

METHODS FOR ESTIMATING THE PROPENSITY SCORE WHEN WEIGHTING
VOLUNTEER WEB SAMPLES: A COMPARISON OF STRATEGIES FOR VARIABLE
CHOICE

A THESIS SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE MASTER OF SCIENCE

BY

RACHEL STENGER

DR. JOCELYN BOLIN – ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2018

ACKNOWLEDGEMENTS

I would like to thank my thesis committee for their support throughout this thesis process. First, a special thank you to my committee chair, Dr. Jocelyn Bolin, for her patience in reading numerous drafts of this document and for providing encouragement. I would also like to thank my other committee members, Dr. Holmes Finch and Dr. Chad Menning, for their feedback and support. This project would not have been completed without all of their expertise and encouragement from beginning to end. Thank you.

TABLE OF CONTENTS

| | |
|---|----|
| INTRODUCTION..... | 5 |
| The Growing Popularity of Web Surveys..... | 5 |
| Current Study..... | 9 |
| LITERATURE REVIEW..... | 11 |
| Bias in Web Survey Estimation..... | 11 |
| Coverage error..... | 11 |
| Sampling error..... | 13 |
| Non-response error..... | 13 |
| Measurement error..... | 15 |
| Approaches to Dealing with Error..... | 15 |
| Weighting..... | 16 |
| Background on Propensity Scores..... | 19 |
| Use of Propensity Scores in Survey Research..... | 20 |
| Estimating the propensity score..... | 21 |
| Reference survey..... | 22 |
| Methods for using the propensity score..... | 23 |
| Determining the best propensity score model..... | 25 |
| Effectiveness of propensity score weighting..... | 26 |
| Variable Selection..... | 29 |
| Variable selection methods in observational research..... | 30 |
| Webographic or attitudinal variables..... | 31 |
| Current Study..... | 37 |

| | |
|---|----|
| Hypotheses..... | 40 |
| METHODS..... | 41 |
| Dataset..... | 42 |
| Study Variables..... | 42 |
| Step 1: Creating the Pseudo Population and Defining the Population Value..... | 43 |
| Step 2: Creating the Volunteer and Reference Samples..... | 44 |
| Step 3: Estimating the Propensity Score..... | 45 |
| Step 4: Calculating the Weighted Volunteer Sample Estimate..... | 49 |
| Step 5: Determining the Best Model..... | 50 |
| RESULTS..... | 53 |
| Models..... | 53 |
| Religiosity Variable..... | 58 |
| Political Views Variable..... | 60 |
| DISCUSSION..... | 62 |
| Limitations and Suggestions for Future Research..... | 67 |
| REFERENCES..... | 69 |
| APPENDIX A: Full R code for project..... | 75 |

Introduction

The Growing Popularity of Web Surveys

Researchers want to know about characteristics of the American population for various reasons—for academic purposes, for budgeting purposes, for pure curiosity, and for reasons like allocating the number of representatives from each state. When researchers want to find out this information, they can attempt to survey the entire American population; this is what the Census Bureau does every ten years. However, censuses take a lot of time and money. Instead, researchers can take a sample of the American population and make estimates of what these characteristics are in the population based on the information that they gather. For survey research purposes, a sample is defined as “all units of the population that are drawn for inclusion in the survey” (Dillman, Smyth, & Christian, 2014; p. 59). Researchers are then faced with the task of selecting the members of their sample.

Over the last 60 plus years, the vast majority of survey researchers have used probability sampling in order to make generalizations about the population of study (AAPOR, 2013). Probability sampling is defined as, “every member of the sampling frame is given a known, nonzero chance of being included in the sample that allows a survey’s results to be generalizable to the full target population” (Dillman, Smyth, & Christian, 2014; p. 75). With probability sampling, survey researchers are able to estimate proportions in the population with a certain level of precision. This precision level is calculated based on the sample size.

In order to have a probability sample, a researcher must have a sampling frame, or, “the list of units in the population that the sample will be drawn from” (Dillman, Smyth, & Christian, 2014; p. 59). The sample is thus the participants drawn from the sampling frame. The sample drawn is a portion of the target population, or, “all of the units (e.g., individuals, households,

organizations) to which one desires to generalize the survey results” (Dillman, Smyth, & Christian, 2014; p. 59). Thus, a probability sample allows researchers to generalize estimates obtained from the sample to the target population within a certain level of confidence, given that the sample is representative of the target population. For example, in election forecasting polls, researchers predict who the winner of an election will be with a certain level of precision; they predict what proportion of the votes each candidate will get with a margin of error, usually within a few percentage points. The sample size needed to obtain a certain level of confidence in the estimates increases as the target population increases, up to a certain point. This is why surveys that only interview 1,000 to 2,000 people are able to estimate proportions in the entire American population (Dillman, Smyth, & Christian, 2014). Probability sampling allows researchers to be confident in their estimates while also saving them time and money, compared to carrying out a census.

Traditionally, probability samples have involved taking a random sample of addresses or landline phone numbers through random digit dialing (RDD) because these sampling frames are readily available. Over time, it has become more and more difficult to convince the American public to participate in surveys, evidenced by declining response rates (Groves, 2006; Keeter, Christian, Dimock, & Gewurz, 2012; Oldendick & Link, 1994). Because of these issues, researchers have started to question these methods and whether they can produce accurate estimates (Groves, 2006; Keeter, Christian, Dimock, & Gewurz, 2012). Declining response rates will result in what is termed non-response error, defined as, “the difference between the estimate produced when only some of the sampled units respond compared to when all of them respond” (Dillman, Smyth, & Christian, 2014; p. 3). If those who respond are systematically different

from those who do not respond, the estimate obtained only from the members of the sample that responded will be systematically different from the population value.

In addition, one of the most prominent methods of contacting participants historically was through the use of RDD, which involves the use of a landline phone. The National Center for Health Statistics estimates that about half of American households don't have a landline phone and there are many demographic differences between those with landline phones and those without (Blumberg & Luke, 2017). In other words, the use of RDD methods to contact participants will face issues of what is termed coverage error, which occurs "when the list from which sample members are drawn does not accurately represent the population on the characteristic(s) one wants to estimate with the survey data" (Dillman, Smyth, & Christian, 2014; p. 3). In this case, those who don't have a landline phone will have a zero probability of inclusion. Therefore, the estimate obtained only from those with a landline phone will be systematically different from the population value.

Because of these issues and with the growing popularity of the Internet, web surveys are becoming more and more popular (AAPOR, 2013; Baker et al., 2010; Couper & Bosnjak, 2010; Pew Research Center, 2015b). The term "web surveys" comprises a host of meanings and these surveys can rely on probability sampling or non-probability sampling methods (AAPOR, 2013; Baker et al., 2010; Couper & Bosnjak, 2010). In contrast to address-based sampling and RDD, there is no list of all available email addresses from which to draw a random sample and if it did, laws prohibiting mass emailing would prevent a sampling frame to be developed from it (AAPOR, 2013; Baker et al., 2010; Dillman, Smyth, & Christian, 2014). Therefore, some non-probability-based web surveys use volunteers, which can be recruited in a variety of ways (AAPOR, 2013). In contrast, survey researchers can also use previously discussed methods of

probability sampling, like using addresses or phone numbers to develop a sample, and then ask participants to go online to take a survey (Baker et al., 2010).

Web surveys also face issues of coverage error and non-response error. Not every member of the American population has Internet access and thus, an estimate obtained only from those with Internet access will be systematically different from the population value. Similar to other types of surveys, an estimate obtained only from respondents to a web survey will be systematically different from the population value, especially when those respondents are volunteers. In both probability and non-probability samples, researchers are faced with estimates that are systematically different from the population value and must attempt to improve the accuracy of the estimates. With probability samples, researchers attempt to improve the accuracy of their estimates by selecting a random sample of the population—a design-based approach. This approach, “while using models to adjust for undercoverage and nonresponse, provide some protection against the risk of sampling bias” (AAPOR, 2013; p. 14). With non-probability samples, researchers attempt to improve the accuracy of their estimates by accounting for inaccurate estimates after receiving the data—a model-based approach. This approach “[relies] more heavily on the appropriateness of the models and, in most cases, on the selection, availability and quality of the variables used for respondent selection and post hoc adjustment” (AAPOR, 2013; p. 14). While some researchers argue that probability samples are the only way of accurately estimating proportions in the population, others argue that non-probability samples can be just as accurate, as long as the model is correctly specified (AAPOR, 2013). This study will focus on a model-based approach, namely, using propensity scores as a weighting method, and methods of correctly specifying this model. Propensity scores can be used to estimate the probability of volunteering for a survey from a selection of known covariates through a logistic

regression. In order to use propensity scores for weighting volunteer web survey data, researchers must first identify the important covariates to use in the model.

Current Study

Because estimates from non-probability samples follow a model-based approach, the most important process for researchers is to thus correctly specify the model. In the use of propensity scores, this means identifying the covariates to use in the model to correct for coverage error and non-response error. While correcting for coverage error is somewhat easier, because this oftentimes means including demographic variables into the model, the more difficult component of the model is correcting for non-response error, or attempting to estimate the probability of volunteering. The purpose of using the covariates chosen is to make the sample as representative of the target population as possible. While there has been literature on variable selection for propensity scores in observational studies, which are the original use of propensity scores (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Myers et al., 2011; Westreich et al., 2011), there has not been as much research done on variable selection for propensity scores in weighting survey estimates beyond Fukuda (2011). Moreover, the method by which to select the covariates used in the model needs more scrutiny, specifically the use of webographic or attitudinal variables (AAPOR, 2013; Couper, 2000; Duffy et al., 2005; Lee, 2006; Schonlau et al., 2007), which are implemented in order to account for some of the differences between Internet users and non-users that are not captured through demographics.

The use of propensity scores in weighting survey estimates was introduced in the early 2000s and work on the topic has continued to improve. Namely, computational methods of using the propensity score in weighting have been studied (Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011). However, the research that has been done during this time has

concluded varying effectiveness in using propensity scores as a weighting method. In addition, this research has not been consistent in the type of variables included in the model. Few studies have been published that have focused on the important covariates to use in the model (Fukuda, 2011; Schonlau et al., 2007) and this research did not focus on the methods of selecting covariates. This study will aim to fill these gaps by answering the questions: (1) What types of webographic variables included in the model are most effective at reducing bias in the estimate? (2) What is the best method of selecting these variables for the propensity score model? This study will answer this question by investigating several methods of variable selection based on previous research.

Literature Review

Bias in Web Survey Estimates

Among the advantages of web surveys is a lower cost, an ability to reach more people in a shorter amount of time, and access to hard-to-reach populations for specialized surveys (AAPOR, 2013; Baker et al., 2010; Couper, 2000; Couper & Bosnjak, 2010; Duffy et al., 2005). These advantages of web surveys are balanced with several other considerations. For example, the probability framework cannot be applied to estimates from non-probability samples and therefore, a level of precision with the estimates cannot be calculated. Other considerations will be discussed next in context of the four areas of total survey error. Survey error can be defined as, “the difference between an estimate that is produced using survey data and the true value or the variables in the population that one hopes to describe” (Dillman, Smyth, & Christian, 2014; p. 3). These sources of error cause biased estimates. Bias is defined as, “a systematic shift in estimates away from the true value” (Dillman, Smyth, & Christian; p. 101) and is the difference between the survey estimate and the population proportion (AAPOR, 2013).

Coverage error. Coverage error is a function of the mismatch between the target population and the frame population (Baker et al., 2010; Bethlehem, 2010; Couper, 2000; Couper, 2001; Couper & Bosnjak, 2010). This type of error occurs “when the list from which sample members are drawn does not accurately represent the population on the characteristic(s) one wants to estimate with the survey data” (Dillman, Smyth, & Christian, 2014; p. 3). Some assert that this problem, prevalent in non-probability samples in general, is more appropriately termed “exclusion bias” because part of the target population has a zero probability of inclusion (AAPOR, 2013).

Although all modes of survey collection are potentially affected by this error, it is more severe in online surveys based on non-probability sampling methods (Baker et al., 2010). Because the web has no list of individuals from which to sample from (like addresses or phone numbers), generalizability with these surveys is questioned. In other words, trying to estimate proportions only using information from Internet users will not necessarily accurately reflect the entire population. In their study comparing participants of their American Trends Panel that responded through the web to those who responded through mail, Pew Research Center (2015b) found that most of the differences between online and offline respondents was small. Notably, the biggest differences were between those 65 and older and among survey items asking about Internet use and other technology-related items. In other words, online and offline populations are different on certain characteristics and these differences will influence some estimates. For example, if one were to estimate the average time spent on the Internet by conducting a web survey, the results would overestimate what the true average time is in the population because the sample included only those who spend time on the Internet.

Coverage error is going to be more of a problem for web surveys when trying to estimate proportions for a larger and more diverse population, like the entire United States population, than smaller populations, like students at a particular university or employees at a company, where every individual has a valid email address and thus a sampling frame can be developed (Couper, 2001; Couper & Bosnjak, 2010; Dillman, Smyth, & Christian, 2014). Coverage error is also more of a problem when those included in the sample differ in meaningful ways from those not included in the sample (Couper, 2001; Dillman, Smyth, & Christian, 2014). While Pew Research Center (2015b) estimated that a vast majority (89%) of the US population uses the Internet, this does not indicate complete coverage. That is, about one in ten Americans would

have a zero probability of inclusion if a survey were only conducted through the Internet. While Internet use has grown very quickly in the US, it is not clear if it will ever reach complete coverage of the entire population and whether or not those who do not have access will continue to be different from those who do. These differences, therefore, threaten estimates for the population when only using data from those with Internet access.

Sampling error. Sampling error is defined as, “the difference between the estimate produced when only a sample of units on the frame is surveyed and the estimate produced when every unit on the list is surveyed” (Dillman, Smyth, & Christian, 2014; p. 3). Because of the cost and time of surveying every member of the population, and thus developing a sampling frame, researchers will always be faced with sampling error unless the survey is a census of the entire population (Dillman, Smyth, & Christian, 2014). In addition, surveys will not be faced with sampling error if the survey is based on volunteers because there is no sampling frame and thus no method of selecting the sample. One could argue that instead of sampling error, non-probability samples face selection error (AAPOR, 2013), discussed below.

Non-response error. In addition to contacting participants, researchers also have to convince them of their participation and respondents, whether from a probability sample or a non-probability sample, have the choice of participating (Dillman, Smyth, & Christian, 2014). Non-response error is defined as, “the difference between the estimate produced when only some of the sampled units respond compared to when all of them respond. It occurs when those who do not respond are different from those who do respond in a way that influences the estimate” (Dillman, Smyth, & Christian, 2014; p. 3).

Not only are there differences between Internet users and non-users, but also between those willing and able to take surveys online and those who are not. Similar to the problem that

telephone surveys once faced of telemarketers oversurveying participants, people might not be willing to take surveys online because they are overwhelmed with the amount of survey requests and/or are not interested enough in the topic to respond (Couper, 2000). Therefore, just because people have access to the Internet does not mean that they are willing to take surveys online; those willing to do so are not necessarily representative for the entire population. For example, Pew Research Center (2015b) compared these groups of people and found that those who have Internet access and occasionally use it, but say that they cannot or will not take surveys online are more like those who are not online than they are to others on the Internet. In other words, estimating characteristics of the population only from those who have Internet access *and* use it more often will not accurately represent the population.

With probability samples, the researcher is in control of who is selected as a member of the sample for a survey. In contrast, estimates from non-probability samples relying on volunteers face self-selection error, which occurs because respondents select themselves for participation (Bethlehem, 2010; Duffy et al., 2005). In this case, respondents' probability of participating is related to the variable of interest, or, the characteristic the survey is attempting to estimate in the population. This relationship between the probability of participating and the variable of interest will bias estimates (Bethlehem, 2010; Groves, 2006). For example, say a researcher is conducting a web survey about the population's interest in travelling and posts the survey on a variety of websites. The estimate obtained from the survey will be biased because those who see the survey (and have Internet access) and have an interest in travelling will be more likely to participate in the survey. Those who see the survey (and have Internet access) and don't have an interest in travelling will be less likely to participate in the survey. The choice of

participating is completely left to the participant, which causes self-selection error. The survey estimate will thus overestimate the population's interest in travelling.

Measurement error. Measurement error is defined as, “the difference between the estimate produced and the true value because respondents gave inaccurate answers to survey questions. It occurs when respondents are unable or unwilling to provide accurate answers” (Dillman, Smyth, & Christian, 2014; p. 3). One issue that can lead to measurement error is providing too many response options and thus, making it difficult for the participant to differentiate between options and provide an accurate response (Schonlau et al., 2003). This type of error can be present with any type of survey.

Another issue that can lead to measurement error is the mode of the survey, whether it is conducted with an interviewer, as in a telephone or face-to-face survey, or conducted without an interviewer, as in a web survey or a mail survey (Couper, 2001). This confounding issue can potentially cause additional differences when estimates are compared from face-to-face surveys and web surveys (Loosveldt & Sonck, 2008). In a study by the Pew Research Center (2015a) on the mode effects of two samples randomly assigned to either a web survey or a telephone survey, they found that most items were not subject to mode effects. Of the items that did have significant differences between the two samples, the mean difference was 5.5 percentage points. Some of these items included levels of life satisfaction and societal discrimination against certain groups. There is sizable literature on mode effects of surveys but this study will not focus on this source of error.

Approaches to Dealing with Error

No matter the mode of the survey, researchers must decide their method of sampling and must deal with biased estimates. Probability samples have increasingly experienced problems of

non-response error, while non-probability samples have problems of coverage error and self-selection error. Researchers are thus faced with balancing these issues and choosing a method, or a combination of methods, that will eliminate the most bias (Couper, 2001; Couper & Bosnjak, 2010; Dillman, Smyth, & Christian, 2014). As discussed in Chapter 1, probability samples often employ a design-based approach while non-probability samples often employ a model-based approach in order to improve accuracy of the estimates obtained. This study will focus on a specific model-based approach—using propensity scores as a weighting method. First, the concept of weighting in general will be explained.

Weighting. With a probability sampling framework, the method by which the researcher chooses the sample from the sampling frame is termed the sample selection and “every unit in the population must have a known chance of being included in the sample, but the rate at which different units are sampled can vary” (Dillman, Smyth, & Christian, 2014; p. 59). If every member of the sampling frame has an equal probability of selection, this is what is referred to as simple random sampling (Dillman, Smyth, & Christian, 2014). Oftentimes, survey researchers do not place equal probabilities of selection for each member of the sample in order to increase precision of estimates for minority groups. One alternative method to simple random sampling is to employ stratification techniques, which “refers to grouping the units on the sample frame into subgroups, called strata, based on certain characteristics, so that sampling can be performed independently for each stratum” (Dillman, Smyth, & Christian, 2014; p. 57). This method assists researchers in ensuring that minority groups are represented at rates similar to the target population. But, in order to account for this unequal probability of selection, sample members must be given a survey weight (Dillman, Smyth, & Christian, 2014). Traditionally, individuals from these surveys would be weighted with the inverse of their probability of selection as a base

weight (AAPOR, 2013; Kalton & Flores-Cervantes, 2003; Lee & Valliant, 2008). This base weight determines how many people from the target population that participant represents so that those with a higher probability of selection represent the correct number of people (Dillman, Smyth, & Christian, 2014). Other forms of weighting can then be applied to account for other sources of error.

With non-probability sampling, researchers don't know the individual's exact probability of selection. One method of estimating this probability that has been suggested in the literature is through the use of propensity scores. Propensity scores estimate an individual's probability of volunteering for a web survey from a selection of known covariates and the inverse of this propensity score is then used to form the base weight as a "pseudo design-based weight" (AAPOR, 2013; p. 67). This method was originally introduced to the literature for use in observational studies, when random assignment was not possible and/or unethical (Rosenbaum & Rubin, 1983), but has gained popularity for use in weighting methods for volunteer surveys (Danielsson, 2004; Lee, 2006; Lee & Valliant, 2009; Schonlau et al., 2003; Schonlau et al., 2007; Schonlau et al., 2009; Taylor, 2000; Valliant & Dever, 2011).

Figure 1, below, summarizes the goals of using propensity scores as a weighting method.

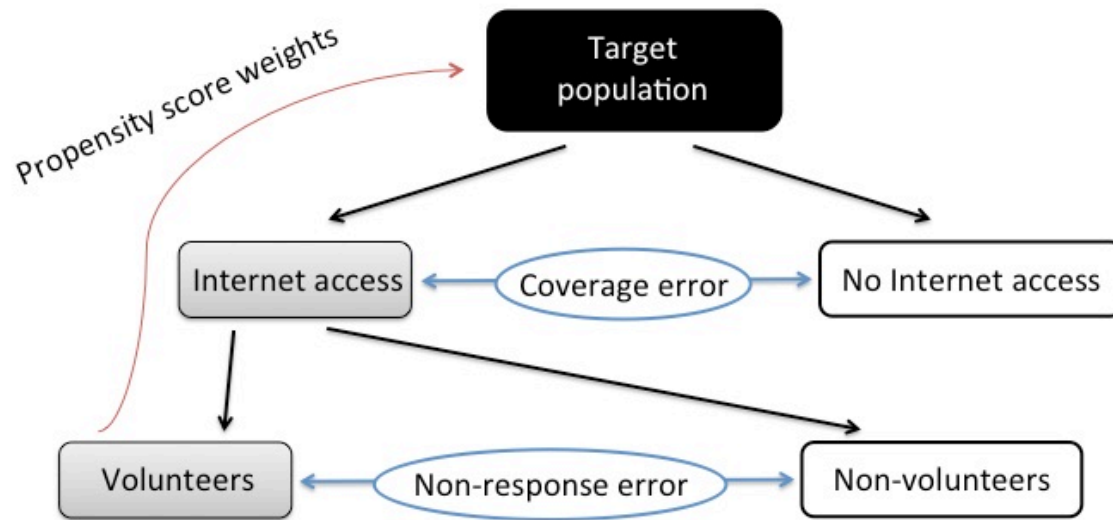


Figure 1. The goal of using propensity scores as a weighting method.

The target population, as defined earlier, is the group of people to which the survey attempts to generalize estimates. By conducting a survey on the Internet, this thus divides the population into those who are able to complete the survey (i.e., those who have Internet access) and those who are not able to complete the survey (i.e., those who do not have Internet access). The differences between these two groups have the potential to cause coverage error. Those who have Internet access thus become the pool from which your sample will be taken and will be divided into those who are willing to respond to the survey (i.e., volunteers) and those who are not willing to respond to the survey (i.e., non-volunteers). The differences between these two groups have the potential to cause non-response error. Phrased a different way, those who choose to respond have the potential to cause self-selection error. The group of people within the target population who have access to the Internet *and* volunteer for the survey become the sample.

In order to generalize estimates from the sample to the target population (the red line in Figure 1), the researcher must account for non-response error and coverage error. Propensity scores attempt to account for both of these types of error and thus decrease the bias in the estimate from the volunteer web sample in order to resemble the value in the target population as

closely as possible. By accounting for demographic differences between those who have Internet access and those who don't, the model will attempt to account for coverage error. By accounting for the reasons why certain people choose to volunteer and others choose not to volunteer, the model will attempt to account for non-response error. This study will focus on the latter, which has been discussed in the literature as "webographic" or attitudinal variables. Propensity scores and methods for calculating them will be discussed next.

Background on Propensity Scores

The use of propensity scores was first introduced to the literature for use in observational research and was popularized by Rosenbaum and Rubin (1983). Observational research is done in situations when the researcher does not randomly assign participants to certain groups (i.e., treatment and control) but when participants either self-select themselves into groups and/or when the groups have been formed prior to research. When random group assignment is not possible and/or unethical (as in, medical research or educational research), comparing results from the two groups will be biased because of confounding factors. For example, if researchers are comparing an outcome (the result of some intervention) between smokers (treatment group) and non-smokers (control group), it is unethical to randomly assign participants to these groups. In addition, the groups will not be balanced on certain characteristics important to the study, like demographic variables. If smokers are more likely to be male than female, then the treatment and control groups are not equal on this measure and therefore, the outcome will be confounded by this variable. In order to balance these groups and thus be able to compare means on the outcome between these groups without these confounding factors, researchers commonly use the propensity score.

A propensity score, $e(x)$, is defined as the conditional probability of being in a group or treatment, given a set of covariates x :

$$(1) \quad e(x) = \text{pr}(z = 1|x)$$

Rosenbaum and Rubin (1983, 1984) suggest estimating the propensity score using a logistic regression, which is the most popular way of estimating the propensity score (Weitzen et al., 2004). The most common method of using propensity scores is to stratify participants into a certain number of groups based on the propensity score (discussed in detail in Rosenbaum & Rubin, 1984; Cochran, 1968) and thus, balancing the two groups. Using the propensity score thus makes comparison between groups unbiased, assuming all relevant covariates are used in the model. Once the researcher controls for confounding factors, then treatment assignment is “strongly ignorable” (Rosenbaum & Rubin, 1983; p. 43) and outcomes can be directly compared without bias.

Use of Propensity Scores in Survey Research

The goal of using propensity scores in observational research is different from that of survey research (AAPOR, 2013; Brick, 2011; Fukuda, 2011; Valliant & Dever, 2011). In observational research, the goal is often to determine a causal effect between treatment and outcome by comparing the means from the two groups on that outcome. This is accomplished by creating a matched sample to eliminate confounding factors, rendering treatment assignment ignorable (Rosenbaum & Rubin, 1983). Therefore, the goal of this balancing process is to *identify* differences between the means of these two groups.

In contrast, survey research has the goal of making point estimates of proportions in the target population. One can think of the volunteers in a web survey as comparable to those in the treatment group in observational studies and those in the target population as comparable to

those in the control group. Propensity scores attempt to estimate the probability of treatment assignment, or, probability of volunteering. After adjusting for this probability of volunteering, the estimate from the volunteer web survey needs to resemble the value in the target population as closely as possible. Therefore, the goal of this balancing process is to *eliminate* the differences between these two groups.

Estimating the propensity score. Because participants who have Internet access and choose to participate in a web survey are not necessarily representative of the target population, the estimate obtained from this sample will be biased. In other words, it will be systematically different from the estimate in the population and this difference is due mainly to coverage error and non-response error, as discussed earlier. A weighting scheme thus needs to be implemented in order to correct for this bias. One weighting method involves the use of propensity scores, in which a participant's probability of volunteering for the survey is estimated using a set of known covariates. The propensity score of volunteering for a web survey, $\pi(x_k)$, can be defined as, modified from Valliant and Dever (2011):

$$(2) \quad \pi(x_k) = \pi(W|x_k)\pi(V|W, x_k),$$

where:

x_k = a vector of covariates for person k that are predictive of participation;

$\pi(W|x_k)$ = probability of having access to the Internet;

$\pi(V|W, x_k)$ = probability of volunteering for the survey given that person k has access to the Internet.

Because researchers do not control participation as they would in a probability sample, $\pi(x_k)$ is estimated and therefore a pseudo-selection probability.

Lee and Valliant (2008) list the five assumptions related to using propensity scores in volunteer web surveys (p. 176): 1. Strong ignorability of treatment assignment given the value of a propensity score. This refers to the assumption that treatment assignment is random, given the value of the propensity score. This implies that all relevant covariates are used in the model. 2. No contamination among study units. This could be violated if there is an independence issue with the data (i.e., data come from multiple people in the same household). 3. Nonzero probability of treatment or nontreatment. This could be violated if there were individuals who were never covered by the sample frame (i.e., those without Internet access could never be included in a web survey). 4. Observed covariates represent unobserved covariates. This could be violated if relevant covariates were not available in the dataset or were not used in the model, but would affect the propensity score. 5. Treatment assignment does not affect the covariates. In this situation, “treatment assignment” would be equivalent to responding to the web survey and covariates oftentimes include demographic variables, like gender or income. Therefore, this is rarely violated in survey research because responding does not affect variables like gender or income.

Reference survey. When using propensity scores with volunteer web surveys, researchers have suggested conducting a reference survey that is a true random sample of the target population (Bethlehem, 2010; Danielsson, 2004; Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Scholnau et al., 2003; Taylor, 2000; Valliant & Dever, 2011). This reference survey will thus be representative of the target population because it is a probability sample. In the same way that using the propensity score in observational research is meant to balance the two groups, using the propensity score as a weighting method in survey research is meant to make the

volunteer web sample distribution resemble the reference sample distribution and thus accurately predict estimates in the target population.

When estimating the propensity score, which involves combining the data from the reference survey and the volunteer survey, it is important to include the original weights from the reference survey (Valliant & Dever, 2011). This will ensure that the reference survey estimates are unbiased. If no weights are used, then the sample will only be representative of the combined reference and volunteer samples, not the full target population. In addition, the sample size of the reference survey will be smaller than that of the volunteer web survey. As the reference sample size increases and the volunteer web sample size decreases, the bias in the estimate obtained tends to increase (Valliant & Dever, 2011).

The reference survey will contain all covariates used in the propensity score model, assuring that the fourth assumption of propensity scores (observed covariates represent unobserved covariates) is not violated. Some researchers have suggested the use of a source like a census or a large survey that has limited coverage errors to serve as the reference survey because this would have accurate estimates of basic demographics (Valliant & Dever, 2011). Others, such as Schonlau et al. (2003) argue that a census could not be used as a reference survey because it does not include “webographic” or attitudinal variables. As discussed earlier, these variables are important to include in order to account for non-response error. A discussion of these attitudinal variables will come later.

Methods for using the propensity score. In order to weight estimates from the volunteer web sample to the target population, researchers must first estimate the propensity score (i.e., the probability of volunteering). Data from the reference sample and from the volunteer web sample will be combined to estimate the propensity score. This assures that the third assumption of

propensity scores (nonzero probability of treatment or nontreatment) is not violated. Importantly, researchers must use the original survey weights for the reference survey sample (Valliant & Dever, 2011). Then, this propensity score is what is used to form the base weight only for the volunteer web sample. Figure 2, below, is a continuation of Figure 1 from Chapter 1, adding in the reference sample.

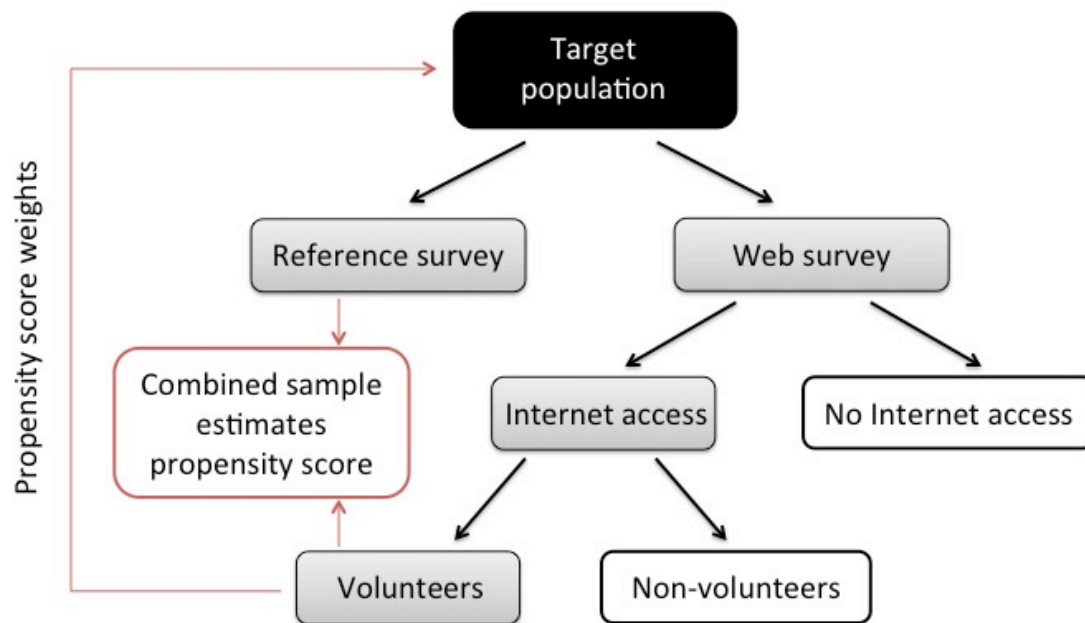


Figure 2. Using the reference sample and volunteers to develop the propensity score weights.

As discussed earlier, the base weight attempts to account for the unequal probabilities of selection into the sample. In the case of volunteer web surveys, this probability of volunteering is estimated through the propensity score. By combining data with the reference sample, the probability of volunteering is treated as a “quasi-random process” (Valliant & Dever, 2011) and each person has a probability of volunteering, estimated from the propensity score, as shown in Equation 2. Thus, estimating the propensity score with only data from the volunteer web sample would not be possible because it would not include the full range of probabilities. Methods for developing this base weight will be discussed next.

In general, there have been three methods of using the propensity score for weighting survey estimates that researchers have found work well. First, using the inverse of the individual's propensity score as the base weight (Valliant & Dever, 2011). Second, dividing the participants into groups based on the propensity score (Cochran, 1968) and using the inverse of the average propensity score for that group as the base weight (Valliant & Dever, 2011). Third, using a combination of the propensity score and what is termed calibration weighting, to be discussed next (Kim & Riddles, 2012; Lee & Valliant, 2009). In theory, the volunteer web sample, when weighted with propensity scores, will resemble the reference survey. If this reference survey is representative of the target population, then the volunteer web sample will thus resemble the target population. If, however, the reference sample has substantial coverage error, then an additional weighting step needs to adjust for this error. Calibration weighting is this second step and it further weights the volunteer web sample to resemble the target population on key demographic variables.

In Valliant and Dever (2011)'s comparison of the above methods, using the inverse of each individual's propensity score and calibration weighting produced the least biased estimates. While it would be beneficial to continue Valliant et al.'s (Lee & Valliant, 2009; Valliant & Dever, 2011) investigation of calibration weighting, this is beyond the scope of this project. Because the focus of this study is on variable selection, the simplest yet most efficient method of using the propensity score will be used. Therefore, this study will use the inverse of each individual's propensity score as a weight.

Determining the best propensity score model. Researchers have differed in how they analyze the effectiveness of the propensity score model. Some researchers have access to actual volunteer web sample data and thus use the propensity score to weight the estimates obtained

from that sample. They then compare the unweighted and weighted proportions of the study variable from the volunteer sample to the estimate obtained from the reference sample in order to determine if there remains any significant difference in the estimates after weighting (Duffy et al., 2005; Loosveldt & Sonck, 2008; Schonlau et al., 2003; Schonlau et al., 2009). Investigating significant remaining differences between the samples after weighting is important because this is the goal of using propensity scores as a weighting method—to eliminate these differences, as discussed earlier.

In contrast, some researchers do not have access to actual volunteer web sample data and thus simulate this data as well as a target population and reference sample. These researchers are then able to calculate the bias in the estimates by comparing the target population value with the weighted estimate from the volunteer web sample. After calculating the bias in the estimates, researchers are able to use this measure to compare different weighting methods in order to determine the method that produces the least biased estimate (Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011). Calculating the bias is important because it quantifies the amount of error in the estimate of the study variable compared to the value in the population. These methods, also used in this study, will be explained in more detail in Chapter 3. As with any weighting method, researchers who have used propensity scores as a weighting method have found that although it might decrease the bias, it usually increases the variance (Kalton & Flores-Cervantes, 2003; Lee, 2004; Lee, 2006; Schonlau et al., 2009). Weights, in general, decrease the precision of estimates and therefore, increase the variance. Therefore, it is important to calculate the variance concurrently with the bias of the estimate when using weighting.

Effectiveness of propensity score weighting. The first mention in the literature of using propensity scores for weighting volunteer web surveys was from Harris Interactive (Taylor,

2000). Since then, in general, researchers have had varying success in decreasing the bias of the estimates obtained from volunteer web samples when using propensity scores as a weighting method. Methods that have defined effective propensity score models as those that eliminate significant differences between volunteer web samples and reference samples have found that propensity score weighting has generally not been effective (Duffy et al., 2005; Loosveldt & Sonck, 2008; Schonlau et al., 2009; Schonlau et al., 2003). In contrast, those that defined the most effective model by calculating the bias in the sample estimate have found more success with using propensity scores as a weighting method (Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011). It is not clear whether researchers have found varying results because of the way that they have defined the most effective model or because of the inconsistencies in the variables used in the model, which will be discussed in the next section.

The effectiveness of the propensity score weighting method sometimes varies depending on the study variable being investigated. The “study variable” is the characteristic of the population that the survey is attempting to estimate. Study variables can be categorized into factual (i.e., age, gender, work status, etc.) or attitudinal (i.e., opinions on politics or immigration) items. Weighting methods like post-stratification (a post hoc weighting method employed to make the sample resemble the target population on key demographic characteristics) are suggested to correct biases for factual variables while weighting methods like propensity scores are suggested for correcting biases for attitudinal variables (Loosveldt & Sonck, 2008).

These types of study variables are also prone to different sources of error, namely self-selection error and measurement error (Duffy et al., 2005; Loosveldt & Sonck, 2008). Self-selection error is more likely to influence estimates in which the sample is more knowledgeable

and/or more interested in the survey topic. Measurement error is more likely to influence estimates when the study variable is more susceptible to social desirability bias. For example, a participant might answer questions pertaining to opinions about immigrants differently when they are asked by an interviewer face-to-face versus when they are asked online. Factual variables are not as susceptible to social desirability bias as attitudinal variables, so differences in factual variables are most likely attributed to selection bias and differences in attitudinal variables are most likely attributed to mode effects, according to Loosveldt and Sonck (2008). While mode effects might explain part of this difference in attitudinal variables, they most likely don't explain all of the difference, as shown by more recent research (Pew Research Center, 2015a), discussed earlier, which showed that mode effects resulted in few differences.

In several studies, there were still differences between the weighted estimate from the volunteer sample and the estimate from the reference sample and these researchers thus concluded that the remaining differences were due to self-selection bias and social desirability bias (Duffy et al., 2005; Loosveldt & Sonck, 2008; Schonlau et al., 2003). In all of these papers, researchers have mentioned that more variables needed to be included in the propensity score model beyond demographics in order to account for these differences. Similarly, Schonlau et al. (2009) conclude that if there is an underlying variable related to self-selection and the study variable of interest that is not included in the propensity score model or is not possible to include in the propensity score model, then no weighting scheme will eliminate the bias in the estimate. The similar results that the studies have found despite the differences in the variables included in the propensity score model begs the question of what variables are important to include and if it is possible to include variables that account for all sources of error. The topic of variable selection will be discussed next.

Variable Selection

Just like in any model-based adjustment, one of the biggest hurdles researchers face in using propensity score methods is correctly specifying the model, or deciding which covariates to use in the model (Bethlehem, 2010; Couper, 2000). In using propensity scores for observational research, there is sizable literature on the methods used for variable selection. Rosenbaum and Rubin (1984) specify that the outcomes being compared will only be unbiased when all variables related to both outcome and treatment are included in the model. The general consensus in previous research is that using variables only related to treatment, but not outcome, will increase bias in the model (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Myers et al., 2011; Westreich et al., 2011). In addition, Brookhart et al. (2006) and Westreich et al. (2011) explain that the goal of propensity scores in observational research is to balance the groups on the propensity score in order to control for confounding.

In contrast, Fukuda (2011) explains that propensity scores for use in survey research have a different goal because survey participation is different from treatment assignment. One issue is that survey participation has more divisions than just treatment and assignment. Included in the target population, there are those who are sampled and those not sampled, those who are a part of the volunteer survey and those a part of the reference survey, and those who respond to their relevant survey and those who do not respond (see *Figure 1* and *Figure 2*). Therefore, covariates will be related to sample inclusion, response probability, and to the study variable.

In order to assess the importance of these three types of variables in propensity score models, Fukuda's (2011) simulations showed that the essential variables were those that are associated with both participation probability and the study variable. Among these, the most important to include are the variables that are simultaneously associated with the study variable,

the sampling inclusion, and the response probability. While some researchers using propensity scores in observational research generally recommend using all available covariates, even if the relationship between the covariate and the study variable is weak (Lee & Valliant, 2008), Fukuda's (2011) simulation found that including variables only relating to the study variable did not change the estimates. In a similar discussion, Kalton and Flores-Cervantes (2003) note that it is important that the variables predict response probabilities. In addition, they note that when benchmarking to external sources, it is important that the variables predict key survey variables. In other words, it is important for variables used in weighting to be both related to response probability and the study variable. Because of this, models will be effective on a case-by-case basis (Fukuda, 2011; Lee & Valliant, 2008).

Variable selection methods in observational research. When propensity scores were first introduced to the literature for observational data, researchers used stepwise regression for covariate selection (Rosenbaum & Rubin, 1983; 1984). While these researchers also suggested that models include all available covariates, research has shown that this impacts bias and efficiency of the propensity score (Franklin, Eddings, Glynn & Schneeweiss, 2015; Shortreed & Ertefaie, 2017). Because a variable selection method is preferable to including all available covariates, especially when there are a large number of covariates available as in health research and social science data (like the GSS), researchers oftentimes use methods like lasso regression (Tibshirani, 1996) or ridge regression in practice (Koch, Vock, & Wolfson, 2017; Vansteelandt, Bekaert, & Claeskens, 2010). These methods select a “best subsets” from a list of variables to use in the regression based on statistical criteria. However, more current research has found that because these methods focus more on the relationship between the covariates and the treatment, these models tend to overlook variables that are strongly related to the study variable and those

related to the study variable through exposure, which are the true confounders and the precise variables most important to include (Koch, Vock, & Wolfson, 2017; Vansteelandt, Bekaert, & Claeskens, 2010). Propensity score methods tend to work best when covariates are related to the study variable under investigation, as using variables only related to treatment will increase variance in the model (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Myers et al., 2011; Westreich et al., 2011). Therefore, researchers have begun to suggest variable selection methods that alleviate this problem by focusing on the relationship between covariates and the outcome as well as covariates and the treatment (Franklin, Eddings, Glynn & Schneeweiss, 2015; Koch, Vock, & Wolfson, 2017; Shortreed & Ertefaie, 2017; Vansteelandt, Bekaert, & Claeskens, 2010). While these methods have proven to work well for the use of propensity scores in observational research, their effectiveness in variable selection for propensity scores as a weighting method has not been investigated.

Webographic or attitudinal variables. In their report on non-probability methods in survey research, AAPOR (2013) said, “To be of value non-probability samples must rely on some form of statistical adjustment to manage this risk of large biases. The effectiveness of those adjustments depends on the *identification* of important covariates, their *availability* and *quality*” (p. 33; emphasis added). As more and more of the population has access to the Internet, we better understand the Internet population (File & Ryan, 2014; Pew Research Center, 2015b) and conducting surveys on the web has become easier. Mirroring this change, more researchers have begun to include items in their surveys asking about technology and Internet use, providing more options for covariates to use in propensity models. One source of a variety of covariates is the General Social Survey, which tracks the American population’s attitudes and behaviors on a wide variety of topics over time. While these covariates are readily available, it is unknown

which variables are most important to use in estimating the propensity score because previous research has varied in the covariates used in this process, as will be discussed next.

There has been some discussion in the literature about inclusion of “webographic” or “attitudinal” variables in propensity score models (Lee, 2004; Lee, 2006; Schonlau, van Soest, & Kapteyn, 2007). Their use in balancing is important because web users and non-web users not only differ in terms of demographic characteristics, but also in attitudes and behavior (AAPOR 2013; Couper, 2000; Duffy et al., 2005; Schonlau et al., 2003). These variables also capture the reasons why some participants choose to volunteer and others do not, as discussed in Chapter 1. In their report on non-probability sampling, AAPOR said in a footnote, “Webographics are attitudinal variables thought to account for the difference between people who do surveys online and those who do not. They generally measure lifestyle issues such as the types of activities people engage and their frequency, media use, attitudes toward privacy, and openness to innovation” (AAPOR, 2013, p. 70).

Beyond this, there is no agreed upon definition of what constitutes a webographic variable and examples used in the literature vary greatly. For example, Lee (2004; 2006) used class, work status, political party, religion, and opinion toward ethnic minorities as “nondemographic” covariates. Schonlau, van Soest, & Kapteyn (2007) defined their webographic variables in terms of four categories: attitudinal variables (i.e., eager to learn, takes chances, often feel alone), factual variables (i.e., in the last month have you traveled, participated in a sport, read a book), privacy variables (i.e., questions about airport searches, cookies, phone calls, AIDS screenings, and credit card storage), and variables related to knowing gay people. Duffy et al., (2005) mentioned Internet usage as a demographic weight for their online UK sample as well as several “propensity score questions,” which included “issues such as online

purchasing behavior, views on the amount of information respondents receive and personal attitudes towards risk, social pressure and rules” (p. 623). Similarly, even though Schonlau et al. (2003) did not specify the covariates used in their propensity score model, they gave examples of webographic questions: “Do you feel alone?” and “On how many separate occasions did you watch news programs on TV during the past 30 days?”

Schonlau, van Soest, and Kapteyn (2007) supported the use of webographic variables in their models because there were still differences between their reference survey and their volunteer web survey even after controlling for demographics. They conclude, therefore, that there is “no downside” (p. 162) to using them in the model when they are available. However, it is unclear whether including these webographic variables improves the model, as researchers have yet to define important attitudinal variables to include in the propensity score model. When researchers have to choose between using certain webographic variables over others, Schonlau, van Soest, and Kapteyn (2007) suggest that the variables that are most imbalanced after controlling for demographics are the most important to include. Lee (2004; 2006) compared propensity score models that included demographic variables that were either highly or weakly related to the study variable and found that the model containing all demographic variables and the model containing only demographic variables highly related to the study variable were more effective in decreasing bias than the model containing only demographic variables weakly related to the study variable. Moreover, the studies compared propensity score models that included demographics and nondemographics variables and found that those including demographics variables and all variables produced less biased estimates than the ones with only nondemographic variables. More specifically, it was shown that the effect of including nondemographic variables in addition to demographic variables was minimal; this was attributed

to the study variables in question, which were more related to demographic variables than nondemographic variables.

The use of webographic variables is not often discussed in detail in the literature; Lee (2006) summarizes the problem, “The importance of including nondemographic variables in PSA [propensity score adjustment] for web surveys is unclear due to two facts: (a) inclusion of more variables automatically increases the predictive power of the model and (b) nondemographic (e.g., attitudinal) covariates can often be explained by demographic variables to a certain degree.”

While the purpose of including webographic variables in the models is clear—to balance differences in online and offline populations (Schonlau, van Soest, & Kapteyn, 2007; Lee, 2004; Lee, 2006)—the method of selecting them is not. Fukuda (2011) discusses the reason why it is important to include variables that are associated with the study variable and the inclusion probability and/or the response probability. For one, the reference sample is going to be different from the volunteer sample in terms of the sampling frame. Second, the participants who respond to the volunteer survey are going to be different from the participants who don’t respond in terms of the response probability. Therefore, Fukuda (2011) suggests that comparing the distributions between these groups will aid the researcher in determining the variables to use in the model. While researchers implementing propensity score weighting in practice would have information on the reference survey respondents, the construction of the important covariates to use on the survey has to be determined beforehand. Therefore, researchers have little guidance thus far as to what covariates to include in their reference survey and/or how to choose the covariates used in the propensity score model. Most of the suggestions in the literature tell researchers to choose covariates known to be related to the inclusion probability (Fukuda, 2011; Kalton and Flores-

Cervantes, 2003; Lee & Valliant, 2008). In addition, it is impossible for a known “set” of covariates to be used for multiple study variables, because the model will depend on the study variable in question (AAPOR, 2013; Fukuda, 2011; Lee & Valliant, 2008).

Little research has mentioned the number of variables that it is important to include in order to decrease bias, most likely because models will be different depending on the study variable (AAPOR, 2013; Fukuda, 2011; Lee & Valliant, 2008). In their models, Schonlau et al. (2009) used a relatively small number of covariates and found that while the propensity score weighting decreased bias, there were still significant differences between the Internet sample and the random sample. Looking at articles using propensity score weighting more closely, it is difficult to tell in some instances the exact variables used in the propensity score model (i.e., Schonlau et al., 2003; Duffey et al., 2005). It is important to know what variables are used in the model so that future researchers looking at similar outcomes will have an idea of what variables are important to include in their reference survey. In addition, transparency is a good standard to uphold not only in social science research in general, but with non-probability samples in survey research specifically (AAPOR, 2013).

Table 1 (below) presents a summary of the articles’ author(s), the number of demographic variables used in the propensity score model (as defined by the author), the number of nondemographic variables used (whether this was attitudinal and/or factual), and the total number of variables used.

Table 1
Number of Variables Used in Previous Propensity Score Models

| Article | Number of Demographic Variables Used | Number of Nondemographic Variables Used | Total Number of Variables Used |
|-------------------------|---|--|---------------------------------------|
| Loosveldt & Sonck, 2008 | 5 | 0 | 5 |

| | | | |
|--------------------------------------|---|----|----|
| Schonlau et al., 2009 | 4 | 4 | 8 |
| Lee, 2004; 2006 | 9 | 5 | 14 |
| Valliant & Dever, 2011 | 6 | 10 | 16 |
| Schonlau, van Soest, & Kapteyn, 2007 | 7 | 12 | 19 |
| Lee & Valliant, 2009* | 8 | 22 | 30 |

Note. *This article did not subdivide their covariates into demographic and nondemographic, so the number of demographic variables was determined by this researcher.

The differing number of variables used in the model might be a reason why results have differed in the effectiveness of using propensity score weighting methods. The more useful covariates that are used in the model, the higher the chance of reducing the differences between volunteer web samples and reference survey samples and thus reducing the bias in the estimate of the study variable.

As presented in the table, researchers vary greatly in the number of demographic variables used. Looking at the articles more closely, it is clear that researchers differ on their definition of a demographic variable. Table 2 (below) presents a summary of the article's author(s) and their list of demographic variables used in the models, if specified.

Table 2
Demographic Variables Used in Previous Propensity Score Models

| Author(s) | Demographic Variables |
|-------------------------|---|
| Duffy et al., 2005 | Reference survey: region, social class, car ownership, and age and work status within gender; Online survey: age within gender, region, education level, income level, and Internet usage |
| Loosveldt & Sonck, 2008 | Gender, age, work status, education, and living area (suburban v. rural) |
| Schonlau et al., 2009 | Race/ethnicity, gender, education level, and age |
| Lee, 2004; 2006 | Age, education, size of residential area, household size, family income, race, gender, marital status, and region of the |

| | |
|--------------------------------------|---|
| | residential area |
| Valliant & Dever, 2011 | Age, race, gender, wireless phone, education, and income |
| Schonlau, van Soest, & Kapteyn, 2007 | Age, income, gender, language is English (yes or no), born in the US (yes or no), education, and self-assessed health |
| Lee & Valliant, 2009* | Age, education, income, gender, household size, work full-time (yes or no), marital status, and race |

Note. *This article did not subdivide their covariates into demographic and nondemographic, so the number of demographic variables was determined by this researcher.

In addition to these variations in number and type of variables, the researchers presented in these tables provided little justification for the use of the variables selected for the model beyond a theoretical justification (i.e., the variables selected beyond demographics are meant to balance the reference and volunteer samples; Duffey et al., 2005). As for a statistical justification of the variables chosen, researchers either presented the relationship between the covariate and the study variable (Lee, 2004; Lee, 2006), showed that the reference sample and volunteer sample differed on key variables even after controlling for demographics (Schonlau et al., 2003; Schonlau, van Soest, & Kapteyn, 2007), or presented both (Loosveldt & Sonck, 2008; Schonlau et al., 2009). Some researchers provided no statistical justification for variables used beyond intuition (i.e., the use of health-related variables when looking at a health outcome) (Valliant & Dever, 2011; Lee & Valliant, 2009).

Current Study

The first mention in the literature of using propensity scores for weighting volunteer web surveys was from Harris Interactive (Taylor, 2000). Since that first mention, there has been substantial research on the computational aspect of propensity score weighting (Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011) but there has yet to be a defined process for selecting covariates used to estimate the propensity score itself. Researchers who have used propensity

scores as a weighting method have done so without a predefined structure and these methods have resulted in varying effectiveness. This study's purpose is to suggest a structure for covariate selection for estimating the propensity score.

It is unknown which variables are most important to use in estimating the propensity score because previous research has varied in the covariates used in this process, as shown in Table 1 and Table 2. Previous research has identified that a certain combination of demographic and webographic variables are important to use in estimating the propensity score (Fukuda, 2011; Lee, 2004; Lee, 2006) but the process of selecting these covariates has been somewhat haphazard. This study will be focusing on the use of webographic variables in estimating the propensity score because they are not precisely defined in the literature and their use in previous research has not been systematic. Thus, this study also aims to answer the call put forth in previous research (Schonlau et al., 2009) for a continued search for suitable webographic variables of quality.

It is unknown based on previous research what types of variables are the most important to include in the propensity score model. First, it has been found that a small number of variables will not be sufficient in reducing the bias (Schonlau et al., 2009). However, it is not known how many variables are needed to sufficiently reduce the bias because previous studies have used various amounts of demographic and nondemographic variables (see Table 1 and Table 2) and because one "set" model will not work for all study variables (AAPOR, 2013; Fukuda, 2011; Lee & Valliant, 2008). Second, it has been shown that the use of webographic variables will depend on two things: (a) their effectiveness in balancing even after controlling for demographics (Duffy et al., 2005; Schonlau, van Soest, & Kapteyn, 2007); and (b) their relationship with the study variable and the inclusion probability (Fukuda, 2011; Lee, 2006; Lee & Valliant, 2008). This

study will focus on the latter and incorporate variables of varying relationships with the study variable in order to make suggestions about what types of variables are most important to include in estimating the propensity score.

Variable selection methods have yet to be investigated for propensity score weighting methods in survey research. Therefore, this study will be investigating several methods suggested for use in observational research. As discussed previously, propensity score methods tend to work best when covariates are related to the study variable under investigation, as using variables only related to treatment will increase variance in the model (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Franklin, Eddings, Glynn & Schneeweiss, 2015; Koch, Vock, & Wolfson, 2017; Myers et al., 2011; Shortreed & Ertefaie, 2017; Vansteelandt, Bekaert, & Claeskens, 2010; Westreich et al., 2011). Because variable selection methods suggested for observational research have yet to be investigated for using propensity scores as a weighting method in survey research, this study will investigate several of these methods by including a propensity score model using stepwise regression and another using lasso regression.

Therefore, the research questions for this study is: (1) What types of webographic variables included in the propensity score model are most effective at reducing bias in the estimate? (2) What is the best method of selecting these variables for the propensity score model? To answer these questions, this study will be building off of the computational work done on propensity scores as a weighting method by Lee et al. (Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011) and off of the simulation work investigating variable selection by Fukuda (2011). In addition, while previous research was successful in identifying what methods of using the propensity score reduce the most bias, most research provided little justification for the variables used in their models. This study will provide justification for each

variable used in the model and aim to determine which variables are most important to include in the propensity score model.

Hypotheses. In order to determine which variable selection method provides the least biased estimate, this study will compare the results of several regression models, including: A. All variables (demographics plus webographics). B. Demographics only. C. Demographics plus only webographics that are significantly related to both the study variable and inclusion probability. D. Demographics plus webographics using stepwise regression. E. Demographics and webographics selected using lasso regression. Studies that have compared models similar to Models A, B, and C (Lee, 2004; 2006) have found that the results from these types of models are similar—the models reduce the bias (but not completely) while also increasing variance in the estimates. Similarly, studies on observational research have found that model selection methods that improve the prediction model of treatment (i.e., Model D and Model E) are not optimal (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Koch, Vock, & Wolfson, 2017; Myers et al., 2011; Vansteelandt, Bekaert, & Claeskens, 2010; Westreich et al., 2011). While variable selection methods like Model D and Model E have not been investigated for propensity score weighting volunteer web surveys, it is hypothesized that they will produce more biased estimates than Models A, B, or C. Based on simulation work done by Fukuda (2011), it is hypothesized that Model C will produce the least biased estimate.

Methods

As discussed in Chapter 2, the research questions for this study are: (1) What type of webographic variables included in the model are most effective at reducing bias in the estimate? (2) What is the best method of selecting these variables for the propensity score model? In order to answer these questions, several propensity score models will be compared and used to weight a volunteer web sample. In order to accomplish this, a volunteer web sample and a reference sample were be created. In addition, in order to calculate the amount of bias in the estimate of the study variable of interest, a target population and an estimate of the study variable from that population were also needed. Therefore, this study used the General Social Survey (GSS) to generate a target population, a reference sample, and a volunteer web sample, following the procedures of Lee et al. (Lee, 2004; Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011).

In general, the procedure included the following steps: 1. Create a pseudo population and define the population value. 2. Create volunteer and reference samples. 3. Combine the reference and volunteer samples to estimate the propensity score. During this step, the five different logistic regression models were used and compared in order to determine which variable selection method produced the least biased estimate. 4. Use the propensity score-weighted volunteer sample to obtain the estimate of the study variable. 5. Calculate bias in volunteer estimate (as compared to the population value defined in Step 1) and variance in the volunteer estimate. This process was repeated 1,000 times and thus, results will be reported as the average volunteer estimate and average bias and variance over the repetitions. All simulations and analyses were conducted in R (2013). See Appendix A for the full code for all simulations and analyses. The specific process of each of these steps will be described in more detail but first, the dataset and study variable will be described.

Dataset

Analyses used data from the 2014 GSS, which is a nationally representative probability sample of the US population, funded by the National Opinion Research Center (NORC). It targets non-institutionalized adults who are at least 18 years old. Conducted since 1972, the GSS monitors trends in attitudes, behaviors, and attributes. The survey uses an address-based sampling frame and since 2002, it has used Computer Assisted Personal Interviewing (CAPI), which is a method employed to help guide the interviewer through the survey questions and to record responses with the use of a computer. The year 2014 was selected because it contained all relevant covariates used in the propensity score models, which will be discussed below. There were 2,538 completed interviews in this year.

Propensity score estimation will not be as accurate if participants do not have data on the full set of covariates, so missing data was handled with a multiple imputation method—Multiple Imputation by Chained Equations (MICE) with recursive partitioning, specifically, random forest (Doove, Van Burren, & Dusseldorp, 2014). With this procedure, random forest develops a prediction model for each of the variables that have missing data in turn using all other variables in the dataset. This prediction model is formed through an algorithm that separates participants into homogenous groups. Random forest does not make assumptions about the distribution of the data, allows for nonlinear relationships between variables, and allows for any type of dependent variable. Imputations are thus random draws of the dependent variable from participants in the same group as those with the missing data on that variable. This method was selected because the variables used in this dataset include continuous, nominal, and categorical and MICE with random forest is flexible when it comes to variable type.

Study Variables

Two study variables were used as outcomes in this study. First, one topic that surveys of the US population often try to estimate is religiosity. While there are many ways of measuring religiosity, the item from the GSS that will be used to measure religiosity is: Respondent considers themselves a religious person (1=Not religious; 2=Slightly religious; 3=Moderately religious; 4=Very religious). Second, another topic that surveys oftentimes ask participants about is political views. On the GSS, this is measured by asking participants how liberal or conservative they consider their political opinions to be (1=Extremely liberal; 2=Liberal; 3=Slightly liberal; 4=Moderate; 5=Slightly conservative; 6=Conservative; 7=Extremely conservative). Religiosity and political attitudes are not only routinely investigated by the GSS, but also by Pew Research Center, the American National Election Studies (ANES), and the National Longitudinal Survey of Youth. Comparing the effectiveness of propensity score weighting with two study variables is important because different types of variables are affected by different types of error (Loosveldt & Sonck, 2008) and thus, propensity score weighting might affect these variables differently.

Step 1: Creating the Pseudo Population and Defining the Population Value

First, the 2,538 respondents from the 2014 GSS were expanded to create the pseudo-population by bootstrapping with simple random sampling. This served as a pseudo target population as well as a pool from which to draw the reference sample. This pseudo population consists of the entire GSS sample plus a random selection of 10% of the GSS sample added to the population until the population size reached 20,192. The percentage in the imputed GSS sample of those with Internet access in their home is 86.33% while this percentage in the pseudo-population is 86.36%. The population value for each of the study variables will thus be the mean for each variable.

Step 2: Creating the Volunteer and Reference Samples

Next, the volunteer sample was created from the pseudo-population. Because the selection of people having Internet access and being a volunteer for a survey is assumed to be not random, this study aims to capture that process with as little simulation as possible. Other methods in previous research have randomly selected participants who have Internet access to create a volunteer web sample (Valliant & Dever, 2011). This method involves simulating the process of volunteering, which is difficult to capture. In the current study, however, subgroups of the created web volunteer sample will resemble subgroups of a real web volunteer sample, following the methods of Lee et al. (Lee, 2004; Lee, 2006; Lee & Valliant, 2009), discussed below. By creating the volunteer web sample for this study to resemble a real web sample, the process of volunteering will not be simulated in a purely random process.

Subgroup proportions from a real volunteer web sample are obtained from a publicly available dataset from the Pew Research Center about cybersecurity knowledge (Pew Research Center, 2016). The sample selection and survey was conducted by the GfK Group using KnowledgePanel, which uses both address-based sampling and random-digit dialing (RDD) to invite participants to panels and take surveys. If participants do not have access to the Internet, they are provided a device that gives them Internet access. Cell proportions are formed from four demographic variables: age, gender, education, and race, shown in Table 3 below. These variables were selected to mirror the work of Lee (2004, 2006) and because Internet access is related to each of these variables (File & Ryan, 2014).

Table 3
Cell Proportions from Pew Research Center (2016)

| | | High School or Less | | Some College or Above | |
|-------|--------|---------------------|----------|-----------------------|----------|
| | | White | Nonwhite | White | Nonwhite |
| 18-39 | Male | 3.31% | 2.46% | 6.16% | 2.55% |
| | Female | 2.27% | 2.74% | 7.96% | 3.79% |

| | | | | | |
|-------|--------|-------|-------|--------|-------|
| 40-64 | Male | 5.21% | 2.65% | 9.66% | 3.60% |
| | Female | 5.59% | 2.84% | 12.70% | 3.41% |
| 65+ | Male | 2.65% | 0.18% | 6.63% | 1.04% |
| | Female | 4.92% | 1.13% | 5.02% | 1.42% |

Members of the volunteer web survey were selected with the probabilities in each cell using `pois.sam` in R, a function created by Lee (2004), to obtain a desired sample size of 1,000. Second, members of the reference survey were selected from those remaining in the pseudo-population with simple random sampling using `ref.sam` in R, a function created by Lee (2004), to obtain a desired sample size of 250.

Step 3: Estimating the Propensity Score

In order to estimate the propensity score, first, the reference sample and the volunteer sample were combined into one dataset. Then, the propensity score for each individual was estimated using logistic regression to predict the probability of volunteering based on a set of covariates, which will be discussed in detail below. As discussed previously, the different logistic regression models were used and compared in order to determine which variable selection method produced the least biased estimate. These regression models included: A. All variables (demographics and webographics). B. Demographics only. C. Variables that are significantly correlated with the outcome and/or having Internet access. D. Demographics and webographics selected using stepwise regression. E. Demographics and webographics selected using lasso regression. Because this process was repeated 1,000 times, this produced different results for Models D and E each time it is ran. Therefore, the list of variables included in Models D and E was saved each time and a frequency table was ran to show how often variables are selected for each of the propensity score models.

Each propensity score model was calculated using the original GSS weights for the reference survey members only, as previous research has shown that the inclusion of this weight will decrease the bias in the estimate (Valliant & Dever, 2011). The volunteer survey members will all have a weight of 1 because in practice, these participants would not have a known selection probability and therefore, no weight assigned.

The goal of estimating the propensity score is to make the volunteer web sample resemble the reference sample, and thus accurately predict estimates in the target population. Therefore, a combination of demographic and nondemographic items will be used as covariates in the model. In their study comparing probability samples and non-probability samples, Yeager et al. (2011) defined primary demographics as “those that were used by some of the survey firms to create weights or to define strata used in the process of selecting people to invite to complete the Internet surveys. Thus, explicit steps were taken by the survey firms to enhance the accuracy of these specific measures in the Internet surveys.” They defined secondary demographics as those “not used to compute weights or to define sampling strata, so no procedures were implemented explicitly to assure their accuracy.”

This study will follow these definitions to categorize primary demographic variables, which will include variables that the GSS used for quota sampling (sex, age, and employment status) and for sampling purposes (region, age, and race). Secondary demographic items will thus be those variables that are factual, but are not used for sampling purposes. The remaining non-demographic variables will be considered “webographic,” based on the discussion in Chapter 2 of the use of these variables.

As Schonlau et al. (2007) suggested, the most important attitudinal or webographic variables to use will be the ones that have the most imbalance between the reference sample and

the volunteer sample after controlling for demographics. In other words, including webographic variables in the propensity score model is meant to account for some of the differences between volunteer samples and the general population beyond demographics, like attitudes and behavior. The Pew Research Center (2015) has found that those willing and able to take surveys online are different from those not using the Internet on a variety of demographic, political, and religious variables. In addition, their Web-only sample differed from the general public mostly on technology-related items. This study will be building off of these ideas to select potential webographics variables to be used in the four types of regression models.

Table 4, below, shows a list of available covariates that can be used in the propensity score models, grouped into categories—primary demographics, secondary demographics, and webographics. As discussed in Chapter 2, AAPOR (2013) provided a definition of webographics items in their report on non-probability samples: “Webographics are attitudinal variables thought to account for the difference between people who do surveys online and those who do not. They generally measure lifestyle issues such as the types of activities people engage and their frequency, media use, attitudes toward privacy, and openness to innovation” (p. 70). Several of the variables in Table 4 reflect this definition. Also among the webographics items selected are items related to technology, politics, and religion—all areas in which Internet and non-Internet populations differ (Pew Research Center, 2015). In addition, items related to religiosity are expected to be highly related to the study variable. Previous research has included items relating to opinions toward ethnic minorities (Lee, 2006) and items relating to knowing gay people (Schonlau et al., 2007) as webographic items. Thus, variables were chosen to reflect these areas as well.

Table 4
Variables Used in the Propensity Score Model

| Variable Name | Type |
|--|----------------------------|
| <i>Primary Demographics</i> | |
| Sex | Nominal (2 categories) |
| Age | Continuous (18-89) |
| Labor force status | Nominal (7 categories) |
| Race | Nominal (3 categories) |
| <i>Secondary Demographics</i> | |
| Marital status | Nominal (5 categories) |
| Number of persons in household | Continuous (1-11) |
| Citizenship status | Nominal (2 categories) |
| Highest year of school completed | Continuous (0-20) |
| Total family income | Ordinal (12 categories) |
| Number of children | Continuous |
| <i>Webographics</i> | |
| Uses computer | Nominal (2 categories) |
| Uses home Internet thru mobile device | Nominal (2 categories) |
| Uses Internet other than email | Nominal (2 categories) |
| Internet use minutes per week | Continuous (0-59) |
| Expressed political views on Internet | Nominal (3 categories) |
| Main source of news is Internet | Nominal (2 categories) |
| Political party affiliation | Nominal (8 categories) |
| Think of self as liberal or conservative | Ordinal (7 categories) |
| Vote in 2012 election | Nominal (3 categories) |
| How often attends religious services | Ordinal (9 categories) |
| How often respondent prays | Ordinal (6 categories) |
| Belongs to a church/religious org. | Categorical (4 categories) |
| Feelings about the Bible | Categorical (4 categories) |
| Spend evening with relatives | Ordinal (7 categories) |
| Spend evening with neighbor | Ordinal (7 categories) |
| Spend evening with friends | Ordinal (7 categories) |
| Spend evening at bar | Ordinal (7 categories) |
| How often volunteer for charity | Ordinal (6 categories) |
| Volunteer in past month | Nominal (2 categories) |
| How often reads newspaper | Ordinal (5 categories) |
| Hours per day watching TV | Continuous (0-24) |
| Subjective class identification | Ordinal (4 categories) |
| How often people take advantage | Ordinal (4 categories) |
| People can be trusted | Ordinal (4 categories) |
| America as melting pot | Categorical (2 categories) |
| Attitudes toward homosexuality | Ordinal (4 categories) |
| Science makes our lives better | Ordinal (4 categories) |

Models C, D, and E will be described in more detail next. For Model C, it is not clear based on previous research whether models should include demographics regardless of their

relationship with the study variable and having Internet access. In addition, it is unclear whether it is more important that variables be significantly related to the study variable as in Lee (2004, 2006) or that they are significantly related to both the study variable and having Internet access as in Fukuda (2011). Therefore, three models based on correlations were included. First, Model C1 included all demographics plus only webographics that are significantly correlated with both the study variable and with having Internet access. Second, Model C2 removed demographic variables that are not significantly correlated with both the study variable and having Internet access. Third, Model C3 included only variables that are significantly correlated with the study variable. There are no guidelines from previous research that indicate how strong these relationships have to be in order to decrease the bias in the estimate, therefore any variable that has a significant correlation will be used in the three models as described.

The covariates used in Model D were selected through backward stepwise regression. In this process, the full model (Model A) is run and then covariates are removed from the model one at a time based on a statistically significant decrease in AIC. The process stops when removing covariates no longer significantly decreases AIC.

The covariates selected for Model E were selected through lasso regression. Lasso regression (Tibshirani, 1996) selects the best subset of the variables by shrinking the beta ($\hat{\beta}$) values for the variables that contribute the least to prediction accuracy, with some betas shrunk completely to zero. This shrinkage is controlled by the tuning parameter (λ), where larger tuning parameters indicate greater shrinkage. The fitting criterion can be written as:

$$(3) \quad e^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j|$$

Step 4: Calculating the Weighted Volunteer Sample Estimate

As discussed in Chapter 2, this study used the inverse of each individual's propensity score ($\frac{1}{\hat{\pi}_k}$) as a weight. The resulting weighted study variable mean for the volunteer sample (S_V) is (Valliant & Dever, 2011):

$$(4) \quad \hat{\bar{y}} = \frac{\sum_{S_V} d_{Vk} y_k / \hat{\pi}_k}{\sum_{S_V} d_{Vk} / \hat{\pi}_k},$$

where:

y_k = the value observed for volunteer sample unit k ,

d_{Vk} = the corresponding base weight for unit k .

The base weight for those in the reference sample is: $(S_R) = N_{\text{pop}}/n_R$, and the base weight for those in the volunteer sample is: $(S_V) = N_{\text{pop}}/n_V$.

Step 5: Determining the Best Model

Several methods were implemented in order to determine the best model. First, calculating the percentage of bias reduction in the estimates after weighting.

The bias can be calculated as (Valliant & Dever, 2011):

$$(5) \quad \text{bias}(\hat{\theta}) = \bar{\hat{\theta}} - \theta,$$

where:

θ = the finite population value;

$\bar{\hat{\theta}}$ = average estimate across the samples.

The percentage of bias reduction ($p.\text{bias}$) can be calculated as (Lee, 2004; Lee & Valliant, 2009):

$$(6) \quad p.\text{bias}(\hat{\theta}^{W.A}) = \left[\frac{|\text{bias}(\hat{\theta}^{W.U})| - |\text{bias}(\hat{\theta}^{W.A})|}{|\text{bias}(\hat{\theta}^{W.U})|} \right] \times 100,$$

where:

$\hat{\theta}^{W.U}$ = the unadjusted estimate;

$\hat{\theta}^{W.A}$ = an adjusted estimate.

As discussed in Chapter 2, calculating the variance is just as important as calculating the bias. One measure of variability that was calculated is the root mean square deviation (RMSD), which is a measure of how deviated the web estimates are from the population value. The RMSD can be calculated as (Lee, 2004):

$$(7) \quad rmsd(\bar{y}^W) = \sqrt{\frac{\sum_{m=1}^M (y_m^W - y^P)^2}{M}},$$

where:

\bar{y}^W = the average unadjusted volunteer sample estimates over all iterations;

M = the number of iterations;

y_m^W = the y estimate for the volunteer sample of the m^{th} iteration;

y^P = the population value.

This formula can thus be applied to the average estimate obtained from each model. In addition, estimates that have smaller RMSD values are less deviated from the population value.

Similarly, the percentage of RMSD reduction can be defined as (Lee, 2004):

$$(8) \quad p.rmsd(\bar{y}^{W.PSA}) = \left[\frac{rmsd(\bar{y}^{W.U}) - rmsd(\bar{y}^{W.PSA})}{rmsd(\bar{y}^{W.U})} \right] * 100,$$

where:

$\bar{y}^{W.PSA}$ = the propensity-score adjusted average volunteer sample estimate;

$\bar{y}^{W.U}$ = the unadjusted average volunteer sample estimate.

Last, another measure of variability that was calculated is the standard error (se), which can be calculated as (Lee, 2004):

$$(9) \quad se(\bar{y}^W) = \sqrt{\frac{\sum_{m=1}^M (y_m^W - \bar{y}^W)^2}{M}},$$

where:

\bar{y}^W = the average unadjusted volunteer sample estimates over all iterations;

M = the number of iterations;

y_m^W = the y estimate for the volunteer sample of the m^{th} iteration.

This formula can thus be applied to the average estimate obtained from each model. This statistic shows how deviated estimates are from the unadjusted volunteer sample estimate.

The best variable selection method will be the regression model specified in Step 3 that has the smallest bias.

Results

Models

As described in the Methods section, the variables used in Models C1-3 were chosen based on correlations with having Internet access and with each of the study variables. Table 5 below shows each of the available covariates and their correlations with having Internet access, religiosity, and political views.

Table 5
Correlations Between Covariates and Internet Access, Religiosity, and Political Views

| Variable Name | Cor(Internet) | Cor(Religiosity) | Cor(Polviews) |
|---|---------------|------------------|---------------|
| Sex | 0.02 | -0.11*** | 0.02 |
| Age | -0.16*** | 0.21*** | 0.11*** |
| Labor force status (working v. not) | 0.17*** | -0.11*** | -0.06** |
| Race (white v. not) | 0.04* | -0.09*** | 0.09*** |
| Marital status (married v. not) | 0.13*** | 0.08*** | 0.10*** |
| No. persons in household | 0.06** | -0.00 | 0.06** |
| Citizen status | 0.11*** | 0.04 | 0.03 |
| Highest year of school completed | 0.30*** | -0.11*** | -0.09*** |
| Total family income | 0.26*** | -0.07*** | 0.03 |
| Number of children | -0.12*** | 0.22*** | 0.14*** |
| Uses computer | 0.48*** | -0.10*** | -0.08*** |
| Uses home Internet thru mobile device | 0.28*** | -0.07*** | -0.02 |
| Uses Internet other than email | 0.10*** | -0.03 | 0.03 |
| Internet use minutes per week | 0.02 | 0.03 | 0.03 |
| Expressed political views on Internet (yes v. no) | 0.13*** | -0.08*** | -0.08*** |
| Main source of news is Internet | 0.17*** | -0.14*** | -0.08*** |
| Political party affiliation | | | |
| Democrat | -0.02 | -0.08*** | -0.40*** |
| Republican | 0.08*** | 0.16*** | 0.44*** |
| Independent | -0.08*** | -0.07*** | -0.02 |

| | | | |
|--|----------|----------|----------|
| Other | 0.04* | -0.03 | 0.01 |
| Think of self as liberal or conservative (higher scores=more conservative) | -0.06** | 0.29*** | --- |
| Vote in 2012 election | 0.14**** | 0.09*** | 0.04* |
| How often attends religious services | -0.01 | 0.60*** | 0.28*** |
| How often respondent prays | -0.05* | 0.60*** | 0.21*** |
| Belongs to a church/religious org. | | | |
| Belongs & participates | 0.01 | 0.44*** | 0.22*** |
| Belongs but doesn't participate | 0.02 | 0.06*** | 0.01 |
| Used to belong | 0.03 | -0.29*** | -0.15*** |
| Never belonged | -0.06** | -0.29*** | -0.11*** |
| Feelings about the Bible | | | |
| Word of god | -0.15*** | 0.39*** | 0.19*** |
| Inspired word | 0.10*** | 0.05* | 0.06** |
| Fables | 0.04* | -0.49*** | -0.28*** |
| Other | 0.03 | -0.06** | -0.06** |
| Religious person (high=more religious) | -0.08*** | --- | 0.29*** |
| Spend evening with relatives | 0.03 | 0.09*** | 0.03 |
| Spend evening with neighbor | -0.02 | 0.06** | 0.00 |
| Spend evening with friends | 0.08*** | -0.05* | -0.03 |
| Spend evening at bar | 0.10*** | -0.20*** | -0.14*** |
| How often volunteer for charity | 0.07*** | 0.10*** | 0.04 |
| Volunteer in past month | 0.11*** | 0.11*** | 0.07*** |
| How often reads newspaper | 0.06** | 0.02 | -0.02 |
| Hours per day watching TV | -0.10*** | 0.03 | -0.02 |
| Subjective class identification | 0.13*** | -0.02 | 0.00 |
| How often people take advantage | 0.03 | -0.05* | -0.03 |

| | | | |
|---|---------|----------|----------|
| (higher=more fair) | | | |
| People can be trusted (higher=more trusting) | 0.03 | -0.00 | -0.01 |
| America as melting pot (yes v. no) | -0.03 | 0.06** | 0.09*** |
| Attitudes toward homosexuality (higher=more accepting) | 0.12*** | -0.39*** | -0.29*** |
| Science makes our lives better | 0.02 | 0.02 | -0.05* |

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

The specific variables included in each model appear in the Table 6 below.

Table 6
Covariates Included in Model C1, Model C2, and Model C3

| Variable | Religiosity | | | Political Views | | |
|---|-------------|---------|---------|-----------------|---------|---------|
| | ModelC1 | ModelC2 | ModelC3 | ModelC1 | ModelC2 | ModelC3 |
| Sex | ✓ | | ✓ | ✓ | | |
| Age | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Labor force status (working v. not) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Race (white v. not) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Marital status (married v. not) | ✓ | ✓ | ✓ | ✓ | ✓ | |
| No. persons in household | ✓ | | | ✓ | ✓ | ✓ |
| Citizen status | ✓ | | | ✓ | | |
| Highest year of school completed | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Total family income | ✓ | ✓ | ✓ | ✓ | | |
| Number of children | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Uses computer | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Uses home Internet thru mobile device | ✓ | ✓ | ✓ | | | |
| Expressed political views on Internet (yes v. no) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Main source of news is Internet | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Political party affiliation | | | | | | |
| Democrat | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Republican | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

| | | | | | | |
|--|---|---|---|---|---|---|
| Independent | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Other (reference group) | | | | | | |
| Think of self as liberal or conservative (higher scores=more conservative) | ✓ | ✓ | ✓ | | | |
| Vote in 2012 election | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| How often attends religious services | | | ✓ | | | ✓ |
| How often respondent prays | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Belongs to a church/religious org. | | | | | | |
| Belongs & participates | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Belongs but doesn't participate | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Used to belong | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Never belonged (reference group) | | | | | | |
| Feelings about the Bible | | | | | | |
| Word of god | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Inspired word | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Fables | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Other (reference group) | | | | | | |
| Religious person (high=more religious) | | | | ✓ | ✓ | ✓ |
| Spend evening with relatives | | | ✓ | | | |
| Spend evening with neighbor | | | ✓ | | | |
| Spend evening with friends | ✓ | ✓ | ✓ | | | |
| Spend evening at bar | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| How often volunteer for charity | ✓ | ✓ | ✓ | | | |

| | | | | | | |
|--|---|---|---|---|---|---|
| Volunteer in past month | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| How often people take advantage (higher=more fair) | | | ✓ | | | |
| America as melting pot (yes v. no) | | | ✓ | | | ✓ |
| Attitudes toward homosexuality (higher=more accepting) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Science makes our lives better | | | | | | ✓ |

In addition, the variables chosen for Model D were selected through backward stepwise regression. Figure 3 below shows the number of times each variable was selected in the resulting model for each of the 1,000 iterations. The two variables selected most often were if the respondent uses the computer and if the respondent uses the web on their mobile device. Other variables that were selected more frequently include years of education, if the respondent is a democrat, and if the respondent is white.

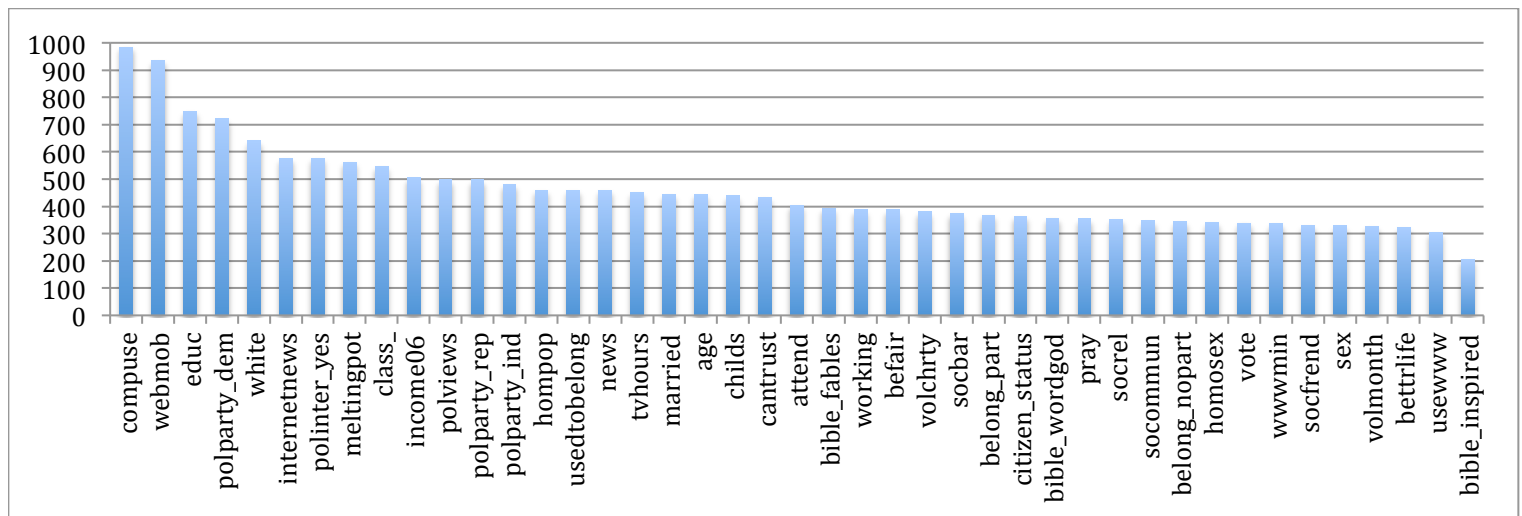


Figure 3. Frequency of selection for each covariate in Model D.

Last, the variables chosen for Model E were selected through lasso regression. The figure below shows the number of times each variable had a beta value not equal to zero, effectively

being “selected,” in the resulting model for each of the 1,000 iterations. The variables that were selected the most often were if the respondent uses the computer, if the respondent uses the web on their mobile device, and years of education. Other variables that were selected more frequently include if the respondent is white, subjective class identification, and income.

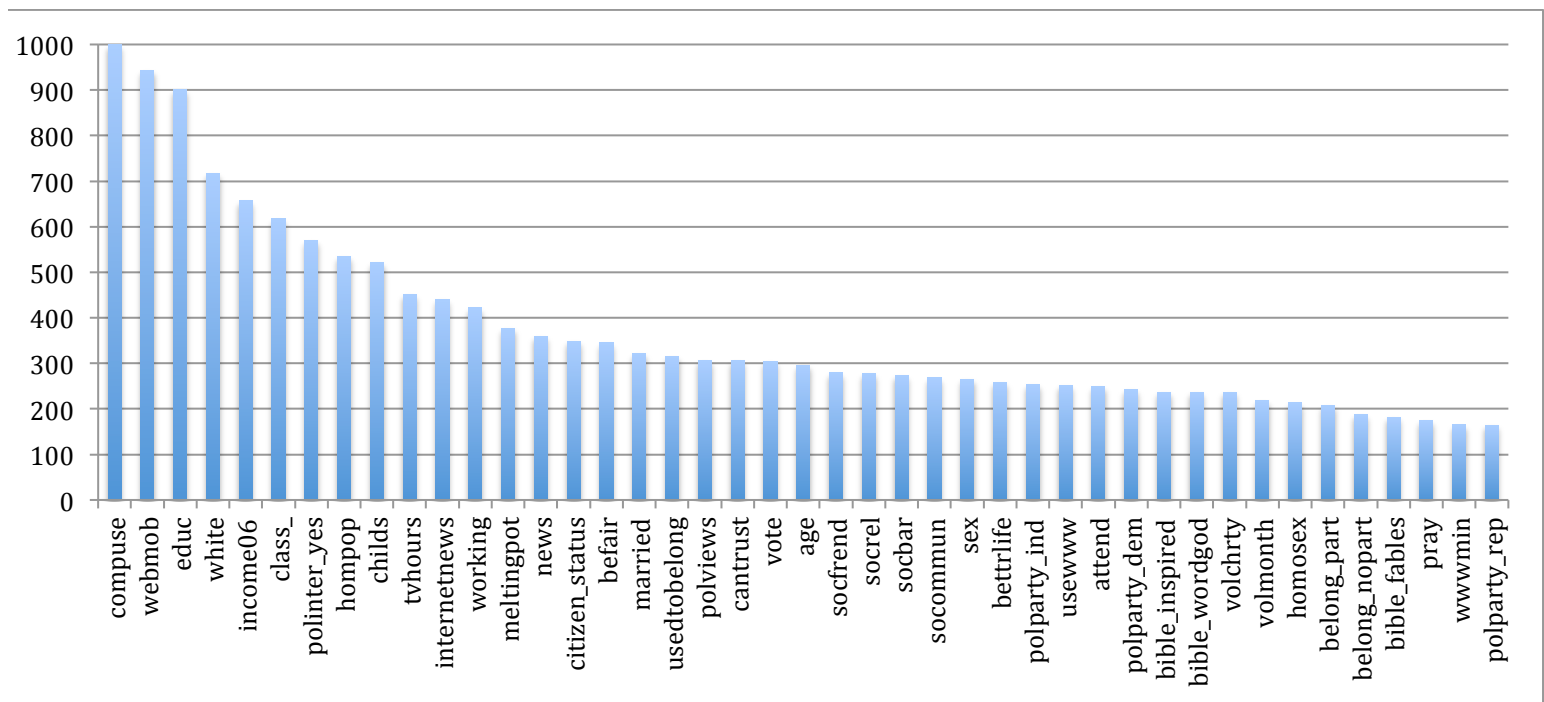


Figure 4. Frequency of selection for each covariate in Model E.

Religiosity Variable

Table 7 below shows the population value, reference sample estimate, unadjusted volunteer sample estimate, and the propensity-score adjusted volunteer estimates from each of the models specified previously for the religiosity variable. The population value was 2.4906. Overall, it is clear that propensity score weighting did not perform well for this variable, indicated by the fact that the largest percent bias reduction was not even 5%. In addition, all of these models except for Model B added variability in the estimates, indicated by negative percent

RMSD reduction. This means that they all produced estimates that were more deviated from the population value than the unadjusted estimate.

Table 7
Results for Religiosity Variable

| | \bar{y} | se | Bias(\bar{y}) | % Bias Reduction | RMSD | % RMSD Reduction |
|------------|-----------|--------|-------------------|------------------|--------|------------------|
| Reference | 2.4651 | 0.0744 | | | | |
| Unadjusted | 2.4341 | 0.0304 | -0.0565 | | 0.0641 | |
| Model A | 2.4344 | 0.0347 | -0.0562 | 0.44 | 0.0661 | -2.96 |
| Model B | 2.4369 | 0.0347 | -0.0536 | 4.97 | 0.0638 | 0.49 |
| Model C1 | 2.4356 | 0.0372 | -0.0549 | 2.67 | 0.0663 | -3.41 |
| Model C2 | 2.4361 | 0.0364 | -0.0545 | 3.42 | 0.0655 | -2.17 |
| Model C3 | 2.4349 | 0.0355 | -0.0557 | 1.28 | 0.0661 | -3.06 |
| Model D | 2.4345 | 0.0357 | -0.0561 | 0.68 | 0.0665 | -3.64 |
| Model E | 2.4076 | 0.0374 | -0.0829 | -46.87 | 0.0871 | -35.72 |

The model that produced the least biased estimate (bias = -0.0536) and the greatest percentage in bias reduction (bias reduction = 4.97%) was Model B, which included demographics variables only. This model also produced the smallest RMSD (0.0638), which means that the estimates from each of the 1,000 iterations were less deviated from the population value. This was the only model to decrease RMSD (RMSD reduction = 0.49%), meaning that this was the only model that produced estimates that were less deviated from the population value than the unadjusted volunteer estimate. Last, Model A (se = 0.0347) and Model B (se = 0.0347) both had the smallest standard errors, indicating that these models produced estimates that were the least deviated from the unadjusted volunteer estimate.

While Model B had the least biased estimate and had the least amount of variability, Model E, which chose variables based on lasso regression, had the most biased estimate (bias = -0.0829). This model actually produced an estimate that was more biased than the unadjusted estimate, indicated by a negative percent in bias reduction (bias reduction = -46.87%). In addition, this model had the largest standard error (se = 0.0374) and the largest RMSD (RMSD =

0.0871), indicating that it produced estimates that were the most varied from both the population value and the unadjusted volunteer estimate.

Political Views Variable

Table 8 below shows the population value, reference sample estimate, unadjusted volunteer sample estimate, and the propensity-score adjusted volunteer estimates from each of the models specified previously for the political views variable. The population value was 4.0761. Overall, it is clear that propensity score weighting was much more effective for this variable than it was for the religiosity variable, indicated by much larger percent bias reduction.

Table 8
Results for Political Views Variable

| | \bar{y} | se | Bias(\bar{y}) | % Bias Reduction | RMSD | % RMSD Reduction |
|------------|-----------|--------|-------------------|------------------|--------|------------------|
| Reference | 4.1050 | 0.1029 | | | | |
| Unadjusted | 4.0685 | 0.0450 | -0.0075 | | 0.0457 | |
| Model A | 4.0784 | 0.0638 | 0.0022 | 69.84 | 0.0630 | -37.95 |
| Model B | 4.0741 | 0.0525 | -0.0019 | 73.78 | 0.0522 | -14.31 |
| Model C1 | 4.0763 | 0.0530 | 0.0001 | 97.92 | 0.0525 | -14.85 |
| Model C2 | 4.0775 | 0.0531 | 0.0013 | 81.84 | 0.0524 | -14.68 |
| Model C3 | 4.0771 | 0.0531 | 0.0009 | 87.02 | 0.0525 | -14.83 |
| Model D | 4.0794 | 0.0544 | 0.0032 | 57.26 | 0.0534 | -16.81 |
| Model E | 4.0762 | 0.0524 | 0.0000 | 98.95 | 0.0518 | -13.47 |

The model that produced the least biased estimate (bias = 0.00) and the largest percent bias reduction (bias reduction = 98.95%) was Model E, which chose variables based on lasso regression. This model also had the smallest standard error (se = 0.0524) and the smallest RMSD (RMSD = 0.0518), which means that this model produced estimates that were the least deviated from both the population value and the unadjusted volunteer estimate. Note that Model C1 (bias = 0.0001), which included all demographics and webographics that were significantly correlated with either Internet access or political views, and Model C3 (bias = 0.0009), which included variables that were significantly correlated with political views, both produced estimates that had

a very small amount of bias. However, both of these models produced estimates that were more varied ($se_{C1} = 0.0530$; $se_{C3} = 0.0531$; $RMSD_{C1} = 0.0525$; $RMSD_{C3} = 0.0525$) than the estimates from Model E, even though their bias was comparable.

The model that produced the most biased estimate (bias = 0.0032) and the smallest percent bias reduction (bias reduction = 57.26%) was Model D, which chose variables based on backwards stepwise regression. The model that produced the most varied estimates was Model A, which included all variables. This model had the largest standard error ($se = 0.0638$) and the largest RMSD ($RMSD = 0.0630$), indicating that it produced estimates that were the most deviated from both the population value and the unadjusted volunteer estimate.

Overall, propensity score weighting was more successful in general for the political views variable than it was for the religiosity variable. The models that produced the least biased and the least varied estimates were Model B (demographics only) for the religiosity variable and Models E (lasso regression) for the political views variable. In addition, Models C and C3 (correlations) produced similarly biased estimates for the political views variable but more varied estimates. Last, the worst performing models were Model E (lasso regression) for the religiosity variable and Model A (all variables) for the political views variable.

Discussion

This study set out to suggest a structure for covariate selection when estimating the propensity score for use as a weighting method for volunteer surveys. Previous research has suggested the best methods of using the propensity score as a weighting method (Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011), but has yet to define a process for selecting covariates used to estimate the propensity score itself. While methods have been suggested for use in observational research, the original use of propensity scores (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Franklin, Eddings, Glynn & Schneeweiss, 2015; Koch, Vock, & Wolfson, 2017; Myers et al., 2011; Shortreed & Ertefaie, 2017; Vansteelandt, Bekaert, & Claeskens, 2010; Westreich et al., 2011), methods for variable selection for propensity scores as a weighting method have been unclear. The majority of the discussion has been focused on the use of webographic variables in these models, whose purpose is to balance the volunteer sample with a representative sample on certain covariates that represent differences in lifestyles and opinions (AAPOR, 2013), but whose effectiveness in decreasing bias in the estimates is unclear (AAPOR, 2013; Duffy et al., 2005; Fukuda, 2011; Lee & Valliant, 2008; Schonlau, van Soest & Kapteyn, 2007; Schonlau et al., 2009).

This study used two study variables in order to compare the effectiveness of propensity score weighting with two variables that are likely to be prone to different sources of error, as discussed previously. It is clear from the results that propensity score weighting was not effective in decreasing the bias in the first study variable, religiosity. While Model B (demographics only) decreased the most bias in the unadjusted estimate, this model decreased 4.97% of the bias in the unadjusted estimate. This is compared to the political views variable, of which the most percent decrease in bias (98.95%) was Model E (lasso regression). This is reflective of previous research

that suggests that one “set model” will not be effective for all study variables (AAPOR, 2013; Fukuda, 2011; Lee & Valliant, 2008). This difference in effectiveness between study variables is present in other research (Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011), with percent reduction in bias for some models as low as 5% and some as high as 94%.

One potential explanation for why results in general from the two study variables were so different might be because the unadjusted estimates were more similar for the religiosity variable than for the political views variable. Propensity score weighting uses information from the reference participants and in the case of the religiosity variable, reference participants did not differ from volunteer participants by very much. Therefore, the propensity score weights would not have an effect on the weighted estimate. Another explanation comes from Schonlau et al. (2009), who have suggested that no weighting scheme will be effective in reducing the bias in the estimate when there is an unobserved characteristic that influences both Internet access and the study variable in a way unrelated to the propensity score model. In other words, if this characteristic is known but not included in the propensity score model or if this characteristic is not known, then weighting will not eliminate the bias in the estimate. In this study, it is possible that there was some unobserved characteristic related to both religiosity and Internet access that was not included in the model and therefore, the propensity-score weighting method was not effective. Because propensity score weighting was more effective for the political views variable than for the religiosity variable, the discussion will focus on results from the political views variable.

To summarize the results for the political views variable, Model E (lasso regression) produced the least biased estimate and produced an estimate that was less varied than the other models. Model C1 (demographics plus webographics significantly related to either Internet

access or political views) and Model C3 (any variable that was significantly correlated with political views) also produced estimates with small biases, but were more varied than Model E.

Although Model E (lasso regression) worked the best for the political views variable, it was the worst model for the religiosity variable. Based on previous literature from observational research, models similar to lasso regression have not been effective because they produce models that are highly predictive of treatment and potentially eliminate confounding variables—those that are related to the study variable through treatment (Koch, Vock, & Wolfson, 2017; Vansteelandt, Bekaert, & Claeskens, 2010). Therefore, observational researchers would not suggest using lasso regression as a variable selection method. This could, however, be a potential explanation for why results were so different between the two variables. It could have been the situation where the lasso regression model ended up selecting more confounding variables for political views than it did for religiosity. However, because the differing results cannot be explained, results from the lasso regression in this study should not be trusted or interpreted.

Therefore, more focus will be put on the results from Model C1 and Model C3 for the political views variable. From these results, it is clear that choosing variables most related to the study variable or to having Internet access is important for propensity score weighting effectiveness. In addition, also including all demographics available was even more effective in reducing the bias in the estimate. Because both of these models had comparable variability in the estimate and because including all demographics in the model improved bias reduction, Model C1 was the best performing model for the political views variable. This is reflective of previous research, which has suggested that the essential variables to include in the model are those that are associated with the participation probability and the study variable (Fukuda, 2011).

One of the research questions posed for this study was: What types of webographic variables included in the propensity score model are most effective at reducing bias in the estimate? From the results of the political views variable, it is clear that including any variable beyond demographics that is significantly related to the study variable or to having Internet access is most important to include in the model. Another research question posed for this study was: What is the best method of selecting these variables for the propensity score model? Because Model C1 was the best performing model for the political views variable, it is suggested that all demographics should be included in the model and therefore, no method needs to be employed. Regarding selection of webographic variables, it is suggested that any variables related to the study variable or to Internet access should be included. However, as discussed in the Methods section, previous research has not specified how strong this relationship needs to be in order to eliminate the most bias in the estimate. In this study, any variable that had a significant correlation with political views or to Internet access was included; however, it is not clear if the model would have been improved if those variables with weak correlations were not included. More research on this topic should investigate this more closely.

In summary, this study investigated the use of webographic variables in the propensity score model when using the propensity score as a weighting method for volunteer surveys. Previous research has shown that the use of webographic variables will depend on two things: (a) their effectiveness in balancing even after controlling for demographics (Duffy et al., 2005; Schonlau, van Soest, & Kapteyn, 2007); and (b) their relationship with the study variable and the inclusion probability (Fukuda, 2011; Lee, 2006; Lee & Valliant, 2008). Based on the results of this study, it is most important to include all demographics and to include webographics that are significantly related to the study variable in question or to having Internet access. However, more

research should be done on how strong these relationships need to be. Results were clear on which models were not effective in reducing the bias in the unadjusted estimate. For both of the variables, Model A (all variables) and Model D (stepwise regression) were unsuccessful in decreasing the bias in the estimate. This is reflective of previous research that showed that models using all available covariates in the propensity model are not effective (Lee, 2004; Lee, 2006) nor are models that are highly predictive of treatment (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006; Koch, Vock, & Wolfson, 2017; Myers et al., 2011; Vansteelandt, Bekaert, & Claeskens, 2010; Westreich et al., 2011). Results from this study should be interpreted with caution because there is not one “set model” that will be effective for all study variables (AAPOR, 2013; Fukuda, 2011; Lee & Valliant, 2008).

Guidelines for Practice

First, as discussed in the literature review, model-based approaches are only effective if all relevant covariates are used in the model (AAPOR, 2013; Lee & Valliant, 2008; Rosenbaum & Rubin, 1983). In this study, propensity score weighting was not effective for the religiosity variable and one possible explanation could be that not all relevant covariates were used in the model. This implies that researchers should understand the relationships between having Internet access, volunteering for surveys, and the study variable in question before attempting to use propensity score weighting. In practice, researchers would not have a population value to compare their results to and therefore, would not know how biased their sample estimate was. Because of this, they would not know if all relevant covariates were used in the propensity score model. Researchers using propensity scores as a weighting method should understand these relationships in order to have an idea of what the relevant covariates are to include in the model and whether or not they have access to them.

Second, another explanation for why propensity score weighting was not effective for the religiosity variable is because the unadjusted estimates from the reference and volunteer samples were very similar. In practice, survey researchers would not have access to the population value and therefore could not determine which variable selection method eliminates the most bias from their unadjusted volunteer estimate. Rather, they would be able to compare the unadjusted estimates from the volunteer and reference samples. Based on results from this study, if the unadjusted estimates from the reference and volunteer groups are similar to each other, then propensity score weighting might not be a good option. In addition, researchers using propensity score weighting in the future would be able to determine whether their adjusted estimates are trustworthy by comparing several different propensity score models, similar to the methods of this study. If the adjusted estimates from the models are all similar to each other, as in the political views variable here, then results would be more trustworthy. However, if the adjusted estimates from the models are different from each other, or at least one model produces an estimate that is very different from the rest of the models, as in the religiosity variable here, then results might not be trustworthy. In summary, it is suggested that researchers first compare the unadjusted estimates and then compare adjusted estimates from several propensity score models to determine if propensity score weighting would produce a trustworthy estimate.

The issue of a reference group. Previous research has asked whether using an existing reference sample, such as a census, is sufficient when calculating the propensity score or if conducting a new reference survey to ensure the inclusion of important covariates is preferable (Lee & Valliant, 2009; Valliant & Dever, 2011). Based on results from this study, it is clear that the answer to this question again depends on the study variable in question. If the study variable in question is more likely to be biased from having volunteers take the survey on the Internet, as

in the political views variable here, then a reference survey that includes related webographics is important to have in order to calculate the propensity score and decrease the bias in the unadjusted estimate. However, if the study variable in question is less likely to be biased from having volunteers take the survey on the Internet, as in the religiosity variable here, then the answer to this question remains unanswered. If all relevant covariates that are needed for the model are present on an existing reference survey, then that would be sufficient; but a researcher must first identify what those relevant covariates are.

Limitations and Suggestions for Future Research

This study has several limitations. First, as stated before, it is unclear why Model E (lasso regression) produced such varied results between the two study variables in question here. More research is needed to investigate this method of variable selection. Variants of lasso regression that take into account not only the relationship between the covariates and the inclusion probability but also between the covariates and the study variable have been suggested in observational research (Koch, Vock, & Wolfson, 2017; Shortreed & Ertefaie, 2017). These methods would be worth investigating for use in developing models for propensity scores as a weighting method.

Second, when selecting covariates in the propensity score model that are significantly related both to having Internet access and to the study variable, previous research provides no guidelines to indicate how strong these relationships have to be in order to decrease the bias in the estimate. In this study, any covariate with a significant relationship was used, regardless of strength. Future research should investigate this topic in more detail and provide suggestions for how strong these relationships should be in order to qualify for inclusion in the propensity score model.

Third, this study used the inverse of each individual's propensity score from each model as a weight. Previous research has had success in using calibration weighting in addition to propensity score weighting with volunteer surveys (Lee, 2004; Lee & Valliant, 2009). This method further weights the volunteer sample to population totals on key covariates. Future research should investigate various covariate selection methods for the propensity score with and without calibration weighting and see if the same conclusions reached here are supported.

Last, it was suggested in this study that the variable selection method used in developing the propensity score model will not only depend on the study variable in question, but it will also depend on how similar or different the unadjusted estimates from the volunteer and reference samples are from each other. Future research should investigate this finding in more detail. It would be worth analyzing study variables that have more different estimates and variables that have more similar estimates than the ones used here in order to see if the same conclusions reached here are supported.

References

- American Association for Public Opinion Research (AAPOR). (2013). *Report of the AAPOR task force on non-probability sampling*. Oakbrook Terrace, IL: Reg Baker, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, & Roger Tourangeau.
- Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, 26, 734-753.
- Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, 74(4), 711-781.
- Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2), 161-188.
- Blumberg, S. J., & Luke, J. V. (2017). *Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2016*. Retrieved from www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201512.pdf.
- Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75(5), 872-888.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149-1156.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24, 295-313.

- Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.
- Couper, M. P. (2001). The promises and perils of web surveys. In A. Westlake, W. Sykes, T. Manners, & M. Riggs (Eds.), *The challenge of the Internet* (35-56). London: Association for Survey Computing.
- Couper, M. P., & Bosnjak, M. (2010). Internet surveys. In J. Wright, P. H. Rossi, & A. B. Anderson (Eds.), *Handbook of survey research* (2nd ed.) (527-550). Bingley, UK: Emerald Group Publishing Limited.
- Danielsson, S. (2004). The propensity score and estimation in nonrandom surveys: An overview. Modern Statistical Survey Methods Project Report No. 18, Department of Statistics, University of Linköping.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis*, 72, 92-104.
- Duffy, B., Smith, K., Terhanian, G., & Bremer, J. (2005). Comparing data from online and face-to-face surveys. *The Market Research Society*, 47(6), 615-639.
- File, T., & Ryan, C. (2014). *Computer and Internet use in the United States: 2013*. American Community Survey Reports, ACS-28. U.S. Census Bureau, Washington, D.C.
- Franklin, J. M., Eddings, W., Glynn, R. J., & Schneeweiss, S. (2015). Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *American Journal of Epidemiology*, 182(7), 651-659.

- Fukuda, M. (2011). Effects of variables in a response propensity score model for survey data adjustment: A simulation study. *Behaviormetrika*, 38(1), 33-61.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646-675.
- Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(2), 81-97.
- Keeter, S., Christian, L. M., Dimock, M., & Gewurtz, D. (2012). *Assessing the representativeness of public opinion surveys*. Retrieved from <http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/>.
- Kim, J. K., & Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38(2), 157-165.
- Koch, B., Vock, D. M., & Wolfson, J. (2017). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*. doi: 10.1111/biom.12736.
- Lee, S. (2004). *Statistical estimation methods in volunteer panel web surveys*. Retrieved from <https://drum.lib.umd.edu/handle/1903/2003>.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2), 329-349.
- Lee, S., & Valliant, R. (2008). Weighting telephone samples using propensity scores. In J. M. Lepowski, C. Tucker, J. M. Brick, E. de Leeuw, L. Japac, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.), *Advances in Telephone Survey Methodology* (170-183). John Wiley & Sons, Inc.

- Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319-343.
- Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, 2(2), 93-105.
- Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, 174(11), 1213-1222.
- Oldendick, R. W., & Link, M. W. (1994). The answering machine generation: Who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58, 264-273.
- Pew Research Center. (2015a). *From telephone to the web: The challenge of mode of interview effects in public opinion polls*. Washington, DC: Scott Keeter & Rachel Weisel.
- Pew Research Center. (2015b). *Coverage error in Internet surveys: Who web-only surveys miss and how that affects results*. Washington, DC: Scott Keeter, Kyley McGeeney, & Rachel Weisel.
- Pew Research Center. (June 17-27, 2016). *Cybersecurity knowledge* [Data file]. Retrieved from <http://www.pewinternet.org/dataset/june-2016-cybersecurity-knowledge/>.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

- Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516-524.
- Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K., Marcus, S., Adams, J., Kan, H., Turner, R., & Berry, S. (2003). A comparison between responses from a propensity-weighted Web survey and an identical RDD survey. *Social Science Computer Review*, 21(10), 1-11.
- Schonlau, M., van Soest, A., & Kapteyn, A. (2007). Are “Webographic” or attitudinal questions useful for adjusting estimates from Web surveys using propensity scoring? *Survey Research Methods*, 1(3), 155-163.
- Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in Web surveys and the use of propensity scores. *Sociological Methods & Research*, 37(3), 291-318.
- Shortreed, S. M. & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4), 1111-1122.
- Taylor, H. (2000). Does internet research work? *Journal of the Market Research Society*, 42(1), 51-63.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288.
- Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-137.
- Vansteelandt, S., Bekaert, M., & Claeskens, G. (2010). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1), 7-30.

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review.

Pharmacoepidemiology and Drug Safety, 13, 841-853.

Westreich, D., Cole, S. R., Jonsson Funk, M., Brookhart, A., & Sturmer, T. (2011). The role of the *c*-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety*, 20, 317-320.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, 75(4), 709-747.

Appendix A

Full R code for project.

```
#####

#Deal with Missing Data#

> library(mice)

> imputed.data <- mice(gss.norecode, print=FALSE, maxit = 20,
  m=10, seed = 24416, method = "rf")

> imputed <- complete(imputed.data)

#####

#Creating Psuedo-Population#

> library(caTools)

> set.seed(sample(1:538, 1))

> gss.sample1 = sample.split(imputed, SplitRatio = .1)

> population = subset(imputed, gss.sample1==TRUE)

> population <- rbind(imputed, population)

> set.seed(sample(1:538, 1))

> gss.sample2 = sample.split(imputed, SplitRatio = .1)

> population2 = subset(imputed, gss.sample2==TRUE)

> population <- rbind(population, population2)

#Repeat until N=20,000

#####
```

```

#Creating volunteer sample#
> wpop = subset(population, population$intrhome==1)
> pois.sam <- function(subpop, pop, ph, str, n)
{
  # Select stratified Poisson sample from pop of size Nh
  # subpop: subpopulation, e.g., web population
  # pop: population, e.g., Entire GSS population
  # ph: vector of proportions in strata that define rates of web
      usage
  # str: column of pop for stratum (can be name or number)
  # n: desired expected total sample size
  h <- subpop[,str]
  N <- nrow(subpop)
  Nh <- table(subpop[,str])
  H <- length(Nh)
  u <- runif(N, min=0, max=1)
  if (any(is.na(h))) {
    stop("stratum vat str missing for some cases. Processing
      stopped.\n")
  }
  if(H != length(ph)) {
    stop("H != length(ph). Processing stopped.\n")
  }
  adjh <- n/ sum(Nh * ph)

```

```

ph <- ph*adjh
ph.pop <-ph[h]
sam <- (u < ph.pop)
sam <- subpop[sam,]
basewgt<-dim(pop)[[1]]/dim(sam)[[1]]
dat <- cbind(sam, basewgt)
}

#####

#Creating reference sample#
> ref.sam <- function (pop, n)
{
  # Select an srs as a reference sample
  # pop: population
  # n: sample size
  N <- nrow(pop)
  sam <- sample(1:N, n, replace = F)
  dat <- pop[sam, ]
  basewgt<-dim(pop)[[1]]/dim(dat)[[1]]
  dat<-cbind(dat, basewgt)
}

#####

#Simulations#

```

```

> library(MASS)
> library(glmnet)

> set.seed(123)

> out.estR <- NULL
> out.estP <- NULL
> modelD.coef <- NULL
> modelE.coef <- NULL
> for(s in 1:1000)
{
  skip <- FALSE
  #skip_ct <- 0
  if (s%%1==0)
#####
  # sample draw
  ref<-ref.sam(population, 250)
  ref$volunteer <- 0
  vol.sam<-pois.sam(wpop, population, ph =
                    c(0.033175355, 0.02464455,
                      0.022748815, 0.027488152,
                      0.052132701, 0.026540284,
                      0.055924171, 0.028436019,
                      0.026540284, 0.001895735,
                      0.0492891, 0.011374408,

```

```

0.061611374, 0.025592417,
0.079620853, 0.037914692,
0.096682464, 0.036018957,
0.127014218, 0.034123223,
0.066350711, 0.01042654,
0.050236967, 0.014218009),
"str", 1000)

vol.sam$volunteer <- 1

vol.sam$WTSS <- 1

# basic estimates

y.popR <- mean(population$relpersn)
y.popP <- mean(population$polviews)
y.wpopR <- mean(vol.sam$relpersn)
y.wpopP <- mean(vol.sam$polviews)
y.refR <- weighted.mean(ref$relpersn, ref$WTSS)
y.refP <- weighted.mean(ref$polviews, ref$WTSS)

#####

# merge reference and web samples

merged<-rbind(ref, vol.sam)

#####

# propensity score adjustment

modelA <- glm(intrhome~sex+age+working+white
+married+hompop+citizen_status+educ+income06+childs

```



```

+compuse+webmob+usewww+wwwmin+polinter_yes+internetnews
+polparty_dem+polparty_rep+polparty_ind+polviews+vote
+attend+pray+belong_part+belong_nopart+usedtobelong+bible_w
ordgod+bible_inspired+bible_fables
+socrel+socommun+socfrend+socbar+volchrty+volmonth+news+tvh
ours+class_

+befair+cantrust+meltingpot+homosex+bettrlfe,

weights = WTSS, family = binomial, data = merged)

modelA.fit <- modelA$fitted.values

merged <- data.frame(merged, modelA.fit)

modelA.weight <- 1/modelA.fit

merged <- data.frame(merged, modelA.weight)


modelB <- glm(intrhome~sex+age+working+white
+married+hompop+citizen_status+educ+income06+childs,

weights = WTSS, family = binomial, data = merged)

modelB.fit <- modelB$fitted.values

merged <- data.frame(merged, modelB.fit)

modelB.weight <- 1/modelB.fit

merged <- data.frame(merged, modelB.weight)


modelCR <- glm(intrhome~sex+age+working+white

```

```

+married+hompop+citizen_status+educ+income06+childs
+compuse+webmob+polinter_yes+internetnews
+polparty_dem+polparty_rep+polparty_ind+polviews+vote
+pray+belong_part+belong_nopart+usedtobelong+bible_wordgod+
bible_inspired+bible_fables
+socfrend+socbar+volchrty+volmonth+homosex,
weights = WTSS, family = binomial, data = merged)
modelCR.fit <- modelCR$fitted.values
merged <- data.frame(merged, modelCR.fit)
modelCR.weight <- 1/modelCR.fit
merged <- data.frame(merged, modelCR.weight)

modelC2R <- glm(intrhome~age+working+white
+married+educ+income06+childs
+compuse+webmob+polinter_yes+internetnews
+polparty_dem+polparty_rep+polparty_ind+polviews+vote
+pray+belong_part+belong_nopart+usedtobelong+bible_wordgod+
bible_inspired+bible_fables
+socfrend+socbar+volchrty+volmonth+homosex,
weights = WTSS, family = binomial, data = merged)
modelC2R.fit <- modelC2R$fitted.values
merged <- data.frame(merged, modelC2R.fit)
modelC2R.weight <- 1/modelC2R.fit

```

```

merged <- data.frame(merged, modelC2R.weight)

modelC3R <- glm(intrhome~sex+age+working+white
               +married+educ+income06+childs
               +compuse+webmob+polinter_yes+internetnews
               +polparty_dem+polparty_rep+polparty_ind+polviews+vote
               +attend+pray+belong_part+belong_nopart+usedtobelong+bible_w
               ordgod+bible_inspired+bible_fables
               +socrel+socommun+socfrend+socbar+volchrtty+volmonth
               +befair+meltingpot+homosex,
               weights = WTSS, family = binomial, data = merged)

modelC3R.fit <- modelC3R$fitted.values

merged <- data.frame(merged, modelC3R.fit)

modelC3R.weight <- 1/modelC3R.fit

merged <- data.frame(merged, modelC3R.weight)

modelCP <- glm(intrhome~sex+age+working+white
               +married+hompop+citizen_status+educ+income06+childs
               +compuse+polinter_yes+internetnews
               +polparty_dem+polparty_rep+polparty_ind+vote
               +pray+belong_part+belong_nopart+usedtobelong+bible_wordgod+
               bible_inspired+bible_fables+relpersn
               +socbar+volmonth+homosex,
               weights = WTSS, family = binomial, data = merged)

```

```

modelCP.fit <- modelCP$fitted.values
merged <- data.frame(merged, modelCP.fit)
modelCP.weight <- 1/modelCP.fit
merged <- data.frame(merged, modelCP.weight)

modelC2P <- glm(intrhome~age+working+white
  +married+hompop+educ+childs
  +compuse+polinter_yes+internetnews
  +polparty_dem+polparty_rep+polparty_ind+vote
  +pray+belong_part+belong_nopart+usedtobelong+bible_wordgod+
  bible_inspired+bible_fables+relpersn
  +socbar+volmonth+homosex,
  weights = WTSS, family = binomial, data = merged)
modelC2P.fit <- modelC2P$fitted.values
merged <- data.frame(merged, modelC2P.fit)
modelC2P.weight <- 1/modelC2P.fit
merged <- data.frame(merged, modelC2P.weight)

modelC3P <- glm(intrhome~age+working+white+hompop+educ+childs
  +compuse+polinter_yes+internetnews
  +polparty_dem+polparty_rep+polparty_ind+vote
  +attend+pray+belong_part+belong_nopart+usedtobelong
  +bible_wordgod+bible_inspired+bible_fables+relpersn
  +socbar+volmonth+meltingpot+homosex+bettrlife,

```

```

    weights = WTSS, family = binomial, data = merged)

modelC3P.fit <- modelC3P$fitted.values

merged <- data.frame(merged, modelC3P.fit)

modelC3P.weight <- 1/modelC3P.fit

merged <- data.frame(merged, modelC3P.weight)


modelD <- stepAIC(modelA, direction = "backward", trace = 0)

modelD.fit <- modelD$fitted.values

merged <- data.frame(merged, modelD.fit)

modelD.weight <- 1/modelD.fit

merged <- data.frame(merged, modelD.weight)

modelD.iteration.coef <- as.matrix(coef(modelD))

modelD.coef <- rbind(modelD.coef, modelD.iteration.coef)


attach(merged)

x <- as.matrix(data.frame(sex, age, working, white, married,
    hompop, citizen_status, educ, income06, childs, compuse,
    webmob, usewww, wwwmin, polinter_yes, internetnews,
    polparty_dem, polparty_rep, polparty_ind, polviews, vote,
    attend, pray, belong_part, belong_nopart, usedtobelong,
    bible_wordgod, bible_inspired, bible_fables, socrel,
    socommun, socfrend, socbar, volchrty, volmonth, news,
    tvhours, class_, befair, cantrust, meltingpot, homosex,
    bettrlfe))

```

```

modelE <- glmnet(x, y=as.factor(intrhome), alpha = 1, family =
  "binomial", weights = WTSS)
cv.modelE <- cv.glmnet(x, y=intrhome, alpha=1)
best.lambda <- cv.modelE$lambda.min
modelE.fit <- predict(modelE, newx = x, type = "response",
  s=best.lambda)
colnames(modelE.fit) <- "modelE.fit"
merged <- data.frame(merged, modelE.fit)
modelE.weight <- 1/modelE.fit
colnames(modelE.weight) <- "modelE.weight"
merged <- data.frame(merged, modelE.weight)
detach(merged)
modelE.iteration.coef <- as.matrix(coef(modelE,
  s=best.lambda))
modelE.coef <- rbind(modelE.coef, modelE.iteration.coef)

#drop reference sample
vol.weight <- merged[251:nrow(vol.sam),]

#adjusted estimates-religiosity
vol.weight$numerator.modelAR <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelA.wei
  ght

```

```

vol.weight$denominator.modelAR <-
  vol.weight$basewgt*vol.weight$modelA.weight
sum.numerator.modelAR <- sum(vol.weight$numerator.modelAR)
sum.denominator.modelAR <- sum(vol.weight$denominator.modelAR)
y.modelAR <- sum.numerator.modelAR/sum.denominator.modelAR

vol.weight$numerator.modelBR <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelB.
  weight
vol.weight$denominator.modelBR <-
  vol.weight$basewgt*vol.weight$modelB.weight
sum.numerator.modelBR <- sum(vol.weight$numerator.modelBR)
sum.denominator.modelBR <- sum(vol.weight$denominator.modelBR)
y.modelBR <- sum.numerator.modelBR/sum.denominator.modelBR

vol.weight$numerator.modelCR <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelCR.w
  eight
vol.weight$denominator.modelCR <-
  vol.weight$basewgt*vol.weight$modelCR.weight
sum.numerator.modelCR <- sum(vol.weight$numerator.modelCR)
sum.denominator.modelCR <- sum(vol.weight$denominator.modelCR)
y.modelCR <- sum.numerator.modelCR/sum.denominator.modelCR

```

```

vol.weight$enumerator.modelC2R <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelC2R.
  weight
vol.weight$denominator.modelC2R <-
  vol.weight$basewgt*vol.weight$modelC2R.weight
sum.numerator.modelC2R <- sum(vol.weight$enumerator.modelC2R)
sum.denominator.modelC2R <-
  sum(vol.weight$denominator.modelC2R)
y.modelC2R <- sum.numerator.modelC2R/sum.denominator.modelC2R

vol.weight$enumerator.modelC3R <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelC3R.
  weight
vol.weight$denominator.modelC3R <-
  vol.weight$basewgt*vol.weight$modelC3R.weight
sum.numerator.modelC3R <- sum(vol.weight$enumerator.modelC3R)
sum.denominator.modelC3R <-
  sum(vol.weight$denominator.modelC3R)
y.modelC3R <- sum.numerator.modelC3R/sum.denominator.modelC3R

vol.weight$enumerator.modelDR <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelD.we
  ight

```



```

vol.weight$denominator.modelDR <-
  vol.weight$basewgt*vol.weight$modelD.weight
sum.numerator.modelDR <- sum(vol.weight$numerator.modelDR)
sum.denominator.modelDR <- sum(vol.weight$denominator.modelDR)
y.modelDR <- sum.numerator.modelDR/sum.denominator.modelDR

vol.weight$numerator.modelER <-
  vol.weight$basewgt*vol.weight$relpersn*vol.weight$modelE.
  ight
vol.weight$denominator.modelER <-
  vol.weight$basewgt*vol.weight$modelE.weight
sum.numerator.modelER <- sum(vol.weight$numerator.modelER)
sum.denominator.modelER <- sum(vol.weight$denominator.modelER)
y.modelER <- sum.numerator.modelE/sum.denominator.modelER

#adjusted estimates-polviews
vol.weight$numerator.modelAP <-
  vol.weight$basewgt*vol.weight$polviews*vol.weight$modelA.
  ight
vol.weight$denominator.modelAP <-
  vol.weight$basewgt*vol.weight$modelA.weight
sum.numerator.modelAP <- sum(vol.weight$numerator.modelAP)
sum.denominator.modelAP <- sum(vol.weight$denominator.modelAP)
y.modelAP <- sum.numerator.modelAP/sum.denominator.modelAP

```

```

vol.weight$enumerator.modelBP <-
  vol.weight$basewgt*vol.weight$polviews*vol.weight$modelB.weight
vol.weight$denominator.modelBP <-
  vol.weight$basewgt*vol.weight$modelB.weight
sum.numerator.modelBP <- sum(vol.weight$enumerator.modelBP)
sum.denominator.modelBP <- sum(vol.weight$denominator.modelBP)
y.modelBP <- sum.numerator.modelBP/sum.denominator.modelBP

vol.weight$enumerator.modelCP <-
  vol.weight$basewgt*vol.weight$polviews*vol.weight$modelCP.weight
vol.weight$denominator.modelCP <-
  vol.weight$basewgt*vol.weight$modelCP.weight
sum.numerator.modelCP <- sum(vol.weight$enumerator.modelCP)
sum.denominator.modelCP <- sum(vol.weight$denominator.modelCP)
y.modelCP <- sum.numerator.modelCP/sum.denominator.modelCP

vol.weight$enumerator.modelC2P <-
  vol.weight$basewgt*vol.weight$polviews*vol.weight$modelC2P.weight
vol.weight$denominator.modelC2P <-
  vol.weight$basewgt*vol.weight$modelC2P.weight

```

```

sum.numerator.modelC2P <- sum(vol.weight$numerator.modelC2P)
sum.denominator.modelC2P <-
    sum(vol.weight$denominator.modelC2P)
y.modelC2P <- sum.numerator.modelC2P/sum.denominator.modelC2P

vol.weight$numerator.modelC3P <-
    vol.weight$basewgt*vol.weight$polviews*vol.weight$modelC3P.
    weight
vol.weight$denominator.modelC3P <-
    vol.weight$basewgt*vol.weight$modelC3P.weight
sum.numerator.modelC3P <- sum(vol.weight$numerator.modelC3P)
sum.denominator.modelC3P <-
    sum(vol.weight$denominator.modelC3P)
y.modelC3P <- sum.numerator.modelC3P/sum.denominator.modelC3P

vol.weight$numerator.modelDP <-
    vol.weight$basewgt*vol.weight$polviews*vol.weight$modelD.we
    ight
vol.weight$denominator.modelDP <-
    vol.weight$basewgt*vol.weight$modelD.weight
sum.numerator.modelDP <- sum(vol.weight$numerator.modelDP)
sum.denominator.modelDP <- sum(vol.weight$denominator.modelDP)
y.modelDP <- sum.numerator.modelDP/sum.denominator.modelDP

```

```

vol.weight$numerator.modelEP <-
  vol.weight$basewgt*vol.weight$polviews*vol.weight$modelE.wei
  ight
vol.weight$denominator.modelEP <-
  vol.weight$basewgt*vol.weight$modelE.weight
sum.numerator.modelEP <- sum(vol.weight$numerator.modelEP)
sum.denominator.modelEP <- sum(vol.weight$denominator.modelEP)
y.modelEP <- sum.numerator.modelEP/sum.denominator.modelEP
#####
# bind all estimates into y.est
y.estR <- cbind (y.popR,
                y.wpopR,
                y.refR,
                y.modelAR,
                y.modelBR,
                y.modelCR,
                y.modelC2R,
                y.modelC3R,
                y.modelDR,
                y.modelER)
out.estR <- rbind(out.estR, y.estR)

y.estP <- cbind (y.popP,
                y.wpopP,

```

```

        y.refP,
        y.modelAP,
        y.modelBP,
        y.modelCP,
        y.modelC2P,
        y.modelC3P,
        y.modelDP,
        y.modelEP)

    out.estP <- rbind(out.estP, y.estP)
} # end of s loop

write.table(out.estR, "/Users/rachelstenger/Documents/Grad
    School/Thesis/resultsR", col.names = TRUE, row.names =
    FALSE)

write.table(out.estP, "/Users/rachelstenger/Documents/Grad
    School/Thesis/resultsP", col.names = TRUE, row.names =
    FALSE)

write.table(modelD.coef, "/Users/rachelstenger/Documents/Grad
    School/Thesis/modelD", col.names = TRUE, row.names = TRUE)
write.table(modelE.coef, "/Users/rachelstenger/Documents/Grad
    School/Thesis/modelE", col.names = TRUE, row.names = TRUE)

#####

#Average estimates#

```

```

> y.popR <- mean(population$relpersn)
> y.wpopR <- mean(resultsr$y.wpopR)
> y.refR <- mean(resultsr$y.refR)
> y.modelAR <- mean(resultsr$y.modelAR)
> y.modelBR <- mean(resultsr$y.modelBR)
> y.modelCR <- mean(resultsr$y.modelCR)
> y.modelC2R <- mean(resultsr$y.modelC2R)
> y.modelC3R <- mean(resultsr$y.modelC3R)
> y.modelDR <- mean(resultsr$y.modelDR)
> y.modelER <- mean(resultsr$y.modelER)

> y.popP <- mean(population$polviews)
> y.wpopP <- mean(resultsp$y.wpopP)
> y.refP <- mean(resultsp$y.refP)
> y.modelAP <- mean(resultsp$y.modelAP)
> y.modelBP <- mean(resultsp$y.modelBP)
> y.modelCP <- mean(resultsp$y.modelCP)
> y.modelC2P <- mean(resultsp$y.modelC2P)
> y.modelC3P <- mean(resultsp$y.modelC3P)
> y.modelDP <- mean(resultsp$y.modelDP)
> y.modelEP <- mean(resultsp$y.modelEP)

#####

#Bias in each estimate and percent bias reduction#

```

```

> bias.wpopR <- y.wpopR-y.popR
> bias.modelAR <- y.modelAR-y.popR
> bias.modelBR <- y.modelBR-y.popR
> bias.modelCR <- y.modelCR-y.popR
> bias.modelC2R <- y.modelC2R-y.popR
> bias.modelC3R <- y.modelC3R-y.popR
> bias.modelDR <- y.modelDR-y.popR
> bias.modelER <- y.modelER-y.popR

> bias.wpopP <- y.wpopP-y.popP
> bias.modelAP <- y.modelAP-y.popP
> bias.modelBP <- y.modelBP-y.popP
> bias.modelCP <- y.modelCP-y.popP
> bias.modelC2P <- y.modelC2P-y.popP
> bias.modelC3P <- y.modelC3P-y.popP
> bias.modelDP <- y.modelDP-y.popP
> bias.modelEP <- y.modelEP-y.popP

> p.bias.modelAR <- ((abs(bias.wpopR)
abs(bias.modelAR))/abs(bias.wpopR))*100
> p.bias.modelBR <- ((abs(bias.wpopR)-
abs(bias.modelBR))/abs(bias.wpopR))*100
> p.bias.modelCR <- ((abs(bias.wpopR)-
abs(bias.modelCR))/abs(bias.wpopR))*100

```

```
> p.bias.modelC2R <- ((abs(bias.wpopR) -  
abs(bias.modelC2R))/abs(bias.wpopR))*100  
  
> p.bias.modelC3R <- ((abs(bias.wpopR) -  
abs(bias.modelC3R))/abs(bias.wpopR))*100  
  
> p.bias.modelDR <- ((abs(bias.wpopR) -  
abs(bias.modelDR))/abs(bias.wpopR))*100  
  
> p.bias.modelER <- ((abs(bias.wpopR) -  
abs(bias.modelER))/abs(bias.wpopR))*100  
  
  
> p.bias.modelAP <- ((abs(bias.wpopP) -  
abs(bias.modelAP))/abs(bias.wpopP))*100  
  
> p.bias.modelBP <- ((abs(bias.wpopP) -  
abs(bias.modelBP))/abs(bias.wpopP))*100  
  
> p.bias.modelCP <- ((abs(bias.wpopP) -  
abs(bias.modelCP))/abs(bias.wpopP))*100  
  
> p.bias.modelC2P <- ((abs(bias.wpopP) -  
abs(bias.modelC2P))/abs(bias.wpopP))*100  
  
> p.bias.modelC3P <- ((abs(bias.wpopP) -  
abs(bias.modelC3P))/abs(bias.wpopP))*100  
  
> p.bias.modelDP <- ((abs(bias.wpopP) -  
abs(bias.modelDP))/abs(bias.wpopP))*100  
  
> p.bias.modelEP <- ((abs(bias.wpopP) -  
abs(bias.modelEP))/abs(bias.wpopP))*100
```



```
#####
```

```
#RMSD#
```

```
> resultsr$rmsd.wpopR <- (resultsr$y.wpopR-resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelAR <- (resultsr$y.modelAR-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelBR <- (resultsr$y.modelBR-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelCR <- (resultsr$y.modelCR-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelC2R <- (resultsr$y.modelC2R-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelC3R <- (resultsr$y.modelC3R-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelDR <- (resultsr$y.modelDR-  
resultsr$y.popR)^2
```

```
> resultsr$rmsd.modelER <- (resultsr$y.modelER-  
resultsr$y.popR)^2
```

```
> resultsp$rmsd.wpopP <- (resultsp$y.wpopP-resultsp$y.popP)^2
```

```
> resultsp$rmsd.modelAP <- (resultsp$y.modelAP-  
resultsp$y.popP)^2
```

```
> resultsp$rmsd.modelBP <- (resultsp$y.modelBP-  
resultsp$y.popP)^2
```

```

> resultsp$rmsd.modelCP <- (resultsp$y.modelCP-
resultsp$y.popP)^2
> resultsp$rmsd.modelC2P <- (resultsp$y.modelC2P-
resultsp$y.popP)^2
> resultsp$rmsd.modelC3P <- (resultsp$y.modelC3P-
resultsp$y.popP)^2
> resultsp$rmsd.modelDP <- (resultsp$y.modelDP-
resultsp$y.popP)^2
> resultsp$rmsd.modelEP <- (resultsp$y.modelEP-
resultsp$y.popP)^2

> rmsd.wpopR <- sqrt((sum(resultsr$rmsd.wpopR))/1000)
> rmsd.modelAR <- sqrt((sum(resultsr$rmsd.modelAR))/1000)
> rmsd.modelBR <- sqrt((sum(resultsr$rmsd.modelBR))/1000)
> rmsd.modelCR <- sqrt((sum(resultsr$rmsd.modelCR))/1000)
> rmsd.modelC2R <- sqrt((sum(resultsr$rmsd.modelC2R))/1000)
> rmsd.modelC3R <- sqrt((sum(resultsr$rmsd.modelC3R))/1000)
> rmsd.modelDR <- sqrt((sum(resultsr$rmsd.modelDR))/1000)
> rmsd.modelER <- sqrt((sum(resultsr$rmsd.modelER))/1000)

> rmsd.wpopP <- sqrt((sum(resultsp$rmsd.wpopP))/1000)
> rmsd.modelAP <- sqrt((sum(resultsp$rmsd.modelAP))/1000)
> rmsd.modelBP <- sqrt((sum(resultsp$rmsd.modelBP))/1000)
> rmsd.modelCP <- sqrt((sum(resultsp$rmsd.modelCP))/1000)

```

```

> rmsd.modelC2P <- sqrt((sum(resultsp$rmsd.modelC2P))/1000)
> rmsd.modelC3P <- sqrt((sum(resultsp$rmsd.modelC3P))/1000)
> rmsd.modelDP <- sqrt((sum(resultsp$rmsd.modelDP))/1000)
> rmsd.modelEP <- sqrt((sum(resultsp$rmsd.modelEP))/1000)

> p.rmsd.modelAR <- ((rmsd.wpopR-rmsd.modelAR)/rmsd.wpopR)*100
> p.rmsd.modelBR <- ((rmsd.wpopR-rmsd.modelBR)/rmsd.wpopR)*100
> p.rmsd.modelCR <- ((rmsd.wpopR-rmsd.modelCR)/rmsd.wpopR)*100
> p.rmsd.modelC2R <- ((rmsd.wpopR-rmsd.modelC2R)/rmsd.wpopR)*100
> p.rmsd.modelC3R <- ((rmsd.wpopR-rmsd.modelC3R)/rmsd.wpopR)*100
> p.rmsd.modelDR <- ((rmsd.wpopR-rmsd.modelDR)/rmsd.wpopR)*100
> p.rmsd.modelER <- ((rmsd.wpopR-rmsd.modelER)/rmsd.wpopR)*100

> p.rmsd.modelAP <- ((rmsd.wpopP-rmsd.modelAP)/rmsd.wpopP)*100
> p.rmsd.modelBP <- ((rmsd.wpopP-rmsd.modelBP)/rmsd.wpopP)*100
> p.rmsd.modelCP <- ((rmsd.wpopP-rmsd.modelCP)/rmsd.wpopP)*100
> p.rmsd.modelC2P <- ((rmsd.wpopP-rmsd.modelC2P)/rmsd.wpopP)*100
> p.rmsd.modelC3P <- ((rmsd.wpopP-rmsd.modelC3P)/rmsd.wpopP)*100
> p.rmsd.modelDP <- ((rmsd.wpopP-rmsd.modelDP)/rmsd.wpopP)*100
> p.rmsd.modelEP <- ((rmsd.wpopP-rmsd.modelEP)/rmsd.wpopP)*100

#####

#Standard Error#

> resultsr$se.wpopR <- (resultsr$y.wpopR-y.wpopR)^2

```

```

> resultsr$se.refR <- (resultsr$y.refR-y.wpopR)^2
> resultsr$se.modelAR <- (resultsr$y.modelAR-y.wpopR)^2
> resultsr$se.modelBR <- (resultsr$y.modelBR-y.wpopR)^2
> resultsr$se.modelCR <- (resultsr$y.modelCR-y.wpopR)^2
> resultsr$se.modelC2R <- (resultsr$y.modelC2R-y.wpopR)^2
> resultsr$se.modelC3R <- (resultsr$y.modelC3R-y.wpopR)^2
> resultsr$se.modelDR <- (resultsr$y.modelDR-y.wpopR)^2
> resultsr$se.modelER <- (resultsr$y.modelER-y.wpopR)^2

> se.wpopR <- sqrt(sum(resultsr$se.wpopR)/1000)
> se.refR <- sqrt(sum(resultsr$se.refR)/1000)
> se.modelAR <- sqrt(sum(resultsr$se.modelAR)/1000)
> se.modelBR <- sqrt(sum(resultsr$se.modelBR)/1000)
> se.modelCR <- sqrt(sum(resultsr$se.modelCR)/1000)
> se.modelC2R <- sqrt(sum(resultsr$se.modelC2R)/1000)
> se.modelC3R <- sqrt(sum(resultsr$se.modelC3R)/1000)
> se.modelDR <- sqrt(sum(resultsr$se.modelDR)/1000)
> se.modelER <- sqrt(sum(resultsr$se.modelER)/1000)

> resultsp$se.wpopP <- (resultsp$y.wpopP-y.wpopP)^2
> resultsp$se.refP <- (resultsp$y.refP-y.wpopP)^2
> resultsp$se.modelAP <- (resultsp$y.modelAP-y.wpopP)^2
> resultsp$se.modelBP <- (resultsp$y.modelBP-y.wpopP)^2
> resultsp$se.modelCP <- (resultsp$y.modelCP-y.wpopP)^2

```

```
> resultsp$se.modelC2P <- (resultsp$y.modelC2P-y.wpopP)^2
> resultsp$se.modelC3P <- (resultsp$y.modelC3P-y.wpopP)^2
> resultsp$se.modelDP <- (resultsp$y.modelDP-y.wpopP)^2
> resultsp$se.modelEP <- (resultsp$y.modelEP-y.wpopP)^2

> se.wpopP <- sqrt(sum(resultsp$se.wpopP)/1000)
> se.refP <- sqrt(sum(resultsp$se.refP)/1000)
> se.modelAP <- sqrt(sum(resultsp$se.modelAP)/1000)
> se.modelBP <- sqrt(sum(resultsp$se.modelBP)/1000)
> se.modelCP <- sqrt(sum(resultsp$se.modelCP)/1000)
> se.modelC2P <- sqrt(sum(resultsp$se.modelC2P)/1000)
> se.modelC3P <- sqrt(sum(resultsp$se.modelC3P)/1000)
> se.modelDP <- sqrt(sum(resultsp$se.modelDP)/1000)
> se.modelEP <- sqrt(sum(resultsp$se.modelEP)/1000)
```