

ABSTRACT

Title of dissertation: SELECTION BIAS IN
NONPROBABILITY SURVEYS: A
CAUSAL INFERENCE APPROACH

Andrew William Mercer
Doctor of Philosophy, 2018

Dissertation directed by: Professor Frauke Kreuter
Joint Program in Survey Methodology

Many in the survey research community have expressed concern at the growing popularity of nonprobability surveys. The absence of random selection prompts justified concerns about self-selection producing biased results and means that traditional, design-based estimation is inappropriate. The Total Survey Error (TSE) paradigm's designations of selection bias as attributable to undercoverage or nonresponse are not especially helpful for nonprobability surveys as they are based on an implicit assumption that selection and inferences rely on randomization.

This dissertation proposes an alternative classification for sources of selection bias for nonprobability surveys based on principles borrowed from the field of causal inference. The proposed typology describes selection bias in terms of the three conditions that are required for a statistical model to correct or explain systematic differences between a realized sample and the target

population: exchangeability, positivity, and composition. We describe the parallels between causal and survey inference and explain how these three sources of bias operate in both probability and nonprobability survey samples. We then provide a critical review of current practices in nonprobability data collection and estimation viewed through the lens of the causal bias framework.

Next, we show how net selection bias can be decomposed into separate additive components associated with exchangeability, positivity, and composition respectively. Using 10 parallel nonprobability surveys from different sources, we estimate these components for six measures of civic engagement using the 2013 Current Population Survey Civic Engagement Supplement as a reference sample. We find that a large majority of the bias can be attributed to a lack of exchangeability.

Finally, using the same six measures of civic engagement, we compare the performance of four approaches to nonprobability estimation based on Bayesian additive regression trees. These are propensity weighting (PW), outcome regression (OR), and two types of doubly-robust estimators: outcome regression with a residual bias correction (OR-RBC) and outcome regression with a propensity score covariate (OR-PSC). We find that OR-RBC tends to have the lowest bias, variance, and RMSE, with PW only slightly worse on all three measures.

SELECTION BIAS IN NONPROBABILITY SURVEYS: A
CAUSAL INFERENCE APPROACH

by

Andrew William Mercer

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Frauke Kreuter, Chair/Advisor
Professor J. Michael Brick
Professor Michael Elliott
Dr. Scott Keeter
Professor Michael Rendall

© Copyright by
Andrew William Mercer
2018

Acknowledgments

First, I would like to express my gratitude to my advisor, Frauke Kreuter, for her guidance and mentorship. She was the first to suggest that I pursue a PhD in survey methodology, and our work together has helped me become a better scholar, speaker, and research professional. I wish to thank my committee – Scott Keeter, Mike Brick, and Mike Elliott – whose help, suggestions, and feedback were always on point. The diversity of their perspectives helped shape this work into something both rigorous and practical. Liz Stuart also deserves special thanks for sharing her expertise in causal inference and helping turn Chapter 2 into a scholarly publication. I am grateful to my colleagues at Pew Research Center, especially Courtney Kennedy, whose support both in terms of data and time made this dissertation possible. I am similarly grateful to the faculty, staff, and fellow students at the Joint Program in Survey Methodology for everything I have learned, and for the community of survey researchers that has developed in JPSM’s orbit over the years. To my parents and siblings, I am thankful for their unwavering confidence in my eventual success. Most of all, I am grateful to Kris, my partner in everything, who’s unflagging support and encouragement kept me going through the long and sometimes overwhelming task of completing this dissertation. This accomplishment would not have been possible without her.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Tables	v
List of Figures	vii
1 Introduction	1
1.1 Problems with the Total Survey Error framework	6
1.2 A framework inspired by causal inference	10
2 A causal inference framework for selection bias	15
2.1 Randomization and unbiased inference in experiments and surveys	17
2.2 Extending the Framework to Non-Random Samples	21
2.3 Mechanics of Selection Bias in Surveys	23
2.4 Current Practices for Managing Bias in Online, Nonprobability Surveys	27
2.5 Discussion	41
3 Decomposing selection bias in nonprobability surveys	43
3.1 Survey estimates and selection bias	46
3.2 Bias decomposition	53
3.3 Estimating selection bias components on measures of civic en- gagement	57
3.4 Results	65
3.5 Discussion	68
4 Doubly-robust inference for nonprobability surveys with BART	73
4.1 Alternative approaches to nonprobability survey inference . . .	78
4.2 Results	87
4.3 Discussion	97
5 Conclusion	101
5.1 Next steps	103
5.2 Revisiting Total Survey Error	105
A Question wording	109
B Variable coding	113

C Code	117
Bibliography	127

List of Tables

3.1	Survey field dates and sample sizes	58
4.1	Average estimator performance on bias, variance, and RMSE.	88

List of Figures

1.1	Components of error under the Total Survey Error framework	7
3.1	Mean signed bias by outcome averaged over all samples. Bars depict 95% credibility intervals.	65
3.2	Mean absolute bias for samples averaged over all six outcome variables. Bars depict 95% credibility intervals.	66
3.3	Scatterplot of average absolute bias due to exchangeability and composition for all samples. Estimates are averaged over all outcome variables.	67
3.4	Bias components for individual questions across samples. Circles are the absolute estimated bias component values for individual variables. Samples are ordered from highest to lowest by mean absolute net bias across all six outcomes.	67
3.5	Scatterplots of estimated bias components by net bias across outcomes. Variables are ordered from highest to lowest by average absolute net bias across all 10 samples.	69
4.1	Absolute bias by sample and outcome variable. Estimates are presented on a percentage point scale. Samples are ordered by unweighted average absolute bias across all six outcome variables. Outcome variables are ordered by unweighted average absolute bias across all 10 samples.	89
4.2	Change in absolute bias relative to unweighted. Estimates are presented on a percentage point scale. Samples are ordered by unweighted average absolute bias across all six outcome variables. Outcome variables are ordered by unweighted average absolute bias across all 10 samples.	90
4.3	Design effect of four estimators relative to unweighted. The design effect is the ratio of the estimate's posterior variance to the unweighted posterior variance. Samples are ordered by the average design effect across all six outcome variables. Outcome variables are ordered by the average design effect across all 10 samples.	92
4.4	RMSE vs. absolute bias for all variables, estimators, and samples. Estimates are presented on a percentage point scale. The enlarged and highlighted points are the Mechanical Turk estimates for frequency of talking to neighbors. They illustrate an instance where bias was largely eliminated with OR and OR-PSC but RMSE remained high.	93
4.5	Position of 95% credibility intervals relative to population value: Participated in a school group, Talk with neighbors weekly.	94

4.6	Position of 95% credibility intervals relative to population value: Participated in a civic association, Participated in a sports/recreational association.	95
4.7	Position of 95% credibility intervals relative to population value: Always votes in local elections, Trusts all/most people in neighborhood	96

Chapter 1: Introduction

Since the mid-to-late 1990s when internet access first became widely available to the general public, the share of survey research conducted online has grown dramatically. Most of these surveys have not relied on random samples of individuals drawn from reasonably complete population frames but rather on samples of individuals who self-selected into eligibility by choosing to join a panel or clicking on an online advertisement. At first, these surveys were primarily used for market research, but as the costs of probability-based surveys have risen and response rates have declined, they have become more and more common in both academic survey research and public-opinion polling (Callegaro et al., 2014).

The American Association for Public Opinion Research (AAPOR) has attempted to address the issue of quality in nonprobability survey samples. It has produced task force reports on online opt-in panels (Baker et al., 2010), nonprobability surveys more broadly (Baker et al., 2013), and a report entitled “Evaluating Survey Quality in Today’s Complex Environment” that included guidance on assessing the quality of nonprobability samples (Baker et al., 2016). The consistent theme of these reports has been that nonprobability methods are comprised of a wide variety of disparate practices, the appropriateness of which will depend on the specific research application. There is little in the way of specific guidance that could be applied broadly to nonprobability research writ large.

Over the years, a number of studies have compared the accuracy of different nonprobability samples to each other and to probability-based surveys across a variety of population benchmarks. A study conducted in 2004 comparing a

random digit dial (RDD) telephone survey, a probability-based web survey, and seven different nonprobability samples found that the nonprobability samples were consistently less accurate than the probability-based surveys and that the overall level of bias varied widely between samples from different vendors (Yeager et al., 2011). A 2013 study conducted by the Advertising Research Foundation compared samples from 17 different online sample vendors and an RDD telephone survey and also found that the RDD sample showed the lowest bias across a variety of benchmarks. Again, the accuracy of the nonprobability samples varied considerably from vendor to vendor (Gittelman et al., 2015). A report by Pew Research Center compared 9 online nonprobability samples and the Center's probability based American Trends Panel (ATP) across 24 different government benchmarks. In this study, one nonprobability sample proved consistently more accurate than both the ATP and the other samples, though again, there was substantial variation in the level of accuracy across the nonprobability sample vendors (Kennedy et al., 2016). While these studies have succeeded at measuring the magnitude of error across nonprobability sample sources and demonstrated that it is at least possible for nonprobability samples to produce accurate survey estimates, they have been less successful at explaining why some samples perform better than others.

A major source of difficulty has been the fact that most of this research has attempted to apply the tools and frameworks that have evolved over time for the study of probability-based methods to the study of nonprobability methods. In general, they have not proven helpful. It is worth considering why this might be the case.

It is certainly not because the field is unfamiliar with the potential risks

associated with the use of nonrandom survey samples. The 1936 *Literary Digest* poll – that famously predicted Alfred Landon would defeat Franklin Roosevelt 54% to 41% – remains the go-to cautionary tale about the dangers of nonrandom sampling (Lusinchi, 2012). Polling by George Gallup that same year proved more accurate. Gallup attributed this success to the ostensibly more scientific method of quota sampling in which interviewers select respondents purposively to obtain a pre-specified number of interviews among certain geographic and demographic groups (Lusinchi, 2017). However quota sampling was also discredited after a report by the Social Science Research Council identified it as one of many causes behind the spectacular failure of the polls to correctly predict the 1948 election – an event memorialized in the famous photos of a newly elected Harry Truman holding up a copy of the Chicago Daily Tribune with the headline “Dewey Defeats Truman” (Mosteller et al., 1949; Mosteller, 2010). Subsequently, most public opinion and social science research transitioned to probability-based methods, for which a more robust statistical theory had already been developed, and which had already been adopted by the federal government for its own research (Converse, 1987; Hansen et al., 1953; Neyman, 1934).

That theory, which has come to be known as design-based inference, requires that every unit in the population has a known, nonzero probability of selection. When survey samples are randomly selected in this way, there are strong mathematical guarantees that over repeated sampling, survey samples will match the population distribution with respect to any population characteristic provided that it can be measured accurately (Horvitz and Thompson, 1952). Over the intervening years, the vast majority of statistical and methodological research into survey sampling and data collection has been premised on the idea of random sampling. Model-based approaches to survey

sampling and inference that do not fundamentally depend on random selection have been developed, but even modelers recommend random sampling because it protects the validity of model-based estimates against misspecification (e.g. [Little and Zheng, 2007](#); [Valliant et al., 2000](#), pp. 19-21).

During the 1980s and 1990s, the Total Survey Error (TSE) framework became the dominant research paradigm in survey methodology. The strength of the TSE approach lies in its recognition that there are multiple sources of error in surveys besides variability from random sampling. It categorizes error in survey estimates according to the stage in the survey process during which the error was introduced, which makes it possible to isolate, identify, and eliminate or reduce specific causes of survey error. It identifies noncoverage, when some members of the population have no chance of being selected, and nonresponse, the failure of some portion of sampled units to complete the survey, as the primary sources of systematic selection bias in survey estimates ([Biemer, 2010](#); [Groves, 1989](#); [Groves et al., 2009](#); [Groves and Lyberg, 2010](#)). Under the TSE framework, data quality depends most on the *process* used to produce a sample rather than characteristics of the sample itself. Noncoverage and nonresponse describe features of the data collection process with the potential to undermine the guarantees provided by random selection - also a characteristic of the data collection process. Although rarely achieved in practice, the goal is to make the survey process as close as possible to the ideal of perfect random selection in order to reduce or eliminate the need to rely on statistical models or untestable assumptions in analysis after the data has been collected.

This digression through the history of survey research is meant to illustrate just how deeply the idea of random sampling is embedded in the field of survey methodology. In fact, the very premise of Jerzy Neyman's foundational

paper describing stratified random sampling is to contrast random and purposive sampling. In it, he concludes “. . . the only method which can be advised for general use is the method of stratified random sampling” (Neyman, 1934, pp. 588). Since the field’s inception the solution to the kinds of problems observed in 1936 and 1948 *is* randomization, not a better form of purposive selection. That would be like trying to improve the design of a two-legged stool. Better to just add a third leg.

The 2013 AAPOR report on nonprobability sampling acknowledges the underlying problem when it states in its conclusion:

If non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality. One of the key advantages of probability sampling is the toolkit of measures and constructs (such as TSE) developed for it that provides ways of thinking about quality and error sources. Using that toolkit to evaluate non-probability samples is not especially helpful because the framework for sampling is different. Arguably the most pressing need is for researched [sic] aimed at developing better measures of the quality of non-probability sampling estimates that include bias and precision (Baker et al., 2013, pp. 109).

This dissertation attempts to address the absence of a coherent framework for evaluating the error properties of nonprobability samples, specifically the matter of selection bias that can arise in the absence of random sampling.

1.1 Problems with the Total Survey Error framework

Figure 1.1 depicts the way in which the TSE framework classifies survey error according to the stage in which it enters the survey process. It distinguishes between errors of measurement which relate to the ability to accurately record the characteristics of individual units, and errors of representation which arise because of differences between the composition of a sample and the target population. TSE further distinguishes between errors that are random (variance) and those that are systematic (bias) (Groves, 1989). This research is focused on systematic errors of representation, more commonly known as selection bias.

The procedure for moving from the full target population to a survey statistic describing that target population consists of a set of sequential and conceptually distinct steps. The first step is constructing a sampling frame that contains as much of the target population as possible. If units are omitted from the frame, the magnitude of bias depends on how different those units are and their share of the population. The second step is selection from the sampling frame. As long as this is done randomly, it should not introduce any new systematic errors, although different procedures will vary with respect to statistical efficiency. Step three involves interviewing the selected units. If the chosen data collection procedure does not successfully interview all of the sampled units, the level of bias in a statistic will depend on how many are missing and how distinct they are with respect to the outcome variable. Finally, there is adjustment error. This step involves any weighting or statistical modeling aimed at correcting biases introduced at earlier phases of the process. The goal of the TSE framework is to minimize the need for post-survey adjustment that requires making assumptions about the nature of

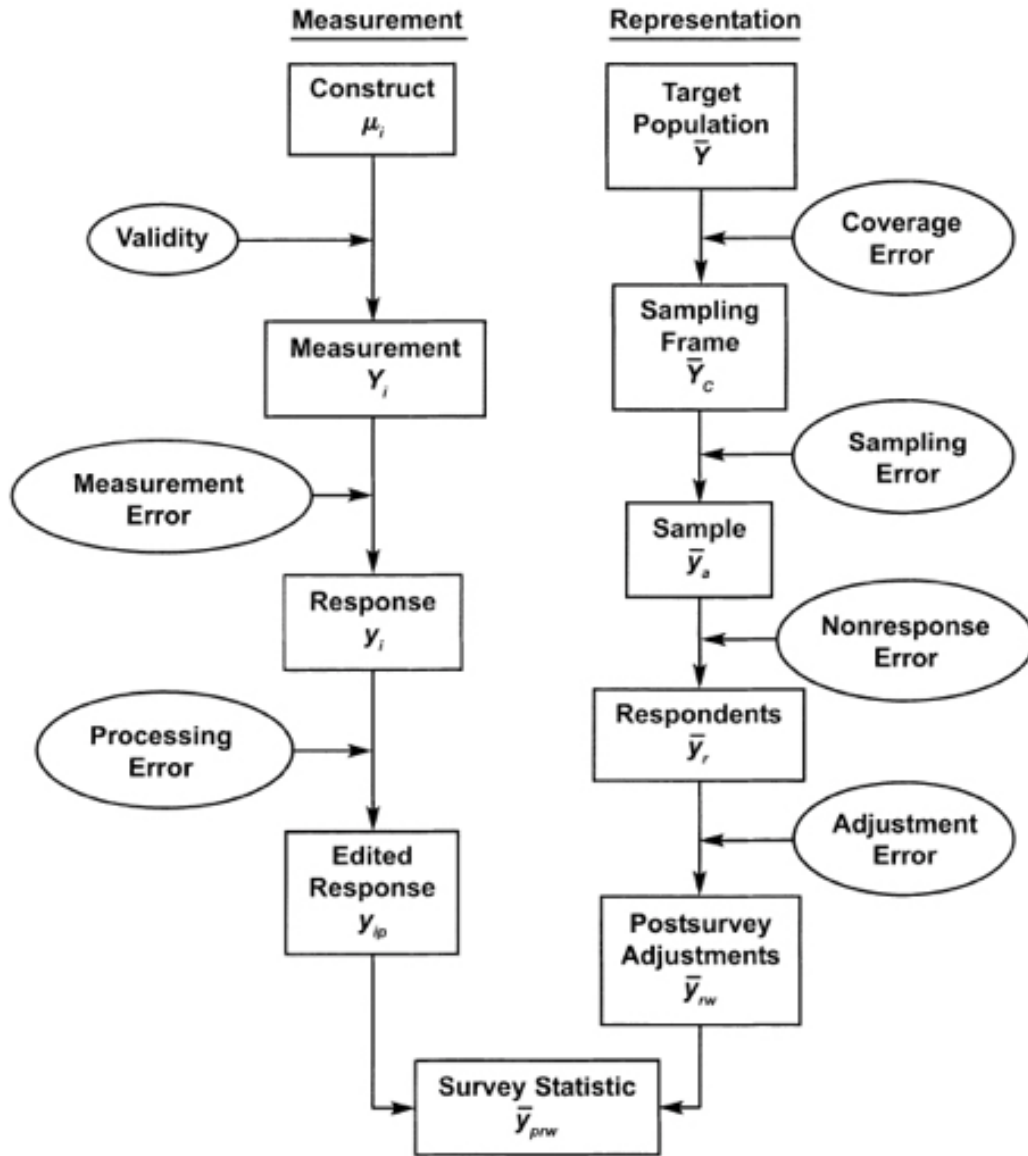


Figure 1.1: Components of error under the Total Survey Error framework (Groves et al., 2009)

the missing data (Kalton and Flores-Cervantes, 2003; Little and Rubin, 2002).

These assumptions are discussed in detail in Chapters 3 and 4 of this dissertation.

The framework is elegant and has proven enormously useful for design-based survey research. In theory, if one were able to devote sufficient resources to obtaining a sampling frame with perfect coverage and implement a data

collection procedure that resulted in perfect response, one could entirely eliminate systematic selection biases. Although this is almost never possible in practice, it sets a standard against which different survey designs can be evaluated and permits survey designers to weigh the advantages and disadvantages of different design features in a principled and coherent way.

Even for model-based sampling procedures such as balanced sampling or cube sampling, the TSE framework is helpful. These methods generally involve the selection of a finite sample from a complete population frame. The selection algorithms are all based on well defined probability models and generally involve a great deal of randomization, even if they are not technically ‘design-based’. Complete coverage and response helps ensure that the modeling assumptions that underpin these methods remain valid (Deville and Tillé, 1998, 2000; Deville, 2004; Little and Zheng, 2007; Särndal et al., 1992; Valliant et al., 2000).

In contrast, the kinds of nonprobability samples that are the subject of this thesis, online opt-in samples in particular, differ from more traditional samples in important ways that make the TSE framework inapplicable. First, there is generally no pretense or even aspiration to complete coverage of the population. Even panels that contain millions of individuals contain only a tiny fraction of the whole population. It is tempting to think of a panel as a sort of sampling frame with very poor coverage (e.g. Fahimi et al., 2015). This is not necessarily wrong, but it is also beside the point. Increasing the fraction of the population that is included in a panel to a level that could be said to reduce the risk of coverage bias, as the TSE framework would recommend, is simply not within the set of possible actions that could be taken to improve data quality. Instead, efforts are made to recruit diverse kinds of individuals

and make sure that there are no specific *types* of person missing from the panel, but this moves away from the TSE conception of coverage and back toward purposive selection and raises the question of how to determine what the relevant types of people would be.

Second, the process by which individuals are selected for a specific survey does not typically resemble the well-defined procedures of design-based or model-based survey sampling. The actual algorithms are usually considered proprietary, and there is no set of standard procedures that are consistently applied across vendors. Moreover, there is often not any sort of finite sample. Usually, there is a set of quotas specifying a desired distribution of respondent characteristics, and panelists are continuously directed to the survey until the criteria are met. Often panelists are not even selected for a specific survey but are routed to one of many currently fielding surveys based on algorithms designed to efficiently allocate sample (Brigham et al., 2014; Gitterman et al., 2015). This effectively negates the conceptual utility of nonresponse error as there are no clear groups of people who can be coherently defined as respondents and nonrespondents. Even in the instances where a finite sample is selected, there is no way to disentangle bias that occurs because some people choose not to participate from the bias that occurs because panel membership is self-selected. The issues of panel recruitment and sampling are discussed more in Chapter 2, Sections 2.4.1 and 2.4.2.

The tidy, linear depiction of representation error in Figure 1.1 becomes an intractable knot when applied to the typical nonprobability survey. It is not possible to separate bias due to coverage, sampling, and nonresponse in any meaningful way. A TSE analysis of nonprobability surveys would effectively recommend that they convert themselves in to probability-based surveys.

Again, this might be desirable from a statistical perspective, but it is beside the point.

Perhaps the most fundamental disconnect between the TSE framework and nonprobability survey sampling is that TSE seeks to eliminate the need to rely on models and assumptions. The power and appeal of random sampling is that, when unthreatened by problems such as nonresponse and noncoverage, one need not know anything in particular about the target population other than that it has been fully enumerated. Data quality is determined by the process that generated the data, and strong assumptions are unnecessary for analysis. Once the data has been collected, analysts are free to conduct their analyses as they see fit as long as they account for design features.

For nonprobability samples, attempting to minimize or eliminate strong assumptions is to make a category error. There is no avoiding strong assumptions either in practice or in principle. While the data collection process is of critical importance, there is no mechanism that provides the same sorts of theoretical guarantees as random selection. Instead, the validity of population inferences depends entirely on a set of modeling assumptions that explain the relationship between the realized sample and the target population. Rather than minimize assumptions, an error framework for nonprobability sampling should instead shine a spotlight on them. It should make it easy for analysts to know what assumptions they are making and give them the analytical tools to assess whether or not those assumptions are reasonable.

1.2 A framework inspired by causal inference

Fortunately, survey research is not the first field that has had to grapple with problems of nonrandom data. The field of causal inference holds a great deal

in common with survey inference. When trying to measure the effect of some treatment such as a drug, medical procedure, or social program, the field of causal inference has also relied on randomization to provide valid statistical inferences. Just as survey research has relied on random selection to ensure that sampled and nonsampled units are comparable, causal inference has relied on random assignment to ensure that treatment and control groups are comparable.

However, there are instances where random treatment assignment is not feasible, either for ethical reasons or because the treatment is outside the control of the researcher. For instance, political scientists who wish to measure the effects of Latino immigration into communities on support for presidential candidates have no way of randomly assigning different levels of immigration to different communities and measuring the difference (e.g. [Newman et al., 2018](#)). For these sorts of instances, statisticians have developed a framework clarifying what conditions must hold in order to make valid causal claims on the basis of observational data. It happens that the conditions that must hold in order to conduct causal inference with observational data are very similar to the conditions that are necessary to draw valid population inferences from nonprobability survey samples.

Building on the similarities between these fields, this dissertation proposes an alternative framework for considering selection bias in nonprobability survey samples and probability-based samples with noncoverage or nonresponse.

Chapter 2 of this dissertation elaborates on the parallels between causal inference and survey inference. Drawing on these parallels, it proposes a typology for different kinds of selection bias that is based on causal inference principles. Specifically, it describes bias not in terms of defects in the data

collection process but in terms of the validity of assumptions about the relationship between a survey sample and the target population conditional upon a proposed model and set of covariates. This proposed framework is appropriate both for nonprobability surveys and probability-based surveys with nonresponse or noncoverage.

Specifically, the framework classifies selection bias as attributable to problems of exchangeability, which requires that a model relating sample to population must condition all of the covariates that are necessary to make sampled and nonsampled units equivalent (Greenland and Robins, 1986, 2009; Little and Rubin, 2002; Rosenbaum, 2002; Rubin, 1974); positivity, which requires that all of the necessary types of respondent defined by the model covariates are represented in the sample (Hernán and Robins, 2006; Petersen et al., 2012); and composition, which requires that the distribution of the model covariates must match their distribution in the larger target population. After providing a conceptual description of these types of bias, Chapter 2 proceeds with a critical review of current practices in nonprobability sampling and estimation from the perspective of the causal framework and discusses their implications for reducing or increasing selection bias in survey estimates.

The causal framework proposed in Chapter 2 has spoken to a clear need in the survey community. It was first presented as part of a special session on the future of survey research at the 2016 AAPOR conference and published in the associated special issue of Public Opinion Quarterly (POQ) (Mercer et al., 2017). It has been subsequently presented at conferences focused on nonprobability survey research sponsored by the European Survey Research Association and the National Institute for Statistical Science. It has been presented to audiences as diverse as the Advertising Research Foundation's

2017 FORECASTxSCIENCE conference and the 2018 Conference of the American Association for the Advancement of Science. It was most recently presented at the DC chapter of AAPOR's POQ Special Issue Conference. In each of these venues these ideas have provoked thoughtful discussion and commentary from statisticians, pollsters, market researchers, and survey methodologists who felt that the standard analytic toolkit did not sufficiently address their needs.

Chapter 3 further develops the causal framework for selection bias by providing a more formal, mathematical description of these biases based on conditional probabilities. It goes on to demonstrate that total selection bias can be decomposed into additive components associated with exchangeability, positivity, and composition respectively. When an appropriate reference sample is available, these components can be estimated conditional on a set of chosen covariates.

Using Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), the magnitude of each type of bias is estimated for six measures of civic engagement that were measured on 10 different nonprobability samples commissioned by Pew Research Center. These include the 9 nonprobability samples analyzed in the study by Kennedy et al. (2016) and a survey that was sampled via Amazon's Mechanical Turk that used the same questionnaire. It finds that exchangeability bias from unobserved confounding variables is generally the largest source of bias for individual estimates.

Chapter 4 compares two doubly-robust approaches to survey estimation with singly-robust approaches based on propensity weighting (PW) and outcome regression (OR). Doubly-robust estimation involves fitting separate regression models predicting both sample inclusion and the outcome variable of interest.

As long as one or the other model is correctly specified, the resulting estimates will be asymptotically consistent (Bang and Robins, 2005; Kang and Schafer, 2007; Robins et al., 1994). The doubly-robust estimators considered are outcome regression with a residual bias correction (OR-RBC) and outcome regression with a propensity score covariate (OR-PSC) (Kang and Schafer, 2007). All four approaches are implemented using BART, making them similar to methods evaluated by Tan et al. (2018).

The analysis is conducted using the same 10 samples and outcome variables as in Chapter 2. Because many of these variables are known to suffer from unobserved confounding, none of the estimators entirely eliminate selection bias. We find that OR-RBC tends to have the lowest bias, variance, and RMSE, with PW only slightly worse on all three measures. OR and OR-PSC also perform similarly but result in larger variances and appear more likely than PW and OR-RBC to inflate rather than reduce bias.

The nonprobability survey data analyzed in Chapters 3 and 4 are publicly available for download from Pew Research Center at <http://www.people-press.org/datasets/2015/>. The code used to conduct these analyses can be found on GitHub at <https://github.com/awmercer/fpbb-inference> and in Appendix C. This repository is accompanied by a preliminary R package called `bestimate` (<https://github.com/awmercer/bestimate>) that aims to permit others to conduct the same kinds of estimation and bias decompositions with BART that are employed in this dissertation. The dissertation itself is written in `bookdown`, and can also be found on GitHub at https://github.com/awmercer/ameracer_dis.

Chapter 2: A causal inference framework for selection bias¹

The growing use of surveys that do not use traditional probability sampling has provoked both interest and concern from the survey community. Rising data collection costs coupled with declining response rates have highlighted the appeal of lower cost, nonprobability surveys that can be fielded rapidly online. However, respondent self-selection into these surveys renders design-based methods of survey inference inapplicable, and raises concerns about the potential for biased results.

Selection bias refers to systematic differences between a statistical estimate and the true population parameter caused by problems with the composition of the sample (rather than errors in measurement). Traditionally, survey researchers think of selection bias as resulting from noncoverage – when the sampling frame omits portions of the target population – or nonresponse – when selected units do not complete the survey. These concepts are tied to a process of starting with a complete population and randomly selecting a subset. These categories may prove limiting when applied in a nonprobability context. Many nonprobability surveys do not originate from anything resembling a sampling frame. Even the idea of a sample as a finite set of units, some of which may fail to respond, does not apply to many nonprobability surveys. For nonprobability surveys, the processes that lead to a respondent

¹This chapter is a pre-copyedited, author-produced version of an article accepted for publication in *Public Opinion Quarterly* following peer review. The version of record: Mercer, Andrew W., Frauke Kreuter, Scott Keeter, and Elizabeth A. Stuart. 2017. “Theory and Practice in Nonprobability Surveys: Parallels Between Causal Inference and Survey Inference.” *Public Opinion Quarterly* 81 (S1): 250–71 is available online at: <https://doi.org/10.1093/poq/nfw060>

being included in a sample are numerous, potentially arbitrary, and may not resemble the traditional probability-based survey process at all.

Rather than evaluate nonprobability surveys using concepts designed for a different inferential framework and different data collection practices, we propose a more general framework that emphasizes the characteristics of the realized sample, regardless of how it was generated. The underpinnings of this framework are not new, but come from research into the estimation of causal effects from experimental and non-experimental data. In fields such as epidemiology, political science and economics where randomized experiments are frequently not possible and observational studies are commonplace, research has focused on identifying the conditions under which valid statistical inferences about causal effects can be made using observational data. In the causal context, the parameter of interest is a contrast between experimental treatments, whereas surveys measure a broad range of estimates including means, totals, correlations and other measures of association. Despite differences, the conditions that produce selection bias in causal analyses also apply in a survey context.

Others have noted similarities between causal inference and survey inference. [Little and Rubin \(2002\)](#) apply many of these same concepts to experiments, observational studies, survey nonresponse, and imputation. [Groves \(2006\)](#) uses a causal framework to describe when nonresponse will produce bias in survey estimates. [Keiding and Louis \(2016\)](#) reviewed the many objectives and challenges shared by both epidemiological studies and surveys, and suggest that both fields could benefit from sharing methodologies.

Drawing on this work, we identify three components that determine whether or not nonrandom selection could lead to biased results:

- Exchangeability – Are all confounding variables known and measured for all sampled units?
- Positivity – Does the sample include all of the necessary kinds of units in the target population, or are certain groups with distinct characteristics missing?
- Composition – Does the sample distribution match the target population with respect to the confounding variables, or can it be adjusted to match?

In this paper, we first describe how this framework applies in the familiar context of randomized experiments and probability based-surveys before demonstrating how it extends to cover observational studies and nonprobability surveys. Second, we demonstrate the mechanics by which each component can produce bias in survey estimates by way of a simplified example. Finally, through the lens of this framework, we provide a critical review of current practices in online, nonprobability data collection and their implications for selection bias.

2.1 Randomization and unbiased inference in experiments and surveys

Questions about causal effects are usually framed in terms of potential outcomes or counterfactuals (Rubin, 1974). A patient's outcome may be different if he is given Treatment A or Treatment B. Prior to choosing a treatment, either outcome is possible, but we observe only the results under the treatment that is actually provided to the patient. We can never observe what would have happened if a different treatment had been applied. The causal effect is the difference between the two potential outcomes. Although

we can never observe both outcomes on a single individual, we can compare the average outcome for people who receive Treatment A to that of people who receive Treatment B to make inferences about which treatment is better. When treatments are assigned randomly, we can be reasonably confident that observed differences in the outcomes across treatment conditions are due to the treatments themselves and not some other difference between the two groups. When treatments are not assigned randomly, these assessments are more difficult. For instance, if patients who receive Treatment A tend to do worse, but Treatment A is usually given to sicker patients, it is difficult to know if the difference is due to the treatment or due to the fact that the patients who received it were in worse shape to begin with. The baseline level of sickness is known as a confounder. Confounders are variables associated with both the choice of treatment and the outcome of interest, and are the primary source of selection bias in causal analyses.

The parallels between causal inference and survey inference are substantial. A probability-based survey is essentially a randomized experiment where the pool of subjects is the set of units on the sampling frame and the treatment is selection into the survey. Unlike experiments where we observe outcomes on both treated and untreated subjects, in surveys we observe outcomes only on the selected units, with the expectation that there should be no difference between selected and non-selected units. The conditions under which causal effects can be estimated without selection bias are analogous to the conditions that produce unbiased estimates in surveys. Before discussing nonprobability surveys, we will first examine how these conditions are met in the context of randomized experiments and probability-based surveys.

2.1.1 Strong Ignorability – Exchangeability and Positivity

Rosenbaum and Rubin (1983b) devised the notion of strong ignorability to describe the conditions under which inferences about causal effects can be estimated without selection bias for a given sample. Strong ignorability consists of two requirements. The first, known as “exchangeability” (Greenland and Robins, 1986, 2009), “ignorability”, “no unobserved confounding,” or “no hidden bias” (Rosenbaum, 2002), requires the mechanism by which subjects are assigned a treatment to be independent of the measured outcome either unconditionally or conditional upon observed covariates. Unconditional exchangeability is analogous to the notion of data that is missing completely at random (MCAR), whereas conditional exchangeability corresponds to data missing at random (MAR) (Little and Rubin, 2002). When unobserved confounders are present, it is not possible to isolate the effect of the treatment from the effect of the confounder without additional assumptions.

Second, it must be possible for any subject to have received any of the treatments. This requirement is called positivity because it requires all subjects have a positive probability of receiving treatment. If certain types of subjects receive only treatment or control, it is not possible to learn about causal effects for those subjects, and the treatment and control groups will have systematic differences that cannot be resolved. In practice, we generally require not just a positive probability but also enough cases to produce sufficiently precise statistical estimates (Hernán and Robins, 2006; Petersen et al., 2012).

In experiments, random treatment assignment guarantees that on average, the exchangeability and positivity conditions will be met. Randomization ensures exchangeability by preventing any relationship between treatment assignment

and unobserved variables and ensures positivity because any subject has a chance of receiving any treatment. In probability-based surveys, random selection functions in much the same way. By randomly selecting a sample from the entire population, there can be no unobserved variables systematically associated with selection, and all members of the population have a chance of being included.

2.1.2 Composition

For experiments, the composition of treatment groups with respect to potential confounders is important in two respects. First, the distribution of potential confounders in the treatment group needs to match the distribution in the control group. Random treatment assignment guarantees that this will occur naturally on average, and this equivalence between treatment groups is implied whenever unconditional exchangeability holds. Second, the composition of the experimental sample affects the degree to which findings can be generalized to an external population.

Strong ignorability guarantees only that the results of an experiment are generalizable to the group of subjects included in an experiment; in other words, it ensures “internal validity” but does not necessarily imply “external validity” (Shadish et al., 2002). It is rare for samples in randomized trials, which have historically prioritized internal validity, to match a larger population. Because of this there has been a growing literature on methods to allow the generalization of experimental results to target populations, including reweighting strategies that aim to equate the experimental sample and the population with respect to observed characteristics (Cole and Stuart, 2010; Kern et al., 2016; Stuart et al., 2015). Pearl and Bareinboim (2014) refer to the transportability of empirical findings from one sample to a separate

target population. They note that generalization requires one to know the distribution of the outcome conditional upon treatment and any confounders, as well as the joint distribution of the confounding variables in the target population. Put simply, to generalize beyond the experimental sample to a target population, the sample needs to look like (or be made to look like) the target population with respect to the distribution of confounding variables.

The situation for surveys is somewhat less complex than for causal analyses. Whereas experiments must be concerned with the comparability of treatment and control as well as sample and population, surveys need be concerned only about sample and population. It is understood that the composition of a sample will match that of the population when all units have an equal probability of selection, implying unconditional exchangeability. When probabilities of selection are unequal but known for every unit in the frame, the situation is equivalent to conditional exchangeability, and weighting observations by the inverse of the probability of selection yields unbiased population estimates (Horvitz and Thompson, 1952). In either case, random selection ensures that on average the sample will match the target population on the distribution of any variables measured on the survey.

2.2 Extending the Framework to Non-Random Samples

For causal analyses and surveys random treatment assignment and respondent selection provide a powerful mechanism for producing the conditions necessary for unbiased estimation of causal effects and population parameters. However, these conditions are guaranteed only when randomization is 100% successful. In practice, this is rarely the case. In experiments subjects drop out of trials or are lost to follow-up. In surveys, the sampling frames may not perfectly cover the target population, and nonresponse means that some share of

sampled units is never observed. When such problems occur, the usual response is to perform statistical adjustments to correct any imbalance. In experiments methods such as matching or propensity score weighting can be used to adjust for imbalances between experimental treatment groups (see [Imbens and Rubin, 2015](#), Part VI). In probability surveys corrections involve nonresponse weighting adjustments for which a variety of techniques exist (see [Kalton and Flores-Cervantes, 2003](#); [Valliant et al., 2013](#)).

When we perform these adjustments to randomized experiments or probability surveys, we are no longer relying solely on randomization to produce unbiased estimates. Rather, these adjusted estimates are conditional upon a model that assumes that positivity and exchangeability hold and that the adjustment reconstructs the correct sample composition for the confounding covariates. Even if we perform no adjustment, we are implicitly assuming a model where the correlation between missingness and the outcome of interest is zero, or unconditional exchangeability.

In the causal world, it is recognized that as long as exchangeability and positivity hold, it is possible to make unbiased inferences about causal effects from non-experimental data ([Greenland and Robins, 1986, 2009](#); [Rosenbaum and Rubin, 1983b](#); [Rubin, 1974, 1978](#)). Quasi-experimental designs such as regression discontinuity and instrumental variables models are techniques that can be used to identify causal effects from non-experimental data when the appropriate conditions are met ([Angrist and Pischke, 2009](#); [West et al., 2008](#)). Methods such as matching, marginal structural models and structural nested models have been developed to estimate causal effects from observational data and have been proven to produce unbiased estimates when their underlying assumptions are met ([Cole et al., 2003](#); [Robins, 1999a,b](#); [Stuart, 2010b](#)).

However, for all of these techniques, one can never be certain if the exchangeability and positivity conditions have been met. Therefore, the bar for accepting results from non-experimental data is much higher than for randomized experiments.

The same is true for surveys that do not use probability sampling. When units are not randomly selected from the target population, researchers must rely on statistical models to generalize back to the target population.

Probability-based surveys with undercoverage or nonresponse must also specify a model that relates the observed units to the unobserved (Brick, 2013; Valliant et al., 2000). For probability samples the initial design performs most of the work in ensuring exchangeability, positivity, and correct sample composition. Statistical models are employed during estimation to correct what are hopefully minor biases. In contrast, nonprobability samples cannot rely on randomization to help meet these requirements, and instead must rely on models at all stages of the survey process from sample selection to estimation. As in causal analyses, researchers can never know with certainty that these requirements have been met.

2.3 Mechanics of Selection Bias in Surveys

In this section we focus specifically on the survey context and demonstrate through a simplified example the mechanics behind each of the components in this framework to show how they can introduce bias into survey estimates.

2.3.1 Exchangeability

Suppose we have a sample intended for estimating what share of the population will vote for the Democratic and Republican candidates in an election, and that we have measured each respondent's candidate preference

and age. Let us also assume that some feature of the recruitment process over-represents older people but that there are no additional unmeasured confounders. Because older people tend to vote Republican more than young people, an estimate of the overall vote using this sample would be biased in favor of the Republican candidate. However, because inclusion depends only on age, estimates of the vote within the younger and older subgroups would still be correct. In this case, the sampled individuals are exchangeable with non-sampled individuals within the same age group. When sampled observations are conditionally exchangeable, subgroups are internally unbiased with respect to the outcome of interest, even if some groups are over or under represented relative to their share of the target population. Because there are no additional confounders the overall proportion of the sample voting Democratic would be biased, but measures of the relationship between age and vote preference would be unbiased prior to any adjustment.

However, if inclusion in the sample depends on an unmeasured characteristic related to the survey outcome, the distribution of the outcome variable within the observed subgroups will no longer match that of the target population. In our voting example, suppose our sample also over-represents respondents who live in big cities but, unlike age, this has not been measured. Because urban dwellers tend to vote Democratic, the Democratic vote share among both young and old respondents will be too high. In this case, young and old respondents in our sample are not exchangeable with their non-sampled counterparts because they are more urban, making urbanicity an unmeasured confounder. The bias in favor of the Democrat due to an excess of city dwellers could actually offset some of the Republican bias produced by having too many older respondents. In this scenario, the estimated vote for the full sample could be close to the true population value while subgroup estimates

would be biased. Note that the crucial aspect of exchangeability is not which cases are included in the sample but what characteristics have been measured. If we knew which cases were urban and which were rural, we could adjust by both age and urbanicity to recover the correct sample composition.

In practice, the biases need not cancel out. The unobserved variable could have opposite effects for young and old respondents, or there could be different unobserved variables affecting different subgroups. Because the confounding variables are unobserved, it is impossible to know from the data alone whether or not the exchangeability requirement is met.

The associations that produce bias need not be direct. If we took this same sample but measured something such as eye color, which is not directly related to either age or urbanicity, we might still achieve biased estimates if eye color is associated with race. Over-representing urban respondents likely also means over-representing racial groups that live in urban areas which could in turn affect the distribution of observed eye colors. The reverse is also true.

Variables that are not confounders themselves but are closely correlated with confounders may help reduce bias by serving as proxies during adjustment.

2.3.2 Positivity

The positivity requirement states that even if we know and have measured all potential confounders, all of the subgroups defined by confounding variables must also be represented in the sample (Hernán and Robins, 2006). Groups that are underrepresented but present can be weighted up. However, it is not possible to weight up groups that were not surveyed. Returning to our example where inclusion depends on age and urbanicity, suppose that there are no older, urban respondents included in the sample. Even if we were able

to record both age and urbanicity, there is no adjustment we can perform that will make up for the absence of older, urban respondents, although subgroup estimates for those groups that were observed would remain unbiased. On the other hand, if older and younger city dwellers are the same with respect to their voting preference, the absence of older urbanites would not introduce bias because younger urbanites could stand in for them in the sample with no change to the estimate. When a group is entirely missing from the set of observed units, the researcher requires a theoretical justification for believing that the missing group is not systematically different from other, superficially similar groups that were surveyed.

2.3.3 Composition

In our example, we have assumed that our sample composition does not match the target population on age and urbanicity. If it can be adjusted to match the distribution in the target population, our estimates of the vote will be unbiased. We have already alluded to the simplest approach, which is to weight each group to be proportional to its share of the target population.

Sample composition can be managed by design as well as through post-hoc adjustment. Random selection yields the correct sample composition in expectation, though individual samples will not match exactly. If the confounders are known in advance, purposive methods such as quota sampling, where we pre-determine the number of interviews required in each group, can be used to produce an exact sample match([Gittelman et al., 2015](#)).

Managing sample composition through design or adjustment rather than random selection requires the researcher to be confident that all confounders are truly known and measured. When exchangeability or positivity does not

hold, bias will not be eliminated and may even be magnified. In our example, if we adjust only for age but not urbanicity, we would eliminate the pro-Republican bias caused by an older sample but not the pro-Democratic bias due to an excess of urban respondents. The biases no longer offset each other and the adjusted estimate would be more biased toward the Democrats than it was before weighting.

2.4 Current Practices for Managing Bias in Online, Nonprobability Surveys

We can use this framework to consider current practices in fielding nonprobability web surveys and producing statistical estimates from the resulting samples. We distinguish between recruitment, whereby an individual becomes eligible for inclusion in one or more surveys (e.g., joining a panel) and sampling, the process by which an individual is selected for a particular survey after recruitment. After reviewing these two features of the data collection process, we discuss alternative approaches to post-survey adjustment and estimation.

2.4.1 Recruitment

The most common form of recruitment involves inviting individuals to join opt-in panels, which are lists maintained by sample providers of individuals who have agreed to participate in surveys on an ongoing basis. Individuals can become empanelled in a variety of ways, such as directly through a panel website, clicking on banner advertisements, or when corporations grant panel vendors access to members of their customer loyalty programs. Panels provide an opportunity to collect a large amount of profile information on their members that can be used in both sampling and adjustment. Maintaining

respondent profiles across many dimensions can aid in providing exchangeability only if the correct variables are measured. On the other hand, some fear that panel conditioning and attrition may mean that panel members may become less reflective of their non-empanelled counterparts over time, potentially reducing exchangeability (Callegaro and DiSogra, 2008; Callegaro et al., 2015; Couper, 2000).

The main alternative to panels is river sampling, in which potential respondents are recruited via similar sources, but are directed to a one-off survey rather than asked to join a long-term panel (Callegaro et al., 2014). River sampling avoids panel attrition and conditioning, but provides no profile data on respondents in advance. Respondent characteristics must be obtained at the time of the survey, limiting the number of characteristics that can be measured. Some online survey providers have begun using a mixture of panel and river respondents (e.g. Lorch et al., 2010; Young et al., 2012).

Both panels and river sampling face an immediate threat to the positivity requirement because individuals who do not use the internet cannot participate. Studies conducted on the Pew Research Center's American Trends Panel and the Dutch LISS panel, two probability-based panels that take steps to cover individuals without internet access, found that the exclusion on non-internet individuals produced only small differences in most survey estimates. However, for outcomes pertaining to technology use, differences in estimates could be large. The Pew Research study also found that indicators of socioeconomic status differed considerably for some subgroups such as the elderly or racial minorities (Eckman, 2016; Keeter et al., 2015).

Obtaining a diverse array of potential respondents is crucial to the success of any recruitment method. Pettit (2015) demonstrated that respondents

recruited via different websites can exhibit dramatically different demographic distributions. Respondents recruited from different sources likely vary on other characteristics as well; for instance, individuals recruited via a website dedicated to video games could differ from those recruited from websites devoted to personal finance with respect to variables such as interest in retirement planning or their use of leisure time. Recruiting from a diverse set of sources necessarily improves the probability of meeting the positivity requirement; however, it also increases the complexity of the recruitment process, potentially creating a trade-off between positivity and exchangeability. As the number of sources increases, it may become more difficult to know which characteristics distinguish between individuals recruited from different sources.

To date, the great majority of research into nonprobability surveys has relied on data from online panels. Many of these studies have compared different panels to one another and found that while some nonprobability surveys compare favorably to probability-based surveys, the same survey fielded on different panels can result in dramatically different results (Callegaro et al., 2014; Craig et al., 2013; Erens et al., 2014; Kennedy et al., 2016; Schnorf et al., 2014; Yeager et al., 2011). However, none of these studies were designed to evaluate alternative methods of panel recruitment or isolate the design features that produce such varying results.

Very little research has directly compared panels to river sampling. One such analysis found that after weighting for demographic characteristics, panel respondents were largely similar to river respondents, although panelists were more likely to be registered to vote and more likely to use Twitter. River respondents were closer to the chosen benchmark on both measures (Clark

et al., 2015). A study performed as part of the Foundations of Quality 2 (FOQ2) initiative compared the demographic composition of surveys using panels and the river sampling. It found that on average, the river samples yielded demographic compositions similar to non-river samples, and required somewhat less extreme weighting when adjusted to match demographics not used in the sampling process (Bremer, 2013). Unfortunately, there was no evaluation of differences in other non-demographic estimates.

At present, there is not enough research to recommend one recruitment method over the other. The availability of profile data on panels offers flexibility and control for the purposes of sampling and adjustment, but the limited empirical research discussed previously does suggest some possible advantages to river samples. Other practices such as profiling, sampling or quota design may also be more important than the recruitment process.

2.4.2 Sampling

Nonprobability surveys generally rely on purposive selection to achieve the desired sample composition while data collection is ongoing. This is commonly achieved through quotas, where the researcher pre-specifies a particular distribution across one or more variables. Usually these are cells defined by a cross-classification of demographic characteristics such as gender by age, with each cell requiring a specified number of completed interviews in that category. The end result is a sample that matches the pre-specified distribution across the chosen variables. The use of quotas relies on the assumption that individuals that comprise each quota cell are exchangeable with non-sampled individuals who share those characteristics. If that assumption is met, the sample will have the correct composition on the confounding variables, allowing for the estimation means and proportions that generalize to the

target population.

Most contemporary web surveys that employ quotas define the cells across no more than a handful of demographic variables. However, there is a growing consensus that basic demographic variables such as age, sex, race, and education are insufficient for achieving exchangeability. A recent study using the FOQ2 data compared three progressively more stringent sets of demographic quotas. Across a range of benchmarks, the application of more stringent quotas did nothing to reduce bias, and post-survey weighting actually increased the average bias for all but five out of seventeen sample providers. The study also evaluated three quota schemes that incorporated additional, non-demographic variables, however their success was mixed. (The details of the methods employed were not specified to avoid identifying the sample providers (Gittelman et al., 2015)). This finding is consistent with research in causal inference suggesting that demographics alone are generally insufficient for eliminating bias in observational studies (Cook et al., 2008).

If traditional quota methods are insufficient for producing strong ignorability, sampling methods that allow researchers to control both more and different dimensions may improve the ability to condition on a more appropriate set of potential confounders. The best documented of these methods is implemented by YouGov on surveys conducted using its panel in the United States.

YouGov first draws a random sample of cases from a high quality data source, such as the American Community Survey (ACS) Public Use Microdata Sample, that is believed to reflect the true joint distribution for a large number of variables in the target population. This subsample is referred to as a synthetic sampling frame (SSF) and serves as a template for the eventual survey sample. Each panelist who completes the survey is matched to a case

in the SSF with similar characteristics using a distance measure such as Euclidean distance. When every record in the SSF has been matched with a suitably similar respondent, the survey is complete (Rivers, 2007).

Because a limited number of covariates are available on any single survey such as the ACS, it is possible to impute additional variables onto the SSF using models built with other data sources. This was the approach taken on the 2008 Cooperative Congressional Election Study which augmented an SSF drawn from the ACS with estimates of voter registration and turnout from the Current Population Survey Voting and Registration Supplement, and of internet use, religion and interest in politics from Pew Research Center surveys. The resulting survey sample produced estimates of the presidential vote that closely matched national exit polls and the American National Election Studies (Ansolabehere and Rivers, 2013).

This approach is appealing in its capacity to flexibly match the target population on a larger number of covariates than is possible with traditional quota methods. For this approach to succeed, the composition of matching variables in the SSF must accurately match the target population, and any models used to combine datasets must be correctly specified. More importantly, the matching variables must be the correct variables for ensuring conditional exchangeability, and the panel must be able to supply respondents that are close matches to each case in the SSF. If there are remaining confounders that are not accounted for, resulting survey estimates will be biased. One side-benefit of this approach is that problems with positivity should be immediately apparent if there are portions of the SSF for which no clear matching respondents can be found.

Another approach to sampling on a higher number of dimensions is the use of

propensity score matching to construct quota cells. Under this approach, a probability survey that is assumed to accurately reflect the target population is fielded in parallel with a nonprobability survey. Using a set of common covariates collected on each survey, a propensity model is estimated by combining the two samples and predicting the probability that each respondent belongs to the probability survey. When subsequent online surveys are fielded, the propensity model is used to calculate a propensity score for each respondent as they are screened for the new survey. Quotas are set not on particular respondent characteristics but are based on quintiles of the propensity score distribution (Terhanian and Bremer, 2012).

As with the SSF used in sample matching, much hinges on how well the parallel reference survey matches the target population. If the reference survey suffers from its own nonresponse or coverage bias, those biases will be transferred into the nonprobability survey. On the other hand, the researcher could tailor the contents of the baseline surveys to include any variables believed to be necessary to ensure conditional exchangeability. Under other approaches, researchers are limited to covariates that are available from preexisting data sources. This method performed well in a simulation, however the data used to construct the propensity model was the same data used to generate the simulated survey. The evaluation also generated only a single simulated dataset (Terhanian and Bremer, 2012). As such, it is difficult to know how this technique performs on new samples and over repeated applications. Dividing the propensity score into quintiles will result in a loss of information contained in the full distribution of propensity scores, though it is also possible that quintiles provide a sufficient foundation of balance and positivity that can be further refined through post-survey adjustment. Additional research comparing this approach with the matching approach

described above would be valuable, particularly if the same survey and set of covariates can be used.

Another, less understood component of the sampling process for many nonprobability surveys is the use of routers. Most nonprobability survey vendors have many surveys fielding simultaneously. When a router is employed, rather than draw separate samples for each survey, respondents are invited to participate in an unspecified survey. The actual survey taken is determined dynamically based on the characteristics of the respondent and the needs of active surveys with respect to quotas or screening criteria. This makes for a more efficient use of sample, but means the sample for any one survey depends on what other surveys are in the field simultaneously. If there are ample respondents and few competing surveys, routers may pose little threat of bias. On the other hand, the presence of surveys focused on rare groups may mean that individuals belonging to those groups are not routed to other surveys. In such an event, the routing process becomes a confounder that would be difficult to observe and account for.

The only empirical study evaluating routers compared the effects of three different routing methods against a non-routed control and found that all four conditions produced similar estimates. In a set of simulations, the authors did find that routing could produce bias for questions that are highly correlated with the selection criteria for other surveys in the field. This study evaluated routing under a narrow set of conditions that the authors recognize may not generalize to many circumstances observed in practice ([Brigham et al., 2014](#)). Additional experiments and simulations testing alternative algorithms and scenarios, or observational studies comparing router performance over time for different vendors would be of substantial benefit.

2.4.3 *Post-survey Adjustment*

Because it may not be feasible to achieve the desired sample composition through sampling alone, post-survey adjustment is still needed. Most of the research on adjusting nonprobability samples has focused on adapting the methods used to perform non-response adjustment with probability samples. Calibration and propensity score weighting are the two most common approaches to weighting.

Calibration methods directly adjust the composition of the sample to match a known distribution of variables in the target population. The simplest form of calibration is post-stratification, in which the sample is divided into mutually exclusive cells that are weighted up or down such that the proportion of each cell in the sample matches the corresponding proportion in the target population. Whereas post-stratification requires knowledge of the joint distribution of the stratification variables in the target population, other calibration methods such as raking and generalized regression estimation require only knowledge of the marginal distribution of any adjustment variables (Deville and Särndal, 1992; Kalton and Flores-Cervantes, 2003).

Calibration methods generally require that the outcome be a linear function of the calibration variables, and may not perform well in the presence of nonlinear relationships between the outcome and adjustment variables or unmodeled interactions (Valliant et al., 2000).

Propensity score weighting involves combining a nonprobability sample with a parallel probability or gold-standard data source as a reference sample. A model predicting sample membership is fit to these combined data, and observations in the nonprobability sample are weighted by the inverse of their probability of appearing in the nonprobability sample (Lee, 2006; Taylor, 2000;

Terhanian and Bremer, 2000; Valliant and Dever, 2011). Valliant and Dever (2011) demonstrated that for propensity score adjustment to be effective, the propensity model must incorporate any nonresponse adjustment and bias correction that has been applied to the reference sample. Otherwise, those biases will be transferred to the nonprobability sample.

Given the same set of covariates, generalized regression estimation (GREG) has been found to perform comparably to propensity score weighting, suggesting that a parallel reference survey may be unnecessary when the requisite population totals are available (Valliant and Dever, 2011).

Propensity score weighting can more easily accommodate nonlinear associations and interactions between confounding variables. If there are a large number of confounders or it is unknown which of the observed covariates are confounders, machine learning methods such as boosting or random forests can fit high dimensional propensity models if a suitable reference sample with common covariates is available (Buskirk and Kolenikov, 2015; Lee et al., 2010).

Some have explored matching as an alternative to weighting for post-survey adjustment of nonprobability surveys. Traditionally, matching is used in causal inference in order to adjust for differences in composition between treatment groups (see Stuart (2010a) for a review of their use in causal inference). With matching, the idea is to create groups containing one or more observations from both a reference sample and a nonprobability sample that are similar on a set of auxiliary variables believed to be associated with selection. Groups in the nonprobability sample are then weighted so that their distribution matches the distribution in the reference sample. For example, a reference sample might be divided into cells based on a set of covariates or a propensity score, while cases in the nonprobability in matching cells would be weighted so

that the proportion in each cell matches the proportion in the reference sample. In this sense, matching is very similar to propensity score weighting or poststratification with one important exception. In many applications, observations for which there is no acceptable match are removed from the final dataset. When this happens, information is lost, and inference is only possible for those portions of the samples that overlap. On the other hand, identifying a lack of overlap forces researchers to evaluate the validity of the positivity assumption in ways that other methods may not. Unlike standard weighting methods that will generally produce a weight for every observation (even if some are quite large), matching software often automatically identifies those observations in a reference sample for which no counterparts exist in the nonprobability sample (e.g. the MatchIt package for the R statistical software platform (Ho et al., 2011)). Dutwin and Buskirk (2017) found that raking to basic demographics was more effective at reducing bias than matching on a more extensive set of demographics; however, a two-stage process of matching followed by raking reduced bias more than raking alone.

A final approach to post-survey estimation is multilevel regression and poststratification (MRP). In traditional poststratification, a sample is divided up into mutually exclusive cells, each of which is weighted to be proportional to their representation in the target population. As the number of cells becomes large, the number of observations in each cell becomes small and estimates become unstable. MRP enables poststratification using a large number of cells by fitting a multilevel model that pools information about cells sharing similar characteristics and allows for the estimation of cell means even when cells are sparse. A weighted mean is then constructed using the estimated cell means (Ghitza and Gelman, 2013; Lax and Phillips, 2009; Park et al., 2004).

This approach performed well when used to predict 2012 presidential election results using a survey conducted via the Microsoft Xbox platform whose sample composition differed radically from the population of voters and for which unadjusted estimates were wildly inaccurate (Wang et al., 2014). Unweighted, the sample was 93% male, only 1% 65 years old or older, and showed Barack Obama losing badly to Mitt Romney. On the surface, it seems unlikely that such a survey could produce accurate estimates. However, the Xbox study enjoyed two benefits not available to many other studies. The first is a very large sample size (345,858 unique respondents), which means that even groups that are dramatically underrepresented in the sample in relative terms still have enough observations in absolute terms to avoid problems with positivity. The 1% of the sample 65 years old or older yields 3,400 observations – more than enough cases to produce stable estimates for that subgroup. The second is that the authors had a very powerful set of covariates, including party identification and ideology, making it much more likely that the exchangeability requirement was satisfied for the purpose of predicting partisan voting behavior.

Another study using only demographic covariates met with less success. It compared MRP based estimates of presidential approval and country direction to estimates from the Pew Research Center’s probability-based telephone surveys over the same time period. For the share of the population that thinks the country is on the right track, the MRP estimates were not different from the estimates obtained using a simple post-stratification adjustment, and lower than the telephone based estimates. On the other hand, Presidential approval changed dramatically, moving from an underestimate to an overestimate relative to the comparison telephone survey (Petrin and El-Dash, 2015).

Although the telephone survey benchmarks are themselves estimates and have

their own biases, if the goal of adjustment was to match that particular benchmark, neither MRP nor traditional post-stratification were successful. Each of these approaches to estimation comes with advantages and disadvantages. When control totals are available for the confounders and their relationship with the survey outcome is linear, calibration methods are quite powerful and easy to apply. Propensity score methods provide a great deal of flexibility at the cost of requiring an auxiliary dataset with a shared set of covariates. It is less clear if matching offers substantial benefits over propensity score weighting or calibration. For approaches that produce weights, there is some indication that methods applied in combination may offer an improvement over the use of a single method (Brick, 2015; Dutwin and Buskirk, 2017; Lee and Valliant, 2009; Mercer et al., 2018). MRP may be most efficient at extracting information from smaller datasets, but at the cost of computational complexity and the fact that a separate model is required for each outcome variable. Additional research directly comparing adjustment methods to one another would be valuable in helping researchers choose the most appropriate tool.

All of these methods will fail if the exchangeability and positivity requirements are not met, or if the model specification does not correctly replicate the target composition on the confounding variables. If exchangeability and positivity are met, the best method is the one that can most closely mirror the correct sample composition using the available data and information. If exchangeability and positivity are not met, there is no a-priori reason to believe that any of these methods will perform better than any other.

2.4.4 Variable Selection

Given the centrality of exchangeability and positivity in achieving unbiased estimates from nonprobability surveys, what variables should practitioners measure and utilize in sampling and adjustment? A number of researchers have attempted to find sets of variables that can reliably serve to achieve at least partial exchangeability for a broad range of survey topics. These include so-called “webographics,” early adopter characteristics and other behavioral and attitudinal factors intended to differentiate between survey participants and the broader population (DiSogra et al., 2011; Fahimi et al., 2015; Schonlau et al., 2004, 2007). While such general-purpose variables may fill a need, their effect will be limited unless they are correlated with the outcome to be measured.

Researchers will be best served if they can identify a likely set of theoretically grounded confounders prior to data collection, and use these as the starting point for a research design. For example, in studies of U.S. politics, many outcome variables of interest will be related to respondents’ underlying political engagement and partisanship. These may be effective confounders to use in sampling and adjustment. In the absence of strong theory regarding the survey topic, achieving exchangeability will prove extremely challenging.

Researchers must also be confident that the variables they have identified can account for any indirect confounding resulting from idiosyncrasies associated with recruitment or sampling. Although some vendors consider sampling practices proprietary, vendors must be fully transparent about any variables used in the selection process to ensure that researchers are aware of any potential for confounding.

2.5 Discussion

Whereas the emphasis in probability based surveys has traditionally been to develop processes that minimize confounding, the emphasis suggested here is to first identify likely confounders and design the data collection and analysis so that they are measured and actively accounted for. To be clear, this is more a shift in emphasis than a full-scale departure. Probability based surveys generally seek to measure and account for specific characteristics that are associated with bias, and we have discussed how data collection practices may introduce or mitigate confounding in nonprobability surveys.

Grounding this framework in causal inference suggests that there may be other techniques from that field that can be applied in a survey context. Testing the sensitivity of findings to unmeasured confounding is another common practice in causal inference whose adoption would likely benefit the survey field ([Rosenbaum, 2005](#)). Unlike probability surveys where the maximum range of bias is bounded by the size of the nonresponding sample, selection bias is unbounded and non-identifiable in nonprobability surveys. Although some methods such as pattern mixture models have been developed to evaluate selection bias under such constraints, they are not widely used in practice ([Andridge and Little, 2011](#)). Other techniques that do not rely on assumptions about the probability of selection may also prove useful for nonprobability surveys (e.g. [Manski, 2007](#); [Robins et al., 1999](#)). Additionally, the use of causal diagrams and other methods of identifying confounders represent another worthwhile area for future research (e.g. [Myers et al., 2011](#); [Pearl, 2009b](#); [Steiner et al., 2010](#)).

Finally, it is one thing to know in principle that exchangeability, positivity and composition must be achieved in order to avoid selection bias in

nonprobability survey estimates. It is another thing to achieve them successfully in practice. Even when the subject matter is well known and many likely confounders are identified, it may prove difficult to have complete confidence that there is not some yet unknown factor quietly introducing bias into survey estimates. Nevertheless, by making explicit a set of assumptions that to date have been largely implicit, the notions of exchangeability, positivity and composition provide a framework by which to evaluate and critique specific research findings and improve methodological practice.

Chapter 3: Decomposing selection bias in nonprobability surveys

For both probability and nonprobability surveys, researchers devote a great deal of effort to identifying and mitigating sources of selection bias in survey estimates. By selection bias, we mean a difference between a survey estimate and the true population quantity that arises because some aspect of the selection process results in samples whose composition does not match that of the population. This is in contrast with sampling error, which is not systematic, and errors of measurement. The Total Survey Error (TSE) framework identifies noncoverage and nonresponse as the primary sources of selection bias in surveys. Noncoverage occurs when units in the population are missing from the sampling frame and have no possibility of selection. Nonresponse is when units that have been selected fail to complete the survey (Biemer, 2010; Groves, 1989; Groves and Lyberg, 2010). Bias results when the noncovered or nonresponding units are systematically different from the observed units with respect to the outcome of interest.

The TSE framework was developed under the probability-based survey paradigm where the validity of statistical inferences comes from the fact that sampled units are randomly selected from the population. When a sample is selected randomly from a frame, undercoverage and nonresponse describe ways in which the actual data collection process differs from the ideal of perfect randomization. In principle, selection bias could be eliminated from probability-based surveys by achieving 100% coverage and response. The same cannot be said for nonprobability surveys. In the absence of random selection,

perfect coverage and response provide no information about the possibility of systematic differences between a sample and the population.

For nonprobability surveys, statistical inferences are not – by definition – based on randomization. Rather, they are based on a model that (explicitly or implicitly) assumes a relationship between the units in the sample and the target population. Bias results when the model used to produce an estimate incorrectly specifies this relationship. Whereas reducing bias in probability-based estimates requires identifying deficiencies in the randomization mechanism (i.e. undercoverage and nonresponse), reducing bias in nonprobability estimates requires identifying deficiencies in the modeling assumptions.

Most previous research into selection bias in nonprobability samples has applied a standard estimation procedure to different nonprobability and probability-based samples and compared the resulting estimates either to each other or to population benchmark values. Some of this research found that nonprobability samples yielded consistently inferior estimates to probability samples (e.g. [Chang and a. Krosnick, 2009](#); [Yeager et al., 2011](#)). Other studies have found nonprobability survey estimates to compare favorably to probability-based estimates (e.g. [Ansolabehere and Rivers, 2013](#); [Ansolabehere and Schaffner, 2014](#); [Wang et al., 2014](#)). Those studies that have evaluated estimates from several different nonprobability sample vendors have generally found that the level of bias varies considerably across nonprobability sample vendors ([Craig et al., 2013](#); [Erens et al., 2014](#); [Gittelman et al., 2015](#); [Kennedy et al., 2016](#); [Yeager et al., 2011](#)).

Most studies at least mention the need to make assumptions about selection being ignorable given some set of adjustment variables, but aside from noting

the presence or absence of biased estimates there is rarely any additional probing into whether these assumptions are violated and how. This may be attributable in part to the fact that under the standard TSE approach to survey error, and design-based inference more broadly, the goal is to minimize reliance on unverifiable assumptions and focus attention on measurable phenomena such as coverage and nonresponse. For nonprobability surveys, however, there is no escaping unverifiable assumptions. A better understanding of the sources of and solutions to selection bias in nonprobability surveys requires a framework that places the assumptions front and center and puts the focus on assessing the degree to which those assumptions are justified.

In the previous chapter, we proposed such an alternative framework focused on the three conditions that must be met in order for nonprobability surveys to produce population estimates that are free from selection bias: exchangeability, positivity, and composition. Exchangeability is the requirement that the researcher has measured any variables necessary to render the survey outcome conditionally independent of sample membership. Positivity is the requirement that there are no portions of the population that are entirely absent from the sample. Composition is the requirement that the joint distribution of any confounding variables either matches or can be made to match the population distribution. When all three of these requirements are met, there can be no systematic selection bias in survey estimates. These ideas are not new, but are drawn from the field of causal inference which has grappled with the problem of estimating causal effects from non-experimental data for decades (e.g. [Rubin, 1974](#)).

The current chapter further develops the proposed framework by providing a more rigorous formulation for the different types of selection bias. In section

3.1, we describe exchangeability, positivity, and composition in detail in the form of conditional probabilities. In section 3.2.1, we show that the net selection bias in a survey estimate can be decomposed into separate, additive components associated with exchangeability, positivity, and composition respectively by taking the differences of several conditional means. We then show how to estimate the values of these components if a satisfactory reference dataset is available. In section 3.3, we provide an empirical example. We estimate bias components for six different outcome variables related to civic engagement for the nine nonprobability samples evaluated by Kennedy et al. (2016) as well as an additional sample collected using Amazon Mechanical Turk. We assess the how the magnitude of each bias component varies both for individual survey outcomes and across sample sources.

3.1 Survey estimates and selection bias

For a target population of size N , let $S = 1$ indicate a set of n units included in a survey sample. For nonsampled units $S = 0$. Assume there is a variable X that is measured for every unit in the population, a variable U that is unknown for every unit in the population, and an outcome variable Y that is measured for the units in the sample but unknown for the nonsampled units. Specific realized values of Y , X and U are indicated in lower case as y , x , and u respectively. For pedagogical simplicity, we will assume that Y is binary and that X and U are univariate and discrete, and that n is small relative to N such that $n/N \approx 0$.

To produce estimates from the sample that generalize to the larger population, the distribution of the observed variable in the sample must match the distribution in the larger population such that

$$\Pr(Y | S = 1) = \Pr(Y). \quad (3.1)$$

Because $n/N \approx 0$, we assume $\Pr(Y) = \Pr(Y | S = 0)$, which means we can formulate bias as a systematic difference between sampled and nonsampled units

$$\delta_Y = \Pr(Y | S = 1) - \Pr(Y | S = 0). \quad (3.2)$$

Under this formulation, δ_Y is analogous to a treatment effect in causal inference where $S = 1$ and $S = 0$ are the treatments. In causal studies, Y is measured on both treated and nontreated units in the population and the groups are compared. In surveys, we generally assume that sampled and nonsampled units are not systematically different and that measurements on the sampled units will also accurately describe the nonsampled units. If we accept that there is no direct causal relationship between inclusion in the sample and Y , then any observed difference between population and sample values must be the result of differences between $\Pr(Y, X, U | S = 1)$ and $\Pr(Y, X, U | S = 0)$.

As in causal inference, the nature of selection bias can be represented in terms of the conditional dependencies between Y , X , U and S (Hernán et al., 2004; Pearl, 2009a; Rosenbaum and Rubin, 1983a). To produce unbiased population estimates from a sample, three conditions must be met: exchangeability, positivity and composition. If we assume knowledge of $\Pr(Y | X, S)$ and $\Pr(X | S)$, it is possible to construct a set of hypothetical survey and population means. By taking the differences of these quantities, we can partition the net bias δ_Y into separate terms associated with each component

such that:

$$\delta_Y = \delta^{\text{exch}} + \delta^{\text{pos}} + \delta^{\text{comp}} \quad (3.3)$$

We will restrict our focus to estimates of population proportions, but this approach can be applied to other types of estimates as well. In this section we will describe the exchangeability, positivity, and composition requirements in terms of conditional probabilities.

3.1.1 Exchangeability

Exchangeability describes the situation where the distribution of Y is the same for both sampled and nonsampled units, either unconditionally or conditional on some set of observed characteristics X . When Y is unconditionally independent of S , denoted $Y \perp\!\!\!\perp S$, we say that the sampled and nonsampled cases are unconditionally exchangeable with respect to Y .

This is the case in expectation under probability sampling when all units have an equal probability of selection, and implies that

$\Pr(Y | S = 1) = \Pr(Y | S = 0)$. If $\Pr(Y) \neq \Pr(Y | S = 1)$ then sampled observations are not exchangeable and sample estimates will be biased unless Y can be made independent of S by conditioning on observed variables.

If $Y \perp\!\!\!\perp S | X$ then nonsampled units are said to be conditionally exchangeable, which implies that $\Pr(Y | X = x, S = 1) = \Pr(Y | X = x, S = 0)$. If all possible values of x in X have been observed in the sample and the population distribution of X is known, then the unconditional population distribution of Y can be recovered as follows:

$$\Pr(Y) = \Pr(Y | X, S = 1) \Pr(X | S = 0). \quad (3.4)$$

If $Y \not\perp S | X$ then sampled and nonsampled units are not exchangeable, meaning that we would need to condition on U in order to make $Y \perp S$. Since U is unobserved for the sampled units, it cannot be conditioned upon and as a result the population distribution of Y cannot be recovered without strong assumptions (Greenland and Robins, 1986, 2009). In the survey context, the key to achieving conditional exchangeability is ensuring that all confounding variables are measured for all sampled units.

Unconditional and conditional exchangeability are often referred to as missingness completely at random (MCAR) and missingness at random (MAR) respectively (Little and Rubin, 2002). We prefer the term exchangeability as it more directly emphasizes the necessity of equivalence between sampled and nonsampled units that share a set of observed characteristics.

If we accept that $Y \perp S | X, U$ it follows that

$$\Pr(Y | X, U) = \Pr(Y | X, U, S). \quad (3.5)$$

We can marginalize over either U or S on the right hand side of (3.5) to demonstrate that

$$\Pr(Y | X, S) = \Pr(Y | X, U), \quad (3.6)$$

thus implying that any observed difference between $\Pr(Y | X, S = 1)$ and $\Pr(Y | X, S = 0)$ is the result of confounding by U . Effectively, S serves as an instrumental variable for measuring the effect of U on Y (Angrist et al., 1996).

The magnitude of the bias due to the confounding influence of U on the population proportion is

$$\delta^{\text{exch}} = \sum_{x \in X} [\Pr(Y | X = x, S = 1) - \Pr(Y | X = x, S = 0)] \Pr(X = x | S = 1). \quad (3.7)$$

This is equivalent to the average effect of treatment on the treated (ATT) in causal inference terminology, and is the weighted sum of conditional average treatment effects on the treated (CATT) over X , where “treatment” is inclusion in the sample. Because there is no true treatment effect of S , we refer to δ^{exch} as a confounding effect. Confounding effects may not be the same for all values of X , and we can think of δ_x^{exch} as a conditional confounding effect for a particular value of X . We can think of δ_i^{exch} as the expected difference between sampled and nonsampled units who share the same value of x as the i th unit in the sample.

3.1.2 Positivity

Conditional exchangeability is a necessary but insufficient condition for producing unbiased population estimates. The positivity requirement states that if Y is conditionally exchangeable given X , all distinct values of X must be observed in the sample, or that $\Pr(S = 1 | X = x) > 0$ for all values of $x \in X$ (Hernán and Robins, 2006; Petersen et al., 2012). To illustrate why this is the case, we can reformulate (3.4) as

$$\Pr(Y) = \sum_{x \in X} \frac{\Pr(Y | X = x, S = 1) \Pr(X = x | S = 1)}{\Pr(S = 1 | X = x)}. \quad (3.8)$$

If there are any instances where $\Pr(S = 1 | X = x) = 0$ (i.e. excluded from

the sample), (3.8) is undefined. Intuitively, groups that are underrepresented but present in the sample provide some information that can be used to generalize back to similar units in the population, but we have no information about groups that are missing from the sample altogether.

The magnitude of bias resulting from a failure of positivity is simply the difference between the population mean for those portions of the population that are represented in the sample and the mean for the full population. Let $\phi_x = 1$ if $\Pr(S = 1 | X = x) > 0$ and 0 otherwise. We can quantify the bias due to a failure of positivity as

$$\delta^{\text{pos}} = \Pr(Y | S = 0, \phi_x = 1) - \Pr(Y | S = 0) \quad (3.9)$$

which is equivalent to

$$\delta^{\text{pos}} = \Pr(\phi_x = 0 | S = 0) [\Pr(Y | S = 0, \phi_x = 1) - \Pr(Y | S = 0, \phi_x = 0)] \quad (3.10)$$

Equations (3.9) and (3.10) are similar to the traditional formulations for coverage and nonresponse error in probability-based surveys with the difference being that ϕ_x is necessarily conditional on an observed X and therefore conditional on a model in which X has been specified. In contrast, noncoverage and nonresponse are not necessarily conditional on any chosen set of covariates but rather on the survey design and execution (Groves, 1989). As with coverage and nonresponse error in probability-based surveys, the magnitude of bias due to a lack of positivity depends on what proportion of the population is systematically excluded and how different the excluded units

are with respect to Y .

When exchangeability and positivity both hold, the condition is known as strong ignorability (Rosenbaum and Rubin, 1983b). Under strong ignorability, the conditional distribution of $Y | X$ in the sample matches that of nonsampled population, and unbiased predictions can be made about any units in the target population for which X is known. In causal studies, strong ignorability of treatment assignment only guarantees internal validity in that it permits unbiased causal inferences about the units included in the study (Shadish et al., 2002). The ability to generalize from an experimental sample to a larger population, or external validity, requires an additional layer of strong ignorability with respect to inclusion in the study (Stuart et al., 2011). The survey context is less complicated in that only inclusion in the sample must be strongly ignorable for the outcomes of interest.

3.1.3 Composition

When the exchangeability and positivity requirements are both met, it is possible to make predictions about individual units in the population if we know their value for X . However to estimate population parameters we must also know the distribution of X in the population, or $\Pr(X | S = 0)$. Bias that results from problems with composition amounts to the error that comes from having the necessary kinds of units in the sample but in the wrong proportions. More formally, it is the difference between the sample mean if there were no confounding and the mean for the set of nonsampled units where $\phi_x = 1$:

$$\delta^{\text{comp}} = \sum_{x \in X} \Pr(Y | X = x, S = 0) [\Pr(X = x | S = 1) - \Pr(X = x | S = 0, \phi_x = 1)]. \quad (3.11)$$

We might be tempted to think of a violation of the positivity requirement as a special case of a composition problem where $\Pr(X = x | S = 1) = 0$, and in a sense this is correct. However, the critical distinction lies in the fact that problems with composition can be corrected if the distribution of X is known, whereas an absence of positivity cannot be rectified without additional strong assumptions. It is worth noting that conventional methods of adjusting for noncoverage and nonresponse used in both probability-based and nonprobability surveys, such as raking and generalized regression estimation (GREG, rely on an assumption of strong ignorability and as such can only correct problems with composition. Additionally, such techniques only work if the population distribution of X is known (Kalton and Flores-Cervantes, 2003). In some instances, it may be possible to achieve strong ignorability only by conditioning on an observed variable whose population distribution is unknown. In this case, it is not possible to estimate population parameters (although it remains possible to make predictions about individual units in the population given).

3.2 *Bias decomposition*

For estimates of a population mean or proportion, it is most straightforward to see how each of these bias terms add up to the net bias if we consider the following conditional means. Let $\bar{y}_{s_1}^{(1)}$ be the mean for the realized survey sample. The superscript (1) indicates that the value is confounded (i.e. based on $\Pr(Y | X, S = 1)$). The subscript s_1 indicates that the value is based on the distribution of units in the in the survey sample $\Pr(X | S = 1)$. Let $\bar{y}_{s_0}^{(0)}$ be the true mean for the population where (0) indicates that the value is unconfounded (i.e. based on $\Pr(Y | X, S = 0)$ and subscript s_0 indicates that the value is based on the distribution of units in the target population, that is

$\Pr(X | S = 0)$. In principle, both of these quantities are observable. Let $\bar{y}_{s_0, \phi=1}^{(0)}$ denote the true mean for the share of the population for which common support exists in the sample. This is not observable unless for some reason ϕ is known. Finally, let $\bar{y}_{s_1}^{(0)}$ denote the counterfactual mean that is based on the observed distribution of X in the sample but is unconfounded. It follows that

$$\delta^{net} = \bar{y}_{s_1}^{(1)} - \bar{y}_{s_0}^{(0)} = \underbrace{(\bar{y}_{s_1}^{(1)} - \bar{y}_{s_1}^{(0)})}_{\delta^{exch}} + \underbrace{(\bar{y}_{s_0, \phi=1}^{(0)} - \bar{y}_{s_0}^{(0)})}_{\delta^{pos}} + \underbrace{(\bar{y}_{s_1}^{(0)} - \bar{y}_{s_0, \phi=1}^{(0)})}_{\delta^{comp}}. \quad (3.12)$$

3.2.1 Estimation

To calculate δ^{exch} , δ^{pos} , and δ^{comp} for a given nonprobability sample, we need to know $\Pr(Y | X, S)$ and $\Pr(X | S)$ as well as ϕ_X . To estimate these quantities, we require data for the nonsampled units in the population in addition to a nonprobability dataset. Since this is impossible in most situations, we employ a reference dataset that is assumed to accurately reflect the population joint distribution $\Pr(Y, X)$. Depending on the application, this could be administrative or census data or a high quality probability-based survey in which both Y and X have been measured. Let Y_s and X_s represent the vectors Y and X for dataset $s \in S = \{0, 1\}$ where 0 and 1 indicate membership in the reference and nonprobability datasets respectively. Let y_{si} and x_{si} represent the realized values of Y and X for unit i in dataset s . For convenience, we will use s_0 and s_1 respectively to refer to the reference and nonprobability datasets in their entirety.

We can estimate the values $\hat{\delta}^{exch}$, $\hat{\delta}^{pos}$, and $\hat{\delta}^{comp}$ by modeling the outcome Y as a function of X and S and calculating the expected counterfactual

outcomes $\hat{y}_{si}^{(1)} = \mathbb{E}(Y | x_{si}, s_1)$ and $\hat{y}_{si}^{(0)} = \mathbb{E}(Y | x_{si}, s_0)$ for each unit in the nonprobability and reference samples. This is paired with a propensity score model for $\Pr(S = 1 | X)$ that is used to estimate $\hat{\phi}_i$ for each observation in the reference dataset. Finally, we rely on the observed distribution of X in the reference and nonprobability datasets themselves for $\Pr(X | S)$.

To model the conditional distribution $\Pr(Y | X, S)$, we fit models to predict $\hat{Y}^{(s)} = f(X_s)$ on the reference and nonprobability datasets separately. By estimating these functions separately, we are implicitly conditioning on S , and we refer to these models as $f_s(\cdot)$. While it would be possible to combine the two datasets and fit a single model estimating $f(X, S)$, if there is substantial covariate imbalance or lack of overlap between the two samples, those regions of X that are highly correlated with S will function as instruments or partial instruments for S . In such situations, [Pearl \(2010\)](#) demonstrated that conditioning on both a treatment variable and an instrument at the same time leads to biased estimates of treatment effects. By fitting models for each dataset separately, we avoid this problem because S is never included in the same model as any potential instruments. For each observation in both datasets, we calculate the values $\hat{y}_{si}^{(0)} = f_0(x_{si})$ and $\hat{y}_{si}^{(1)} = f_1(x_{si})$.

For ϕ_X , we need to identify those observations in the reference dataset for which no comparable units exist in the nonprobability dataset. If X is high dimensional, sparsity makes it likely that there will be many observations for which no exact match exists in the nonprobability dataset. The causal inference literature contains many approaches to identifying the region of X for which common support exists (e.g. [Crump et al., 2009](#); [Dehejia and Wahba, 1999, 2002](#); [Heckman et al., 1997, 1998](#); [Hill and Su, 2013](#); [King and Zeng, 2006](#); [Lechner, 2008](#); [Porro and Iacus, 2009](#)). Here we opt for the simple

approach used by Dehejia and Wahba (1999) which identifies units lacking common support as those control units with propensity scores outside the range of scores observed on the treated units. Evaluating alternative methods for defining the area of common support for nonprobability samples may be a useful area of future research.

To estimate ϕ_X , the reference and nonprobability datasets are stacked into a single dataset, and we fit a propensity model $\hat{\pi} = g(X)$ where $\hat{\pi}_x \approx \Pr(S = 1 | X = x)$. We can then define $\hat{\phi}$ as follows:

$$\hat{\phi}_{si} = \begin{cases} 1, & \text{if } \hat{\pi}_{si} < \min(\hat{\pi}_{1i}) \\ 0, & \text{otherwise} \end{cases} \quad (3.13)$$

Note that this implies that $\hat{\phi}_x = 1$ for all observations in the nonprobability dataset. This is primarily for completeness as $\hat{\phi}_x$ is only needed on the reference dataset for this exercise.

With these estimated values, we can now calculate $\hat{\delta}^{\text{exch}}$ as

$$\hat{\delta}^{\text{exch}} = \frac{1}{n_1} \sum_{i \in s_1} (\hat{y}_i^{(1)} - \hat{y}_i^{(0)}) \quad (3.14)$$

where n_1 is the number of observations in the nonprobability sample.

We calculate $\hat{\delta}^{\text{pos}}$ as

$$\hat{\delta}^{\text{pos}} = \frac{\sum_{i \in s_0} \hat{y}_i^{(0)} \hat{\phi}_i}{\sum_{i \in s_0} \hat{\phi}_i} - \frac{\sum_{i \in s_0} \hat{y}_i^{(0)}}{n_0}, \quad (3.15)$$

and $\hat{\delta}^{\text{comp}}$ as

$$\hat{\delta}^{\text{comp}} = \frac{\sum_{i \in s_1} \hat{y}_i^{(0)}}{n_1} - \frac{\sum_{i \in s_0} \hat{y}_i^{(0)} \hat{\phi}_i}{\sum_{i \in s_0} \hat{\phi}_i}. \quad (3.16)$$

3.3 *Estimating selection bias components on measures of civic engagement*

3.3.1 *Data*

To demonstrate an empirical application of this framework, we estimate the components of selection bias for several questions related to civic engagement. These questions originally come from the 2013 Current Population Survey (CPS) Civic Engagement Supplement, which we treat as the reference dataset representing the true population distribution of outcomes and covariates. To minimize the potential effects of measurement error due to proxy reporting we use the supplement self-respondent weight (PWSRWGT) which yields an unweighted sample size of 27,566. These questions were also asked on a set of nine parallel nonprobability surveys conducted by Pew Research Center in 2015 and described in a report by [Kennedy et al. \(2016\)](#). These surveys were conducted with eight different online sample vendors and are labeled Samples A through I. We also include a survey fielded using Amazon’s Mechanical Turk, a crowdsourcing platform where individuals are paid to complete online tasks. The MTurk survey, also sponsored by Pew Research Center, used the same questionnaire as the the surveys examined in [Kennedy et al. \(2016\)](#). The field dates and sample sizes are listed in [Table 3.1](#). We do not report response rates as they are not substantively meaningful for nonprobability surveys.

We do not apply any weighting adjustments to the nonprobability samples. As such, the estimated bias components will not necessarily reflect a final

estimate, but instead provide a sense of the distribution of bias prior to any attempt to correct it. Because (3.14), (3.15), and (3.16) are based on predicted values for individual units, calculating weighted versions of these quantities is straightforward.

Table 3.1: Survey field dates and sample sizes

Survey	Field dates	Sample size
Sample A	Feb. 25, 2015	1,022
Sample B	Feb. 26 - Mar. 3, 2015	1,049
Sample C	Feb. 25-27, 2015	1,178
Sample D	Feb. 25-27, 2015	1,005
Sample E	Feb. 24 - Mar. 8, 2015	1,022
Sample F	Feb. 25-26, 2015	1,008
Sample G	Oct. 1-6, 2015	1,010
Sample H	Oct. 2-8, 2015	1,007
Sample I	Aug. 19-31, 2015	1,000
MTurk	Sep. 25 - Oct. 2, 2015	1,017

For each of these surveys, we are interested in estimating population percentages for the following six survey outcomes (Y) coded as binary indicators:

- Always votes in local elections.
- Trusts all or most people in their neighborhood.
- Typically talk to their neighbors every day or a few times a week.
- In the last twelve months, participated in a school, neighborhood or community group.
- In the last twelve months, participated in a civic or service organization.

- In the last twelve months, participated in a sports or recreation organization.

For adjustment covariates, we use sex, age, race and Hispanic ethnicity, educational attainment, and the Census Bureau's administrative region. Where categories or scales differed between the CPS and the comparison surveys, variables were recoded into a common set of categories. The question wording for all variables used in this analysis can be found in Appendix A. A description of variable coding can be found in Appendix B along with a description of the process used to singly impute missing values for the demographic variables. With the exception of Census region, none of the demographic variables on any of the samples were missing for more than 2% of the cases. For region, missingness ranged from 1% to 5%. Imputing the few missing values in these variables permits us to retain all of the interviews and avoid any additional biases that would be introduced by performing a complete case analysis. The cost is that variance estimates will be slightly underestimated, though this effect should be largely negligible.

In the study by [Kennedy et al. \(2016\)](#), the weighting also incorporated population density and cellular telephone usage. These variables are not included in the CPS Civic Engagement Supplement and so cannot be included in this analysis. However, even in their absence, the remaining variables represent a standard set of demographics that are often used in quotas and weighting adjustments for both probability and nonprobability surveys. These were the primary demographics used in weighting by [Yeager et al. \(2011\)](#) and to compare quota schemes by [Gittelman et al. \(2015\)](#). They are also the variables used in nonresponse adjustment for a number of major media surveys conducted with both probability and nonprobability samples (e.g. [GfK Public](#)

Affairs, 2016; Ipsos Public Affairs, 2016; The Washington Post and ABC News, 2016). As a result, this analysis speaks directly to current scholarship and practice in the area of nonprobability surveys.

Because the CPS Civic Engagement Supplement is an interviewer-administered telephone survey while the comparison surveys are self-administered, it is possible that some portion of observed differences are the result of mode differences, particularly if measures of civic engagement are socially desirable (Kreuter et al., 2008). To the extent such measurement differences are present in the outcome variables, they will affect the exchangeability component, δ^{exch} . This makes sense in that the factors that cause differential measurement are not observed on both samples, making them effectively unobserved confounders that are associated with measurement rather than selection. Given that Kennedy et al. (2016) found that the nonprobability samples exhibited higher levels of civic engagement than the benchmarks, the presence of social desirability bias in the CPS would imply that the true level of selection bias is greater than our estimates would suggest. Although we expect differential measurement to be minimal for the chosen demographic variables, to the extent that it is present its impact will vary depending on the the nature of the error and its correlation with Y and S . Another possible reason for differences between the CPS and nonprobability estimates would be if the true population value changed substantially between 2013, when the supplement was fielded, and 2015 when the nonprobability surveys were conducted. This is impossible to know for certain. To the extent that there is true population change, this would also manifest as bias due to a lack of exchangeability.

Finally, although the supplement is a high quality, government sponsored

survey with a high response rate, it is subject to sampling and nonresponse error of its own. As such, estimated bias components should be viewed as comparisons to the best available measurements of the outcomes of interest rather than deviations from a hypothetical “true value”.

3.3.2 Estimating bias components with BART

In principle, we could estimate the functions $f_s(\cdot)$ and $g(\cdot)$ using any kind of regression method (Snowden et al., 2011). However, we often lack knowledge of the correct functional form, and linear models can be misspecified if interactions or nonlinearities are not accounted for in the model. In such instances, machine learning methods can help us avoid this kind of model misspecification as they permit us to fit models using a potentially large number of covariates while automatically detecting non-linear associations and complex interactions. In particular, for this sort of exercise we are not interested in drawing inferences about these model parameters but rather about their predicted values. As such, the black-box nature of many such techniques does not pose a problem.

In particular, algorithms that use ensembles of classification and regression trees (CART) such as random forests, bagging, and boosting have attracted interest because of their flexibility, ease of use, and generally high predictive performance (Breiman, 1996, 2001; Friedman, 2002). Lee et al. (2010) found that for propensity score estimation, boosted regression trees (Friedman, 2002) performed almost as well as linear models with respect to bias and mean squared error when the associations between predictors and treatment were linear and additive and performed much better when the associations involved nonlinearities and complex interactions. Austin (2012) had similar results for boosted regression trees when they were used to model the outcome directly.

In this paper, we use Bayesian additive regression trees (BART) to estimate both $f_s(\cdot)$ and $g(\cdot)$ (Chipman et al., 2010). Hill (2011) proposed the use of BART for estimating causal effects by directly modeling the outcome as a function of treatment assignment and baseline covariates, and demonstrated its efficacy in a simulation. Green and Kern (2012) also demonstrated BART’s utility in estimating heterogeneous treatment effects, while Kern et al. (2016) found BART to be particularly effective for generalizing causal effects from experimental samples to larger populations. Hill et al. (2011) also found BART to outperform other machine learning and parametric approaches to estimating propensity scores.

Similar to boosted regression trees, BART approximates a function $f(\cdot)$ with an additive model consisting of m trees where

$$y_i = f(x_i) \approx \bar{y} + \sum_{j=1}^m h(x_i; T_j, M_j) + \epsilon_i \quad (3.17)$$

where T_j refers to the splitting rules and structure of tree j , M_j refers to the set of expected values for each terminal node in the tree, and $\epsilon \sim N(0, \sigma^2)$. To prevent overfitting, BART employs a regularization prior for T_j , M_j and σ that keeps individual trees small in terms of the number of splits, and shrinks the values of M_j toward 0. The hyperparameters that define the prior distribution can be tuned to provide more or less regularization. One appealing attribute of BART is that the default values of these hyperparameters have been found to perform very well on a wide variety of problems. Here we use the default values, though it is also possible to find optimal values via cross-validation (Chipman et al., 2010).

The model is fit using a Gibbs sampling algorithm where the structure of each

T_j is randomly perturbed over many iterations according to a procedure described by [Chipman et al. \(1998\)](#). The series of iterations is a Markov chain, draws from which are used to approximate the posterior distribution of $f(X)$ once it has converged.

For dichotomous outcomes of the sort considered here, BART fits a probit model. We use this probit implementation of BART estimate both $f_s(\cdot)$ and $g(\cdot)$ using the BART package for the R statistical computing platform ([McCulloch et al., 2018](#); [R Core Team, 2017](#)). After 1,000 burn-in iterations of the Markov chain, we estimate $\hat{\delta}_k^{\text{exch}}$, $\hat{\delta}_k^{\text{pos}}$, and $\hat{\delta}_k^{\text{comp}}$ over $k = 1 \dots 1000$ draws from the posterior distributions of $f_s(\cdot)$ and $g(\cdot)$ to quantify the uncertainty of the estimates. For point estimates, we report the posterior mean value for a statistic and 95% credibility intervals as measures of uncertainty.

Because BART is not compatible with the complex sample design features of the CPS, we use the finite population Bayesian bootstrap (FPBB) to create a synthetic population based on the weighted distribution of observations in the reference sample ([Cohen, 1997](#); [Dong et al., 2014](#); [Ghosh and Meeden, 1997](#); [Zhou et al., 2016](#)). To create the synthetic population, we follow the procedure described by [Dong et al. \(2014\)](#) using the CPS supplement as the reference sample. First, the weights for each observation are scaled so that they sum to the sample size which we denote n_r . Next, we resample a total of $N - n_r$ observations from the reference sample using a weighted Pólya urn scheme, where N is the size of the full target population. In practice, the size of the synthetic population only needs to be many times larger than the reference sample. In this case, we create a synthetic population that is 100 times larger than the original CPS dataset. These resampled units are then combined with the original n_r units to create a synthetic population of size N . Effectively,

this procedure is imputing the $N - n_r$ unobserved units in the population based on a posterior predictive distribution generated from the weighted reference sample. This process creates a dataset that retains the same joint distribution as the weighted CPS sample but can be used with procedures that do not accommodate survey weights. To fully incorporate the sampling variance from the CPS into our estimates of uncertainty we would create a large number of synthetic populations and replicate the analysis on each of them. Through experimentation we found that very little changed with multiple synthetic populations, so for simplicity we use only a single synthetic population in this analysis.

We do not fit the BART models for the outcomes or inclusion propensities with the entire synthetic population dataset but rather with a subsample equal in size to the nonprobability sample. This serves two purposes. First, fitting the models to such a large dataset would be computationally intractable. Second, this creates a balanced number of reference and survey cases when we fit the inclusion propensity models. This improves the quality of the estimated inclusion propensities. When the two datasets are substantially imbalanced (as would be the case if we combined the entire synthetic population with the survey sample), the estimated probabilities tend to be biased in favor of predicting membership in the larger group. We subsample the synthetic population rather than oversample the survey data so that we do not artificially increase the level of precision for our estimates (Wallace et al., 2011).

The code used to fit these models and estimate the conditional mean values used to calculate the bias components can be found in Appendix C.

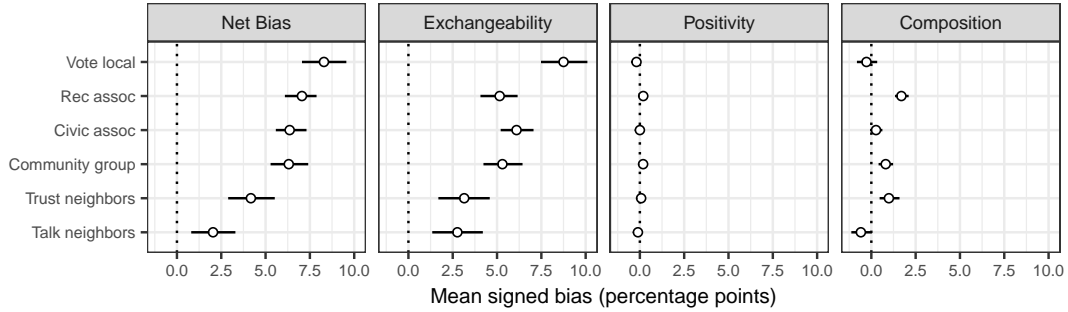


Figure 3.1: Mean signed bias by outcome averaged over all samples. Bars depict 95% credibility intervals.

3.4 Results

Having estimated $\hat{\delta}^{\text{exch}}$, $\hat{\delta}^{\text{pos}}$, and $\hat{\delta}^{\text{comp}}$ for each outcome across all 10 nonprobability datasets, we have several primary research questions. First, in the aggregate, do individual outcomes exhibit different patterns with respect to the relative contribution of each component, and are these patterns consistent in magnitude across vendors? Similarly, is the variability in average bias across sample sources disproportionately attributable to specific bias components, and are the patterns consistent across individual outcomes?

Figure 3.1 depicts the net bias for each question and the values for the three bias components averaged over all samples. The exchangeability component is clearly the primary source of selection bias at the question level. Positivity, on the other hand, contributes almost nothing to bias. Bias attributable to incorrect composition, while present, tends to be small, with the largest component estimated at just under two percentage points for the share that participated in a recreational association in the past year.

However, Figure 3.2 suggests a more complicated pattern when it comes to individual samples. Samples displaying the highest mean absolute net bias tend to have correspondingly high values for the exchangeability component.

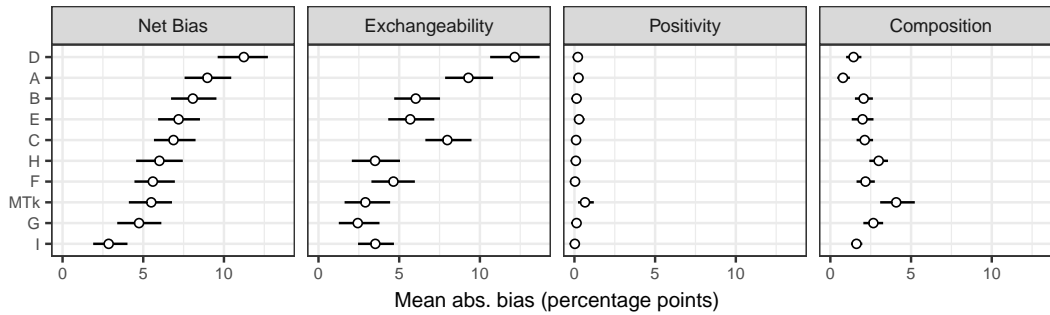


Figure 3.2: Mean absolute bias for samples averaged over all six outcome variables. Bars depict 95% credibility intervals.

Once again, the contribution of positivity to net selection bias appears minimal. This is likely because the share of the population for which common support holds tends to be quite high, ranging from 96% to nearly 100% for the conventional nonprobability samples. The exception is Mechanical Turk, which only covers an estimated 88% of the target population. Even there, the mean absolute bias attributable to positivity is estimated at under 1 percentage point. For composition, the estimated mean absolute bias ranges from 1 to 3 percentage points for the conventional samples and reaches a maximum of 4 points for Mechanical Turk.

There also appears to be an inverse relationship between composition and exchangeability. Figure 3.3 plots this relationship directly at the sample level. The pattern suggests that, on average, the samples with demographic distributions that most closely match the population also tend to suffer from a higher degree of confounding for these outcome variables.

Figure 3.4 shows the absolute values of bias components for each outcome within samples, which are sorted from left to right by average absolute net bias across all six outcome variables. While it is clear that there is a great deal of variability within samples when it comes to the level of exchangeability bias for individual estimates, the loess regression line shows clearly that the level of

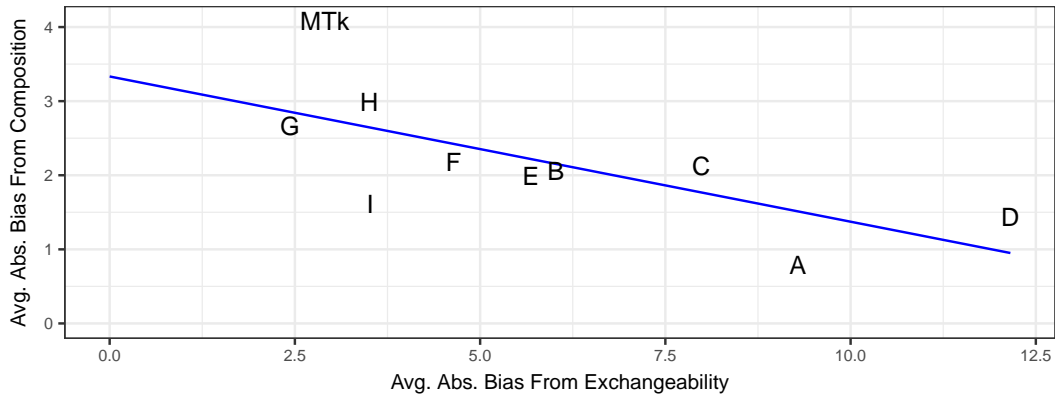


Figure 3.3: Scatterplot of average absolute bias due to exchangeability and composition for all samples. Estimates are averaged over all outcome variables.

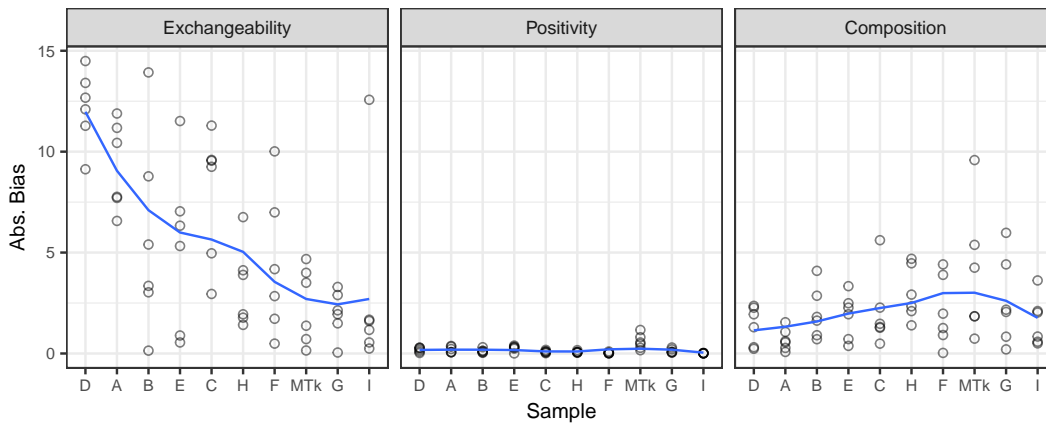


Figure 3.4: Bias components for individual questions across samples. Circles are the absolute estimated bias component values for individual variables. Samples are ordered from highest to lowest by mean absolute net bias across all six outcomes.

exchangeability bias is, to a large degree, a function of the sample. We can also see that the average absolute exchangeability bias for Sample I is disproportionately affected by one particularly large outlier – always voting in local elections – that is roughly 9 percentage points higher than the next highest item.

If we organize the data by individual outcome variables as in Figure 3.5 we can see that for individual estimates, exchangeability is by far the strongest contributor to net bias despite variability in the level of exchangeability bias

within samples. Here the circles reflect the estimated value of a specific bias component on the vertical axis and the net bias for the item on the horizontal axis. For all but voting in local elections and trusting neighbors, the regression lines for positivity and composition are flat, meaning that differences between samples in the net bias for individual survey outcomes are not strongly associated with either of these error sources. For trusting neighbors, the two samples with the highest net biases show a roughly equal mix of bias from exchangeability and composition, while the remaining samples generally only suffer from one or the other. Voting in local elections is particularly interesting. One sample – Mechanical Turk – shows a large, negative bias that is almost entirely attributable to composition. The conventional survey samples all have positive bias due to confounding, although we also see several instances where negative biases from composition offset this effect and reduce the overall net bias. In fact, Sample C has the lowest absolute net bias at 4 percentage points, but only because a large bias term for exchangeability (10 points) is offset by a sizable bias term in the opposite direction (-6 points) for composition.

3.5 Discussion

In this chapter, we have demonstrated mathematically how exchangeability, positivity, and composition relate to the total selection bias for a survey estimate and shown how these individual components can be estimated when a reference dataset of sufficient quality and containing the necessary variables is available. The empirical example nicely illustrates the framework’s potential utility.

Given the overall high level of confounding bias relative to positivity and composition, the most obvious finding is that the basic demographics we have conditioned on in this study are generally poor covariates for explaining

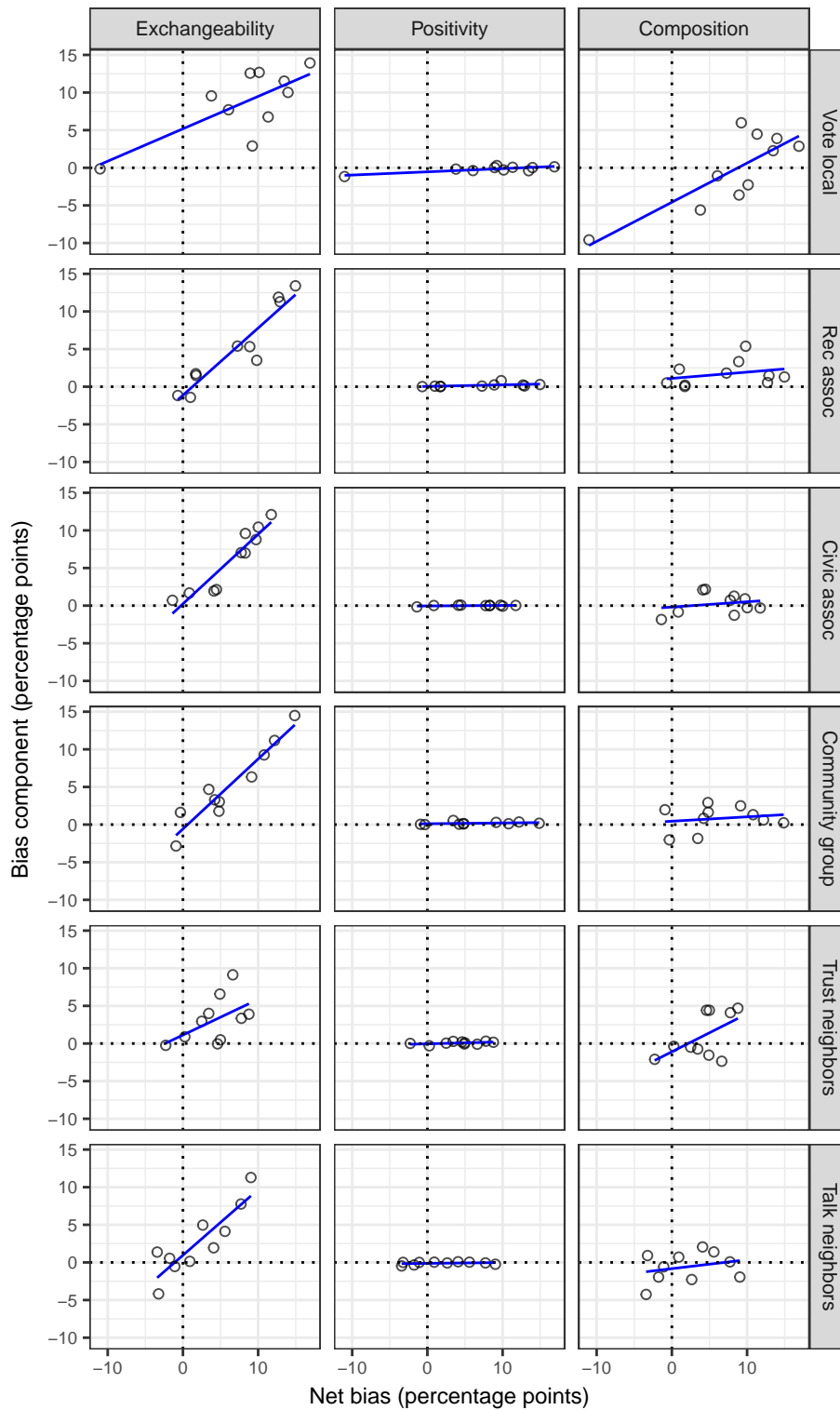


Figure 3.5: Scatterplots of estimated bias components by net bias across outcomes. Variables are ordered from highest to lowest by average absolute net bias across all 10 samples.

differences between these samples and the larger population. That bias due to a lack of positivity is negligible could be considered a positive result. On the other hand, the fact that bias due to composition also tended to be low suggests that statistical adjustments that condition on these demographics should be expected to have limited success. While this study can only speak to these six civic engagement variables, other studies have found that demographic variables are often insufficient for the purposes of correcting selection bias (e.g. [Lee, 2006](#); [Mercer et al., 2018](#); [Schonlau et al., 2007](#)). Of course finding that they are insufficient tells us little about which variables would improve things.

However, the ability to disentangle problems of exchangeability from positivity and composition opens up a variety of new paths for identifying possible solutions. For example, replicating this analysis among subgroups would make it easier to identify specific problem groups and to determine if the possible solutions involve adjusting weighting (composition), recruiting additional types of respondents (positivity), or soliciting expert help to identify possible confounders (exchangeability). If additional confounders were found and incorporated into this kind of analysis, we would see the bias shift from exchangeability into either positivity or composition depending on whether the additional detail identifies certain types of units as missing entirely or simply out of proportion. Repeating this kind of analysis with different sets of covariates may also help identify variables that are either ineffective or actually amplify bias when used for weighting ([Kreuter and Olson, 2011](#); [Pearl, 2010](#)). When estimates appear to have little to no bias, estimating these bias components could help determine if the estimate really is free of bias or if there are offsetting biases that cancel each other out.

Perhaps more interesting is the possibility of an inverse relationship between exchangeability and composition bias. It is plausible that efforts to force a sample to meet a rigid demographic profile could result in something that matches the population very closely on demographics but introduces new confounders. For instance, if an online panel went out of its way to recruit Hispanics by partnering with a corporation whose customers were all located in California, it might have an easier time meeting quotas, but the empanelled Hispanics would not be representative of the broader Hispanic population. While one study involving a limited set of outcome variables is in no way definitive, it does suggest a potentially fruitful avenue for future research. It might be argued that this kind of analysis is only possible with a sufficiently detailed reference dataset that already contains the true population values. This is certainly true, but this criticism is equally true for research into nonresponse or coverage error in probability-based surveys. Even when this sort of reference data is unavailable, the causal framework provides a ready set of tools for hypothesizing about problems and reasoning about potential solutions. Most importantly, it makes it easier for researchers to identify and scrutinize their own assumptions during survey design and analysis.

Chapter 4: Doubly-robust inference for nonprobability surveys with BART

In Chapter 2, we described a variety of methods for correcting selection bias in nonprobability samples, all of which depend on an assumption of strong ignorability (Rosenbaum and Rubin, 1983b). All of these methods involve the creation of a statistical model that induces conditional independence between survey outcomes and inclusion in the sample, although they go about it in different ways. Some, such as propensity weighting and sample matching, do this by modeling the probability of inclusion in the sample (Lee, 2006; Rivers, 2007; Rivers and Bailey, 2009; Valliant and Dever, 2011). Others, such as calibration methods and multilevel regression and poststratification (MRP) do this by modeling the outcome variable (Ghitza and Gelman, 2013; Park et al., 2004). Elliott and Valliant (2017) describe these two approaches as quasi-randomization and superpopulation inference respectively.

Doubly-robust estimation constitutes a third approach. Doubly-robust estimation involves fitting both a propensity model and an outcome regression model and requires that only one or the other be correctly specified to produce consistent population estimates (Bang and Robins, 2005; Kang and Schafer, 2007; Robins et al., 1994).

In this chapter, we compare two approaches to doubly-robust estimation to singly-robust estimation using propensity weighting (PW) and outcome regression (OR). The specific doubly-robust estimators are outcome regression with residual bias correction (OR-RBC) and outcome regression with a propensity score covariate (OR-PSC). Each of these is described in detail in

Section 4.1. As in Chapter 3, we use Bayesian additive regression trees (BART) to construct all four of these estimators (Chipman et al., 2010). For details on the BART algorithm see Chapter 3, Section 3.3.2.

For online nonprobability samples, doubly-robust estimation has an intuitive appeal. Given the general lack of visibility into the recruitment and sampling process, having two chances to correctly specify a model seems like a potentially useful hedge against the inherent uncertainty about the selection mechanism. That said, there are disadvantages as well. A doubly-robust estimator will usually be less efficient than a correctly specified estimate based on outcome regression. Bang and Robins (2005) suggest that additional bias robustness is worth some loss of efficiency given that all models are likely to suffer from some degree of misspecification. In contrast, Kang and Schafer (2007) evaluated a variety of different doubly-robust estimators and found that when both models were misspecified, a singly-robust estimate based only on outcome regression had lower bias and root mean squared error (RMSE) than all of the doubly-robust alternatives. In his commentary, Tan (2007) demonstrated that this finding is by no means universal and the relative performance of singly or doubly-robust estimators will vary depending on the situation and the specific estimator.

The findings of both Bang and Robins (2005) and Kang and Schafer (2007) are based on simulations. In practice, researchers will rarely have the necessary information required to determine the optimal type of estimator. For online nonprobability surveys, there is some empirical evidence that doubly-robust estimation may be more helpful than not. Studies comparing propensity weighting to other techniques have generally found it to be less effective than calibration for reducing bias on a variety of benchmarks.

However, a first-stage propensity adjustment followed by a second stage of calibration does appear to yield somewhat more bias reduction than calibration on its own (Dutwin and Buskirk, 2017; Mercer et al., 2018). Brick (2015) described this as a compositional approach and showed that it is a form of doubly-robust estimation. Lee and Valliant (2009) proposed a similar two-stage procedure but did not describe it in terms of double-robustness.

These empirical studies and the aforementioned simulations all rely on parametric linear models for both propensity and outcome estimation. This means that a failure to correctly capture interactions or nonlinearities in either the outcome or propensity models remains a potential source of error in addition to omitted confounders or lack of common support. The partial exception is the study by Mercer et al. (2018) which used random forests, a tree-based machine learning algorithm, to estimate propensity scores (Breiman, 2001).

Such flexible machine learning algorithms are appealing in that they can automatically detect interactions and nonlinearities, and readily accommodate a large number of covariates. For outcome regression, a number of studies have found BART performs particularly well in a variety of applications including imputation (Tan et al., 2018; Xu et al., 2016), estimating causal effects (Green and Kern, 2012; Hill, 2011; Hill et al., 2011), and generalizing from experimental samples to target populations (Kern et al., 2016).

Tree-based methods such as random forests and boosted regression trees have been found to be effective for the purpose of estimating propensity scores (Buskirk and Kolenikov, 2015; Kern et al., 2016; Lee et al., 2010; McCaffrey et al., 2004). We are aware of only one study to use BART for propensity weighting. It found that propensity scores estimated with BART were less

variable than scores estimated using any of boosted regression trees, logit, and probit regression. At the same time, the propensity weights estimated with BART also produced better covariate balance than the alternatives. Even so, the same study found that an outcome regression model using BART was preferable to estimates based on propensity scores (Hill et al., 2011).

In a simulation study focused on imputation of missing data, Tan et al. (2018) used BART to extend two doubly-robust estimators: augmented inverse probability weighting (AIPW) (Robins et al., 1994) and penalized spline of propensity prediction (PSPP) (Zhang and Little, 2009). They found that estimators that replaced linear propensity and outcome models with BART, which they called AIPW with BART and BARTps respectively, generally resulted in estimates with lower bias and RMSE than standard AIPW and PSPP when both models were misspecified. The added robustness came with only a minimal loss of efficiency relative to linear models when both outcome and propensity models were correctly specified. BARTps proved the most effective method under dual misspecification when the mean and propensity functions involved complex nonlinearities and interactions. An estimator based only on outcome regression with BART was close to but not quite as robust as BARTps. The authors did not evaluate a pure propensity weighting estimator.

In this chapter, we assess whether doubly-robust estimation with BART can be similarly useful for online nonprobability surveys and measure the extent to which the resulting estimates differ from singly-robust approaches based purely on propensity scores or prediction. Because all of our estimators rely on BART, we adopt a different naming scheme from that used by Tan et al. (2018). We compare estimates produced using propensity weighting (PW), outcome regression (OR), and two doubly-robust estimators. The first is OR

estimation with a residual bias correction (OR-RBC) in which the mean of propensity weighted residuals is added to the OR estimate similar to an AIPW estimator (Kang and Schafer, 2007; Robins et al., 1994). The second is an outcome regression estimator that includes the propensity score as a covariate (OR-PSC) similar to the BARTps estimator from Tan et al. (2018).

We compare the performance of these estimators for six binary measures of civic engagement taken from the 2013 Current Population Survey (CPS) Civic Engagement Supplement. The estimates are calculated using 10 different nonprobability surveys commissioned by Pew Research Center in 2015. In the previous chapter, we demonstrated that these items suffer from nonignorable selection bias conditional on demographics. Consequently, we do not expect any of these methods to produce entirely unbiased estimates. Instead, we wish to see if any of these approaches produces consistently superior results in terms of bias, variance, and mean squared error across a diverse set of samples from different vendors on a set of outcomes that, while focused on the topic of civic engagement, serve as exemplars of the kind of problems that can occur in practice.

This chapter proceeds as follows: in section 4.1, we describe each of these estimators in detail and consider their advantages and disadvantages when exchangeability and positivity assumptions are violated. In section 4.2, we compare the performance of each estimator with respect to bias, variance, and root mean squared error (RMSE) for five measures of civic engagement on 10 different nonprobability samples. In section 4.3 we discuss the extent to which these results may generalize to other situations and conclude with suggestions for future research.

4.1 Alternative approaches to nonprobability survey inference

Each of the estimators considered in this study requires unit level microdata that reflects the true joint distribution of the covariates to be used in estimation. As in Chapter 3, we use the 2013 CPS Civic Engagement Supplement microdata as a reference dataset that is assumed to accurately reflect the true population distribution for sex, age, race and Hispanic ethnicity, educational attainment, and the Census Bureau’s administrative region for U.S. adults ages 18 or older. Because BART is not compatible with the complex sample design features of the CPS, we use the same finite population Bayesian bootstrap (FPBB) procedure as in Chapter 3, Section 3.3.2 to create a synthetic population based on the weighted distribution of observations in the reference sample (Cohen, 1997; Dong et al., 2014; Ghosh and Meeden, 1997; Zhou et al., 2016).

Beyond “undoing” the survey weights, the FPBB can also permit us to account for the CPS’s complex design in measures of uncertainty for the nonprobability estimates. Although Dong et al. (2014) and Zhou et al. (2016) describe FPBB methods that account for clustering and stratification in addition to unequal weights, these techniques require the original cluster and strata variables, which are not available for CPS public use data. Instead, we treat the CPS reference sample as if it were a single stage survey with unequal probabilities of selection.

In most applications, one would create a large number of synthetic populations to capture the sampling variance associated with the complex design. This is of particular importance when the synthetic populations are to be used for primary analysis as described by Dong et al. (2014) or for multiple imputation as in Zhou et al. (2016). In this case, some experimentation demonstrated that

measures of variability were almost entirely unaffected by the use of multiple synthetic populations. This is likely due to the large sample size of the reference sample (27,566 adults) both in absolute terms and relative to the nonprobability samples which have sample sizes of approximately 1,000. This is similar to the situation that arises when ignoring the additional variance attributable to the use of estimated control totals in calibration weighting. This added variance is generally minimal when the estimated control totals are very precise and the benchmark sample is many times larger than the analytic sample (Dever and Valliant, 2016). Therefore, for ease of explanation we present results using only a single synthetic population. Though the resulting variance estimates may be smaller than if the CPS’s complex design was fully accounted for, the differences appear to be minimal.

To create the synthetic population, we follow the procedure described by Dong et al. (2014) using the CPS supplement as the reference sample. First, the weights for each observation are scaled so that they sum to the sample size (as opposed to the population size as is the case for most government surveys). Next, we resample a total of $N - n_r$ observations from the reference sample using a weighted Pólya urn scheme, where N is the size of the full target population and n_r is the size of the CPS sample. In practice, the size of the synthetic population only needs to be many times larger than the reference sample. In this case, we create a synthetic population that is 100 times larger than the original CPS dataset. These resampled units are then combined with the original n_r units to create a synthetic population of size N . Effectively, this procedure is imputing the $N - n_r$ unobserved units in the population based on a posterior predictive distribution generated from the weighted reference sample.

For the purposes of this study we proceed as if this synthetic population reflects the true population distribution for both the outcome variables and the demographic covariates. It is important to note that the synthetic population is itself derived from a survey and as such suffers from sampling, nonresponse, and other survey errors. As a result, the measures of bias discussed in subsequent sections of this chapter are most appropriately understood as approximations.

For the remainder of this analysis, we will refer to the synthetic reference population size as N with reference units indexed with $j = 1 \dots N$, and the survey sample size as n with survey respondents indexed as $i = 1 \dots n$. Let \bar{Y} be the synthetic population mean for outcome variable Y which we assume to be the true population value, and \bar{y} denote a sample estimate.

4.1.1 Quasi-randomization inference with propensity weighting

Quasi-randomization inference assumes that each unit i in the target population has an unknown, nonzero probability of inclusion in the sample denoted π_i . If π_i were known, then weighting each case in the sample by its inverse would correct any selection bias. Because π_i is unknown, we rely on an estimate based on a statistical model denoted $\hat{\pi}_i$ (Elliott and Valliant, 2017).

The most common concern with propensity weighting is the possibility of a few cases having extremely large weights which can result in highly unstable estimates. This can occur when there is a high degree of covariate imbalance between the reference and survey samples, even if strong ignorability holds. When the positivity assumption is violated the result is not only extreme weights and large variance but also bias as the weights will not be capable of producing covariate balance (Cole and Hernán, 2008).

Variable selection for propensity weighting is of particular importance. If variables in the model are strongly correlated with inclusion but not the outcome, the best case result will be an increase in variance without any bias reduction. However if there are omitted confounders associated with both inclusion and the outcome, weighting on variables that are only predictive of inclusion can magnify confounding bias considerably (Kreuter and Olson, 2011; Myers et al., 2011; Pearl, 2010). Therefore, simply selecting variables that are strongly predictive of inclusion without consideration of their association with the outcome can backfire. If there are no omitted confounders and the weighting variables are strongly correlated with both inclusion and the outcome variable, the result can be a decrease in both bias and variance (Little and Vartivarian, 2005).

Here, we estimate response propensities by combining the observations from the nonprobability sample with those in the synthetic population and using BART to estimate the function $\hat{\pi}_i = g(x_i)$, where x_i is the vector of demographic covariates measured on each unit in both the reference and opt-in samples.

Rather than use all of the observations in the synthetic population, we take a subsample with the same number of observations as the opt-in sample. The subsampling serves two purposes. First, because the synthetic populations of size N may be quite large, subsampling reduces the computational burden considerably. More importantly, having an equal number of population and sampled units greatly improves the performance of many machine learning methods used to estimate the propensity scores (Wallace et al., 2011).

Because the propensities are estimated relative to an equal sized subsample from the synthetic population, the weights are calculated based on the odds of

the propensity score rather than its inverse:

$$w_i = \frac{1 - \hat{\pi}_i}{\hat{\pi}_i}. \quad (4.1)$$

Weighting by the odds in this way treats the covariate distribution in the reference sample as the “correct” distribution and attempts to mirror that distribution in the nonprobability sample (Hirano et al., 2003; Schonlau et al., 2007). The more familiar approach of weighting by the inverse attempts to reproduce the covariate distribution for the union of the nonprobability and reference samples which is not generally the desired outcome.

For making inferences about the posterior distribution of propensity weighted estimates, we adopt a procedure for Bayesian propensity score estimation similar to that of Kaplan and Chen (2012). They propose creating M sets of propensity weights based on the posterior predictive distribution of the propensity model and then aggregating the M weighted point and variance estimates to capture the variance associated with both the propensity model and the estimated quantity. To account for uncertainty attributable to the fact that the propensities are estimated, we take $m = 1 \dots M$ draws of $\hat{\pi}_{im}$ from the posterior distribution for $\hat{\pi}_i$ returned by BART and create M sets of propensity weights. To account for the sampling variance that would be implied by the differential inclusion probabilities, we apply a weighted finite population Bayesian bootstrap to create $b = 1 \dots B$ synthetic populations of size N^* for each set of propensity weights (Dong et al., 2014). Because the focus here is on estimating proportions, we convert each of these synthetic populations into a set of frequency weights where each unit’s weight w_{imb}^* is equal to the number of times that unit was selected for synthetic population mb .

For each propensity weighted synthetic population

$$\bar{y}_{mb}^{(pw)} = \frac{\sum_{i=1}^n w_{imb}^* y_i}{N^*} \quad (4.2)$$

and we treat the set of MB estimates of $\bar{y}_{mb}^{(pw)}$ as an approximation to the posterior distribution of $\bar{y}^{(pw)}$.

4.1.2 Superpopulation inference with outcome regression

Whereas quasi-randomization relies on a statistical model to predict sample inclusion, superpopulation inference relies on a statistical model for the outcome Y . The model is then used to predict values for the unobserved units in the population. The frequentist theory behind this approach is detailed by [Valliant et al. \(2000\)](#). The Bayesian version of this approach, which relies on a prior distribution rather than a hypothetical superpopulation, has been described in several papers by Rod Little (see [Little, 2004](#); [Little and Zheng, 2007](#); [Little, 2012, 2015](#)).

When ignorability requirements do not hold, error manifests differently for estimates based on outcome regression. When positivity is violated, predictions from a model based on the survey sample may not generalize to units in the population with no representation in the sample. Unlike propensity weighting, where a lack of common support will result in large weights or a failure to balance covariates, outcome regression will not give any signs that anything is wrong. On the other hand, if the model does correctly generalize to the non-covered portion of the population, then this approach can produce efficient, unbiased estimates when propensity weighting would at best be highly variable and at worst both biased and variable. One attractive

feature of BART is the fact posterior predictive intervals are automatically wider for units with poor representation in the survey sample, dynamically inflating measures of uncertainty if there are a large number of such units in the population. Hill and Su (2013) proposed taking advantage of this feature to detect violations of common support in causal inference questions.

Adopting a similar approach in a survey setting could be a valuable piece of future research.

If there are confounding variables associated with both the outcome and selection that are not accounted for in the model, the associations between the model covariates and the outcome may differ from the associations that would be observed for the whole population. For example, if the young people in a sample are more liberal than young people in the overall population, but a measure of ideology is omitted from the model, the predicted values for young people will reflect this liberal bias. This in turn will carry through to the estimated quantity. As a result, simply selecting variables that are highly predictive of the outcome in the sample can lead to bias if the strength of the association is an artifact of selection. It is possible that when used for outcome regression with nonprobability samples, machine learning methods such as BART will detect correlations or interactions that are present in the sample due to confounding. These kinds of models may be maximally tuned to pick up omitted variable biases in ways that simpler linear models are not.

Here we use BART to estimate $\hat{y}_i = f(x_i)$ using the units in the nonprobability sample. Next, we use this model to generate M posterior draws of the predicted \hat{y}_{im} for each unit in the reference population. For each set of predicted values m

$$\bar{y}_m^{(or)} = \frac{\sum_{j=1}^N \hat{y}_{jm}}{N}, \quad (4.3)$$

and these M estimates reflect the posterior distribution for $\bar{y}^{(or)}$. Note that in this case, the nonprobability sample is only used to estimate the conditional distribution $\Pr(Y | X)$, and we rely exclusively on the synthetic population for the marginal distribution $\Pr(X)$.

4.1.3 Doubly-robust inference

Outcome regression with residual bias correction

Perhaps the most basic doubly-robust estimator is the AIPW of [Robins et al. \(1994\)](#). Because the propensity weights used in this study are based on the odds rather than the inverse of the propensity, we refer to this estimator as OR-RBC using the more general terminology proposed by [Kang and Schafer \(2007\)](#). This type of estimator is closely related to model-assisted estimators for probability-based surveys ([Särndal et al., 1992](#)) which have been readily adapted to incorporate machine learning approaches to prediction ([Breidt and Opsomer, 2017](#)).

The OR-RBC estimator is simply the basic OR estimator described above plus the mean of propensity weighted residuals from the nonprobability sample.

Because we have M draws of $\bar{y}_m^{(or)}$ but MB sets of weights w_{mb}^* , we approximate the posterior distribution of $\bar{y}^{(rbc)}$ by calculating

$$\bar{y}_{mb}^{(rbc)} = \bar{y}_m^{(or)} + \frac{\sum_{i=1}^n w_{imb}^* (y_i - \hat{y}_{im})}{N^*}. \quad (4.4)$$

Thus for each of the M instances of $\bar{y}_m^{(or)}$ there we calculate the second term B

times for a total of MB posterior draws.

Because the propensity weights are based on the odds of $\hat{\pi}$ rather than its inverse, this is a bias-corrected estimate for the synthetic population mean. In a survey nonresponse setting, this estimator is described by [Kang and Schafer \(2007, p. 532\)](#) as a bias-corrected estimate of the nonrespondent mean. It is doubly-robust in the sense that if the outcome model is correct, then the second term will equal 0 in expectation. If the outcome model is incorrect but the propensity model is correct, then the second term is equivalent in expectation to $\bar{Y} - \bar{y}^{(or)}$, thus negating any bias in $\bar{y}^{(or)}$.

Outcome regression with a propensity score covariate

This approach is an extension of the PSPP model in which a penalized spline of the propensity score is included in an outcome regression model along with the model covariates ([Little and An, 2004](#); [Zhang and Little, 2009](#)). A variant of this approach, which used piecewise constant coefficients for a binned propensity score in place of a spline, was found by [Kang and Schafer \(2007\)](#) to be more robust under dual-misspecification than the other doubly-robust estimators in their study.

This version, also described as BARTps by [Tan et al. \(2018\)](#) involves first fitting the propensity model with BART as before and then including the posterior mean propensity score as a covariate in the outcome regression model such that $\tilde{y}_i = f(x_i, \hat{\pi}_i)$. The estimate for the population mean is then the same as for the basic OR estimate but substituting \tilde{y} for \hat{y} for each unit in the synthetic population

$$\bar{y}_m^{(psc)} = \frac{\sum_{j=1}^N \tilde{y}_{jm}}{N} \quad (4.5)$$

and make posterior inferences over the M values of $\bar{y}_m^{(psc)}$. [Tan et al. \(2018\)](#) found that this estimator performed best when both the mean and propensity functions were particularly complex, although a less complex PSPP approach that only used BART to estimate the propensity score

4.2 Results

For each of the 10 online nonprobability samples, we estimate the population percentage for six measures of civic engagement using each of these four estimators. We set the number of posterior draws $M = 1000$, and to keep computation manageable, we set $N^* = 20 \times n$ and set $B = 25$. For purposes of comparison, we also include an unweighted estimate of each population percentage, and estimate its variance using a standard Bayesian bootstrap ([Rubin, 1981](#)). Thus, for the unweighted, OR, and OR-PSC estimates we have a total of $M = 1000$ posterior draws for each estimate of \bar{y} , while for PW and OR-RBC we have $M \times B = 1000 \times 25 = 25000$. We compare each estimator with respect to absolute bias, posterior variance, and root mean squared error (RMSE).

The code used to fit these models and generate the posterior draws for each estimate can be found in [Appendix C](#).

[Table 4.1](#) displays the measures of performance averaged over all samples and outcome variables. Because all of the estimates are percentages, and the measures are on common scales, we simply average them without additional standardization. While the unweighted estimates have the lowest average

Table 4.1: Average estimator performance on bias, variance, and RMSE.

Estimator	Avg. Absolute Bias	Avg. Posterior Var.	Avg. RMSE
Unweighted	7.8	1.9	8.0
PW	6.6	2.7	7.0
OR	7.3	2.9	7.6
OR-RBC	6.4	2.5	6.7
OR-PSC	7.4	3.5	7.8

Note:

Estimates are averaged over all 10 samples and six outcome variables.

variance, all of the modeled estimates are preferable in terms of both bias and RMSE. With respect to bias and RMSE, all of the methods are preferable to unweighted estimates. In contrast to [Tan et al. \(2018\)](#) we see that the OR-RBC estimator (similar to their AIPW with BART estimator) has the lowest bias, variance, and RMSE on average, while OR-PSC (analogous to their BARTps) has the highest. PW performs nearly as well as OR-RBC on all three measures. Likewise, OR-PSC has slightly higher variance than OR, but does not offer any added benefit with respect to bias.

When broken out by sample and outcome variable, a more complex picture emerges. [Figure 4.1](#) shows the absolute bias for each estimate broken out by sample and outcome variable. For none of the samples is it the case that a particular estimator is always preferable. The closest is sample E where the lowest bias always belongs to either OR-RBC or PW, both of which reduce bias relative to no adjustment. For sample I, which has the lowest average unweighted bias to begin with, nearly all of the options increase bias relative to doing nothing. More typically, the option with the least bias varies by outcome. Of all 60 items, the unweighted estimate has the lowest bias for 16.

[Figure 4.2](#) presents the same information somewhat differently. It shows the change in absolute bias relative to no weighting. Several patterns become



Figure 4.1: Absolute bias by sample and outcome variable. Estimates are presented on a percentage point scale. Samples are ordered by unweighted average absolute bias across all six outcome variables. Outcome variables are ordered by unweighted average absolute bias across all 10 samples.

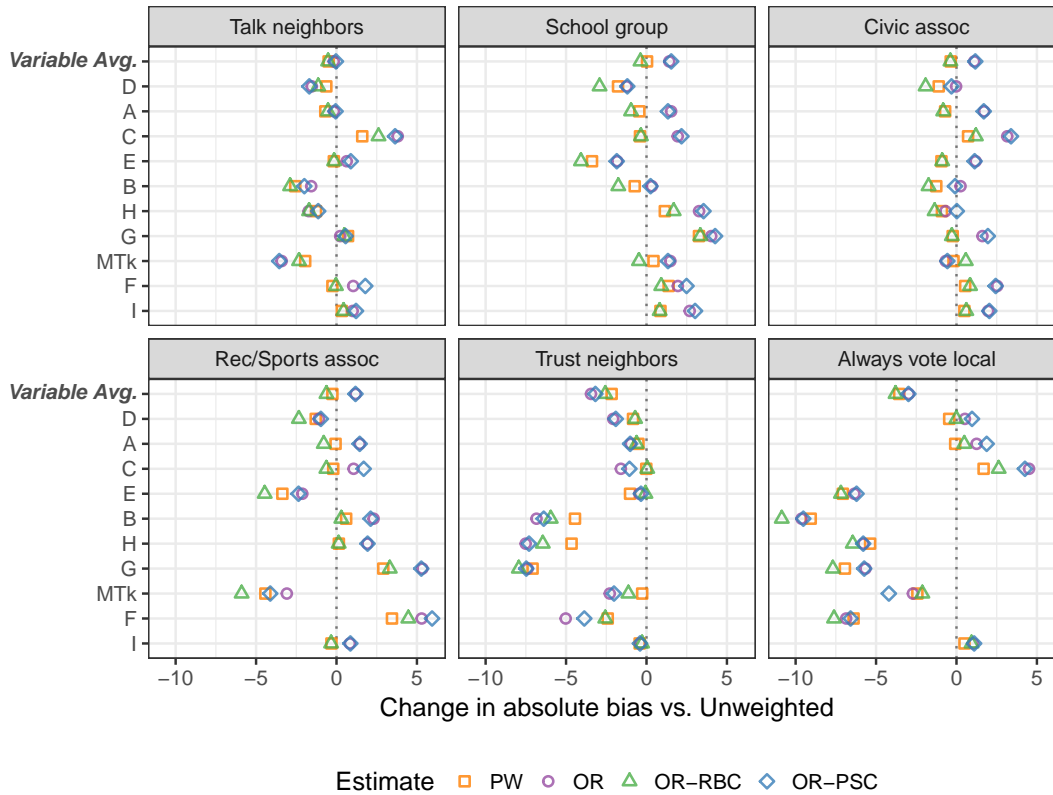


Figure 4.2: Change in absolute bias relative to unweighted. Estimates are presented on a percentage point scale. Samples are ordered by unweighted average absolute bias across all six outcome variables. Outcome variables are ordered by unweighted average absolute bias across all 10 samples.

clear. Overall, PW and OR-RBC tend to produce very similar estimates, and with few exceptions, one or the other is most often the estimate with the lowest bias. When there is bias reduction, OR-RBC almost always performs somewhat better than PW. When there is bias amplification, PW tends to perform somewhat better. The same does not appear to hold for OR and OR-PSC. The differences between the two tend to be smaller and their relative performance is not clearly related to the presence of bias reduction or amplification.

The exceptions to this pattern are also notable. For trusting neighbors, OR and OR-PSC both consistently outperform PW and OR-RBC across samples with respect to bias, even if only slightly in some samples. Additionally, OR

and OR-PSC do well on the Mechanical Turk sample when others do not, particularly on talking to neighbors and voting in local elections. The Mechanical Turk sample is not a traditional survey sample and did not employ any sort of quotas or other demographic controls during data collection. Additionally, if we follow the same procedure as [Chapter 3](#) and define the covered region of common support as those units in the synthetic population with a propensity score higher than the minimum score in the survey sample, the coverage rate for Mechanical Turk is only 88% ([Dehejia and Wahba, 1999](#)). The other samples all have estimated coverage rates between 96% and 100%. It is likely that this lower common support contributes to the worse performance for PW and OR-RBC in the Mechanical Turk sample.

[Figure 4.3](#) shows the design effect ($deff$) for each estimate relative to the unweighted estimate. The design effect is equal to the posterior variance for the estimate divided by the unweighted posterior variance. While the rank ordering of the estimators with respect to design effect is mostly consistent across samples with OR-RBC the lowest followed by PW, OR and OR-PSC, the magnitude of the differences between the estimators clearly depends on the sample. In particular, the $deff$ s for OR and OR-PSC varies to a much greater degree than OR-RBC and PW. For example, the average $deff$ over the six variables for OR-RBC is under 1.5 for all but two samples and only rises as high as 2.1 for Mechanical Turk. For OR-PSC, only three samples are under 1.5, four are over 2, with Mechanical Turk at 3.7. The patterns for PW and OR are similar but less extreme, with PW closer to OR-RBC and OR closer to OR-PSC. Once again, Mechanical Turk is notably different from the other samples, having the highest average $deff$ for all four estimators.

In terms of overall error, [figure 4.4](#) makes clear that RMSE is almost entirely a

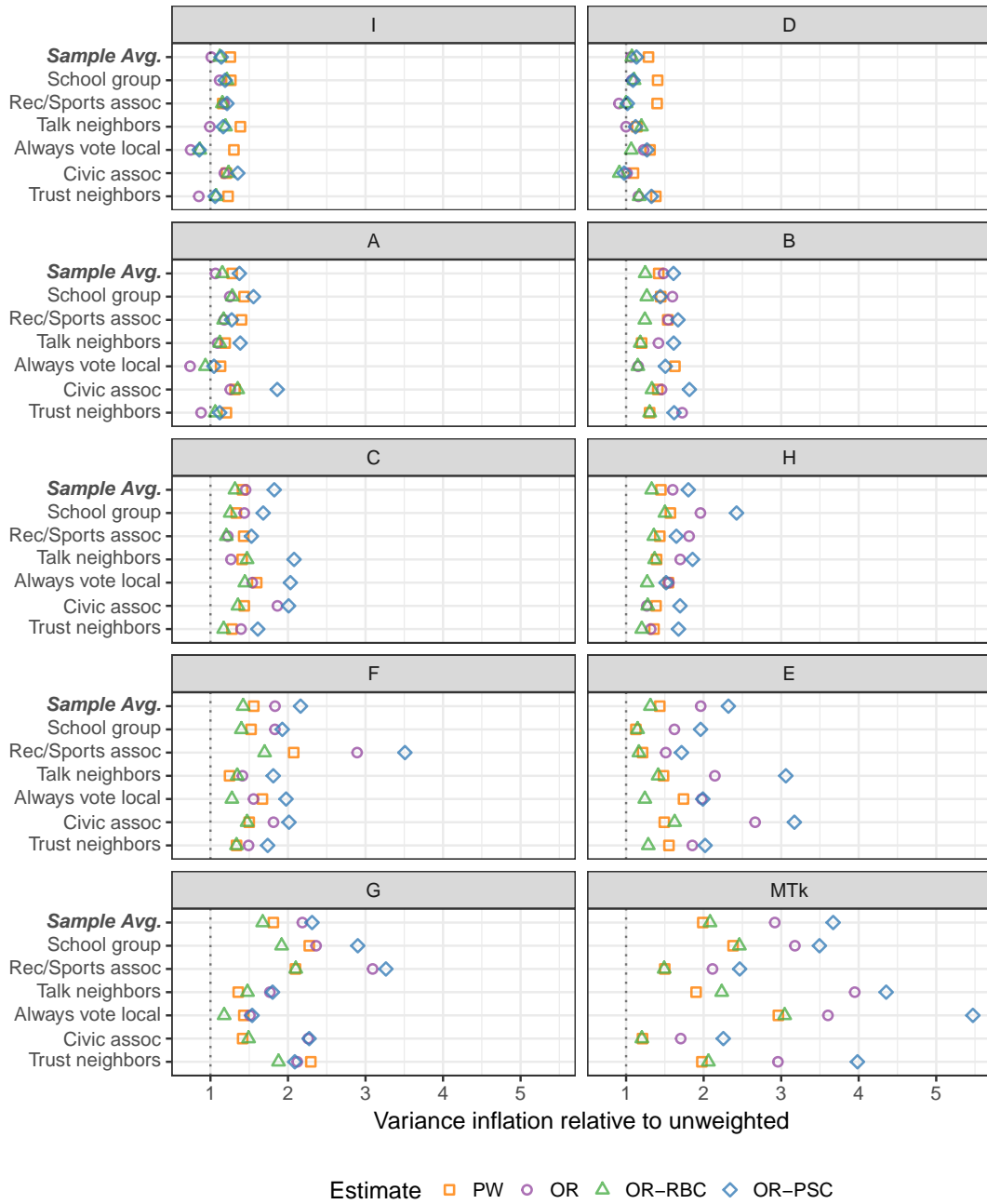


Figure 4.3: Design effect of four estimators relative to unweighted. The design effect is the ratio of the estimate’s posterior variance to the unweighted posterior variance. Samples are ordered by the average design effect across all six outcome variables. Outcome variables are ordered by the average design effect across all 10 samples.

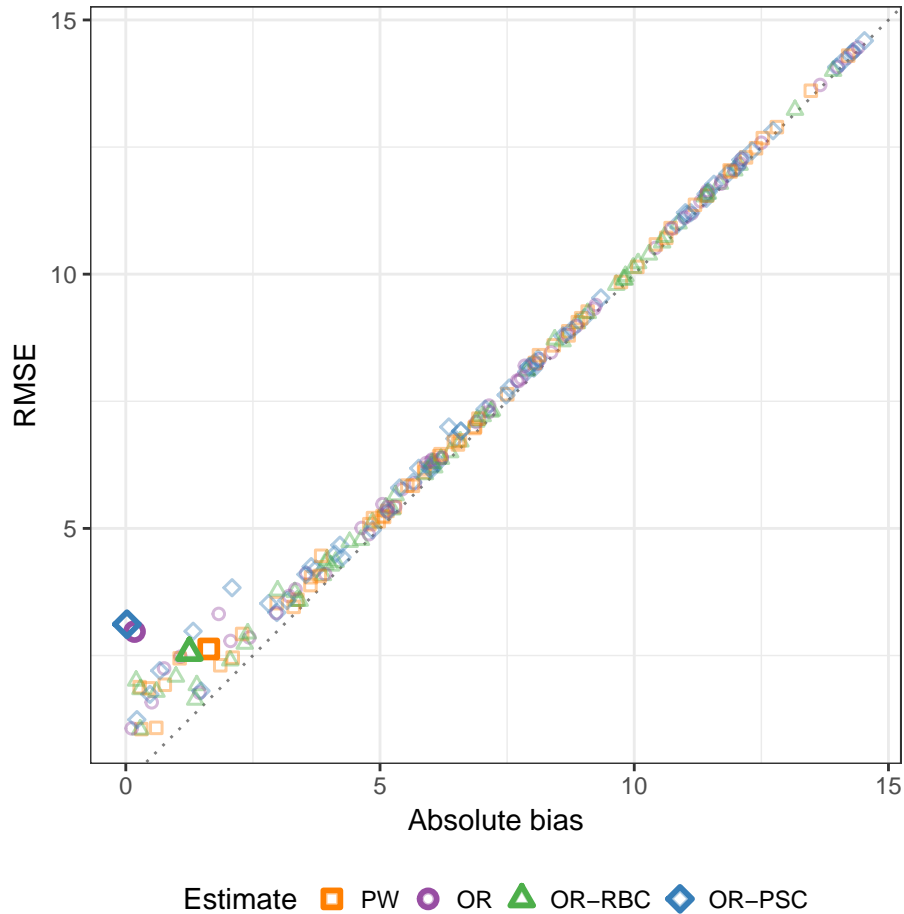


Figure 4.4: RMSE vs. absolute bias for all variables, estimators, and samples. Estimates are presented on a percentage point scale. The enlarged and highlighted points are the Mechanical Turk estimates for frequency of talking to neighbors. They illustrate an instance where bias was largely eliminated with OR and OR-PSC but RMSE remained high.

function of bias. Because bias is generally so large for these items, error from estimates that fall higher or lower than the posterior mean largely cancel out. There are a few exceptions such as frequency of talking to neighbors in the Mechanical Turk sample where OR and OR-PSC successfully eliminated nearly all of the bias but the relatively high variance no longer cancels. The resulting RMSE is actually slightly higher than the more biased estimates based on PW and OR-RBC. These sorts of exceptions only occur in instances where the bias was relatively low to begin with.

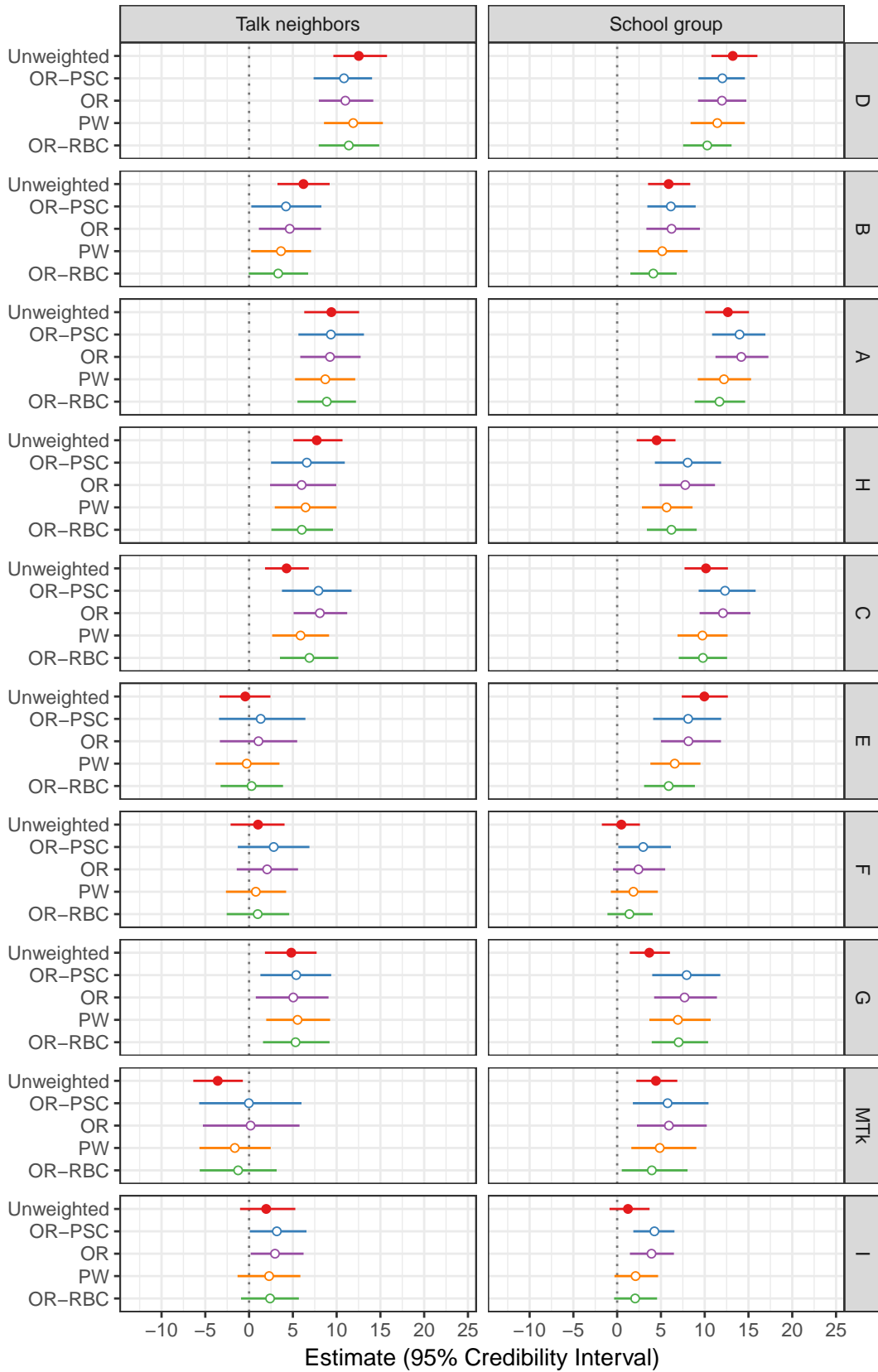


Figure 4.5: Position of 95% credibility intervals relative to population value: Participated in a school group, Talk with neighbors weekly.

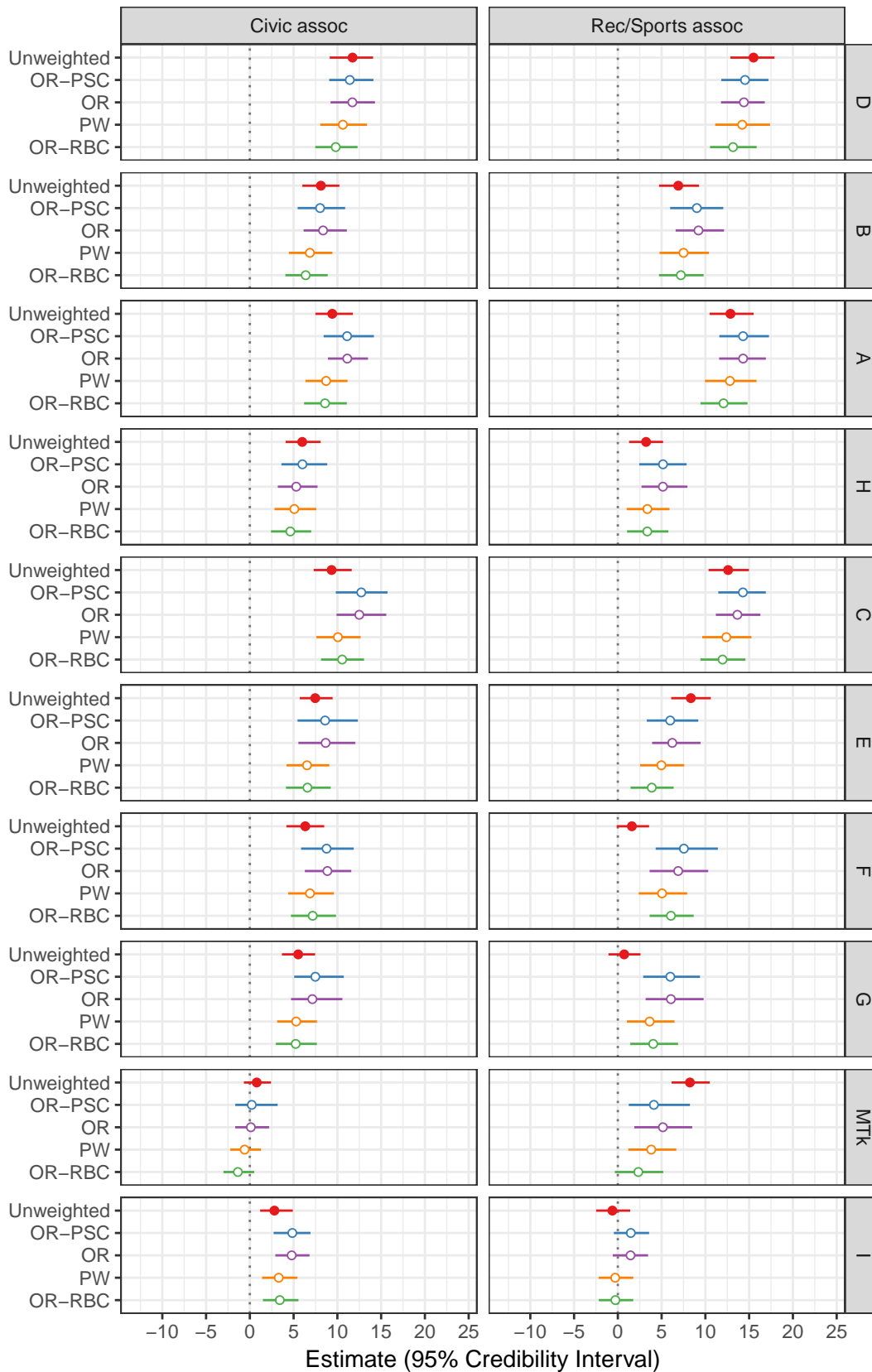


Figure 4.6: Position of 95% credibility intervals relative to population value: Participated in a civic association, Participated in a sports/recreational association.

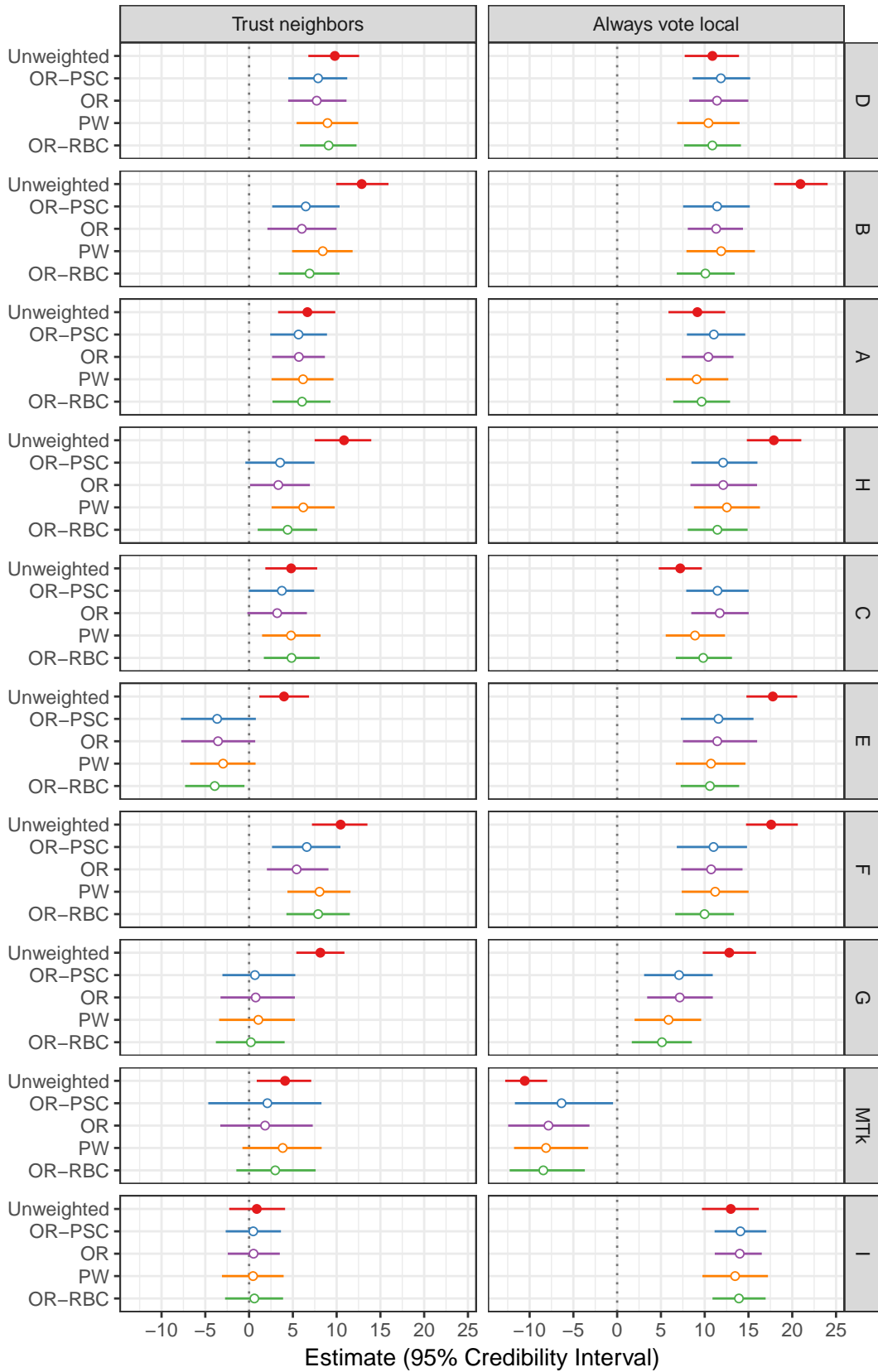


Figure 4.7: Position of 95% credibility intervals relative to population value: Always votes in local elections, Trusts all/most people in neighborhood

While lower variance is usually considered a desirable property of an estimator, a larger posterior variance could prove beneficial if it results in higher coverage of the true population value within the credibility or confidence interval. However, for these estimates, the frequently wider intervals for OR and OR-PSC do not appear to offer such an advantage when compared to OR-RBC and PW. Figures 4.5, 4.6, and 4.7 show the position of 95% credibility intervals relative to the benchmark value. Again, the differences between the different estimators are not dramatic, but there is a consistent pattern across samples and outcome variables where the intervals for the OR-RBC and PW estimates tend to be closer to benchmark value, even when the OR or OR-PSC intervals are wider. This is perhaps most clearly visible for participation in a recreational or sports association depicted in Figure 4.6. Here we again see close similarity between the OR-RBC and PW intervals on one hand and the OR and OR-PSC intervals on the other. Even in instances where the OR and OR-PSC intervals are wider (as they are for samples E, F, G, and Mechanical Turk), the OR-RBC and PW intervals are still closer to the benchmark value. Trust in neighbors is again the exception where the intervals for OR-PSC and OR tend to be closer to the benchmark.

4.3 Discussion

In this chapter we have compared the performance of four approaches to estimation in nonprobability surveys using BART: singly-robust PW and OR and doubly-robust OR-RBC and OR-PSC. As expected, given the high degree of nonignorable selection bias, none of the methods was entirely successful at eliminating bias. Overall, OR-RBC tended to perform best with respect to bias, variance, and RMSE, although PW performed nearly as well. Given that OR-RBC requires an outcome model for each variable while a single set of

propensity weights can be used for for multiple variables, it would be reasonable to weigh the added analytical complexity of the doubly-robust estimator against the modest improvement it yielded over PW alone. OR and OR-PSC also performed similarly and tended to exhibit higher bias.

In particular, it would seem that OR and OR-PSC had a greater tendency to inflate bias than PW and OR-RBC. Both OR and OR-PSC resulted in higher bias for 31 out of 60 estimates across the samples. This is in contrast to 17 and 19 for PW and OR-RBC respectively. Additionally, while OR and OR-PSC produced point estimates that were nearly identical, OR-PSC had consistently higher variance and RMSE.

The differences between the relative performance of the two singly-robust estimators is surprising given PW's relatively poor showing in other studies (e.g. [Dutwin and Buskirk, 2017](#); [Mercer et al., 2018](#); [Valliant and Dever, 2011](#)). While the differences were not usually large, often less than a percentage point, they were consistent across samples and outcome variables (with the important exceptions of Mechanical Turk and trust in neighbors). One possible explanation is that BART (and likely other machine learning algorithms) fit models that are very well tuned for the sample but reflect a spurious conditional distribution for X when exchangeability does not hold. The better performance of calibration methods relative to propensity weighting in other studies may be due to the fact that their comparatively simple functional forms serve to prevent this kind of overfitting.

The differences between the two doubly-robust estimators was similarly notable. The relatively high variance of both OR and OR-PSC suggests that the demographic covariates used in this analysis are not highly predictive of the outcomes. Given that the propensity score is in a sense a univariate

summary of the covariate distribution, it is perhaps not surprising that its inclusion as a covariate added little information to the basic OR estimator given BART's already powerful ability to approximate $\Pr(Y | X)$. In contrast, the OR-RBC involves greater separation between the propensity scores and the outcome regression model. In the presence of confounding, the propensity weights and outcome regression may be more successful at offsetting each other's weaknesses.

Additional research comparing singly and doubly-robust estimators using different combinations of more and less complex outcome and propensity models would be greatly beneficial. In particular, additional evaluation of model performance in the presence of exchangeability violations seems particularly important, and should be done using both simulated and real survey data. In this study, we saw that the usual patterns of estimator performance were largely reversed, likely due to the lack of exchangeability for the civic engagement measures that we identified in Chapter 3. These patterns may not hold for other variables with different confounders, but some degree of confounding is likely to be the norm for nonprobability survey samples. A fuller understanding of its impact on different estimation approaches may go a long way toward the improving the quality of survey estimates under less than ideal conditions.

Chapter 5: Conclusion

In this dissertation, we proposed an alternative framework for describing how survey estimates from nonprobability samples can be affected by selection bias. In the absence of randomization, any statistical inferences are based on an implicit or explicit model that explains the relationship between a sample and the target population. When models are implicit, it is more likely that they have not been subject to scrutiny and are at a greater likelihood of being inappropriate. To date, the tools available to survey researchers have not offered a simple and coherent way to think about and analyze the assumptions that they make, consciously or not, when they make inferences from data with nonexistent or imperfect randomization. The framework proposed here, based on principles from causal inference, offers a simple checklist of the three conditions that must be true for such inferences to be valid: exchangeability, positivity, and correct composition.

In Chapter 2 we examined the theoretical similarities between causal inference and survey inference and showed how principles from the former can be applied fruitfully to the latter. We demonstrated the conceptual utility of the causal framework for thinking about selection bias in surveys and showed how it can be used to reason about why some methods and practices seem to work better than others. The goal of this Chapter was to describe these concepts in an accessible and nontechnical manner that can be readily understood by practitioners as well as methodologists and statisticians.

Chapter 3 went a step further and provided the mathematical details for how exchangeability, positivity, and composition bias affect survey estimates. The

total bias can be decomposed into separate additive components associated with each error source. This permits researchers to target specific kinds of error and develop research methods that are focused on eliminating specific kinds of modeling error in much the same way that the TSE framework helps designers of probability-based surveys eliminate specific threats to randomization inference introduced at each stage of the data collection process. Moreover, these components can be estimated given the appropriate reference data. Although such data may be unavailable for many survey outcomes, this is also true for those trying to study coverage or nonresponse error in probability-based surveys.

When such data are available, there are clear practical benefits. We were able to see with the civic engagement items that there were clear patterns with respect to the average level of exchangeability bias between samples despite high within-sample variability. It is easy to see how this analysis could be extended to explain not only differences between a nonprobability sample and the target population but also differences between samples from different sources. To date, studies comparing data quality from different samples have had little success in explaining why data quality is so variable ([Gittelman et al., 2015](#); [Kennedy et al., 2016](#); [Yeager et al., 2011](#)). In such instances, researchers would not be limited to only those variables that are available on a reference sample but could use many different questions from parallel surveys to diagnose and explain differences between different nonprobability sample sources. This sort of approach could also be used to identify instances where it might be fruitful to combine data from different sources and when to avoid doing so.

In Chapter 4 we compared the performance of four approaches to estimation

under conditions of nonignorable selection. In Chapter 2 the civic engagement items were shown to suffer from high levels of confounding bias in general. In such instances, any model that assumes ignorability will be automatically misspecified. We saw that for this application, the doubly-robust outcome regression with residual bias correction (OR-RBC) generally performed best for bias, variance, and RMSE while outcome regression with a propensity score covariate (OR-PSC) performed worst on all three. Likewise, there were clear similarities between OR-RBC and propensity weighting (PW) and between OR-PSC and outcome regression (OR). That these findings differ from those of Tan et al. (2018) suggest that studying the performance of various estimators when exchangeability and positivity assumptions are violated may be as important as studying their performance under ignorability. A fruitful avenue for future research in this vein would be to extend the analysis from Chapter 3 and derive the bias decompositions for these different kinds of estimators in order to better understand the conditions under which one approach or another should be preferred.

5.1 *Next steps*

There are many directions in which research could proceed from here. Some of the most immediate would include the extensions described above as well as replications of these analyses using different sample sources, reference samples, and outcome variables. Re-analyses of earlier comparative studies of nonprobability samples could help uncover explanations for the variation in data quality that has been observed across sample sources. It would also be worthwhile to experiment with different types of machine learning procedures and try to find the most effective way to estimate these bias components.

We can also see how this framework could be applied in the development of a

variety of diagnostic procedures. For instance, when reference data is available, it is possible to estimate $\hat{\delta}^{\text{exch}}$ for individual cases using BART or other machine learning methods. These values could then be analyzed using procedures such as classification and regression trees. These estimated values could also be treated as outcome variables. This would make it easy to see how $\hat{\delta}^{\text{exch}}$ is affected by different estimation procedures without necessarily having to derive a new formula. One can imagine that an analytic formulation for a method such as OR-PSC with BART, where estimates from one complex model are used as inputs to a second, could be difficult or impossible to derive, but applying OR-PSC to estimated values of $\hat{\delta}^{\text{exch}}$ would make evaluation straightforward.

In many instances, a reference sample with the covariate distribution will be available but not the outcome variable of interest. This prevents the estimation of these bias components, but approaches to sensitivity analysis such as the version proposed by [Robins et al. \(1999\)](#) – in which a hypothetical confounding effect is added to the value of an outcome variable for each unit in a sample – could prove powerful in conjunction with a propensity model to measure the robustness of results to different levels of unobserved confounding. More broadly, the 2013 AAPOR task force report called for both a framework and standard metrics that can be used to evaluate the quality of estimates from nonprobability samples ([Baker et al., 2013](#)). This dissertation has proposed a framework. Going forward it will be important to take the next step and create metrics that can be used as measures of data quality. While response and completion rates are not especially meaningful for many nonprobability surveys, it would certainly be possible to develop measures summarizing what share of the population is missing relative to a reference dataset. Another

possibility would be to develop a family of indicators measuring balance relative to a reference sample – much like the R indicators that have been developed for probability-based samples where auxiliary data is available for nonrespondents (Schouten et al., 2009, 2011, 2012).

One problem that has plagued public-opinion polling in particular is the inadequacy of the “margin of sampling error” as a measure of data quality. It may be that this framework could be employed, perhaps in conjunction with methods such as those developed by Manski (2007) to calculate error bounds that incorporate both bias and variance based on estimated common support, compositional differences, and some reasonable assumptions about potential confounding.

The fact that all these diagnostics are themselves model-based and involve their own sets of assumptions should not be considered problematic. As we have stated repeatedly, there is no escaping assumptions; evaluating our assumptions requires making further assumptions about assumptions. What is important is that the assumptions are reasonable, useful, and above all transparent.

5.2 Revisiting Total Survey Error

We began this dissertation by outlining the ways in which the Total Survey Error (TSE) framework falls short as an approach for researching error in nonprobability survey samples and proposed the causal framework as a more appropriate alternative. While TSE attempts to isolate sources of error that results from defects in the sampling and data collection processes, the causal framework proposed here is focused on defects in the statistical model that is used to relate sample to population. From an inferential perspective, it makes

sense for surveys that aspire to base inferences on randomization to prioritize the TSE framework. Likewise, for surveys where randomization plays no meaningful inferential role, it makes sense to focus on modeling assumptions and interrogating the manner in which those assumptions could be incorrect. In practice though, probability-based surveys suffer from undercoverage and nonresponse, forcing researchers to rely on statistical models and assumptions. Likewise, users of nonprobability surveys still need to worry about data collection.

There are a variety of ways in which the two approaches can be complementary. The bias components and estimation procedures described in this dissertation are equally applicable to probability-based surveys. Given high rates of nonresponse, probability-based surveys have become more and more reliant upon models and statistical adjustment to correct for problems with coverage and nonresponse. Additional tools for evaluating these models can only help. Even though coverage and nonresponse do not hold the same sort of inferential significance for nonprobability samples as they do for probability-based samples, the use of recruitment strategies that appeal to a more diverse set of potential panelists or survey designs that are more likely to result in respondent participation are important both for efficiency and to ensure that survey designs produce samples that are consistent with the models used for estimation and do not introduce additional confounds. These are the types of problems where TSE's focus on the data collection process can be helpful no matter what inferential framework a survey uses.

In conclusion, it is our hope that this dissertation has raised more questions than it has answered. When it comes to studying error in nonprobability survey samples much of the difficulty has been figuring out the right questions

to ask. The error framework and analyses included in this dissertation are only a first step. With luck, they will be further refined and built upon both in our own future work and in that of other statisticians, methodologists, and practitioners who can use the framework to improve the quality of their own research.

Appendix A: Question wording

Below is the question wording for the civic engagement and demographic items as it appeared in the 10 nonprobability surveys analyzed in this dissertation. For the text of the full questionnaire, see [Appendix F](#) of the original report by [Kennedy et al. \(2016\)](#). The variable label for the corresponding item from the 2013 Current Population Survey Civic Engagement Supplement (CPS) is included in brackets after the question number. Question wording for the civic engagement items matches the wording that was used in the CPS.

Q0002 [PES15] During a typical month in the past year, how often did you talk with any of your neighbors?

1. Basically every day
2. A few times a week
3. A few times a month
4. Once a month
5. Not at all

Q0004 [PES18] How much do you trust the people in your neighborhood? In general, do you trust...

1. All of the people in your neighborhood
2. Most of the people in your neighborhood
3. Some of the people in your neighborhood
4. None of the people in your neighborhood

Below is a list of types of groups or organizations in which people sometimes participate. Have you participated in any of these groups during the last 12 months, that is since February 2014?

- Q0009 [PES5a] A school group, neighborhood, or community association such as PTA or neighborhood watch group?
 1. Yes
 2. No
- Q0010 [PES5b] A service or civic organization such as American Legion or Lions Club?

1. Yes
2. No

- Q0011 [PES5c] A sports or recreation organization such as a soccer club or tennis club?

1. Yes
2. No

Q0029 [PES1] The next question is about LOCAL elections, such as for mayor or a school board. Do you. . .

1. Always vote in local elections
2. Sometimes vote in local elections
3. Rarely vote in local elections
4. Never vote in local elections

And finally, a few questions about yourself and your household.

Q0042 [PESEX] What is your gender?

1. Female
2. Male

Q0043 [PRTAGE] What is your age?

[PROGRAMMING NOTE: Numeric text box, 5 characters wide, range 18-120] _____years

Q0044 [PRDTHSP] Are you of Hispanic, Latino, or Spanish origin, such as Mexican, Puerto Rican or Cuban?

1. Yes, Hispanic or Latino
2. No, not Hispanic or Latino

Q0045 [PTDTRACE] Which of the following describes your race?

[You can select as many as apply]

1. White
2. Black of African-American
3. Asian
4. American Indian or Alaska Native
5. Native Hawaiian or other Pacific Islanders
6. Some other race, specify: _____

Q0050 [PEEDUCA] What is the highest grade or year of school you completed?

1. Never attended school or only attended kindergarten
2. Grades 1 through 8 (Elementary School)

3. Grades 9 through 11 (Some High School)
4. Grade 12 or GED (High School Graduate)
5. Completed some college
6. Completed technical school
7. Associate degree
8. Bachelor's degree
9. Completed some postgraduate
10. Master's degree
11. Ph.D., law, or medical degree
12. Other advanced degree beyond a Master's degree

Q0055 [GEREG] What is your zip code?

[PROGRAMMING NOTE: Numeric text box, 5 characters wide, range 0-99999]

Appendix B: Variable coding

The data from the nonprobability and CPS samples was recoded and processed as follows:

1. Each of the six measures of civic engagement was coded as a binary variable. The category or categories chosen as the outcome variable were coded as 1 and all other responses (including item nonresponse) were coded as 0. These reflect the original variable codings used in the report by [Kennedy et al. \(2016\)](#). These were:
 - Always votes in local elections.
 - Trusts all or most people in their neighborhood.
 - Typically talk to their neighbors every day or a few times a week.
 - In the last twelve months, participated in a school, neighborhood or community group.
 - In the last twelve months, participated in a civic or service organization.
 - In the last twelve months, participated in a sports or recreation organization.
2. The demographic variables in both the nonprobability and CPS reference samples were recoded into the categories listed below.
3. For both the nonprobability and CPS datasets, item nonresponse to the demographic items was imputed using the `mice` package and a version of the random forest imputation algorithm described by [Doove et al. \(2014\)](#) implemented with the `ranger` package ([van Buuren and Groothuis-Oudshoorn, 2011](#); [Wright and Ziegler, 2017](#)). While `mice` is a

procedure for performing multiple imputation, we use it more for its ability to jointly impute several variables at once. We use only a single imputed dataset to avoid overly complicating the analysis. With the exception of Census region, none of the demographic variables on any of the samples were missing for more than 2% of the interviews. For region, missingness ranged from 1% to 5% of interviews.

The demographic variables were coded as follows:

- **Sex**
 1. *Male*
 2. *Female*
- **Age**
 - *18 through 85*: Age was left continuous. Respondents who reported being more than 85 years old were topcoded to 85 in order to be consistent with the CPS coding.
- **Race/ethnicity**
 1. *Non-Hispanic White*: Respondents were coded as non-Hispanic white if they only selected white as their race and did not identify as Hispanic or Latino in the ethnicity question.
 2. *Non-Hispanic Black*: Respondents were coded as non-Hispanic black if they only selected black as their race and did not identify as Hispanic or Latino in the ethnicity question.
 3. *Hispanic*: Respondents were coded as Hispanic if they identified as Hispanic or Latino in the ethnicity question. This coding was unaffected by responses to the race question.
 4. *Other*: Respondents who did not identify as Hispanic or Latino in the ethnicity question and selected multiple races or a race other

than white or black.

- **Education**

1. *High school or less:* Respondents were coded as high school or less if they indicated that their highest level of education was Grade 12 or a GED or below.
2. *Some college:* Respondents were coded as some college if they reported completing some college, technical school, or an associate's degree.
3. *College graduate:* Respondents were coded as college graduates if they reported completing a bachelor's degree or higher.

- **Census region**

Census region was coded according to state based on respondent reported zip code. See the U.S. Census Bureau's [webpage](#) for details on the states included in each region.

1. *Northeast*
2. *Midwest*
3. *South*
4. *West*

Appendix C: Code

The following code estimates all of the conditional means that are used in Chapter 3 to estimate bias components. It also produces the estimates used in Chapter 4.

```
library(BART)
library(tidyverse)
library(bayesboot)
library(bestimate)
library(timefactory)
library(stringr)

## NOTE: timefactory and bestimate can be installed with:
## devtools::install_github("awmercer/timefactory")
## devtools::install_github("awmercer/bestimate")

# timefactory is for timing code
# bestimate contains functions to make working with BART easier

get_estimate_posteriors = function(samp_id,
                                   samp,
                                   ref,
                                   synth_pop_ids,
                                   x_vars,
                                   y_vars,
                                   draws,
                                   pweight_synth_pops,
                                   cores) {

  from_start = timefactory()

  ## Convert synthetic population indices into frequency weights for each
  ## record in the synthetic population
  sp_wts = tibble(ids = synth_pop_ids) %>% group_by(ids) %>%
    summarise(wt = n()) %>%
    pull(wt)
```

```

# List containing output
res = list()

## Convenience data structures
x_ref = ref[, x_vars]
x_samp = samp[, x_vars]
n_samp = nrow(samp)

# Get subsample of reference
ref_subsamp_ids = synth_pop_ids[sample(seq_along(synth_pop_ids),
                                       n_samp,
                                       replace = FALSE)]

ref_subsamp = ref[ref_subsamp_ids, ]
x_ref_subsamp = ref_subsamp[, x_vars]

## Estimate response propensities
origin = c(rep(1, n_samp), rep(0, n_samp))
comb = bind_rows(x_samp, x_ref_subsamp)

# Pre-filled in BART call with standard parameters
bart_partial = partial(
  pbart2,
  ndpost = draws,
  verbose = FALSE,
  keeptrainfits = FALSE,
  mc.cores = cores,
  nskip = 1000
)

cat("Fitting propensities ")
propensity_timer = timefactory()
propensity_fit = bart_partial(x.train = comb,
                              y.train = origin)

sample_propensities = pbart_posterior(propensity_fit,
                                       newdata = x_samp,
                                       mc.cores = cores)
cat(sprintf("%.1f\n", propensity_timer()))

# Fit OR models - confounded and unconfounded

y_fits_timer = timefactory()

```

```

cat("Fitting y models ")
y_fits_confounded = y_vars %>%
  map(~ bart_partial(x.train = x_samp, y.train = samp[[.x]]))

y_fits_unconfounded = y_vars %>%
  map(~ bart_partial(x.train = x_ref_subsamp, y.train = ref_subsamp[[.x]]))

cat(sprintf("%.1f\n", y_fits_timer()))

# Add posterior mean propensity score to x_samp for OR-PSC
x_samp_prop = x_samp %>%
  mutate(pi_hat = rowMeans(sample_propensities))

dr_fits_timer = timefactory()
cat("Fitting OR-PSC models ")
y_psc_fits = y_vars %>%
  map(~ bart_partial(x.train = x_samp_prop, y.train = samp[[.x]]))
cat(sprintf("%.1f\n", dr_fits_timer()))

cat("Saving BART fits ")
save_timer = timefactory()
# Save BART fits to file for reuse later
saveRDS(
  file = sprintf("data/bart_models/bart_fits_%s.RDS", samp_id),
  object = list(
    sample_id = samp_id,
    propensity_fit = propensity_fit,
    y_fits_confounded = y_fits_confounded,
    y_fits_unconfounded = y_fits_unconfounded,
    y_psc_fits = y_psc_fits,
    synth_pop_ids = synth_pop_ids
  )
)
cat(sprintf("%.1f\n", save_timer()))

## Estimate posteriors and other quantities

est_timer = timefactory()
cat("Starting estimates:\n")

# Calculate weights as odds of being in the population over sample
sample_weights = map(sample_propensities, ~ (1 - .x) / .x)

```

```

# For each propensity weight create a set of FPBB weights
cat("Creating fpbb propensity weights ")
fpbb_timer = timefactory()
sink("/dev/null")
sample_weight_synth_pops = map(
  sample_weights,
  ~ fpbb_synth_pops(
    weights = .x,
    L = pweight_synth_pops,
    N = length(.x) * 20
  )
)
sink()
cat(sprintf("%.1f\n", fpbb_timer()))

# Get "true" population means
res$y_bar_pop = y_vars %>%
  map_dfc(function(y_var) {
    weighted.mean(ref[[y_var]], sp_wts)
  })

# Estimate propensity weighted means
res$y_bar_propwt = y_vars %>% map_dfc(function(y_var) {
  map(sample_weight_synth_pops, function(sp_wts) {
    map_dbl(sp_wts, function(wt) {
      weighted.mean(samp[[y_var]], wt)
    })
  }) %>% unlist()
})

cat(sprintf("finished propensity means %.1f\n", est_timer()))

# Bayesian bootstrap weights to simulate SRS sampling variance
bb_weights = t(rudirichlet(draws, n_samp) * n_samp) %>%
  as_tibble() %>%
  as.list() %>%
  set_names(sprintf("bb_wt_%s", seq_along(.)))

# Estimate simple unweighted bayes bootstrap means
res$y_bar_samp_bayesboot = map_dfc(y_vars, function(y_var) {
  map_dbl(bb_weights, ~ weighted.mean(samp[[y_var]], .x))
})
cat(sprintf("finished bayesboot means %.1f\n", est_timer()))

```

```

# Estimate basic OR means
res$y_bar_pred = map_dfc(y_fits_confounded, function(y_fit) {
  y_hat_pos = pbart_posterior(y_fit,
                              newdata = x_ref,
                              mc.cores = cores)

  map_dbl(y_hat_pos, ~ weighted.mean(.x, sp_wts))
})

cat(sprintf("finished pred means %.1f\n", est_timer()))

# Estimate DR-RBC means
res$y_bar_drrbc = map_dfc(y_vars, function(y_var) {
  # Get the posterior distribution for the OR mean based on ref
  y_bar_pred_pos = res$y_bar_pred[[y_var]]

  # Get OR model for y_var
  pred_fit = y_fits_confounded[[y_var]]

  # Get posterior predicted values for sample based on OR model
  y_hat_pos_samp = pbart_posterior(pred_fit,
                                   newdata = x_samp,
                                   mc.cores = cores)

  # Calculate the DR-RBC mean for each sp weight associated with each
  # posterior draw
  pmap(list(y_bar_pred_pos, y_hat_pos_samp, sample_weight_synth_pops),
        function(y_bar, y_hat, sp_wts) {
          resid = samp[[y_var]] - y_hat

          # For each sp_weight associated with the draw
          # calculate a weighted mean residual and add it
          # to the predicted mean for that draw
          map_dbl(sp_wts, function(wt) {
            y_bar + weighted.mean(resid, wt)
          }) %>% unlist()

        }) %>% unlist()
})

cat(sprintf("finished DR RBC means %.1f\n", est_timer()))

# Get propensities for reference sample
ref_propensities = pbart_posterior(propensity_fit,

```

```

newdata = x_ref,
mc.cores = cores)

# Estimate DR-PSC means
x_ref_prop = x_ref %>%
  mutate(pi_hat = rowMeans(ref_propensities))

res$y_bar_drpsc = map_dfc(y_psc_fits, function(y_fit) {
  pos = pbart_posterior(y_fit, newdata = x_ref_prop, mc.cores = cores)
  map_dbl(pos, ~ weighted.mean(.x, sp_wts))
})

cat(sprintf("finished DR PSC means %.1f\n", est_timer()))

# Estimate quantities for bias decomposition
ref_phi = map2(sample_propensities, ref_propensities,
  function(s_prop, ref_prop) {
    min_s_prop = min(s_prop)
    phi = ref_prop >= min_s_prop
  })

res$y_bar_samp_confounded = y_vars %>%
  map_dfc(function(y_var) {
    y_pos = pbart_posterior(y_fits_confounded[[y_var]],
      newdata = x_samp,
      mc.cores = cores)
    y_bar_samp_confounded = colMeans(y_pos)
  })

res$y_bar_samp_unconfounded = y_vars %>%
  map_dfc(function(y_var) {
    y_pos = pbart_posterior(y_fits_unconfounded[[y_var]],
      newdata = x_samp,
      mc.cores = cores)
    y_bar_samp_unconfounded = colMeans(y_pos)
  })

# Unconfounded estimates for full population
y_bar_pop = y_vars %>%
  map(function(y_var) {
    y_pos = pbart_posterior(y_fits_unconfounded[[y_var]],
      newdata = x_ref,
      mc.cores = cores)

    list(

```

```

    # Posterior for for unconfounded population mean
    y_bar_pop_unconfounded = colMeans(y_pos),

    # Posterior for unconfounded population mean among region
    # of common support
    y_bar_pop_unconfounded_cs = map2_dbl(y_pos, ref_phi,
                                          function(y, phi) {
                                            weighted.mean(y, phi)
                                          })
  )

  }) %>%
  transpose()

res = c(res, y_bar_pop)

  cat(sprintf("Finished everything %.1f\n", from_start()))
  return(bind_rows(res, .id = "est"))
}

np = readRDS("data/cleaned/cleaned_np_civic_data.RDS")
cps = readRDS("data/cleaned/cps_civic_full_edited.RDS")
draws = 1000
pweight_synth_pops = 25
save_output = TRUE

x_vars = c("age", "sex", "racethn", "educat", "fcregion")
y_vars = str_subset(names(np), "y_") %>% set_names()
np_samples = unique(np$sample_id) %>% set_names()

## Comment out when not testing
# np = filter(np, sample_id %in% c("A", "B")) %>% sample_n(400)
# cps = sample_n(cps, 500)
# draws = 10
# pweight_synth_pops = 10
# save_output = FALSE
# np_samples = np_samples[1:2]
# y_vars = y_vars[1:2]
# save_output = FALSE
# np_samples = "A"
# y_vars = y_vars[1]

```



```

# Create synthetic population for use as reference sample
set.seed(1234)
synth_pop_ids = fpbb_synth_pops(
  weights = cps$pwsrwtg,
  L = 1,
  N = nrow(cps) * 100,
  return_weights = FALSE
)

## Loop over each sample and estimate all of the necessary conditional means
start_timer = timefactory()
est_pos_full = np_samples %>%
  set_names() %>%
  map(function(samp_id) {
    est_pos = get_estimate_posteriors(
      samp_id = samp_id,
      samp = filter(np, sample_id == samp_id),
      ref = cps,
      synth_pop_ids = synth_pop_ids[[1]],
      x_vars = x_vars,
      y_vars = y_vars,
      draws = draws,
      pweight_synth_pops = pweight_synth_pops,
      cores = 10
    )
    if (save_output) {
      saveRDS(est_pos,
              sprintf("data/posteriors/est_pos_%s.RDS", samp_id))
    }
    est_pos
  }) %>% bind_rows(.id = "sample_id")

if (save_output) {
  saveRDS(est_pos_full, "data/posteriors/est_pos_full.RDS")
}
cat(sprintf("Whole thing took %.1f seconds.\n", start_timer()))

### Get minimum inclusion propensities for each sample and
### calculate the portion of the population with common support

synth_pop_wts = synth_pop_ids %>% group_by(sp_idx_1) %>%
  arrange(sp_idx_1) %>%
  summarise(sp_wt = n()) %>% pull(sp_wt)

```

```

fit_files = list.files("data/bart_models", full.names = TRUE)

pop_common_support = map(np_samples, function(samp_id) {
  samp = np %>% filter(sample_id == samp_id)

  fits = readRDS(sprintf("data/bart_models/bart_fits_%s.RDS", samp_id))

  # Get % phi for population based on propensity model
  samp_propensities = pbart_posterior(fits$propensity_fit,
                                     newdata = samp,
                                     mc.cores = 10)

  samp_mins = map_dbl(samp_propensities, min)
  pop_propensities = pbart_posterior(fits$propensity_fit,
                                     newdata = cps,
                                     mc.cores = 10)

  pop_phi = map2_dfc(pop_propensities, samp_mins, ~ .x < .y) %>%
    colMeans()

  tibble(pct_common_support = pop_phi,
         samp_min_pi = samp_mins)
}) %>% bind_rows(.id = "sample_id")

saveRDS(pop_common_support, "data/posteriors/common_support.RDS")

```


Bibliography

- Andridge, R. R. and R. J. A. Little (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics* 27(2), 153.
- Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91(434), 444–455.
- Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Ansolabehere, S. and D. Rivers (2013). Cooperative Survey Research. *Annual Review of Political Science* 16(1), 307–329.
- Ansolabehere, S. and B. F. Schaffner (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis* 22(3), 285–303.
- Austin, P. C. (2012). Using Ensemble-Based Methods for Directly Estimating Causal Effects: An Investigation of Tree-Based G-Computation. *Multivariate Behavioral Research* 47(1), 115–135.
- Baker, R., S. Blumberg, J. M. Brick, M. P. Couper, M. Courtright, M. Dennis, D. Dillman, M. R. Frankel, P. Garland, R. M. Groves, C. Kennedy, J. Krosnick, S. Lee, P. J. Lavrakas, M. Link, L. Piekarski, K. Rao, D. Rivers, R. K. Thomas, and D. Zahs (2010). *AAPOR Report on Online Panels*. American Association for Public Opinion Research.
- Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau (2013). *Report of The AAPOR Task Force on Non-Probability Sampling*. American Association for Public Opinion Research.
- Baker, R., J. M. Brick, S. Keeter, P. Biemer, C. Kennedy, F. Kreuter, A. Mercer, and G. Terhanian (2016). *Evaluating Survey Quality in Today's Complex Environment*. American Association for Public Opinion Research.
- Bang, H. and J. M. Robins (2005, December). Doubly Robust Estimation in Missing Data and Causal Inference Models. 61(4), 962–973.
- Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly* 74(5), 817–848.
- Breidt, F. J. and J. D. Opsomer (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science* 32(2), 190–205.

- Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2), 123–140.
- Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5–32.
- Bremer, J. (2013). Research quality: The interaction of sampling and weighting in producing a representative sample online: An excerpt from the ARF's "Foundations of Quality 2" initiative. *Journal of Advertising Research* 53(4), 363–371.
- Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics* 29(3), 329–353.
- Brick, J. M. (2015). Compositional model inference. *JSM Proceedings (Survey Research Methods Section)*, 299–307.
- Brigham, N., M. Fallig, and C. Miller (2014). The Impact of Survey Routers On Sampling and Surveys: Unraveling the Mysteries of Survey-Router Design and Deployment. *Journal of Advertising Research* 54(4), 381–388.
- Buskirk, T. D. and S. Kolenikov (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field (SMIF)*.
- Callegaro, M., R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, and P. J. Lavrakas (2014). Online panel research: History, concepts, applications and a look at the future. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, and P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective*, pp. 1–22. New York: John Wiley and Sons.
- Callegaro, M. and C. DiSogra (2008). Computing Response Metrics for Online Panels. *Public Opinion Quarterly* 72(5), 1008–1032.
- Callegaro, M., K. L. Manfreda, and V. Vehovar (2015). *Web Survey Methodology*. Los Angeles: Sage.
- Callegaro, M., A. Villar, D. Yeager, and J. A. Krosnick (2014). A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, J. A. Krosnick, and P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective*, pp. 23–53. West Sussex: Wiley and Sons.
- Chang, L. and J. a. Krosnick (2009). National surveys via RDD telephone interviewing versus the internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly* 73(4), 641–678.

- Chipman, H. A., E. I. George, and R. E. McCulloch (1998). Bayesian CART Model Search. *Journal of the American Statistical Association* 93(443), 935–948.
- Chipman, H. A., E. I. George, and R. E. McCulloch (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics* 4(1), 266–298.
- Clark, J., C. Young, and R. Petrin (2015). Meta-Analysis of Online Panel and Non-Panel Sampling: Electoral and Non-Electoral Behavior Metrics. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.
- Cohen, M. P. (1997). The Bayesian Bootstrap and Multiple Imputation for Unequal Probability Sample Designs. *JSM Proceedings, Section on Survey Research Methods*, 635–638.
- Cole, S. R., M. a. Hernán, J. M. Robins, K. Anastos, J. Chmiel, R. Detels, C. Ervin, J. Feldman, R. Greenblatt, L. Kingsley, S. Lai, M. Young, M. Cohen, and A. Muñoz (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology* 158(7), 687–694.
- Cole, S. R. and M. A. Hernán (2008, September). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology* 168(6), 656–664.
- Cole, S. R. and E. A. Stuart (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology* 172(1), 107–115.
- Converse, J. M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.
- Cook, T. D., W. R. Shadish, and V. C. Wong (2008, June). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management* 27(4), 724–750.
- Couper, M. P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly* 64(4), 464–494.
- Craig, B. M., R. D. Hays, A. S. Pickard, D. Cella, D. A. Revicki, and B. B. Reeve (2013). Comparison of US panel vendors for online surveys. *Journal of medical Internet research* 15(11), e260.
- Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009). Dealing with limited overlap in estimation of average treatment effects. 96(1), 187–199.

- Dehejia, R. H. and S. Wahba (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448), 1053–1062.
- Dehejia, R. H. and S. Wahba (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Review of Economics and Statistics* 84(1), 151–161.
- Dever, J. A. and R. Valliant (2016). General Regression Estimation Adjusted for Undercoverage and Estimated Control Totals. *Journal of Survey Statistics and Methodology* 4(3), 289–318.
- Deville, J.-C. (2004). Efficient balanced sampling: The cube method. 91(4), 893–912.
- Deville, J.-C. and C.-E. Särndal (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* 87(418), 376–382.
- Deville, J.-C. and Y. Tillé (1998). Unequal Probability Sampling Without Replacement Through a Splitting Method. 85(1), 89–101.
- Deville, J.-C. and Y. Tillé (2000, April). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference* 86(1), 215–227.
- DiSogra, C., C. Cobb, E. Chan, and J. M. Dennis (2011). Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. Alexandria, VA, pp. 4501–4515. American Statistical Association.
- Dong, Q., M. R. Elliott, and T. E. Raghunathan (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology* 40(1), 29–46.
- Doove, L., S. Van Buuren, and E. Dusseldorp (2014, April). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis* 72, 92–104.
- Dutwin, D. and T. D. Buskirk (2017). Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples. *Public Opinion Quarterly* 81(S1), 213–239.
- Eckman, S. (2016). Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias? *Social Science Computer Review* 34(1), 41–58.
- Elliott, M. R. and R. Valliant (2017). Inference for Nonprobability Samples. *Statistical Science* 32(2), 249–264.

- Erens, B., S. Burkill, M. P. Couper, F. Conrad, S. Clifton, C. Tanton, A. Phelps, J. Datta, C. H. Mercer, P. Sonnenberg, P. Prah, K. R. Mitchell, K. Wellings, A. M. Johnson, and A. J. Copas (2014). Nonprobability web surveys to measure sexual behaviors and attitudes in the general population: A comparison with a probability sample interview survey. *Journal of medical Internet research* 16(12), e276.
- Fahimi, M., F. M. Barlas, R. K. Thomas, and N. Buttermore (2015). Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice* 8(6).
- Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38(4), 367–378.
- GfK Public Affairs (2016). The AP-GfK Poll, October 20-24, 2016.
- Ghitza, Y. and A. Gelman (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science* 57(3), 762–776.
- Ghosh, M. and G. Meeden (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman & Hall.
- Gittelman, S. H., R. K. Thomas, P. J. Lavrakas, and V. Lange (2015). Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples. *Journal of Advertising Research* 55(4), 368–379.
- Green, D. P. and H. L. Kern (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly* 76(3), 491–511.
- Greenland, S. and J. M. Robins (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology* 15(3), 413–419.
- Greenland, S. and J. M. Robins (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic perspectives & innovations* 6(4).
- Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* 70(5), 646–675.
- Groves, R. M. (1989). *Survey Errors and Survey Costs*. Hoboken, New Jersey: John Wiley & Sons.
- Groves, R. M., Floyd J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology* (4 ed.). Hoboken: John Wiley & Sons.

- Groves, R. M. and L. Lyberg (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly* 74(5), 849–879.
- Hansen, M. H., W. N. Hurwitz, and W. G. Maddow (1953). *Sample Survey Methods And Theory*. New York: John Wiley & Sons.
- Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing Selection Bias Using Experimental Data. *Source: Econometrica* *Econometrica* 66(5), 1017–1098.
- Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching Evidence Job As An Econometric Estimator: Evidence from Evaluating a Job Training Programme. *Review of Economic Studies* 64(4), 605–654.
- Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004). A structural approach to selection bias. *15(5)*, 615–625.
- Hernán, M. A. and J. M. Robins (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 60(7), 578–586.
- Hill, J. and Y. S. Su (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes. *Annals of Applied Statistics* 7(3), 1386–1420.
- Hill, J., C. Weiss, and F. Zhai (2011). Challenges With Propensity Score Strategies in a High-Dimensional Setting and a Potential Alternative. *Multivariate Behavioral Research* 46(3), 477–513.
- Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* 20(1), 217–240.
- Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *71(4)*, 1161–1189.
- Ho, D. E., K. Imai, G. King, and E. A. Stuart (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42(8), 1–28.
- Horvitz, D. G. and D. J. Thompson (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* 47(260), 663–685.
- Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social and Biomedical Sciences*. New York, NY: Cambridge University Press.
- Ipsos Public Affairs (2016). Reuters/Ipsos Poll October 13-17, 2016.

- Kalton, G. and I. Flores-Cervantes (2003). Weighting methods. *Journal of Official Statistics* 19(2), 81–97.
- Kang, J. D. Y. and J. L. Schafer (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* 22(4), 523–539.
- Kaplan, D. and J. Chen (2012, July). A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and Case Study. 77(3), 581–609.
- Keeter, S., K. McGeeney, A. Mercer, N. Hatley, E. Patten, and A. Perrin (2015). Coverage Error in Internet Surveys: Who Web-Only Surveys Miss and How That Affects Results. *Pew Research Center*.
- Keiding, N. and T. A. Louis (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 179(2), 319–376.
- Kennedy, C., A. Mercer, S. Keeter, N. Hatley, K. Mcgeeney, and A. Gimenez (2016). Evaluating Online Nonprobability Surveys. *Pew Research Center*.
- Kern, H. L., E. A. Stuart, J. Hill, and D. P. Green (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness* 9(1), 103–127.
- King, G. and L. Zeng (2006). The dangers of extreme counterfactuals. *Political Analysis* 14(2), 131–159.
- Kreuter, F. and K. Olson (2011, May). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research* 40(2), 311–332.
- Kreuter, F., S. Presser, and R. Tourangeau (2008, December). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72(5), 847–865.
- Lax, J. R. and J. H. Phillips (2009). Gay Rights in the States: Public Opinion and Policy Responsiveness. *American Political Science Review* 103(03), 367–386.
- Lechner, M. (2008). A Note on the Common Support Problem in Applied Evaluation Studies. *Annales d'Économie et de Statistique* (91/92), 217–235.
- Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29(3), 337–346.
- 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. and R. Valliant (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research* 37(3), 319–343.
- Little, R. and H. An (2004). ROBUST LIKELIHOOD-BASED ANALYSIS OF MULTIVARIATE DATA WITH MISSING VALUES. *Statistica Sinica* 14(3), 949–968.
- Little, R. J. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association* 99(466), 546–556.
- Little, R. J. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics* 28(3), 309–334.
- Little, R. J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the IAOS* 31(4), 555–563.
- Little, R. J. and S. L. Vartivarian (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology* 31(2), 4–11.
- Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Hoboken: John Wiley and Sons.
- Little, R. J. A. and H. Zheng (2007). The Bayesian Approach to the Analysis of Finite Population Surveys. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 8*, pp. 283–302. Oxford University Press.
- Lorch, J., K. Cavallaro, and R. van Ossenbruggen (2010). Sample Blending: $1+1 > 2$. *Survey Sampling International*.
- Lusinchi, D. (2012, April). "President" Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame? *Social Science History* 36(1), 23–54.
- Lusinchi, D. (2017). The Rhetorical Use of Random Sampling: Crafting and Communicating the Public Image of Polls as a Science (1935-1948). *Journal of the History of the Behavioral Sciences* 53(2), 113–132.
- Manski, C. F. (2007). *Identification for Prediction and Decision*. Cambridge: Harvard University Press.
- McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 9(4), 403–425.

- McCulloch, R., R. Sparapani, R. Gramacy, C. Spanbauer, and M. Pratola (2018). *BART: Bayesian Additive Regression Trees*. R package version 1.5.
- Mercer, A., A. Lau, and C. Kennedy (2018). For Weighting Online Opt-In Samples, What Matters Most? *Pew Research Center*.
- Mercer, A. W., F. Kreuter, S. Keeter, and E. A. Stuart (2017). Theory and Practice in Nonprobability Surveys: Parallels Between Causal Inference and Survey Inference. *Public Opinion Quarterly* 81(S1), 250–271.
- Mosteller, F. (2010). *The Pleasures of Statistics: An Autobiography of Frederick Mosteller*. New York: Springer. OCLC: ocn456903796.
- Mosteller, H., F., H. McCarthy, P., E. S. Marks, D. B. Truman, L. W. Doob, D. MacRae, F. F. Stephan, S. A. Stouffer, and S. Wilks (1949). *The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts*. New York: Social Science Research Council.
- Myers, J. A., J. A. Rassen, J. J. Gagne, K. F. Huybrechts, S. Schneeweiss, K. J. Rothman, M. M. Joffe, and R. J. Glynn (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174(11), 1213–22.
- Newman, B. J., S. Shah, and L. Collingwood (2018). Race, Place, and Building a Base. *Public Opinion Quarterly* 82(1), 122–134.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* 97(4), 558–625.
- Park, D. K., A. Gelman, and J. Bafumi (2004). Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis* 12(4), 375–385.
- Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys* 3(0), 96–146.
- Pearl, J. (2009b). *Causality: Models, Reasoning and Inference (2nd Edition)*. Cambridge, UK: Cambridge University Press.
- Pearl, J. (2010). On a Class of Bias-Amplifying Variables that Endanger Effect Estimates. *Proceedings of UAI*, 417–424.
- Pearl, J. and E. Bareinboim (2014). External validity: From do-calculus to transportability across populations. *Statistical Science* 29(4), 579–595.
- Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21(1), 31–54.

- Petrin, R. A. and N. El-Dash (2015). Reaching Wider, Going Deeper: Incorporating Sample Source Variation and Other Considerations Into MRP Adjustments of Polling Estimates from Blended River Samples. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.
- Pettit, A. (2015). Building a quality nonprobability panel: Methods, problems, and being innovative. *Presentation at the 2015 Conference of the American Association for Public Opinion Research*.
- Porro, G. and S. M. Iacus (2009). Random Recursive Partitioning : A Matching Method for the Estimation of the Average Treatment Effect. *Journal of Applied Econometrics* 24(1), 163–185.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
- Rivers, D. (2007). Sampling for web surveys. In *JSM Proceedings (Survey Research Methods Section)*, Alexandria, VA. American Statistical Association.
- Rivers, D. and D. Bailey (2009). Inference from matched samples in the 2008 US national elections. *Presented at the 2009 American Association for Public Opinion Research Annual Conference, Hollywood, Florida*.
- Robins, J. M. (1999a). Association, causation, and marginal structural models. *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science* 121(1-2), 151–179.
- Robins, J. M. (1999b). Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, Volume 116, pp. 95–134. New York, NY: Springer-Verlag.
- Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, pp. 1–92. New York: Springer-Verlag.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association* 89(427), 846–866.
- Rosenbaum, P. R. (2002). *Observational Studies*. Springer Series in Statistics. New York: Springer.

- Rosenbaum, P. R. (2005). Sensitivity Analysis in Observational Studies. In B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*, pp. 1809–1814. Chichester, UK: John Wiley & Sons.
- Rosenbaum, P. R. and D. B. Rubin (1983a). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* 45(2), 212–218.
- Rosenbaum, P. R. and D. B. Rubin (1983b). The Central Role of the Propensity Score in Observational Studies for Causal Effects. 70(1), 41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5), 688–701.
- Rubin, D. B. (1978, January). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics* 6(1), 34–58.
- Rubin, D. B. (1981, January). The Bayesian Bootstrap. *The Annals of Statistics* 9(1), 130–134.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer.
- Schnorf, S., A. Sedley, M. Ortlieb, and A. Woodruff (2014). A Comparison of Six Sample Providers Regarding Online Privacy Benchmarks. *Symposium on Usable Privacy and Security Workshop on Privacy Personas and Segmentation*.
- Schonlau, M., A. V. Soest, and A. Kapteyn (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., K. Zapert, L. P. Simon, K. H. Sanstad, S. M. Marcus, J. Adams, M. Spranca, H. Kan, R. Turner, and S. H. Berry (2004). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review* 22(1), 128–138.
- Schouten, B., J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner (2012). Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators. *International Statistical Review* 80(3), 382–399.
- Schouten, B., F. Cobben, and J. Bethlehem (2009). Indicators for the representativeness of survey response. *Survey Methodology* 35(1), 101–113.

- Schouten, B., N. Shlomo, and C. Skinner (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics* 27(2), 231–253.
- Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont: Wadsworth Cengage Learning.
- Snowden, J. M., S. Rose, and K. M. Mortimer (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology* 173(7), 731–738.
- Steiner, P. M., T. D. Cook, W. R. Shadish, and M. H. Clark (2010). The Importance of Covariate Selection in Controlling for Selection Bias in Observational Studies. *Psychological Methods* 15(3), 250–267.
- Stuart, E. A. (2010a). Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1), 1–21.
- Stuart, E. A. (2010b). The use of propensity scores to assess generalizability. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 174(2), 369–386.
- Stuart, E. A., C. P. Bradshaw, and P. J. Leaf (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science* 16(3), 475–485.
- Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (2011). The use of propensity scores to assess the generalizability of results from randomized trials: Use of Propensity Scores to Assess Generalizability. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2), 369–386.
- Tan, Y. V., C. A. C. Flannagan, and M. R. Elliott (2018, January). "Robust-squared" Imputation Models Using BART. *arXiv:1801.03147 [stat]*.
- Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science* 22(4), 560–568.
- Taylor, H. (2000). Does internet research work? *International Journal of Market Research* 42(1), 51–63.
- Terhanian, G. and J. Bremer (2000). Confronting the Selection-Bias and Learning Effects Problems Associated With Internet Research. Technical report, Harris Interactive.
- Terhanian, G. and J. Bremer (2012). A smarter way to select respondents for surveys? *International Journal of Market Research* 54(6), 751–780.

- The Washington Post and ABC News (2016). Washington Post-ABC News Poll, October 10-13, 2016.
- Valliant, R. and J. A. Dever (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research* 40(1), 105–137.
- Valliant, R., J. A. Dever, and F. Kreuter (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.
- Valliant, R., A. H. Dorfman, and R. M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons.
- van Buuren, S. and K. Groothuis-Oudshoorn (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3), 1–67.
- Wallace, B. C., K. Small, C. E. Brodley, and T. A. Trikalinos (2011). Class Imbalance, Redux. *Proceedings of the 11th IEEE International Conference on Data Mining*, 754–763.
- Wang, S., J. Shao, and J. K. Kim (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* 24.
- Wang, W., D. Rothschild, S. Goel, and A. Gelman (2014). Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting* 31(3), 980–991.
- West, S. G., N. Duan, W. Pequegnat, P. Gaist, D. C. D. Jarlais, D. Holtgrave, J. Szapocznik, M. Fishbein, B. Rapkin, M. Clatts, and P. D. Mullen (2008). Alternatives to the Randomized Controlled Trial. *American Journal of Public Health* 98(8), 1359–1366.
- Wright, M. N. and A. Ziegler (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77(1), 1–17.
- Xu, D., M. J. Daniels, and A. G. Winterstein (2016, July). Sequential BART for imputation of missing covariates. *Biostatistics (Oxford, England)* 17(3), 589–602.
- Yeager, D. S., J. a. Krosnick, L. Chang, H. S. Javitz, M. S. Levendusky, A. Simpser, and R. Wang (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.

- Young, C., J. Vidmar, J. Clark, and N. El-Dash (2012). Our Brave New World: Blended Online Samples and the Performance of Nonprobability Approaches. *Ipsos Public Affairs*.
- Zhang, G. and R. Little (2009). Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. *65*(3), 911–918.
- Zhou, H., M. R. Elliott, and T. E. Raghunathan (2016, January). Synthetic Multiple-Imputation Procedure for Multistage Complex Samples. *Journal of Official Statistics* *32*(1).