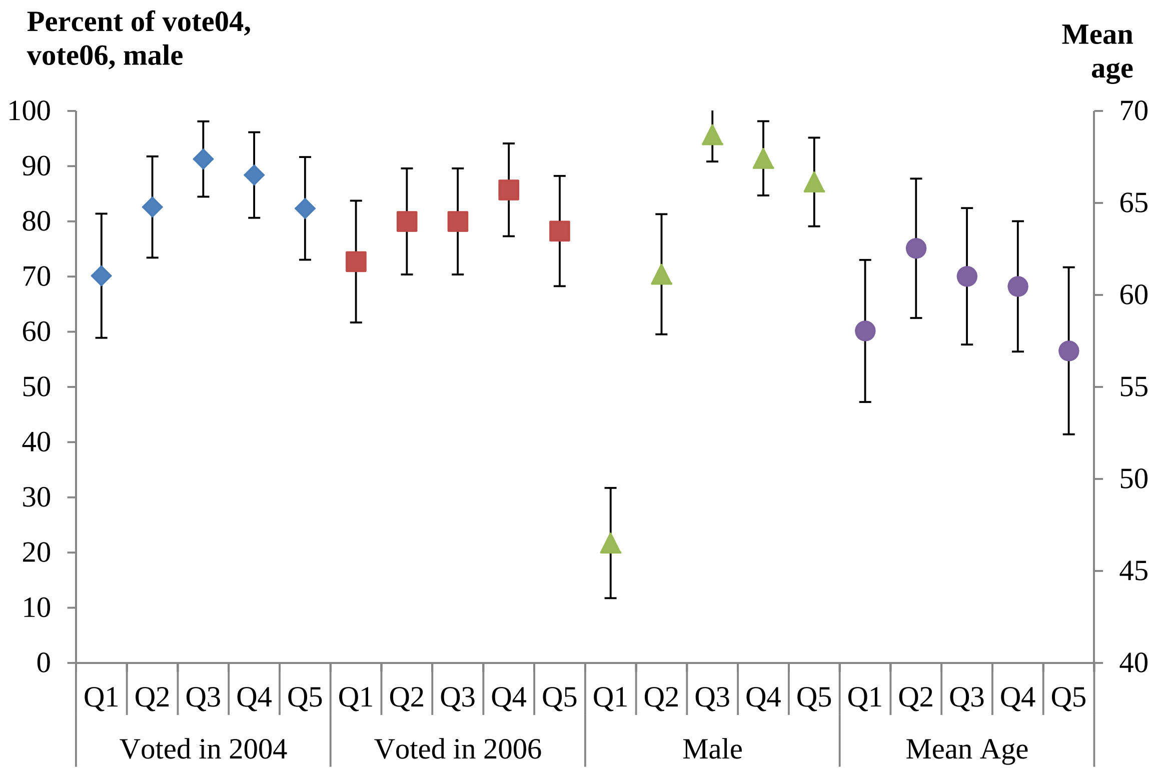

Figure 4.1. Proportions voting in 2004 and 2006, percent of male respondents, and mean age by quintile of weights, telephone sample

Figure 4.2 shows the proportions voting in 2004 and 2006, the percent of respondents who are male, and mean age by weight quintiles for the mail sample. As before, the weights increase from left to right within each variable. The mean for each group of respondents who voted in 2004 is indicated by a diamond and the confidence intervals are shown as vertical lines. As Figure 4.2 shows, there are no clear patterns for all four variables, although there is a lot more variation across weight quintiles for the demographic variables than for the voting variables. More specifically, the percent male is lower in the lowest weight quintile than in the other quintiles.

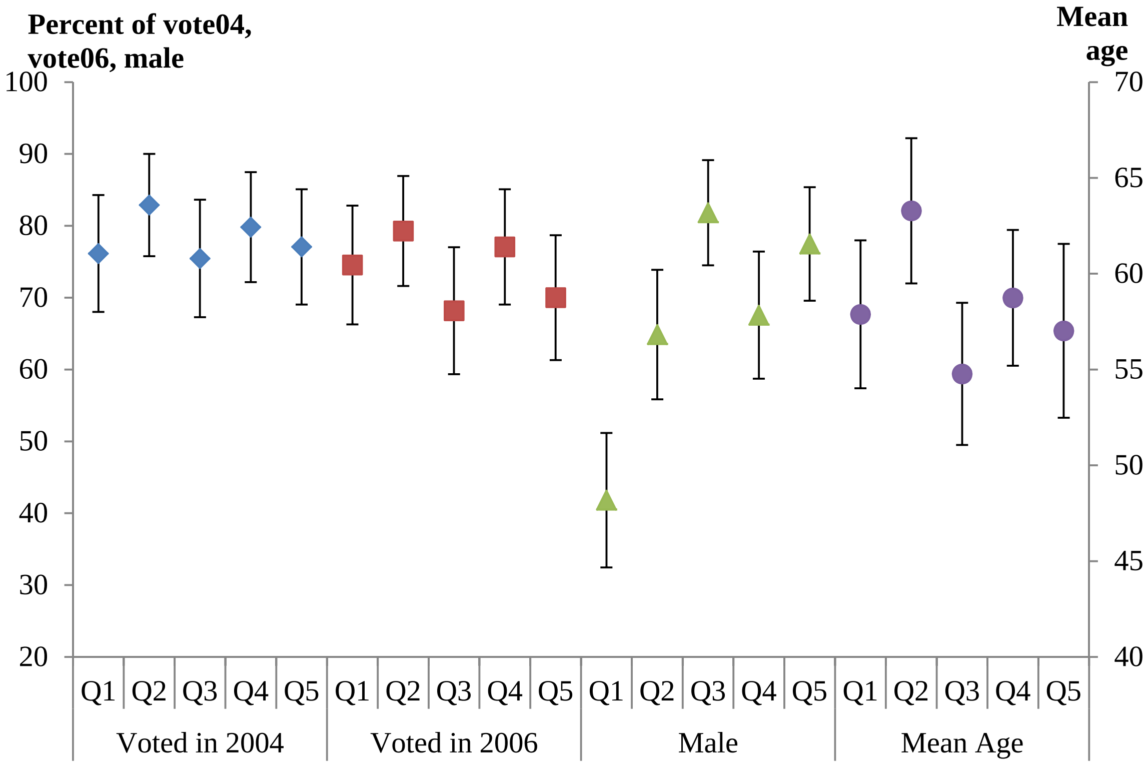

Figure 4.2. Proportions voting in 2004 and 2006, percent of male respondents, and mean age by quintile of weights, mail sample

4.2.4 Comparing the Survey Errors

Table 4.3 shows that there are large positive biases in the two estimates of interest (the proportions voting in 2004 and 2006) in the responding sample. The estimated proportion of sample members who voted in the 2004 election is off by 13.6 percentage points for the telephone sample and by 9.6 percentage points for the mail sample. These numbers represent the difference between frame values for the samples and frame values for the respondents. These differences are our estimates of nonresponse bias. The nonresponse biases are similar for the estimated turnout in 2006—14.2 percentage points for the telephone sample and 12.4 percentage points for the mail sample. This reflects the fact that voters were more likely to be respondents to the surveys than nonvoters were

(41.3% response rate among voters vs. 25.9% response rate among nonvoters). The nonvoters were the persons who did not vote in both the 2004 and 2006 elections.

Overreporting of voting has been documented in previous studies (Belli, Traugott, and Beckmann 2001; Locander, Sudman and Bradburn 1976; Parry and Crossley 1950; Traugott and Katosh 1979). This is thought to reflect the social desirability of behaviors such as voting. Our analysis shows that the measurement biases are also positive and roughly double the size of the nonresponse biases. The same conclusion is reached in a previous analysis of the data (Tourangeau, Groves, and Redline 2010). For the telephone sample, the measurement biases in the estimated proportions who voted in 2004 and 2006 are 21.6 and 22.0 percentage points, respectively. These numbers represent the difference between reported and frame values for the respondents. The numbers for the mail sample are 21.2 and 17.5 for the proportion who voted in 2004 and 2006, respectively.

The nonresponse and measurement biases are cumulative. For the telephone sample, the total bias in the estimated proportion who voted in 2004 is 35.6 percentage points and for the proportion who voted in 2006 it is 36.3 percentage points; for the mail sample, the biases are 30.8 and 30.7 percentage points, respectively. This amounts to an average relbias in the estimates of these two variables of 79.6% for the telephone sample, and 68.1% for the mail sample. Sampling error is much smaller than nonresponse bias and measurement bias.

In practice, weights are used to reduce such biases. However, the standard weighting procedures do not produce much reduction in the errors found in this study. Although the total error after adjustments is smaller than for the unadjusted estimates, the reductions in the total error are all less than 4 percentage points, representing an average relative reduction of 9.9% for the telephone sample, and hardly any (0.1%) for the mail sample.

The biases in the estimates for the two demographic variables are much smaller than those for the two voting variables. Biases due to nonresponse and measurement in both sex and age are positive for the telephone sample. This indicates men were more likely to respond to the telephone survey than women were and that older people were more likely to be respondents to the survey than younger people. However, because these biases are small, nonresponse does not appear to have affected the demographic composition of the sample. The same conclusion holds for the mail sample, although there was a negative bias (5.1 percentage points) in the estimated proportion male, probably due to the wrong person filling out the questionnaire. The adjustments effectively corrected this bias.

Table 4.3. True status in percent, sample bias, nonresponse bias, measurement bias, sampling error, total error, total error after adjustments, and amount of reduction from the adjustments

Variable	Frame Value	Sample bias	Nonresponse bias	Measurement bias	Sampling error	Total error	Total error (adjusted)	% reduction from the adjustments
Telephone								
Voted in 2004	47.5	0.4	13.6	21.6	2.0	35.6	31.8	-10.7
Voted in 2006	43.1	0.1	14.2	22.0	2.2	36.4	33.1	-9.1
average	*45.3*	*0.2*	13.9	21.8	*2.1*	*36.0*	*32.4*	*-9.9*
Male	70.6	1.7	0.3	0.9	2.4	3.7	6.1	64.5
Mean age	56.0	0.7	2.3	1.0	0.9	4.0	3.6	-8.6
N	*49,000*	*1,020*	*350*[§]	*350*[§]	*350*[§]			
Mail								
Voted in 2004	47.5	0.1	9.6	21.2	1.8	30.9	31.3	1.4
Voted in 2006	43.1	0.8	12.4	17.5	1.9	30.7	30.2	-1.6
average	*45.3*	*0.4*	11.0	19.3	*1.8*	*30.8*	*30.8*	*-0.1*
Male	70.6	0.6	0.6	-5.1	2.0	4.3	2.0	-53.8
Mean age	56.0	-0.4	2.6	0.1	0.7	2.4	2.1	-15.0
N	*49,000*	*1,669*	*554*[§]	*554*[§]	*554*[§]			

Note: [§] N varies because of item nonresponse.

Sample bias refers to the difference between mean frame values for this specific sample and mean frame values for the population.

Nonresponse bias was estimated by the difference between frame values for the samples and frame values for the respondents.

Measurement bias was estimated by the difference between reported and frame values for the respondents.

Adjustment bias was estimated by the difference between unweighted and weighted reported values for the respondents.

Total error was estimated by the sum of total absolute bias and sampling error.

Total error after adjustments was estimated by the sum of total absolute bias after adjustments and sampling error.

4.3 Study 2

4.3.1 Study Dataset

The alumni data. A description of this data set can be found in Section 1.7.2 of Chapter One. Because the sample members needed to complete the screener before they could be assigned to each one of the three modes (telephone, Web, and IVR), the coverage bias referred to here is the same for all three modes and is based on all the sample cases for which a telephone number was available.

Table 4.4 lists the key variables and gives the wording of the survey questions. These variables could be checked against records available from the Registrar's Office or the Alumni Association. The items are grouped into two categories: undesirable characteristics and desirable characteristics. One socially desirable item (GPA above 3.5) and one undesirable item (GPA below 2.5) were created from the GPA variable. Other socially undesirable items include dropping a class, getting an unsatisfactory grade, and receiving an academic warning or being placed on probation; the other socially desirable items include receiving academic honors, being a member of the Alumni Association, and donating money to the University of Maryland after graduation or in last year (2004). Some demographic information was also available on the frame. As in Study 1, to examine the errors in the estimates of these variables, we picked two stable characteristics—sex and age. There were used in the postsurvey adjustments.

After excluding cases that were used for the pretest, duplicate numbers, and with telephone numbers outside continental U.S., the sample size for the frame was 17,266;

7,591[3] of these were fielded to get the screener. Out of the 7,591 cases, 1,501 responded to the screener. The telephone survey had 320 respondents, the Web survey had 363, and the IVR survey also had 320.

Table 4.4. Variables of interest from the alumni dataset

Variable (short name)	Question asked in the survey
Undesirable characteristics	
Percent with GPA <2.5 (*GPA below 2.5*)	What was your cumulative overall undergraduate grade point average or GPA at the time you received your undergraduate degree?
Percent with at least one D or F (*F or D*)	Did you ever receive a grade of "D" or "F" for a class?
Percent who dropped a class (*withdraw*)	During the time you were an undergraduate at the University of Maryland, did you ever drop a class and receive a grade of "W"?
Percent getting warning or on probation (*probation*)	Were you ever placed on academic warning or academic probation?
Desirable characteristics	
Percent with GPA > 3.5 (G*PA above 3.5*)	(see GPA<2.5 above)
Percent getting honors (*Honors*)	Did you graduate with cum laude, magna cum laude, or summa cum laude?
Percent who ever donated (*donated*)	Since you graduated, have you ever donated financially to the University of Maryland?
Percent donating in last year (*donated in last year*)	Did you make a donation to the University of Maryland in calendar year 2004?
Percent who are members of Alumni Association (*member*)	Are you a dues-paying member of the University of Maryland Alumni Association?
Demographics	
Sex (*sex*)	What is your gender?
Age in years (*age*)	In what year were you born?

4.3.2 Computing the Survey Errors

[3] The original study (Kreuter, Presser, and Tourangeau 2008) reported that 7,591 cases were fielded, but a re-analysis of the data (Sakshaug, Yan, and Tourangeau 2010) reported the results on 7,535 cases based on some exclusion criteria. Here we used the 7,591 cases reported in the original study.

Following the discussion in Section 4.2.2, we estimated the total survey error as follows:

$$Error_{total} = \sqrt{variance + \left(bias_{measurement} + bias_{nonresponse} + bias_{coverage}\right)^2}$$
$$= \sqrt{\frac{1}{n_{sR} * (n_{sR} - 1)} \sum_{i=1}^{n_{sR}} \left(y_{iR}^O - \bar{y}_{sR}^O\right)^2 + \left[(\bar{y}_{sR}^O - \bar{y}_{sR}^A) + (\bar{y}_{sR}^A - \bar{y}_s^A) + (\bar{Y}_U^c - \bar{Y}_U)\right]^2} \tag{4.7}$$

where n_R is the total number of respondents in the sample s, $\bar{y}_{sR}^O$ is the mean for respondents *R* in the sample *s* based on reported values, $\bar{y}_{sR}^A$ is the mean for respondents R in the sample s based on actual values, $\bar{y}_s^A$ is the mean for the full sample s (including respondents and nonrespondents) based on actual values, $\bar{Y}_U^c$ is the mean for the coverage population, and $\bar{Y}_U$ is the population mean. All coverage cases were contacted for the screener survey. Therefore, $\bar{y}_s^A$ is equal to $\bar{Y}_U^c$ is the formula. The total error is also computed after all adjustments. The adjustment procedures are explained in the next section. As discussed in Kreuter, Presser, and Tourangeau (2008), there were a lot of false positive and false negative in the reports from the respondents, and the false positive and false negative rates were different for different items. We did not consider this measurement variance in the analysis.

4.3.3 Adjustment Procedure

The sample members with a telephone number available were contacted for a screener and then assigned to one of the three modes for the main survey. Therefore, there were two types of nonresponse—screener and main interview nonresponse. Response at the screener stage was modeled using call records and frame variables as candidate independent variables. Auxiliary variables with missing values were imputed using the multivariate sequential regression imputation method (Raghunathan, Lepkowski, Van Hoewyk and Solenberger 2001). Only one imputed dataset was created. The weights were estimated by the inverse of response propensities. The response propensity model was constructed using stepwise selection with the standard 0.05 significance level. The final model is shown in Table 4.5.

Table 4.5. Logistic regression coefficients from the response propensity model

	Coefficient	SE
Intercept	-2.823	0.239
State	-0.185	0.060
Age	-0.055	0.007
Number of call attempts	-0.064	0.035
Age*number of call attempts	0.004	0.001

Note: Coefficients based on 7,591 cases for whom screeners were attempted.

A set of nonresponse weights was estimated from this model. Because none of the auxiliary variables was predictive of the nonresponse status to the main surveys, we inflated the weights uniformly in each mode by a factor of number of assigned cases to number of respondents. Finally, the inflated weights were calibrated to the population totals on sex and age, and the calibration weights were used for the analyses. The total error after adjustments were computed in the way as defined in Equation 4.7, but $\operatorname{var}\left(\bar{y}_{sR}^{0}\right)$ and $\bar{y}_{sR}^{0}$ were estimated with the final weights.

Figure 4.3 shows the proportion who ever withdrew from a class, the proportion who donated in the last year, the percent of respondents who are male, and the mean age by weight quintile for the telephone survey. We chose to show the proportion who ever withdrew from a class because the relative total error in this estimate was largest among the undesirable characteristics. We had the same reason for showing the proportion who donated in the last year, which had the largest relative total error in the estimate among the desirable characteristics. As before, the weights increase from left to right within each variable. For each group of respondents, the proportion who ever withdrew from a class is indicated by a diamond and the confidence interval by the vertical line. As Figure 4.3 shows, the proportion of respondents who are male is higher in the lowest three weight quintiles higher than in the highest two quintiles. As a result, the weighted estimate will show a lower percent of male than the unweighted estimate. The same can be said about mean age—the weighted estimate is expected to have a lower mean age than the unweighted estimates. There are no clear patterns for the other two variables, which means that the weights will not affect those estimates very much. The patterns for the Web survey and the IVR survey are not discussed in detail here. The results are shown in Figure A1 and Figure A2, for the Web survey and the IVR survey, respectively.

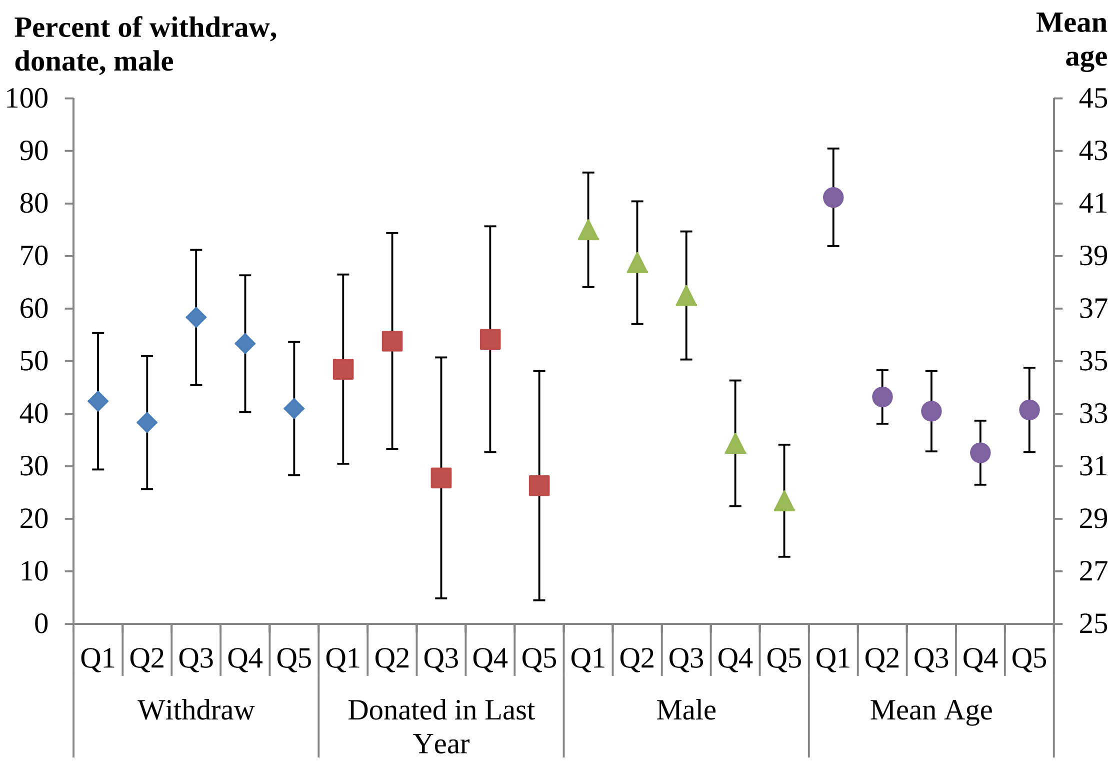

Figure 4.3. Proportion who ever withdrew from a class, proportion who donated in last year, percent of male respondents, and mean age by quintile of weights, telephone survey

4.3.4 Comparing the Survey Errors

Table 4.6 shows the error estimates for the telephone survey (for the point estimates themselves, see Table A2). As the table shows, there are few coverage biases in the estimates for undesirable characteristics and the demographic variables. However, both nonresponse bias and measurement bias generally show large negative biases for the undesirable characteristics and large positive biases for the desirable characteristics. That is, cases with positive characteristics were more likely to respond and the measurement biases tend to go in the expected directions, with overreporting of positive characteristics and underreporting of negative ones. One exception involves the estimated proportion ever receiving an academic warning or placed on academic probation, which shows a large positive measurement bias. The reason for this may be that this characteristic was quite rare. The estimated proportion who ever received a failing grade and who withdrew from a class show large negative measurement biases: -18.8 percentage points for the former and -21.3 percentage points for the latter. On average, absolute nonresponse bias (2.5) is 12.5 times larger than coverage bias (0.2 percentage point) in magnitude, and absolute measurement bias 14.2 percentage points) is 5.7 times larger than nonresponse bias. The cumulative negative biases due to nonresponse and measurement lead to large total biases for the estimates involving undesirable characteristics, and the adjustments basically do not change the total bias. On the other hand, the biases for the estimates of the desirable characteristics due to coverage, nonresponse and measurement are positive. There is a large coverage bias and a large nonresponse bias in the estimated proportion who ever donated to the University of Maryland. The overall amount of bias is 14.1 and 15.0 percentage points, which represents a relbias of 127.4% and 59.7% for coverage and nonresponse, respectively. However, the measurement bias in this estimate is relatively small, only 1.9 percentage points. The average nonresponse bias (7.2 percentage points) is 1.8 times larger than the average measurement bias (3.9 percentage points) for the desirable characteristics, and the average measurement bias is smaller than the average coverage bias (4.9 percentage points) for these five estimates. The findings on nonresponse bias and measurement bias are consistent with those of the analysis in Sakshaug, Yan, and Tourangeau (2010).

The adjustments generally help reduce the biases, but not much. As in Study 1, the biases in the estimates for the demographic variables are small, with nonresponse biases being the largest of the three types of bias, and sampling errors are not trivial any more. Measurement biases for estimates of the demographic variables are essentially 0. The adjustments effectively reduced the biases in the estimates of the demographic variables by about50 percent. The two demographic variables were used in the adjustment procedure.

The results for the Web survey and the IVR survey are similar to those from the telephone survey and are not discussed in detail here. One obvious difference was that nonresponse biases in the estimates of undesirable characteristics were much smaller in the Web survey than in the telephone and IVR survey. Table A5 and Table A6 show the corresponding error estimates for the Web survey and the IVR survey, respectively (for the point estimates themselves, see Table A3 and Table A4).

Table 4.6. True status in percent, coverage bias, nonresponse bias, measurement bias, sampling error, total error, and total error after adjustments, telephone survey

Variable	Frame value	Coverage bias	Nonresponse bias	Measurement bias	Sampling error	Total error	Total error (adjusted)	% reduction from the adjustments
Undesirable characteristics								
GPA below 2.5	15.6	-0.3	-4.6	-8.2	0.9	13.2	13.0	-1.7
F or D	63.0	-0.3	-1.6	-18.8	2.8	21.0	20.8	-1.0
Withdraw	70.7	0.1	-2.8	-21.3	2.9	24.2	23.9	-1.2
Probation	2.7	0.0	-0.7	8.3	1.7	7.7	7.8	1.9
Average (absolute)	*38.0*	*0.2*	*2.5*	*14.2*	*2.1*	*16.5*	*16.4*	-0.9
Desirable characteristics								
GPA above 3.5	17.5	1.1	3.6	1.4	2.5	6.6	7.2	7.7
Honors	8.9	0.5	3.0	3.9	2.1	7.7	7.6	-0.9
Donated	11.1	14.1	15.0	1.9	2.8	31.1	30.0	-3.6
Donated in last year	3.7	4.7	5.6	3.7	2.2	14.2	13.1	-7.5
Member	3.1	3.9	9.0	8.7	2.5	21.8	21.2	-2.9
Average (absolute)	*8.9*	*4.9*	*7.2*	*3.9*	*2.4*	*16.3*	*15.8*	-2.9
Demographics								
Male	48.3	2.6	1.9	0.0	2.8	5.3	2.9	-45.9
Mean age	33.9	-0.5	1.1	0.1	0.4	0.7	0.4	-51.1

Note: Coverage bias was estimated by the difference between frame values for cases with and without a telephone number.
Nonresponse bias was estimated by the difference between frame values for the samples and frame values for the respondents.
Measurement bias was estimated by the difference between reported and frame values for the respondents.
Adjustment bias was estimated by the difference between unweighted and weighted reported values for the respondents.
Total error was estimated by the sum of total absolute bias and sampling error.
Total error after adjustments was estimated by the sum of total absolute bias after adjustments and sampling error.

Table 4.7 presents absolute average errors in the estimates for the undesirable and desirable characteristics by interview mode. Averaging across the three interview modes, sampling error and coverage bias contribute the smallest amount of error to the total error—14.3 percent for sampling error. Coverage error contributes the second smallest amount, 16.2 percent. Nonresponse bias accounts for 31.2 percent of the total error. Measurement bias alone makes up more than a half of the total error, 53.6 percent to be exact. The relative magnitude of each type of error is similar in each of the three interview modes. This is a surprise given that the existing literature suggests that the three interviewing modes have different performance on measurement bias for sensitive questions (see Tourangeau and Yan 2007, for a review). On average, the adjustments resulted in a 2.4 percent reduction in the total error.

Table 4.7. Average absolute coverage bias, nonresponse bias, measurement bias, sampling error, total error, and total error after adjustments

	Coverage bias	Nonresponse bias	Measurement bias	Sampling error	Total error	Total error (adjusted)	% reduction from the adjustments
CATI	2.5	4.9	9.0	2.3	16.4	16.1	-1.9
WEB	2.5	4.1	8.1	2.1	14.5	14.2	-1.8
IVR	2.5	5.9	7.9	2.3	16.0	15.4	-3.4
Average	*2.5*	*4.9*	*8.4*	*2.2*	*15.6*	*15.3*	*-2.4*

4.4 General Discussion

We examined total error in estimates from two surveys, in which the key survey variables are potentially sensitive characteristics—whether people voted and their academic performance as undergraduates. In both studies, measurement biases tend to swamp all other forms of error for the undesirable characteristics, such as failing to vote or getting an unsatisfactory grade as an undergraduate. In both studies, nonresponse tended to produce next-largest errors (although for desirable characteristics in Study 2, nonresponse seems to produce the largest errors). This pattern—with measurement biases and nonresponse biases producing the largest problems—was apparent for both modes of data collection in Study 1 and for all three in Study 2 (see Table 4.7). In both studies, we also examined estimates for two demographic variables. For these variables, the overall errors are much smaller and measurement bias is no longer the main source of error. In Study 2, although estimates of undesirable characteristics were higher in the Web survey than the telephone and IVR surveys as found in Kreuter, Presser, and Tourangeau (2008), nonresponse was the main contributor to this difference. The magnitudes of measurement biases in the estimates of undesirable characteristics were similar under the three modes.

Nonresponse bias is bigger than sampling error and coverage bias in Study 2, although the coverage phenomenon is different from what we usually talk about. In common language in survey research, telephone coverage refers to the population with telephone service; while in Study 2, we refer to the members with telephone information in the database. In a study of the National Intimate Partner and Sexual Violence Survey (NISVS) pilot study, a national RDD survey of adults, Peytchev, Carley-Baxter, and Black (2011) found larger coverage biases than nonresponse biases, although they did not have a "gold standard" for the comparison. Bias is the dominating factor in the total survey error, which means the efforts we see in practice to reduce bias are worth it; however, it is difficult to reach this goal by weighting adjustments. Alternative methods or more effective weighting methods are needed.

Results from the analysis of the two datasets show that nonresponse bias is relatively smaller, compared to measurement bias. This is true for undesirable characteristics. However, for desirable characteristics in Study 2, nonresponse bias is larger than measurement bias. This point— measurement bias tended to be larger for estimates of undesirable characteristics and nonresponse bias tended to be larger for estimates of desirable characteristics—is also emphasized in Sakshaug, Yan, and Tourangeau (2010). For desirable behavior in Study 1, even though questions about voting are at the low end of the Bradburn et al.'s (1979) acute anxiety scale, and not as an uneasy topic as compared to other sensitive questions such as bankruptcy, we found measurement bias in the estimates of voting doubled the size of nonresponse bias. The same conclusion is reached in a previous analysis of the data (Tourangeau, Groves, and Redline 2010).

We see in this analysis that the nonresponse biases are the main component of the total errors in the estimates of the demographic variables. We suspect that nonresponse bias is probably the biggest type of error when it comes to neutral factual variables. Because neutral and factual questions are commonly seen in surveys, we should pay more attention to potential nonresponse errors and efforts to reduce nonresponse errors should be encouraged.

Many studies have found that respondents tend to underreport socially undesirable behaviors (see Tourangeau, Rips, and Rasinski 2000, chap. 9, for a review). In our analysis, measurement biases in the estimates of the undesirable characteristics are much bigger than other types of errors, and because weighting methods are not meant to correct this type of error, we need to devote more resources to develop a valid measurement for them. It may be also help to implement some of the techniques for eliciting sensitive information reviewed in Tourangeau and Yan (2007). Previous studies have shown that the technique of changing the mode of survey administration can be employed to encourage social undesirable reporting (e.g., Tourangeau and Smith 1996). Switching interviewing mode from interviewer-mediated (telephone) to self-administered (Web or IVR) seems to help a little bit in Study 2, but the reductions were limited.

Measure errors in the estimates of the voting items are also very large, although the questions used in the survey already used an alternative version with a softened tone. Rather than asking directly whether the respondent voted or not, the surveys asked whether something came up that kept the respondent from voting. However, this did not help much. Offering a forgiving preamble to the question about voting was also proved to not have any effect by Abelson, Loftus, and Greenwald (1992). Presser (1990) varied the prior questions to set up different contexts for the voting question, but did not find the method made a difference. More research on creative techniques to avoid the large social desirability effect on voting are needed.

Sampling error accounts for a small proportion of total error. However, when the sample size is small or when we are conducting subgroup analysis, sampling error may play a more important role.

There are some limitations in this analysis. The target populations for both datasets examined here are not the general U.S. population, but only a fraction of it. Therefore, the results may not apply to the general population. The key survey variables examined here were potentially sensitive characteristics that were subject to social desirability effects. Therefore, the results should not be generalized to the non-sensitive questions without further investigation.

Chapter 5: Conclusions

5.1 Summary

This dissertation was motivated by concern about declining response rates and the increased risks these pose for nonresponse bias. It is difficult to estimate nonresponse bias in practice. However, the studies presented here were able to estimate nonresponse bias in some key survey variables because they used two datasets that included records data for all sample members. Taking advantage of this, the studies assessed the effectiveness of two commonly used weighting methods for correcting nonresponse errors, examined the performance of the R-indicators for predicting bias in survey estimates, and evaluated the importance of nonresponse error in a total survey error context. Given the potential problem of nonresponse error, how effective are the remedies? Can we effectively monitor the quality of the responding sample? How much effort should be devoted to addressing this potential error? The conclusions are as follows.

First, many weighting methods have been developed and used in nonresponse adjustments. Chapter Two examined two common model-based postsurvey weighting strategies—response propensity weighting and GREG weighting. The results showed that both response propensity weighting and GREG weighting can lead to bias reduction, but the reductions are limited in the data sets used here. Under the same model, the size of the reductions is similar under the two weighting methods, as well as the variation in the weights produced from the two weighting methods. The comparison between response propensity weighting and GREG weighting shows that with the same set of auxiliary variables, the choice between response propensity weighting and GREG weighting makes little difference. When there are "informative" variables that are highly correlated with the outcome variables, both weighting methods are powerful in bias reduction. However, when there is a limited set of auxiliary variables, little to no gain can be achieved. In this situation, neither response propensity weighting nor GREG weighting is effective in reducing bias or mean squared error. In summary, weighting is only useful when there is a sizable set of auxiliary information available and these variables correlate with the outcome variables.

Chapter Three proposed a modified R-indicator, which is labeled "the penalized R-indicator," since the indicator penalizes for low response levels. Chapter Three assessed the effectiveness of the R-indicators in predicting biases in two settings. First it used two datasets with records available to evaluate the performance the R-indicators as call attempts increased (and more respondents were brought into the respondent pool). In these analyses, relative bias in the estimates decreased as the number of call attempts increased. The R-indicator, which takes higher values when the sample is more representative, decreased as the relative bias got smaller. In contrast, the penalized R-indicator increased as the relative bias went down. The results suggest that the penalized R-indicators show patterns that correspond more closely with the bias in estimated means for the survey variables than the R-indicators. Next, Chapter Three reported a simulation that compared the performance of the R-indicator and the penalized R-indicator. The simulation study shows that the penalized R-indicator had a better correlation with bias in estimates than the R-indicator.

Chapter Two and Chapter Three assessed how well we can cope with potential nonresponse bias. Chapter Four turned to the question of how much effort should be put into dealing with the problem. It examined total error in estimates from the two datasets. In the analyses of both datasets, measurement biases tended to swamp all other forms of error for the undesirable characteristics, such as failing to vote or getting an unsatisfactory grade as an undergraduate. In both analyses, nonresponse tended to produce the next-largest error. This pattern—with measurement biases and nonresponse biases producing the largest problems—was apparent for all modes of data collection in the two studies. In the analyses of both data sets, we also examined estimates for two demographic variables. For these variables, the overall errors are much smaller and measurement bias is no longer the main source of error.

To conclude, this dissertation demonstrated that nonresponse error is an important source of error in sample surveys. Nonresponse bias and measurement bias produce larger problems than other sources of error, such as coverage and sampling. Efforts put into dealing with nonresponse error are warranted. The effectiveness of weighting adjustments for nonresponse depends on the availability and quality of the auxiliary variables. The penalized R-indicator may be more helpful in monitoring the quality of the survey than the R-indicator.

5.2 Limitations

5.2.1 Limitation of the Datasets

The target populations for both datasets examined here are not the general U.S. population, but only a fraction of it. Therefore, the results may not apply to the general population. The datasets were not specifically collected for this research, and thus have imposed some restrictions on the analyses.

The Maryland registered voters dataset comes from a survey of registered voters. Whether the results from analyses of this dataset can be generalized to the general population is unknown. Voters are likely to be overrepresented in many surveys, suggesting this limitation may not be all that serious. However, the survey variables examined here involve voting behaviors. Some of the results probably cannot be generalized to non-sensitive factual questions. Moreover, the frame values in the dataset may not be perfect and may be subject to errors themselves, although the responses from the respondents on some demographic variables (e.g. age and sex) are close to identical to the values on the frame.

The University of Maryland alumni dataset comes from a survey of a more specific population—the University of Maryland undergraduate degree recipients from 1989 to 2002. The results from the analysis of this dataset may not be applicable to other populations. As in the Maryland registered voters dataset, the survey variables are potentially sensitive questions that are subject to social desirability effects, which makes it difficult to generalize the results to non-sensitive questions. Although the academic records about the graduates should be accurate, there may be some errors in the variables involving the graduate's relationship with the University. However, neither can be assessed using an independent source.

5.2.2 Limitation of the Research Methods

The two weighting methods examined in Chapter Two used explicit models. As with any models, these models have their limitations. In addition, the available auxiliary variables were limited. Some key auxiliary variables for effective nonresponse adjustments are certainly missing in the sets of variables available in the two datasets. Second, even if we have the right set of the auxiliary variables, which is unlikely, there is a risk of model misspecification. The uncertainty about model specification is not considered in the research.

The empirical studies in Chapter Three also used explicit models. They are subject to the same limitations as the models examined in Chapter Two. The empirical studies evaluated the performance of the R-indicator and the penalized R-indicator by call attempts. Because the final response rates for the studies were just above 30%, we do not know how the R-indicators might have performed had the response rate gone up (moved closer to 100%).

In Chapter Four, we examined nonresponse in the total error context. The key survey variables were potentially sensitive characteristics that were subject to social desirability effects. Therefore, the results should not be generalized to the non-sensitive questions without further investigation. In addition, the target populations for both datasets were not the general U.S. population and this may affect the relative sizes of the different types of errors.

5.3 Future Research

Response propensity weighting and GREG weighting as methods of nonresponse adjustment have many appealing features for survey practitioners. The two weighting methods assessed here can incorporate both continuous variables and interaction terms in the models. Future research should assess the effectiveness of these weighting methods using data with non-sensitive survey questions. Ideally, the target population should be the general population and records should be available for all sample members. More research is needed to identify useful auxiliary variables for nonresponse weighting. As discussed in Chapter Two, it is important to collect a rich set of auxiliary variables and equally important that the auxiliary variables correlate both with the response probability and the survey variables. Research on the consequences of misspecification of the response propensity model is also needed.

We compared nonresponse bias in the estimates of potentially sensitive characteristics to other sources of survey errors. Future research should explore this relationship using data containing records on non-sensitive characteristics. It will be more helpful if the cost of error reduction for each source of errors is also considered.

More empirical research is needed to provide guidance on using the R-indicators and the penalized R-indicators. The selection of auxiliary variables for the response propensity model is open for discussion. Finally, more theoretical research is needed to explore the properties of the penalized R-indictors.

5.4 Final Remarks

As nonresponse rates continue to climb, the demand for efforts to reduce potential nonresponse errors will become increasingly strident. This research is only a small part of a larger effort to develop a better set of tools for the problem. It may not be possible to find a perfect solution, but that may be a good reason to continue working on it.

Appendix A

This appendix contains some additional tables and figures for the preceding analyses. All tables and figures listed in this appendix are based on the University of Maryland alumni dataset. Table A1 lists the final dispositions for all cases. Tables A2-A4 show the estimates at different stages for the telephone, Web, and IVR surveys, respectively. Table A5 and A6 show the errors for the Web and IVR surveys, respectively. Figures A1 and A2 show the proportion who ever withdrew from a class, the proportion who donated in the last year, the percent of respondents who are males, and the mean age by weight quintiles, for the Web and IVR surveys, respectively.

Table A1. Final disposition codes, the University of Maryland alumni dataset

	Total	All Cases (%)	Screener (%)	Mode (%)
Seemingly usable phone numbers fielded	7,591	100	–	–
Not eligible and deceased	2,889	38.1	–	–
Eligible cases and unknown eligibility	4,702	61.9	100	–
Unknown eligibility	1,914	–	40.7	–
Eligible, no-interview				
Language barrier	33	–	0.7	–
Physically/mentally unable	7	–	0.1	–
Noncontact	797	–	17.0	–
Refusal	441	–	9.4	–
Partial screener completion	9	–	0.2	–
Screener completed and assigned to mode	1,501	–	31.9	
Initially assigned to CATI	338	–	–	100
Completes in CATI	320	–	–	94.7
Initially assigned to Web	639	–	–	100
Completes in Web	363	–	–	56.8
Initially assigned to IVR	524	–	–	100
Completes in IVR	320	–	–	61.1

Table A2. Proportions at different stages by item, telephone survey

Variable	Frame value	Coverage mean	Respondent mean	Respondent mean (reported)	Adjusted respondent mean (reported)
N	17,266	7,591	320§	320§	320§
Undesirable characteristics					
GPA below 2.5	15.6	15.4 (0.4)	10.8 (1.9)	2.5 (0.9)	2.7 (1.1)
F or D	63.0	62.6 (0.6)	61.0 (2.8)	42.2 (2.8)	42.4 (2.9)
Withdraw	70.7	70.8 (0.5)	68.0 (2.7)	46.7 (2.9)	47.0 (3.0)
Probation	2.7	2.6 (0.2)	1.9 (0.8)	10.2 (1.7)	10.3 (1.8)
Desirable characteristics					
GPA above 3.5	17.5	18.6 (0.4)	22.2 (2.5)	23.7 (2.5)	24.2 (2.7)
Honors	8.9	9.4 (0.3)	12.4 (1.9)	16.3 (2.1)	16.2 (2.2)
Donated	11.1	25.1 (0.5)	40.1 (2.8)	42.1 (2.8)	40.9 (2.9)
Donated in last year	3.7	8.5 (0.3)	14.0 (2.0)	17.7 (2.2)	16.7 (2.2)
Member	3.1	7.0 (0.3)	16.1 (2.1)	24.8 (2.5)	24.1 (2.5)
Demographics					
Male	48.3	50.9 (0.6)	52.8 (2.8)	52.8 (2.8)	48.3 (2.9)
Mean age	33.9	33.4 (0.1)	34.5 (0.4)	34.5 (0.4)	33.9 (0.4)

Note: § N varies because of item nonresponse.

The statistics in the last two columns were estimated with respondent reported values, and others with frame values.

The statistics in the last column were estimated with weights.

Numbers in parentheses are standard errors.

Table A3. Proportions at different stages by item, Web survey

Variable	Frame value	Coverage mean	Respondent mean	Respondent mean (reported)	Adjusted respondent mean (reported)
N	17,266	7,591	363§	363§	363§
Undesirable characteristics					
GPA below 2.5	15.6	15.4 (0.4)	14.6 (1.9)	6.2 (1.3)	6.0 (1.3)
F or D	63.0	62.6 (0.6)	62.3 (2.5)	50.7 (2.6)	50.3 (2.7)
Withdraw	70.7	70.8 (0.5)	70.7 (2.4)	50.6 (2.6)	50.7 (2.7)
Probation	2.7	2.6 (0.2)	2.2 (0.8)	13.8 (1.8)	14.1 (1.9)
Desirable characteristics					
GPA above 3.5	17.5	18.6 (0.4)	20.8 (2.2)	24.2 (2.3)	23.9 (2.3)
Honors	8.9	9.4 (0.3)	9.9 (1.6)	15.5 (1.9)	15.7 (2.0)
Donated	11.1	25.1 (0.5)	42.7 (2.6)	41.3 (2.6)	40.1 (2.7)
Donated in last year	3.7	8.5 (0.3)	16.4 (2.0)	16.7 (2.0)	15.6 (1.9)
Member	3.1	7.0 (0.3)	17.5 (2.0)	23.6 (2.2)	22.1 (2.2)
Demographics					
Male	48.3	50.9 (0.6)	51.8 (2.6)	51.2 (2.6)	47.8 (2.7)
Mean age	33.9	33.4 (0.1)	34.6 (0.4)	34.7 (0.4)	34.0 (0.4)

Note: § N varies because of item nonresponse.
The statistics in the last two columns were estimated with respondent reported values, and others with frame values.
The statistics in the last column were estimated with weights.
Numbers in parentheses are standard errors.

Table A4. Proportions at different stages by item, IVR survey

Variable	Frame value	Coverage mean	Respondent mean	Respondent mean (reported)	Adjusted respondent mean (reported)
N	17,266	7,591	320§	320§	320§
Undesirable characteristics					
GPA below 2.5	15.6	15.4 (0.4)	9.6 (1.8)	3.7 (1.2)	4.2 (1.3)
F or D	63.0	62.6 (0.6)	58.6 (2.8)	44.3 (2.8)	44.5 (2.9)
Withdraw	70.7	70.8 (0.5)	64.4 (2.7)	45.6 (2.8)	46.9 (2.9)
Probation	2.7	2.6 (0.2)	2.9 (0.9)	13.4 (1.9)	13.6 (2.0)
Desirable characteristics					
GPA above 3.5	17.5	18.6 (0.4)	23.3 (2.6)	20.4 (2.5)	19.8 (2.5)
Honors	8.9	9.4 (0.3)	15.1 (2.0)	19.9 (2.3)	18.5 (2.2)
Donated	11.1	25.1 (0.5)	38.6 (2.8)	40.5 (2.8)	40.5 (2.9)
Donated in last year	3.7	8.5 (0.3)	15.7 (2.1)	16.4 (2.1)	15.3 (2.1)
Member	3.1	7.0 (0.3)	14.4 (2.0)	21.5 (2.3)	20.9 (2.3)
Demographics					
Male	48.3	50.9 (0.6)	47.8 (2.8)	47.2 (2.8)	47.6 (2.9)
Mean age	33.9	33.4 (0.1)	34.5 (0.5)	34.9 (0.5)	34.1 (0.4)

Note: § N varies because of item nonresponse.
The statistics in the last two columns were estimated with respondent reported values, and others with frame values.
The statistics in the last column were estimated with weights.
Numbers in parentheses are standard errors.

Table A5. Status in percent, coverage bias, nonresponse bias, measurement bias, sampling error, total error, and total error after adjustments, Web survey

Variable	Frame value	Coverage bias	Nonresponse bias	Measurement bias	Sampling error	Total error	Total error (adjusted)	% reduction from the adjustments
Undesirable characteristics								
GPA below 2.5	15.6	-0.3	-0.8	-8.4	1.3	9.6	9.7	1.5
F or D	63.0	-0.3	-0.4	-11.6	2.6	12.6	13.0	3.4
Withdraw	70.7	0.1	-0.1	-20.2	2.6	20.3	20.2	-0.6
Probation	2.7	0.0	-0.4	11.6	1.8	11.3	11.6	2.7
Average	*38.0*	*0.2*	*0.4*	*12.9*	*2.1*	*13.4*	*13.6*	1.4
Desirable characteristics								
GPA above 3.5	17.5	1.1	2.2	3.4	2.3	7.0	6.8	-2.5
Honors	8.9	0.5	0.6	5.5	1.9	6.9	7.1	3.4
Donated	11.1	14.1	17.5	-1.4	2.6	30.3	29.2	-3.7
Donated in last year	3.7	4.7	7.9	0.3	2.0	13.1	12.1	-7.8
Member	3.1	3.9	10.5	6.1	2.2	20.6	19.2	-7.2
Average	*8.9*	*4.9*	*7.7*	*3.3*	*2.2*	*15.6*	*14.9*	-4.6
Demographics								
Male	48.3	2.6	0.9	-0.6	2.6	3.9	2.7	-30.2
Mean age	33.9	-0.5	1.1	0.1	0.4	0.9	0.4	-56.5

Note: Coverage bias was estimated by the difference between frame values for cases with and without a telephone number.
Nonresponse bias was estimated by the difference between frame values for the samples and frame values for the respondents.
Measurement bias was estimated by the difference between reported and frame values for the respondents.
Adjustment bias was estimated by the difference between unweighted and weighted reported values for the respondents.
Total error was estimated by the sum of total absolute bias and sampling error.
Total error after adjustments was estimated by the sum of total absolute bias after adjustments and sampling error.

Table A6. Status in percent, coverage bias, nonresponse bias, measurement bias, sampling error, total error, and total error after adjustments, IVR survey

Variable	Frame value	Coverage bias	Nonresponse bias	Measurement bias	Sampling error	Total error[‡]	Total error (adjusted)	% reduction from the adjustments
Undesirable characteristics								
GPA below 2.5	15.6	-0.3	-5.8	-5.9	1.2	12.0	11.6	-3.6
F or D	63.0	-0.3	-4.0	-14.3	2.8	18.9	18.7	-1.1
Withdraw	70.7	0.1	-6.4	-18.8	2.8	25.3	24.0	-4.8
Probation	2.7	0.0	0.2	10.5	1.9	10.9	11.1	2.2
Average (absolute)	*38.0*	*0.2*	*4.1*	*12.4*	*2.2*	*16.8*	*16.4*	-2.4
Desirable characteristics								
GPA above 3.5	17.5	1.1	4.7	-3.0	2.5	3.8	3.4	-10.7
Honors	8.9	0.5	5.7	4.8	2.3	11.3	9.9	-12.5
Donated	11.1	14.1	13.4	2.0	2.8	29.6	29.6	-0.2
Donated in last year	3.7	4.7	7.2	0.7	2.1	12.8	11.8	-8.3
Member	3.1	3.9	7.4	7.1	2.3	18.5	18.0	-3.1
Average (absolute)	*8.9*	*4.9*	*7.7*	*3.5*	*2.4*	*15.2*	*14.5*	-4.6
Demographics								
Male	48.3	2.6	-3.1	-0.6	2.8	3.0	3.0	-1.9
Mean age	33.9	-0.5	1.1	0.3	0.5	1.1	0.5	-53.9

Note: Coverage bias was estimated by the difference between frame values for cases with and without a telephone number.
Nonresponse bias was estimated by the difference between frame values for the samples and frame values for the respondents.
Measurement bias was estimated by the difference between reported and frame values for the respondents.
Adjustment bias was estimated by the difference between unweighted and weighted reported values for the respondents.
Total error was estimated by the sum of total absolute bias and sampling error.
Total error after adjustments was estimated by the sum of total absolute bias after adjustments and sampling error.

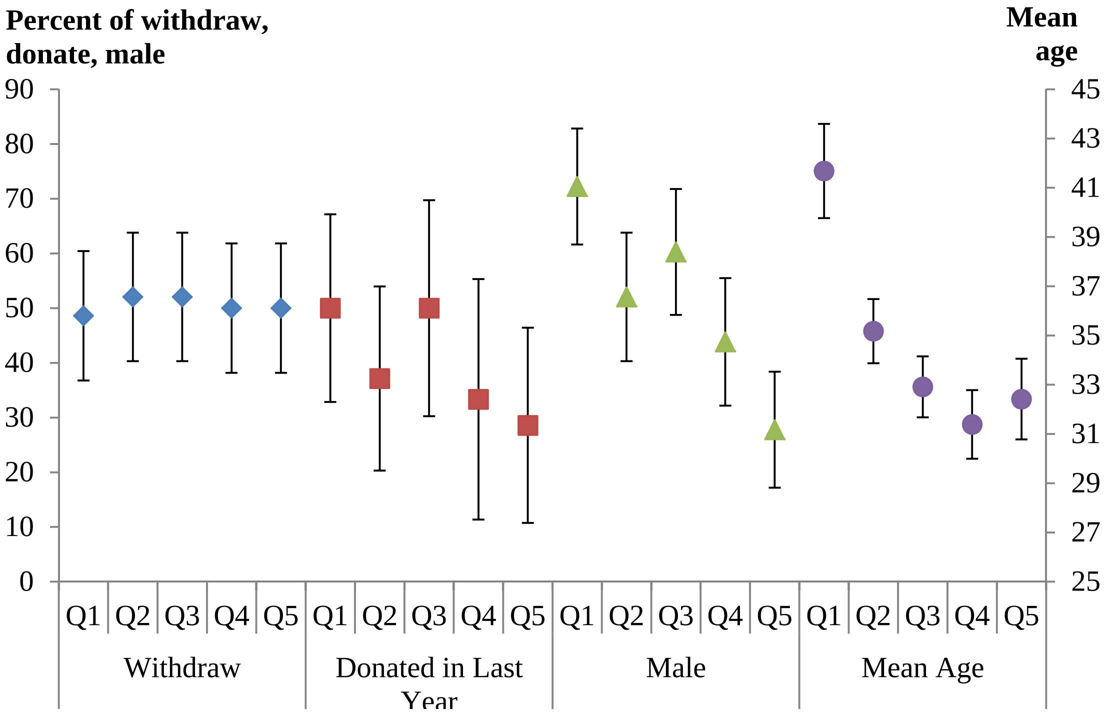

Figure A1. Proportion who ever withdrew from a class, proportion who donated in last year, percent of male respondents, and mean age by quintile of weights, Web survey

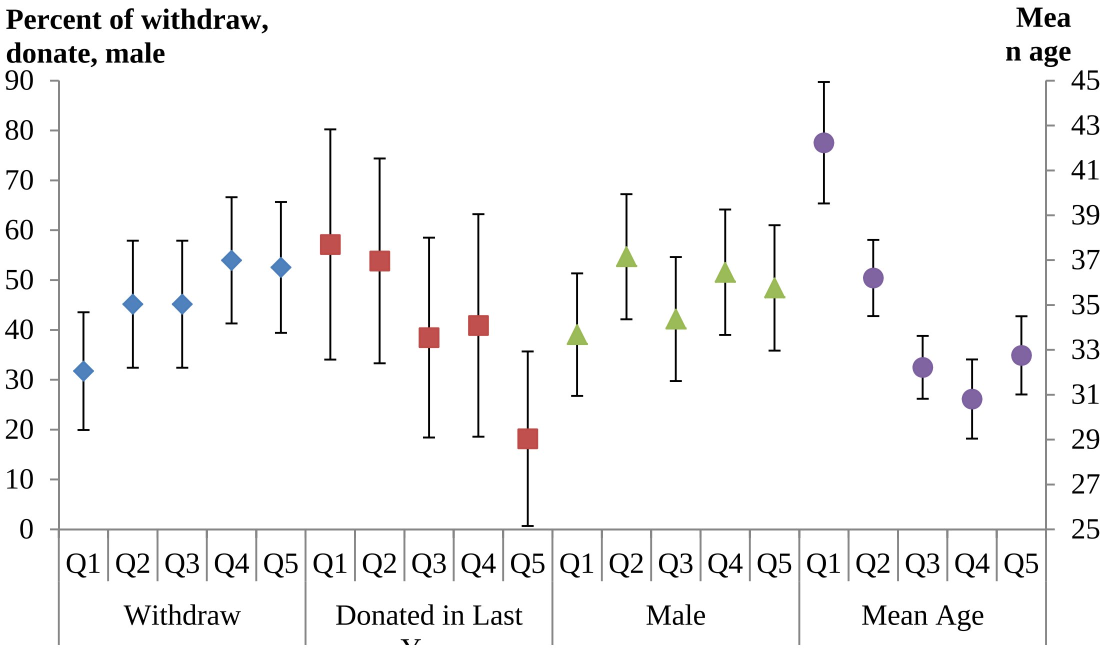

Figure A2. Proportion who ever withdrew from a class, proportion who donated in last year, percent of male respondents, and mean age by quintile of weights, IVR survey

Bibliography

Abelson, Robert P., Elizabeth F. Loftus, and Anthony G. Greenwald. 1992. "Attempts to Improve the Accuracy of Self-Reports of Voting." In *Questions about Questions*, ed. Judith M. Tanur, pp. 138–153. New York: Russell Sage Foundation.

Alreck, Pamela L., and Robert B. Settle. 1995. *The Survey Research Handbook*. New York: Mcgraw-Hill.

Bankier, Michael, Anne-Marie Houle, and Manchi Luc. 1997. "Calibration Estimation in the 1991 and 1996 Canadian Censuses." Proceedings of the Survey Research Methods Section, American Statistical Association: 66–75.

Bankier, Michael, and Darryl Janes. 2003. "Regression Estimation of the 2001 Canadian Census." Proceedings of the Survey Research Methods Section, American Statistical Association: 442–449.

Bankier, Michael, Stephen Rathwell, and Mark Majkowski. 1992. "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census." Proceedings of the Survey Research Methods Section, American Statistical Association: 764–769.

Bates, Nancy, James Dahlhamer, and Eleanor Singer. 2008. "Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse." *Journal of Official Statistics*, 24: 591–612.

Battaglia, Michael P., Donald J. Malec, Bruce D. Spencer, David C. Hoaglin, and Joseph Sedransk. 1995. "Adjusting for Noncoverage of Nontelephone Households in the National Immunization Survey." Proceedings of the Section on Survey Research Methods, American Statistical Association: 678–683.

Belli, Robert F., Michael W. Traugott, and Matthew N. Beckmann. 2001. "What Leads to Vote Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies." *Journal of Official Statistics,* 17: 479–498.

Bethlehem, Jelke. 2002. "Weighting Nonresponse Adjustments Based on Auxiliary Information." In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 275–287. New York: Wiley.

Bethlehem, Jelke, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Bethlehem, Jelke, and Barry Schouten. 2004. "Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS)." Discussion Paper 0230. Statistics Netherlands, Voorburg, The Netherlands.

Biemer, Paul P. 2001. “Nonresponse Bias and Measurement Bias in a Comparison of Face-to-face and Telephone Interviewing.” *Journal of Official Statistics*, 17: 295–320.

Biemer, Paul P., and Michael W. Link. 2008. “Evaluating and Modeling Early Cooperator Effects in RDD Surveys.” In *Advances in Telephone Survey Methodology*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, pp. 587–617. Hoboken, NJ: Wiley.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. New York: Wiley.

Bradburn, Norman M., Seymour Sudman, and Associates. 1979. *Improving Interview Method and Questionnaire Design*. San Francisco: Jossey-Bass.

Brick, J. Michael, and Michael E. Jones. 2008. “Propensity to Respond and Nonresponse Bias.” Metron—*International Journal of Statistics*, 66: 51–73.

Brick, J. Michael, Joseph Waksberg, and Scott Keeter. 1996. “Using Data on Interruptions in Telephone Service as Coverage Adjustments.” *Survey Methodology*, 22: 185–197.

Carlson, Barbara L., and Stephen Williams. 2001. “A Comparison of Two Methods to Adjust Weights for Nonresponse: Propensity Modeling and Weighting Class Adjustments.” Proceedings of the Survey Research Methods Section. American Statistical Association.

Chang, Ted, and Phillip S. Kott. 2008. “Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model.” *Biometrika*, 95: 557–571.

Cochran, William G. 1968. “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies.” *Biometrics*, 24: 205–213.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. “The Effects of Response Rate Changes on the Index of Consumer Sentiment.” *Public Opinion Quarterly*, 64: 413–428.

de Leeuw, Edith D. and Wim de Heer. 2002. “Trends in Household Survey Nonresponse: A Longitudinal and International Comparison.” In *Survey nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 41–54. New York, NY: John Wiley & Sons.

de Leeuw, Edith D., Joop Hox, Elly Korendijk, Gerty Lensvelt-Mulders, and Mario Callegaro. 2007. “The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis.” *Public Opinion Quarterly*, 71: 413–443.

Davern, Michael, James Lepkowski, Kathleen T. Call, Noreen Arnold, Tracy L. Johnson, Karen Goldsteen, April Todd-Malmlov, and Lynn A. Blewett. 2004. “Telephone

Service Interruption Weighting Adjustments for State Health Insurance Surveys." *Inquiry*, 41: 280–290.

Deming, W. Edwards, and Friderick F. Stephan. 1940. "On a Least Squares Adjustment of a Sample Frequency Table When the Expected Marginal Totals Are Known." *Annals of Mathematical Statistics*, 11: 427–444.

Deville, Jean-Claude, and Carl Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association*, 87:3 67–82.

Dillman, Don A. 2007. *Mail and Internet Surveys: The Tailored Design Method*, 2nd ed., 2007 update. Hoboken, NJ: Wiley.

Duncan, Kristin B., and Elizabeth A. Stasny. 2001. "Using Propensity Scores to Control Coverage Bias in Telephone Surveys." *Survey Methodology*, 27: 121–130.

Ekholm, Anders, and Seppo Laaksonen. 1991. "Weighting via Response Modeling in the Finnish Household Budget Survey." *Journal of Official Statistics*, 7: 325–377.

Fay, Robert E. 2005. "Model-Assisted Estimation for the American Community Survey." Proceedings of the Survey Research Methods Section, American Statistical Association: 3016–3023.

Fay, Robert E. 2006. "Using Administrative Records with Model-Assisted Estimation for the American Community Survey." Proceedings of the Survey Research Methods Section, American Statistical Association: 2995–3001.

Folsom, Ralph E. and Michael B. Witt. 1994. "Testing a New Attrition Nonresponse Adjustment Method for SIPP." Proceedings of the Section on Survey Research Methods, American Statistical Association: 428–433.

Folsom, RE, and Singh, AC. 2000. "The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification." Proceedings of the Survey Research Methods Section, American Statistical Association: 598–603.

Frankel, Martin R., K. P. Srinath, David C. Hoaglin, Michael P. Battaglia, Philip J. Smith, Robert A. Wright, and Meena Khare. 2003. "Adjustments for non-telephone bias in random-digit-dialing surveys." *Statistics in Medicine*, 22: 1611–1626.

Fuller, Wayne A, Marie M. Loughin, and Harold D. Baker. 1994. "Regression Weighting for the 1987-88 National Food Consumption Survey." *Survey Methodology*, 20: 75–85.

Garren, Steven T., and Ted C. Chang. 2002. "Improved Ratio Estimation in Telephone Surveys Adjustment for Noncoverage." *Survey Methodology*, 27: 63–76.

Groves Robert M. 1989. *Survey Costs and Survey Errors*. New York: Wiley.

Groves, Robert, J. Michael Brick, M. Couper, William D. Kalsbeek, Brian Harris-Kojetin, Frauke Kreuter, Beth-Ellen Pennell, Trivellore E. Raghunathan, Barry Schouten, Tom W. Smith, Roger Tourangeau, Ashley Bowers, Matt Jans, Courtney Kennedy, Rachel Levenstein, Kristen Olson, Emilia Peytcheva, Sonja Ziniel, and James Wagner. 2008. "Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools." *Survey Practice*: 14–22.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer and Roger Tourangeau. 2009. *Survey Methodology*, 2nd ed. Hoboken, NJ: Wiley.

Groves, Robert M., and Steven Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society Series A*, 169: 439–457.

Groves, Robert M. and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly*, 72: 167–189.

Groves, Robert M., Stanley Presser, and Sarah Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly,* 68: 2–31.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. "Leverage-Salience Theory of Survey Participation: Description and an Illustration." *Public Opinion Quarterly*, 64: 288–308.

Hansen, Morris H., and William N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys." *Journal of the American Statistical Association,* 41: 517–529.

Harris-Kojetin, Brian, and Clyde Tucker. 1999. "Exploring the Relation of Economic and Political Conditions with Refusal Rates to a Government Survey." *Journal of Official Statistics*, 15: 167–184.

Heberlein, Thomas A., and Robert Baumgartner. 1978. "Factors Affecting Response Rates to Mailed Questionnaires: A Quantitative Analysis of the Published Literature." *American Sociological Review*, 43: 447–462.

Hoaglin, David C., and Michael P. Battaglia. 1996. "A Comparison of Two Methods of Adjusting for Noncoverage of Nontelephone Households in A Telephone Survey." Proceedings of the Section on Survey Research Methods, American Statistical Association: 497–502.

Holt, D. and T. M. F. Smith. 1979. "Post-Stratification." *Journal of the Royal Statistical Society Series A*, 142: 33–46.

Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: Wiley

Hox, Joop J., and Edith D. de Leeuw. 1994. “A Comparison of Nonresponse in Mail, Telephone, and Face-to-face Surveys: Applying Multilevel Models to Meta-analysis.” *Quality and Quantity*, 28: 329–344.

Ireland, C. T. and Kullback, S. 1968. “Contingency Tables with Given Marginals.” *Biometrika*, 55: 179–188.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. “Weighting Methods.” *Journal of Official Statistics*, 19: 81–97.

Keeter, Scott, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. 2006. “Gauging the Impact of Growing Nonresponse on Estimates from a National RD Telephone Survey.” *Public Opinion Quarterly*, 70: 759–779.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. “Consequences of Reducing Nonresponse in a National Telephone Survey.” *Public Opinion Quarterly*, 64: 125–148.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley and Sons.

Kish, Leslie. 1987. *Statistical Design for Research*. New York: Wiley.

Kreuter, Frauke, Kristen M. Olson, James Wagner, Ting Yan, Trena M. Ezzati-Rice, Carolina Casas-Cordero, Michael Lemay, Andy Peytchev, Robert M. Groves, and Trivellore E. Raghunathan. 2010. “Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples From Multiple Surveys.” *Journal of the Royal Statistical Society, Series A*, 173: 389–407.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau. 2008. “Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity.” *Public Opinion Quarterly*, 72: 847–865.

Kreuter, Frauke, Ting Yan, and Roger Tourangeau. 2008. “Good Item or Bad—Can Latent Class Analysis Tell? The Utility of Latent Class Analysis for the Evaluation of Survey Questions.” *Journal of the Royal Statistical Society, Series A*, 171: 723–738.

Kulka, Richard, Nicholas Holt, Woody Carter, and Kathryn L. Dowd. 1991. “Self Reports of Time Pressures, Concerns for Privacy and Participation in the 1990 Mail Census.” In Proceedings of the Annual Research Conference. Washington, DC: Bureau of the Census.

Laflamme, François, and Milana Karaganis. 2010. “Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada.” Paper presented at the Symposium on Recent Advances in the Use of Paradata (Process Data) in Social Survey Research, London.

Lee, Sunghee, and Richard Valliant. 2007. "Weighting Telephone Samples Using Propensity Scores." In *Advances in Telephone Survey Methodology*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith D. de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link, and Roberta L. Sangster, pp. 170–183. New York: Wiley.

Lepkowski, James, Graham Kalton, and Daniel Kasprzyk. 1989. "Weighting Adjustments for Partial Nonresponse in the 1984 SIPP Panel." Proceedings of the Section on Survey Research Methods, American Statistical Association: 296–301.

Lessler, Judith T., and William D. Kalsbeek. 1992. *Nonsampling Error in Surveys*. New York: Wiley.

Lin, I-Fen, and Nora C. Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly*, 2: 236–258.

Link, Michael W., and Jennie Lai. 2011. "Cell-Phone-Only Households and Problems of Differential Nonresponse Using an Address-Based Sampling Design." *Public Opinion Quarterly*, 75: 613–635.

Little, Roderick J. A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review*, 54: 139–157.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd Ed. New York: Wiley & Sons.

Little, Roderick J. A., and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, 31: 161–168.

Locander, William, Seymour Sudman, and Norman Bradburn. 1976 "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association*, 71: 269–275.

Merkle, Daniel, and Murray Edelman. 2002. "Nonresponse in Exit Polls: A Comprehensive Analysis." In *Survey Nonresponse*, edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J. A. Little, pp. 243–257. New York: John Wiley & Sons.

Mohl, Chris, and François Laflamme. 2007. "Research and Responsive Design Options for Survey Data Collection at Statistics Canada." Proceedings of the Survey Research Methods Section, American Statistical Association: 2962–2968.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Methods: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society, Series A*, 97: 558–606.

Office of Management and Budget. 2006. Standards and Guidelines for Statistical Surveys. Available at http://www.whitehouse.gov/omb/inforeg/statpolicy/standards_stat_surveys.pdf [Last retrieved in October, 2012].

Oh, H. Lock, and Frederick J. Scheuren. 1983. “Weighting Adjustment for Unit Nonresponse.” In *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliographies*, edited by William G. Madow, Ingram Olkin, and Donald B. Rubin, pp. 143–184. New York: Academic Press.

Olson, Kristen. 2006. “Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias.” *Public Opinion Quarterly*,70: 737–758.

Olson, Kristen. Forthcoming. “Paradata in Nonresponse Adjustment.” *The ANNALS of the American Academy of Political and Social Science* (Special Issue: The Non-Response Challenge to Measurement in Social Science, Editors: Douglas S. Massey and Roger Tourangeau), Volume 645.

Parry, Hugh J., and Helen M. Crossley. 1950. “Validity of responses to survey questions.” *Public Opinion Quarterly*, 14: 61–80.

Peytchev, Andy, Lisa R. Carley-Baxter and Michele C. Black. 2011. “Multiple Sources of Nonobservation Error in Telephone Survey: Coverage and Nonresponse.” *Sociological Methods and Research*, 40: 138–168.

Peytchev, Andy, Emilia Peytcheva, and Roberts M. Groves. 2010. “Measurement Error, Unit Nonresponse and Self-reports of Abortion Experiences.” *Public Opinion Quarterly*,74: 319–327.

Peytchev, Andy, Sarah Riley, Jeffrey Rosen, Joe Murphy, and Mark Lindblad. 2010. “Reduction of Nonresponse Bias in Surveys through Case Prioritization.” *Survey Research Methods*, 4: 21–29.

Peytcheva, Emilia, and Robert M. Groves. 2009. “Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates.” *Journal of Official Statistics*, 25: 193–201.

Presser, Stanley. 1990. “Can Changes in Context Reduce Vote Overreporting in Surveys?” *Public Opinion Quarterly*, 54: 586–593.

Purdon, Susan, Pamela Campanelli, and Patrick Sturgis. 1999. “Interviewers’ Calling Strategies on Face-to-Face Interview Surveys.” *Journal of Official Statistics*,15: 199–219.

Raghunathan, Trivellore E, James M. Lepkowski, John Van Hoewyk, and Peter Solenberger . 2001. “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” *Survey Methodology*, 27: 85–95.

Rizzo, Lou, Graham Kalton, J. Michael Brick, and Rita Petroni. 1994. “Adjusting for Panel Nonresponse in the Survey of Income and Program Participation.” Proceedings of the Survey Research Methods Section, American Statistical Association: 422–427.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika*, 70: 41–55.

Sakshaug Joseph W., and Frauke Kreuter. 2011. “Using Paradata and Other Auxiliary Data to Examine Mode Switch Nonresponse in a ‘Recruit-and-Switch’ Telephone Survey.” *Journal of Official Statistics*, 27: 339–357.

Sakshaug, Joseph W., Ting Yan, and Roger Tourangeau. 2010. “Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-sensitive Items.” *Public Opinion Quarterly*,74: 907–933.

Särndal, Carl-Erik. 2011. “The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation.” *Journal of Official Statistics,* 27: 1–21.

Särndal, Carl-Erik, and Sixten Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: Wiley.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Schaeffer, Nora C., Judith A. Seltzer, and Marieka Klawitter. 1991. “Estimating Nonresponse and Response Bias: Resident and Nonresident Parents’ Reports about Child Support.” *Sociological Methods and Research*, 20: 30–59.

Schouten, Barry, Jelke Bethlehem, Koen Beullens, Øyvin Kleven, Geert Loosveldt, Annemieke Luiten, Katja Rutar, Natalie Shlomo, and Chris Skinner. 2012. “Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators.” *International Statistical Review*, 80: 1–18.

Schouten, Barry, Fannie Cobben, and Jelke G. Bethlehem. 2009. “Indicators for the Representativeness of Survey Response.” *Survey Methodology* 35: 101–113.

Schouten, Barry, Natalie Shlomo, and Chris Skinner. 2011. “Indicators for Monitoring and Improving Representativeness of Response.” *Journal of Official Statistics*, 27: 1–24.

Shlomo, Natalie, Chris Skinner, and Barry Schouten. 2012. “Estimation of an Indicator of the Representativeness of Survey Response.” *Journal of Statistical Planning and Inference,* 142: 201–211.

Siegel, Peter, James Chromy, and Elizabeth Copello. 2005. “Propensity Models versus Weighting Class Approaches to Nonresponse Adjustment: A Methodological

Comparison.” Proceedings of the Survey Research Methods Section, American Statistical Association: 3560–3565.

Singer, Eleanor, and Cong Ye. Forthcoming. “The Use and Effects of Incentives in Surveys." *The ANNALS of the American Academy of Political and Social Science* (Special Issue: The Non-Response Challenge to Measurement in Social Science, Editors: Douglas S. Massey and Roger Tourangeau), Volume 645.

Singleton Jr., Royce A., and Bruce C. Straits. 2005. *Approaches to Social Research*, 4th ed. New York: Oxford University Press.

Skinner, Chris. 1999. “Calibration Weighting and Non-sampling Errors.” *Research in Official Statistics*, 2: 33–43.

Smith, Philip J., J.N.K. Rao, Michael P. Battaglia, Trena M. Ezzati-Rice, Danni Daniels, and Meena Khare. 2001. “Compensating for Provider Nonresponse Using Response Propensities to Form Adjustment Cells: The National Immunization Survey.” *Vital and Health Statistics*, Series 2, No. 133.

Tourangeau, Roger. 2004. “Survey Research and Societal Change.” *Annual Review of Psychology*, 55: 775–801.

Tourangeau, Roger, Robert M. Groves, and Cleo Redline. 2010. “Sensitive Topics and Reluctant Respondents: Demonstrating a Link between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly*, 74: 413–432.

Tourangeau, Roger, Lance Rips, and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge, UK: Cambridge University Press.

Tourangeau Roger, Tom W. Smith. 1996. “Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context.” *Public Opinion Quarterly*, 60: 275–304.

Tourangeau, Roger, and Ting Yan. 2007. “Sensitive Questions in Surveys.” *Psychological Bulletin*, 133: 859–883.

Traugott, Michael W., and Kenneth Goldstein. 1993. “Evaluating Dual Frame Samples and Advance Letters as Means of Increasing Response Rates.” Proceedings of the Survey Research Methods Section, American Statistical Association: 1284–1286.

Traugott, Michael W., Robert M. Groves, and James M. Lepkowski. 1987. “Using Dual Frame Designs to Reduce Nonresponse in Telephone Surveys.” *Public Opinion Quarterly*, 51: 522–539.

Traugott, Michael W., and John P. Katosh. 1979. “Response Validity in Surveys of Voting Behavior.” *Public Opinion Quarterly*, 43: 359–377.

Valliant, Richard, and Jill A. Dever. 2011. “Estimating Propensity Adjustments for Volunteer Web Surveys.” *Sociological Methods Research*, 40: 105–137.

Wagner, James. 2012. “A Comparison of Alternative Indicators for the Risk of Nonresponse Bias.” *Public Opinion Quarterly* Advance Access published September 10, 2012.

Weeks, Michael F., Richard A. Kulka, and Stephanie A. Pierson. 1987. “Optimal Call Scheduling for a Telephone Survey.” *Public Opinion Quarterly,* 51: 540–549.

Williams, E. J. 1959. “The Comparison of Regression Variables.” *Journal of the Royal Statistical Society, Series B*, 21: 396–399.