Title
Validating Causal Estimates in Experimental and Observational Designs

Author
Hartman, Erin Kristine

Publication Date
2013

Peer reviewed|Thesis/dissertation

Validating Causal Estimates in Experimental and Observational Designs

By

Erin Kristine Hartman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Political Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jasjeet Sekhon, Chair
Professor Eric Schickler
Professor Sandrine Dudoit

Fall 2013

Validating Causal Estimates in Experimental and Observational Designs

Abstract

Validating Causal Estimates in Experimental and Observational Designs

by

Erin Kristine Hartman

Doctor of Philosophy in Political Science

University of California, Berkeley

Professor Jasjeet Sekhon, Chair

Social scientists and policy makers continue to put increased emphasis on identifying causal effects in their research, employing the myriad of novel approaches that have been developed in recent years. With this rise in the use of causal analysis tools, the importance of understanding the assumptions underlying such methods has increased. Methods and estimates are occasionally stretched beyond the limits of the underlying assumptions. This has increased the need for a thorough understanding of what assumptions are necessary to identify causal estimates outside of a purely experimental framework. Additionally, it has lead to a need for validation tests that allow researchers to provide sound evidence that estimates are unbiased.

This dissertation focuses on one area, methods for increasing the external validity of experimental estimates, of interest to researchers and policy makers alike. This problem is addressed from multiple angles, both theoretical, practical, and from a design perspective. First, I outline the assumptions necessary to identify population treatment effects from experimental data. I then discuss a new estimating strategy for testing the key identifying assumption central to most causal methods, the notion of exchangeability. This new method leverages tests of equivalence to redefine the approach to validation tests such as balance and placebo tests. Finally, a new approach to improving the underlying data from which we explore questions and estimate quantities of interest is provided. This new method, response rate sampling, allows researchers to collect more representative survey data. Combined, these tools will allow researchers to move more fluidly between experimental and observational estimates.

For my family.

The one I inherited and the one I found along the way.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the support and advice I've received from numerous people over the years.

First and foremost, I thank my adviser Jas Sekhon. His guidance throughout this process has been invaluable. Jas has pushed me to be a better methodologist since the moment I walked onto campus at Berkeley. He encouraged me from early on to explore and present new ideas. He has provided guidance and support both for my academic endeavors and on my chosen career path. Through Jas, I have learned to approach each problem with an eye towards clever and sound design.

I am grateful to Mike Alvarez for pushing me to pursue political science, and for his guidance in navigating a successful career path. I thank Betsy Sinclair for being a mentor and friend over the past years, serving as a sounding board for everything from trivial stats questions to career changing decisions.

The work I have done has been made possible by countless conversations with my fellow graduate students at Berkeley, a group of intelligent and giving compatriots on this journey. In particular, I'd like to thank Adrienne Hosek. Abby Wood, Devin Caughey, Danny Hidalgo, and Rocio Titiunik have also helped me at many steps in this process. Additionally, many members of the Berkeley faculty have provided feedback and support. In addition to Jas, I thank my committee members, Eric Schickler and Sandrine Dudoit, for their feedback. I also thank Sean Gailmard, Henry Brady, and Bin Yu. Berkeley has allowed me to present at numerous colloquia, workshops, and conferences. I owe a debt of gratitude to the discussants and audience members at these events, particularly those at the UC Berkeley Research Workshop in American Politics, the UC Berkeley IGERT seminar, as well as numerous MPSA, APSA, and Political Methodology conferences.

In the middle of my studies, I took an unusual detour and worked for President Obama's reelection campaign. I am grateful to my committee for their support and encouragement on this path, without which I would not be where I am today. I thank Amy Gershkoff for inviting me to work on the campaign, and for introducing me to an amazing team of individuals. I thank Nate Sterken for being a rock during my time on the campaign. I am grateful to Elan Kriegel, who took a chance on me when I joined the Analytics team. He let me take over the world's largest political polling operation, creating an environment where I could test and perfect new methods. I also thank Matt

Holleque, who's insights and support have challenged me to make my work continually better. Finally, I thank everyone at BlueLabs, a group of people that constantly amaze me. I couldn't be prouder of the work that we do.

I am grateful for the financial support of the UC Berkeley Travers Department, the National Science Foundation Integrated Graduate Research Traineeship fellowship, Jas Sekhon, the UC Berkeley Institute for Governmental Studies, and the UC Berkeley Deans Normative Time Fellowship.

I have a network of amazing friends that have supported me throughout my studies. I am forever indebted to Becky Rutherford, Abby Crites, Julia Ma, and Dustin Nest. You all have contributed to making my life better, and I am constantly impressed by everything you accomplish.

Finally, I thank my family. I would not be here without the constant support you have provided. Your unending love and encouragement has allowed me to become the independent, curious person I am today. You have instilled in me a lifelong love of learning, without which this would not have been possible. While I may finally be done with school, thanks to you, I will never be done learning.

Chapter 1

Introduction

The rise of causal inference in the social sciences has lead to many sophisticated and novel estimation strategies. These new methods have greatly improved the ability of social scientists to provide valid estimates of causal quantities of interest. However, potential weaknesses in key identifying assumptions can undermine the validity of those estimates. While experiments provide internally valid sample average treatment effects, often practitioners are more interested the treatment effect on larger populations. For many experiments, the external validity of the estimates depends on the representativeness of the underlying experimental data. In observational methods, a key assumption is the exchangeability of treatment and control groups.

1.1 Estimands of Interest

As the focus of social scientists has turned towards causal inference, randomized trials continue to hold priority in the chain of causal identification strategies. Fisher (1925) first showed the statistical properties of randomized trials, most importantly that they provide an unbiased estimate of the average treatment effect. By allowing experimenters to control the manipulation of a small number of factors, randomized trials can allow, under minimal assumptions, for the identification of causal effects of interest to policy makers and social scientists. Whereas treatment can be endogenous to numerous factors in non-random studies (NRSs), randomized trials ensure that treatment are exogenous. As Imbens (2009) points out, the minimal assumptions required for identification of average effects in randomized controlled trials (RCTs) are more likely to be satisfied than the assumptions necessary to identify similar effects in observational studies because there are no assumptions on the treatment assignment mechanism. Specifically, he states, "it seems difficult to argue that, in a setting where it is possible to carry out a randomized experiment, one would ever benefit from giving up control over the assignment mechanism, by allowing individuals to choose their own treatment status." Freedman agrees, saying, "[t]he moral is, do the experiment; be wary of model-based interpretations" (Freedman 2006). The biggest advantage of RCTs is their strong internal validity, and thus their unbiased estimates of the average treatment effect. However, the validity of the study is limited to the population of people in the experiment (Imbens 2009; Banerjee and Duflo 2008; Heckman and Vytlacil 2005). RCTs identify, based on the sample population, the

Sample Average Treatment Effect (SATE).

Despite the advantages of RCTs, they have many short comings, especially with regards to the policies that policy makers wish to implement and the quantities social scientists wish to estimate. Most importantly, the external validity of a randomized trial is limited (Imbens 2009; Heckman and Urzua 2009; Deaton 2009; Heckman and Vytlacil 2005). In order to extend the findings of an RCT to a more general population some assumptions must be made. As Banerjee and Duflo note, "without assumptions, results from experiments cannot be generalized beyond their context; but with enough assumptions, observational data may be sufficient" (Banerjee and Duflo 2008). The necessity of these assumptions questions the cost-bias trade off of each method. Nonrandom studies can provide information about how best to generalize the findings of an experiment, such as by providing information about who generally receives treatment in the population when the usual selection mechanisms are in place (Heckman 1992). For example, in political science experiments, NRSs provide information about which voters turnout to vote under normal campaign contexts. NRSs also provide a wealth of information about the population of interest above and beyond information about who receives treatment.

Observational studies, however, are often subject to internal validity problems due to confoundedness of treatment assignment. RCTs are usually quite expensive, therefore NRSs typically are larger and cover more of the population of interest. This might insulate them from some of the external validity issues that RCTs face since the treatment is exposed to more subjects in many different settings. Banerjee and Duflo suggest this may be a reason for the "tradeoff between the more 'internally' valid randomized studies and the more 'externally' valid observational studies". This tradeoff is important because we are often interested in an estimand that RCTs cannot identify without strong assumptions, namely the Population Average Treatment Effect (PATE). Policy makers are often more specifically interested in the Population Average Treatment Effect on the Treated (PATT), or the treatment effect for those who are likely to receive treatment in practice. The goal of much of this dissertation is to identify the potential biases and assumptions necessary to move fluidly between experimental and observational estimates. This is done both through theoretical work on the necessary identifying assumptions and by identifying methods that increase the representativeness of experimental data using both methods for improving the representativeness of samples on the front end in survey data collection methods, and through new weighting methods for *post hoc* adjustment of return data.

The work done in this dissertation helps identify, test, and minimize the key identifying assumptions that allow practitioners to go from sample average treatment effect estimates to population estimates of interest. Chapter 2 outlines the assumptions and a new estimating strategy. Chapter 3 focuses on a new approach to validity checks such as balance and placebo tests. This new approach aids practitioners in validating estimates, however the use cases extend beyond extrapolation methods and can be used in either pure experimental or pure observational designs. Chapter 4 focuses on a design based ap-

proach for collecting more representative survey data. This minimizes some of the biases described in the following section related to sample representativeness with regards to the target population, however the methods described in Chapter 4 can be used generally in surveys, extending the reach and use cases of the methods beyond survey experiments.

1.2 Bias Decomposition

Going from the SATE to the estimand of interest, PATT, can lead to estimation error and bias. This bias can be decomposed into two separate factors, sample selection bias and bias of the estimated treatment effect, referred to here as treatment bias (Imai, King and Stuart 2008). This is a theoretical decomposition that serves to show what inherent biases arise when estimating population treatment effects from sample data. This decomposition outlines the problems that the following chapters will address. Chapter 2 outlines the underlying assumptions necessary to alleviate such biases, Chapter 3 focuses on validity checks for identifying when such biases are impacting population estimates, and Chapter 4 addresses design based approaches for alleviating biases due to representativeness in the context of survey data. Some aspects of the bias can be remedied by choosing specific estimands, such as limiting the quantity of interest to the treatment effect of the treated, others can be overcome by employing estimating strategies.

Using the Neyman-Rubin potential outcomes framework (Holland 1986; Rosenbaum 2002; Rubin 1974, 2006), let Y_{i1} denote the potential outcome when unit i receives treatment, and let Y_{i0} denote the potential outcome under control. Therefore, the treatment effect for unit i is defined as $\tau_i = Y_{i1} - Y_{i0}$. Only one of these potential outcomes can be observed at a time, depending on if unit i is assigned to the treatment or control regime. If we define T_i as a treatment indicator, taking value 1 when unit i is in the treatment regime, and 0 when unit i is in the control group, then we can define the observed outcome for unit i as[1]:

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$$

Let I_i denote a sample selection indicator where $I_i = 1$ indicates that unit i is included in the RCT sample, and $I_i = 0$ indicates that unit i did not participate in the randomized controlled trial. Let n denote the number of units in the RCT sample. Finally, let X represent a set of observable covariates in the RCT and let W represent a set of observable covariates for which moments can be calculated in both the sample and population of interest. X and W are not necessarily the same set of covariates, although they can be. X is a set of covariates that potentially confound treatment assignment in the RCT and W is a set of covariates that may contribute to sample selection bias.[2]

[1]For an extension to non-bivariate treatments see (Sekhon 2010) page 491.

[2]While the value of W must be observed for all units in the sample, only the moments need to be known for the full population in order for the MaxEnt weighting method to be applied.

In order to decompose the biases, we must define the support regions for the conditioning covariates X and W. A support region is a set of X values whose complement has zero probability, and the common support regions the overlap of two support regions. Define the support regions are as follows:

- S_{I1} as the common support for the treated and controls in the RCT.
- S_{T1} as the common support of the treated in the RCT and the treated in the NRS
- S_{1i} as the support for the treated / control in the RCT for $i \in (0, 1)$, resp.
- S_{01} as the support for the treated in the NRS

where common support is defined as the region of overlap in the support of covariate distributions for two groups. In the case of common support for the treated and control groups within the RCT, the covariate distributions for the conditioning covariates X must overlap. Similarly, there must be overlap in the covariate distributions of the conditioning covariates W in the common support region between the RCT population and the NRS population.

We can think of the bias in our estimate of the PATT as a combination of bias due to possible lack of internal validity of the estimate of SATE and bias due to poor external validity of the SATE estimate from the RCT to the population of interest.

$$B = B_I + B_E$$

The following theoretical decomposition extends the Heckman et al. (1998) decomposition of bias to the case of population and sample differences in addition to issues arising from the non-random assignment of treatment in a sample. Imai, King and Stuart (2008) provide an alternative discussion of how treatment bias and randomization bias contribute to estimation bias, due both to observable and unobservable characteristics.

The internal validity problem is due to treatment imbalance in the sample population. In the case of an RCT, this may be due to finite sample imbalance in the treatment and control group and in non-randomized samples it may be due to forms of selection bias.

$$\begin{aligned} B_I = {} & \mathbb{E}_{S_{11} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T = 1, I = 1)\} && (1.1) \\ & - \mathbb{E}_{S_{10} \setminus S_{I1}} \{\mathbb{E}(Y_0|X, T = 0, I = 1)\} && (1.2) \\ & + \mathbb{E}_{dF(T=1)-dF(T=0), S_{I1}} \{\mathbb{E}(Y_0|X, T = 0, I = 1)\} && (1.3) \\ & + \mathbb{E}_{dF(T=1), S_{I1}} \{\mathbb{E}(Y_0|X, T = 1, I = 1) - \mathbb{E}(Y_0|X, T = 0, I = 1)\} && (1.4) \end{aligned}$$

Equations (1.1) and (1.2) refer to bias that arises from a lack of common support of the treatment and control groups in the sample. The outer expectation of equation (1.1) is taken with respect to those treated units who are not in the over lap region, meaning

there is no good counterfactual for these units.[3] $S_{11} \backslash S_{I1}$ is defined as the support for the treated units in the RCT sample population not contained in the common support region for the treated and control units in the RCT population. The outer expectation of equation (1.2) is taken with respect to those control units who are not in the overlap region for the sample population, where the notation is analogous to that for equation (1.1). In a Get-Out-The-Vote experiment, this could occur if targets who were either very likely or very unlikely to vote were excluded from either the treatment or control group, but not both. In the case of a randomized trial, lack of common support should be minimal problem because observable and unobservable characteristics are independent of treatment assignment and thus the common support region will be large because there are no observable characteristics that contribute to treatment selection bias, especially if there are no compliance issues. However, it remains an issue in finite samples. These biases can be bounded by using the SATT estimand. SATT only considers the distribution of the treated covariates, and thus it renders the bias in (1.2) irrelevant because those control units who are not in the overlap region of the sample population are not considered.

Equation (1.3) arises because of bias due to the differential weighting of the the Y_0 outcome in the treated and control groups due to the differential densities in observable covariates in these two groups. The outer expectation in equation (1.3) is taken with respect to the difference between the distribution of covariates for treated and control units in the overlap region. This could arise in a political persuasion experiment if door canvassers were less enthusiastic or less likely to converse with supporters of the opposite party despite treatment assignment. Equation (1.3) deals with selection bias within the overlap region. This bias can be bounded by using matching methods and the SATT estimand, which yields equivalent covariate densities for those units in the overlap region. Finally, equation (1.4) is selection bias as typically defined. This bias occurs when the potential outcomes are different for treated and control units even when conditioning on the distribution observable characteristics X. This can happen, for example, if there is imbalance in unobservables. This bias can be of any magnitude and direction (Heckman et al. 1998), however in the case of a randomized trial the problem should be minimal if units do not have control over their selection into treatment regime. This bias might be present in an RCT if treatment was considered receipt of treatment, traditionally referred to as the treatment-on-treated estimator (ToT), rather than intention to treat (ITT). In political persuasion or Get-Out-The-Vote experiments conducted by campaigns, the difference between ToT and ITT estimates can be significantly different, raising concern of bias in the ToT estimator.

Outlining these biases in the context of internal validity provides an intuitive sense for the different sources of bias. The focus of this dissertation, though, lies in methods for

[3]By not being in the overlap region, this means a treated unit is not similar, based on the distributions of X, to any control units. This may be because they were never likely to receive control, and thus there is no unit in the control unit who serves as a good counterfactual for them.

transitioning from sample average treatment effects to estimates among a target population, as well as validating those estimates. The external validity bias is a problem because it limits what the sample estimates can say about treatment in the population of interest. For instance, this problem can be due to RCT populations that are not representative of the population of interest. Unit non-response in survey data, as discussed in Chapter 4, can lead to non-representative samples for survey based experiments. When an RCT sample is unrepresentative of the population of interest, Heckman (1996) refers to this as randomization bias. The following equations decompose the bias due to lack of external validity of the RCT estimate.

$$
\begin{aligned}
B_E = \mathbb{E}_{S_{11}\setminus S_{T1}}\left\{\mathbb{E}(Y_i|W, T=1, I=1)\right\} & \quad (1.5)\\
- \mathbb{E}_{S_{01}\setminus S_{T1}}\left\{\mathbb{E}(Y_i|W, T=1, I=0)\right\} & \quad (1.6)\\
+ \mathbb{E}_{dF(I=1)-dF(I=0),S_{T1}}\left\{\mathbb{E}(Y_i|W, T=1, I=0)\right\} & \quad (1.7)\\
+ \mathbb{E}_{dF(I=1),S_{T1}}\left\{\mathbb{E}(Y_i|W, T=1, I=1) - \mathbb{E}(Y_i|W, T=1, I=0)\right\} & \quad (1.8)\\
\text{for } i \in (0,1) &
\end{aligned}
$$

The biases described in the the equation for B_E are analogous to the equations in B_I. Equations (1.5) and (1.6) refer to biases that arise from lack of common support. The outer expectation in equation (1.5) is taken with respect to the units who are in the sample population but not in the overlap region based on observable characteristics W, similar for equation (1.6) for control units not in the overlap region. In this case, the bias arises from the fact that some units in the sample population are not similar to anyone in the target population based on observable characteristics W and vice versa. The units who are not in the overlap region do not have valid counterfactuals. Given the rise of experiments run by political campaigns and the focus on in-cycle models and analysis, the external validity of estimates is important. Biases such as those described in (1.5) and (1.6) can arise if experiments are run among populations never likely to receive treatment in practice, such as strong partisans of the opposing candidate. This bias can be bounded by using the PATT estimand because then the bias from equation (1.5) is the only one that contributes to the bias since control units not in the overlap region are not a part of the PATT estimand. Only bias arising from those who are in the population of interest who do not have a valid counterfactual in the sample contribute to bias of the PATT estimand.

Equation (1.7) is bias due to the differential weighting of the treatment effect because of differential densities of covariates in the sample population and population of interest. The outer expectation in equation (1.7) is taken with respect to the difference in densities of observable covariates for units in the overlap region. This type of bias can be due to sample selection bias, analogous to the treatment selection bias describe in equation (1.3), or due to heterogenous treatment effects. If heterogenous treatment effects exist, then

treatment effects are likely to vary systematically with underlying covariates (Banerjee and Duflo 2008), and if there are differential densities for those underlying covariates then bias will be induced in estimates of average treatment effects.

The biases described in (1.7) arise, in part, by having non-representative samples. In surveys, certain groups are less likely to respond to survey attempts than others. If not accounted for, this known differential response rate leads to non-representative final data returns. The methods described in Chapter 4 provide a way of increasing the representativeness of surveys, often used for data collection in experiments, which can alleviate some of the bias due to (1.7).

Chapter 2 discusses a placebo test to determine if the identifying assumptions are met, lending credibility to the fact that weighting truly minimizes the bias in equation (1.7). Chapter 3 outlines an estimating strategy for conducting such validity checks. The bias in equation (1.8) is bias that arises because of selection issues, in this case the sample selection issue. This bias could, for example, be problematic if the conditioning characteristics, W, are insufficient for overcoming the sample selection issue. There is no way to bound this problem with an estimand or estimating strategy, and the bias can be of any magnitude and size.

The equations presented above decompose the various forms of bias that can arise when estimating the population average treatment effect. Some of these biases can be bounded using certain estimands, such as those that focus on the distribution of the treated populations. Other parts can be minimized by implementing certain estimating strategies, such as matching and weighting methods. The important aspect of this decomposition is that it describes the inherent biases that underlie estimates of population effects and they exist regardless of estimand or estimating strategy. This dissertation addresses some of these biases, allowing practitioners to more fluidly move between sample and population estimates while providing them with the tools for validating such estimates.

1.3 Organization of This Study

This dissertation focuses on methods that increase the external validity of estimates and provide tests of validity for key identifying assumptions. Chapter 2, co-authored with Richard Grieve, Roland Ramashai, and Jas Sekhon, presents a formal proof for the conditions under which internally valid sample average treatment effects can be extrapolated to target populations of interest. An observable implication of the necessary assumptions for extrapolation provides a natural validity test of the identifying assumptions. Chapter 3, co-authored with Daniel Hidalgo, discusses a new approach to balance and placebo tests by reframing the tests to be tests of *equivalence* instead of tests of *difference*, thus providing statistical evidence justifying the design. Chapter 4 discusses a new sampling method, along with improved weighting methods, that can be used to increase the representativeness of surveys, a common data collection method used in experimental and

observational studies. Combined, these chapters outline methods that will improve and validate the identifying assumptions commonly used in experimental and observational methods. Chapter 5 concludes with a discussion of the importance of a focus on validation methods, and designs that lend themselves to clean validity checks, given the continued emphasis on causal analysis in the social sciences. The following sections outline the approach of each substantive chapter.

1.3.1 Identifying Externally Valid Experimental Estimates

Experiments provide unbiased estimates of sample average treatment effects. A significant amount of work has been put into increasing the internal validity of experimental designs. However, a common concern is that RCTs may fail to provide unbiased estimates of population average treatment effects. In the healthcare literature, for example, policy makers are concerned about valid cost-effectiveness estimates for a treatment used in practice among a more general population than experiments provide. Due to differences in populations, treatment protocols, and general equilibrium effects, the effect of a treatment observed in an experimental setting may differ from effects that would be seen for a treatment in practice. Past research has focused on using non-random studies to investigate population effects. Chapter 2 adds to this literature by deriving the assumptions required to identify population average treatment effects from RCTs aided by NRS data. Additionally, an observable implication of the assumptions allows for a set of placebo tests. Finally, the chapter outlines a new research design for estimating population effects that use non-randomized studies to adjust the RCT data. This design employs a matching approach to create matched strata within the RCT, and then considers two weighting methods for adjusting the experimental data using population information from the NRS.

Following the derivation of the identifying assumptions and the associated placebo tests, Chapter 2 applies the new estimation strategy to an example. This approach is considered in a cost-effectiveness analysis of a clinical intervention, Pulmonary Artery Catheterization (PAC). The literature on PAC is, thus far, mixed in the findings of its effectiveness. A commonly used procedure, experimental findings indicated no overall benefits on measures of survival, whereas NRS methods showed possible negative effects of the intervention. Using a new estimation strategy, we recover a null effect in the population on survival, cost, and cost-effectiveness measures. However, we find subgroups for which the intervention proves to be cost-effective, such as for patients in non-teaching hospitals, or those undergoing elective surgery. More importantly, we show that while one weighting method, MaxEnt, is able to pass a set of placebo tests for most subgroups, inverse propensity score weighting is unable to pass such validity checks. Additionally, using the tests derived in Chapter 3, we are able to discern groups for which the placebo tests mail fail due to lack of power from those where unobservable differences persist. The importance of a set of validity checks for key identifying assumptions cannot be un-

derstated, for it allows practitioners to determine which extrapolated estimates are valid. The assumptions for extrapolation outlined in Chapter 2, as well as the justification for the validity checks, are agnostic to estimation strategy. No one estimation strategy works for every problem, and these validity tests help determine which estimation strategies work best for a given problem.

1.3.2 An Equivalence Approach to Balance and Placebo Tests

In both experimental and non-random studies, an assumption on the exchangeability of the treatment and control groups is necessary for causal estimates. This assumption is met by design in experiments for sample treatment effect estimates, however, for population treatment effect estimates, or in observational methods, this assumption should be validated. Recent debates over the difficulty of causal inference in the social sciences has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests. Yet some methodologists have severely criticized the current status quo of relying on hypothesis tests to assess balance on pre-treatment covariates and placebo outcomes because designs can "pass" traditional tests due solely to small sample size, thus conflating statistical insignificance with substantive insignificance. Chapter 3 shows that these problems are due to the use of an inappropriate null hypothesis, which can result in the equating of non-significant differences with significant homogeneity. When the hypothesis test is correctly specified so that *difference* is the null and *equivalence* is the alternative, the problems afflicting traditional tests are alleviated. Leveraging the statistical literature on tests of equivalence to provide an alternative framework for testing covariate and placebo balance, this chapter provides a new approach to validity checks. In addition to their superior statistical properties, this chapter argues that equivalence tests are better able to incorporate substantive considerations about what constitutes good balance on covariates and placebo outcomes than traditional tests. To demonstrate these advantages, equivalence tests are applied to eleven natural experiments in economics and political science.

1.3.3 Increasing Representativeness of Survey Data for Externally Valid Estimates

In addition to understanding the underlying assumptions necessary to provide valid population estimates, this dissertation explores methods for collecting representative data that increases the external validity of estimates. One common method of data collection used in social science experiments and observational methods is surveys. In Chapter 4, unit non-response is decomposed in to two sources of bias: imbalance on observable characteristics and unobservable non-representativeness of survey respondents. Chapter

4 outlines a design based approach, response rate sampling, to mitigate the impact of unit non-response. This method uses response data from previous surveys to estimate the the likelihood an individual will complete a survey. Sampling can then be done inversely proportional to these known response rates, increasing the likelihood that final survey response data will be representative.

In addition to outlining response rate sampling, this chapter applies this method in the context of a large scale presidential campaign, Obama for America's (OFA) internal analytics polling operation. The 2012 presidential election saw an explosion of publicly available polling, conducted by a wide variety of organizations ranging from university survey centers to large news organizations. Simultaneously, we have seen a decrease in the probability that people respond to political surveys. Due both to more restrictive laws on predictive dialing of cell phones, the rise of a cell phone only population, and behavioral changes in voters' willingness to respond to such polls, the bias from unit non-response has the potential to severely bias the findings of these polls. By incorporating previous data, response rate sampling, and a novel approach to raking weighting methods, the OFA analytics polls allowed the campaign to have accurate and stable estimates of voter opinion, even in times that public polls showed wild swings. Combined with careful analytical approaches to a consistent and accurate target election day electorate, these methods allowed the campaign to have highly accurate measures of public opinion.

Chapter 2

From SATE to PATT: Combining Experimental with Observational Studies to Estimate Population Treatment Effects

Randomized controlled trials (RCTs) can provide unbiased estimates of the relative effectiveness of alternative interventions within the study sample. Much attention has been given to improving the design and analysis of RCTs to maximise internal validity. However, policy-makers require evidence on the relative effectiveness and cost-effectiveness of interventions for target populations that usually differ to those represented by RCT participants (Willan and Briggs 2006; Nixon and Thompson 2005; Mitra and Indurkhya 2005; Willan, Briggs and Hoch 2004; Mojtabai and Zivin 2003; Hoch, Briggs and Willan 2002). A key concern is that estimates from RCTs and meta-analyses may lack external validity (Imbens 2009; Heckman and Urzua 2009; Deaton 2009; Heckman and Vytlacil 2005). In RCTs treatment protocols and interventions differ to those administered routinely, and trial participants, for example, individuals, hospitals, or schools, tend to be atypical of those in the target population, which can limit the generalisability of results (Gheorghe et al. 2013). These concerns pervade RCTs across different areas of public policy, and are key objections to undertaking RCTs, or to using the results in policy-making (Deaton 2009). It is therefore important to establish the conditions under which RCTs can identify population treatment effects, and to develop methods to test if these conditions hold in a given application.

Previous research has proposed using non-randomized studies (NRSs) to assess whether RCT-based estimates apply to a target population (Stuart et al. 2011; Green and Kern 2012; Imai, King and Stuart 2008; Kline and Tamer 2011; Cole and Stuart 2010; Shadish, Cook and Campbell 2002; Greenhouse et al. 2008). A common concern is that there may be many baseline covariates, including continuous measures, which differ between the RCT and target population, and modify the treatment effect. In these situations simple post-stratification approaches for reweighting the treatment effects from the RCT to the target population may not fully adjust for observed differences between the settings (Stuart et al. 2011). Furthermore, there may be unobserved differences between the RCT participants and the target population, and the form of treatment or control, for example the dose of a drug or the rigor of a protocol, may differ between the settings (Cole and Frangakis 2009). Hence the RCT may provide biased estimates of the effectiveness and

cost-effectiveness of the routine delivery of the treatment in the target population.

Imai, King and Stuart (2008) introduced a framework for decomposing the biases that arise when estimating population treatment effects. Stuart et al. (2011) proposed the use of propensity scores to assess the generalizability of RCTs. We extend this literature by defining the assumptions that are sufficient to identify population treatment effects from RCTs, and providing accompanying placebo tests to assess whether the assumptions hold. These tests can use observational studies to establish when treatment effects for the target population can be inferred from a given RCT. Such tests have challenging requirements: they have to follow directly from the identifying assumptions, be sensitive to key design issues, and have sufficient power to test the assumptions– not just for overall treatment effects, but also for subgroups of prime interest. The formal derivations and the placebo tests allow for a number of research designs for estimating population treatment effects. These research designs can be used with a variety of different estimation techniques, and the best estimation approach for a given problem will depend on the application in question.

We illustrate our approach in an evaluation of the effectiveness and cost-effectiveness of Pulmonary Artery Catheterization (PAC), an invasive and controversial cardiac monitoring device used in critical care. While the evidence from RCTs and meta-analyses suggests that PAC is not effective or cost-effective (Harvey et al. 2005), concerns have been raised about the external validity of these findings (Sakr et al. 2005). For this empirical application, we employ an automated matching approach, Genetic Matching (GenMatch) (Diamond and Sekhon 2013; Sekhon and Grieve 2012), to create matched strata within the RCT. We then consider two alternative techniques for weighting the individual RCT strata according to the observed characteristics in the target population: inverse propensity score weighting (IPSW) and maximum entropy (MaxEnt) weighting. Each technique aims to reweight the individual strata in the RCT so that they resemble the distribution of characteristics in the target population.

The chapter proceeds as follows. Section 2.1 introduces the motivating example and the problem to be addressed. Section 2.2 derives the assumptions required for identifying population average treatment effects. Section 2.3 describes the placebo tests for checking the underlying assumptions, while Section 2.4 outlines estimation strategies. In Section 2.5, we illustrate the approach with the PAC case study. Section 2.6 discusses an alternative design identified by the main theorem, and Section 2.7 concludes.

2.1 Motivating Example

Pulmonary Artery Catheterization (PAC) is a cardiac monitoring device used in the management of critically ill patients (Dalen 2001; Finfer and Delaney 2006). The controversy over whether PAC should be used was fuelled by NRSs that found PAC was associated with increased costs and mortality (Connors et al. 1996; Chittock et al. 2004).

These observational studies encouraged RCTs and subsequent meta-analyses, all of which found no statistically significant difference in mortality between the randomized groups (Harvey et al. 2005). The largest of these RCTs was the UK publicly funded PAC-Man Study, which randomized individual patients to either monitoring with a PAC, or no PAC monitoring (no PAC). (Harvey et al. 2005). This RCT had a pragmatic design, with broad inclusion criteria and an unrestrictive treatment protocol, which allowed clinicians to manage patients as they would in routine clinical practice. The study randomized 1,014 subjects recruited from 65 UK hospitals during 2000-2004, and reported that overall, PAC did not have a significant effect on mortality (Harvey et al. 2005), but that there was some heterogeneity in the effect of PAC according to patient subgroup (Harvey et al. 2008). An accompanying CEA used mortality and resource use data directly from the RCT, and reported that PAC was not cost-effective (Stevens et al. 2005). However, despite the pragmatic nature of the RCT, commentators suggested that the patients and centres differed from those where PAC was used in routine clinical practice (Sakr et al. 2005). The major concern was that subgroups for which PAC might be relatively effective (e.g. elective surgical patients), were underrepresented in the RCT, and the unadjusted estimates of effectiveness and cost-effectiveness from the RCT, might not apply to the target population.

To consider the costs and outcomes following PAC use in routine clinical practice, a prospective NRS was undertaken using data from the Intensive Care National Audit Research Centre (ICNARC) Case Mix Program (CMP) database. The ICNARC CMP database contains information on case-mix, patient outcome and resource use for 200 critical care units in the United Kingdom (Harrison, Brady and Rowan 2004). A total of 57 units from the CMP collected additional prospective data on PAC use for consecutive admissions between May 2003 and December 2004.[1] The NRS applied the same inclusion and exclusion criteria for individual patients as the corresponding PAC-Man Study, which resulted in a sample of 1,052 PAC cases and 31,447 potential controls. The overall control group is not exchangeable with those who received PAC in practice (Sakr et al. 2005; Sekhon and Grieve 2012). Hence we only use information from the 1,052 patients who received PAC in routine clinical practice, and from 1,013 RCT participants.

We assume throughout that the patients who received treatment in the NRS represent the target population of interest, as these are the patients who receive PAC in routine clinical practice. Therefore, as is common, the estimand of policy interest is the population average treatment effect on the treated (PATT)—i.e. the average treatment effect of PAC on those individuals in the target population who receive PAC. Information is available on baseline prognostic covariates common to both the RCT and NRS settings, and includes those covariates anticipated to modify the effect of PAC. For a center to participate in

[1]Over this time period, 10 units recorded no PAC use and were excluded from this analysis, as were units participating in the RCT (PAC-Man Study). The RCT data used, excludes one participant for whom no endpoint data were available.

the PAC-Man Study required that local clinicians were in equipoise about the potential benefits of the intervention (Harvey et al. 2005), and the patients randomized had to meet the inclusion criteria. The net effect is that the baseline characteristics of the RCT participants differed somewhat from those who received PAC in routine clinical practice (Table 2.1). The baseline prognosis of the RCT patients was more severe, with a higher mean age, a higher proportion of patients admitted following emergency surgery and a higher proportion having mechanical ventilation. The RCT patients were less likely to be admitted to teaching hospitals than those who received PAC in the target population. For both studies the main outcome measure was hospital mortality, which was higher in the RCT, than for the PAC patients in the NRS. The studies reported similar hospital costs. The effect of PAC on costs and mortality can be incorporated into a measure of cost-effectiveness such as the incremental net monetary benefit (INB) (Willan et al. 2003; Willan and Lin 2001).[2]

This study is an example of where estimates of effectiveness and cost-effectiveness from an RCT may not be directly externally valid for a target population, but there is information from an NRS on the baseline characteristics and outcomes that can inform the estimation of population treatment effects. The next section defines the assumptions required for estimating PATT in this context.

2.2 Identifying PATT from an RCT

For simplicity we consider those circumstances where data come from a single RCT and a single NRS. It is assumed that the treatment subjects in the NRS represent those in the target population of interest. This section outlines sufficient assumptions for identification of PATT.

Let Y_{ist} represent potential outcomes for a unit i assigned to study sample s and treatment t, where $s = 1$ indicates membership of the RCT and $s = 0$ the target population, such as those in the NRS. For simplicity, we assume that in either study a unit is assigned to treatment ($t = 1$) or control ($t = 0$), and that there is no crossover. We define S_i as a sample indicator, taking on value s, and T_i as a treatment indicator taking on value t. For subjects receiving the treatment, we define W_i^T as a set of observable covariates related to the sample selection mechanism for membership in the RCT versus the target population. Similarly $W_i^{C_T}$ is a set of observable covariates related to the sample assignment for

[2]Net Monetary Benefits can be calculated by weighting each life year using a quality adjustment anchored on a scale from 0 (death) to 1 (perfect health), in order to report quality-adjusted life years (QALYs) for each treatment. Then net monetary benefits for each treatment group can be calculated by multiplying the QALY by an appropriate threshold willingness to pay for a QALY gain (e.g. the threshold recommended by NICE in England and Wales is £20,000 to £30,000 to gain a QALY), and subtracting from this the cost. Finally, the INB of the new treatment can be estimated by contrasting the mean net benefits for each alternative.

Table 2.1: Baseline characteristics and endpoints for the PAC-Man Study, and for patients in the NRS who received PAC. Numbers are N (%) unless stated otherwise

	RCT		NRS
	No PAC	PAC	PAC
	n=507	n=506	n=1052
Baseline Covariates			
Admitted for elective surgery	32 (6.3)	32(6.3)	98 (9.3)
Admitted for emergency surgery	136 (26.8)	142 (28.1)	243 (23.1)
Admitted to teaching hospital	108 (21.3)	110 (21.7)	447 (42.5)
Mean (SD) Baseline probability of death	0.55 (0.23)	0.53 (0.24)	0.52 (0.26)
Mean (SD) Age	64.8 (13.0)	64.2 (14.3)	61.9 (15.8)
Female	204 (40.2)	219 (43.3)	410 (39.0)
Mechanical Ventilation	464 (91.5)	450 (88.9)	906 (86.2)
ICU size (beds)			
5 or less	57 (11.2)	59 (11.7)	79 (7.5)
6 to 10	276 (54.4)	272 (53.8)	433 (41.2)
11 to 15	171 (33.7)	171 (33.8)	303 (28.8)
Endpoints			
Deaths in Hospital	333 (65.9)	346 (68.4)	623 (59.3)
Mean Hospital Cost (£)	19,078	18,612	19,577
SD Hospital Cost (£)	28,949	23,751	24,378

controls included in the RCT versus the target population.

The sample average treatment effect (SATE) in the RCT sample is defined as:

$$\tau_{SATE} = \mathbb{E}(Y_{i11} - Y_{i10}|S_i = 1)$$

Randomization ensures that the difference in the mean outcome for the treated versus control units in the RCT is an unbiased estimate of SATE.

Other estimands include the average treatment effect on the treated in the sample (SATT), and the average treatment effect on the controls in the sample (SATC), which in finite samples, may differ from SATE. When treatment assignment is ignorable, they are:

$$\tau_{SAT*} = \mathbb{E}(Y_{i11}|S_i = 1, T_i = t) - \mathbb{E}(Y_{i10}|S_i = 1, T_i = t),$$

where $t = 0$ for τ_{SATC} and $t = 1$ for τ_{SATT}. SATT estimates the average treatment effect conditional on the distribution of potential outcomes under treatment, and SATC

estimates the average treatment effect conditional on the distribution of potential outcomes under control. Randomization implies that the potential outcomes in the treatment and control groups are exchangeable or $(Y_{i11}, Y_{i10}) \perp\!\!\!\perp T_i = 1 | S_i = 1$, and the alternative estimands are asymptotically equivalent.[3]

The Population Average Treatment Effect (PATE) is defined as the effect of treatment in the target population, the Population Average Treatment Effect on Controls (PATC) as the treatment effect conditional on the distribution of potential outcomes under control, and the Population Average Treatment Effect on Treated (PATT) as the treatment effect conditional on the distribution of potential outcomes under treatment:

$$\begin{aligned} \tau_{PATE} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0) \\ \tau_{PATC} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0, T_i = 0) \\ \tau_{PATT} &= \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0, T_i = 1). \end{aligned} \tag{2.1}$$

Our main quantity of interest is (2.1). Because treatment in the target population is not randomly assigned, these three population estimands differ even asymptotically, and they may be difficult to estimate without bias.

The following proof outlines the conditions under which population treatment effects can be identified from RCT data. The following assumptions are necessary to derive the identifiable expression for τ_{PATT} in Theorem 1. Figure 2.1 represents the assumptions, and demonstrates the result of Theorem 1.

Assumption 1: Consistency under Parallel Studies

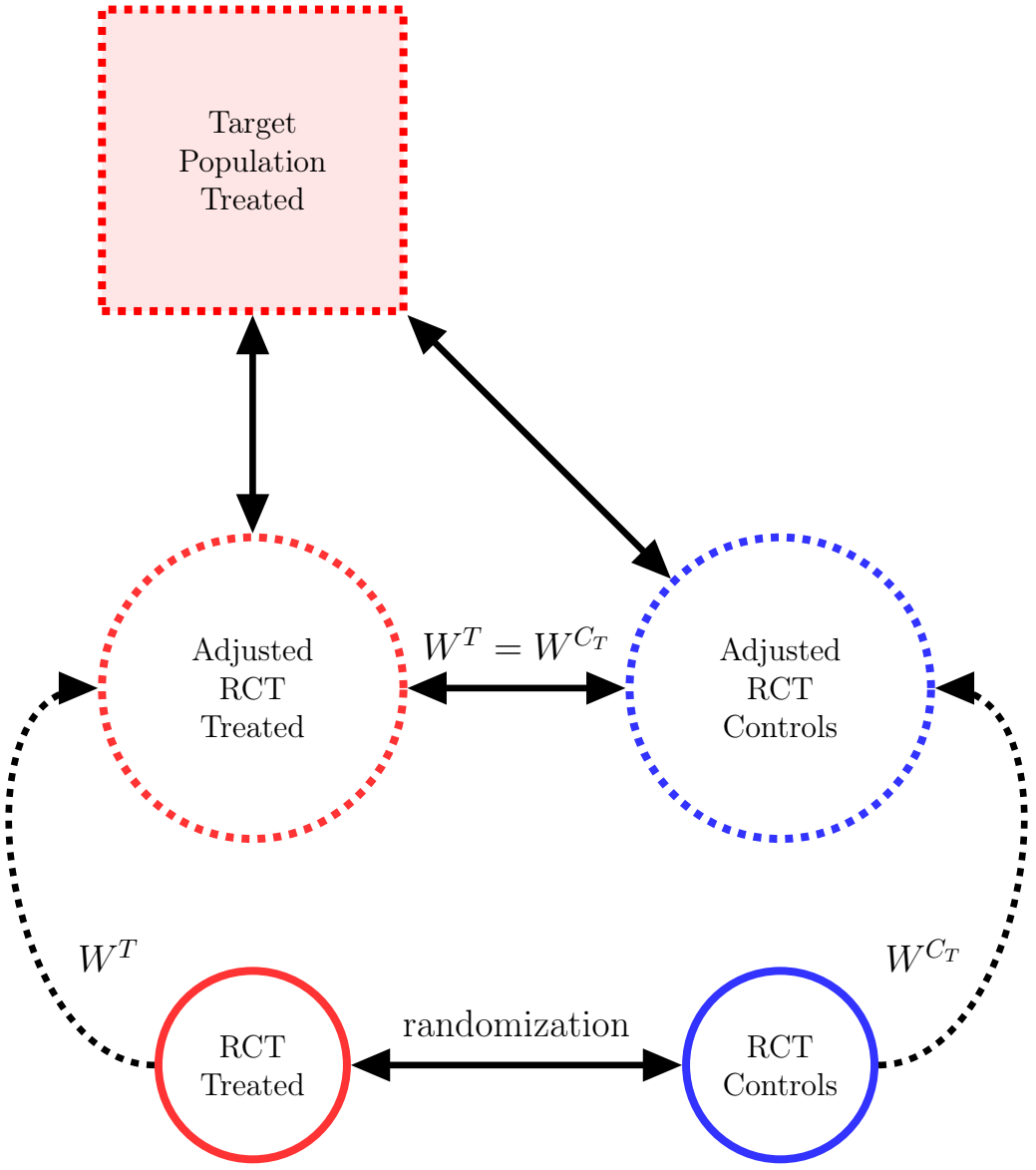
$$Y_{i01} = Y_{i11} \tag{2.2}$$

$$Y_{i00} = Y_{i10} \tag{2.3}$$

For either the treatment or control group, assumption 1 restricts an individual's potential outcomes for the RCT and the target population. Intuitively, it is assumed that if units in the target population were assigned their observed treatment randomly, then their outcome would be the same as if they were assigned that particular treatment in the RCT. This essentially ensures that any differences in the treatment between the RCT and the NRS, for example, in a clinical protocol, do not effect the outcome. Assumption 1 is similar to the assumption of consistency under the parallel experiment design in Imai, Tingley and Yamamoto (2013). Assumption 1 may be violated for example if the clinical protocol for insertion of the PAC differs between the RCT and the NRS. The pragmatic design of the PAC-Man Study helped ensure that this assumption was met. Further examples of violation of the consistency assumption are given in Cole and Frangakis (2009).

[3]The treatment effects discussed here refer to infinite populations and samples, whereas Imai, King and Stuart (2008) refers to treatment effects in infinite populations as super population effects.

Figure 2.1: Schematic of adjustment of sample effect to identify population effect. Double arrows indicate exchangeability of potential outcomes and dashed arrows indicate adjustment of the covariate distribution.



Assumption 2: Strong Ignorability of Sample Assignment for Treated

$$(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i^T, T_i = 1), \qquad 0 < Pr(S_i = 1 | W_i^T, T_i = 1) < 1.$$

Assumption 2 states that the potential outcomes for treatment are independent of sample assignment, for treated units with the same W^T. Assumption 2 implies that

$$\mathbb{E}(Y_{is1} | S_i = 0, T_i = 1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{is1} | W_i^T, S_i = 1, T_i = 1)\}, \tag{2.4}$$

for $s = 0, 1$. The expectation $\mathbb{E}_{01}\{\cdot\}$ is a weighted mean of the W_i^T specific means, $\mathbb{E}(Y_{is1} | W_i^T, S_i = 1, T_i = 1)$, with weights according to the distribution of W_i^T in the treated target population, $Pr(W_i^T | S_i = 0, T_i = 1)$. Essentially, on the right side of Equation (2.4), the characteristics of the treated units in the RCT, W_i^T, are adjusted to match those of the treatment group in the target population. Figure 2.1 illustrates this process with the single arrow from the RCT treated in the solid red circle, to the adjusted group in the dashed red circle. The adjustment can be performed with the weighting methods discussed in Section 2.4.

The right side of Equation (2.4) is the expectation in the adjusted RCT treated group, depicted as the dashed red circle in Figure 2.1. The left side of Equation (2.4) is the expectation in the treatment group in the target population, depicted as the dashed red square in Figure 2.1. Thus by Equation (2.4) the adjusted treatment group in the RCT replicates the Y_{is1} potential outcomes of the treatment group in the target population. This is depicted as the double arrow, representing equivalence between the dashed red circle and the dashed red square in Figure 2.1.

Assumption 3: Strong Ignorability of Sample Assignment for Controls

$$(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i^{C_T}, T_i = 1), \qquad 0 < Pr(S_i = 1 | W_i^{C_T}, T_i = 1) < 1.$$

Assumption 3 states that the potential outcomes for control are independent of sample assignment, for treated units with the same $W_i^{C_T}$.

Assumption 3 implies that

$$\mathbb{E}(Y_{is0} | S_i = 0, T_i = 1) = \mathbb{E}_{01}\{\mathbb{E}(Y_{is0} | W_i^{C_T}, S_i = 1, T_i = 0)\}, \tag{2.5}$$

for $s = 0, 1$, since treatment assignment is random in the RCT, i.e. $Y_{is0} \perp\!\!\!\perp T_i | (W_i^{C_T}, S_i = 1)$. Similarly to Equation (2.4), on the right side of Equation (2.5), the characteristics of the units in the control group in the RCT, $W_i^{C_T}$, are adjusted to match those of the treatment group in the target population. This process is depicted in Figure 2.1 as the single arrow from the RCT control in the solid blue circle to the adjusted group in the dashed blue circle. Again this adjustment can be performed by the various weighting methods discussed in Section 2.4.

The right side of Equation (2.5) is the expectation in the adjusted RCT control group, which is depicted as the dashed blue circle in Figure 2.1. The left side of Equation (2.5) is the expectation in the treated group in the target population, which is depicted as the dashed red square in Figure 2.1. Thus it follows by Equation (2.5) that the adjusted control group in the RCT replicates the Y_{is0} potential outcomes of the treated group in the target population. This is depicted as the double arrow representing equivalence between the dashed blue circle and the dashed red square in Figure 2.1.

Assumption 4: Stable Unit Treatment Value Assumption (SUTVA)

$$Y_{ist}^{L_i} = Y_{ist}^{L_j} \qquad \forall i \neq j,$$

where L_j is the treatment and sample assignment vector for unit j. This is a stable unit treatment value assumption (SUTVA), which states that the potential outcomes of unit i are constant regardless of the treatment or sample assignment of any other unit.

Theorem 1 follows from Assumptions 1-4, with the proof given in Appendix A.1.

Theorem 1. *Assuming consistency and SUTVA hold, if*

$$\begin{aligned}&\mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 0, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 0, T_i = 1)\} \\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 1, T_i = 1)\},\end{aligned} \tag{2.6}$$

or sample assignment for treated units is strongly ignorable given W_i^T and sample assignment for controls is strongly ignorable given $W_i^{C_T}$ then

$$\tau_{PATT} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T}, S_i = 1, T_i = 0)\},$$

where $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$ *denotes* $\mathbb{E}_{W_i^T|S_i=0,T_i=1}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$
and $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W_i^{C_T}, \ldots)\}$ *denotes* $\mathbb{E}_{W_i^{C_T}|S_i=0,T_i=1}\{\mathbb{E}(\cdot|W_i^{C_T}, \ldots)\}$.

From the expression for τ_{PATT} in Theorem 1, it is possible to identify τ_{PATT} from the adjusted RCT data alone. In Figure 2.1, the adjusted experimental controls and treated are only exchangeable if $W_i^T = W_i^{C_T}$. As Figure 2.1 makes plain, in identifying τ_{PATT}, the adjusted RCT controls are being used in place of the subset of population controls who have the same distribution as the treated units in the target population. The adjusted RCT controls are not a substitute for all population controls, since the controls and treated in the target population are not assumed to be exchangeable.

By randomization, the RCT treatment and control groups are exchangeable. Therefore adjusting both groups by the same observable characteristics will asymptotically yield exchangeable groups. This implies that if $W_i^T = W_i^{C_T}$ then the adjusted RCT treated and controls are exchangeable with each other and can replace their counterparts in the

target population. In order to gain precision, matching or stratifying between the treated and control units within the RCT according to baseline characteristics, can be undertaken prior to adjustment to the target population (Miratrix, Sekhon and Yu 2013). Hence τ_{PATT} can be estimated by reporting the treatment effect for each matched pair from the RCT, and then adjusting these unit level treatment effects according to the characteristics of the treatment group in the target population. The corresponding estimate of the SATT is given by the average of the unadjusted unit level effects from the RCT.

2.3 Placebo Tests for Checking Assumptions

Placebo tests are generally used to assess the plausibility of a model or identification strategy when the treatment effect is known, from theory or design (Sekhon 2010). This section describes placebo tests for checking the identifiability assumptions of Theorem 1, regardless of the estimation strategy subsequently chosen. From Section 2.2, if Equation (2.2) in assumption 1 and assumptions 2 and 4 hold, then the Y_{s1} potential outcomes of the adjusted RCT treated group, and the target population are exchangeable, i.e. Equation (A.1) holds. Since the potential outcomes Y_{i01} are observed in the treated group of the target population, then $\mathbb{E}(Y_{i01}|S_i = 0, T_i = 1)$ is equal to $\mathbb{E}(Y_i|S_i = 0, T_i = 1)$ and

$$\mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} = 0, \tag{2.7}$$

from Equation (A.1) in Appendix A.1. Therefore if Equation (2.2) in assumption 1, and assumptions 2 and 4 hold, the expected outcomes in the adjusted RCT treatment group and the target population will be the same. A placebo test can be used to check whether there is a difference in average outcomes between the adjusted RCT treatment group after adjustment versus the treatment group in the target population. If the placebo test detects a significant difference in the above outcomes, then either Equation (2.2) in assumption 1, assumption 2 or assumption 4 is violated.[4] If Equation (2.3) in assumption 1, and assumptions 3 and 4 hold, then the Y_{s0} potential outcomes of the adjusted RCT treated group and the target population are exchangeable, i.e. Equation (A.2) in Appendix A.1 holds. However, since Y_{i00} is not observed in those treated in the target population, then $\mathbb{E}(Y_{i00}|S_i = 0, T_i = 1)$ is not necessarily equal to $\mathbb{E}(Y_i|S_i = 0, T_i = 0)$. Therefore the mean outcome in the adjusted RCT control group is not necessarily the same as the mean outcome in the target treated population. This implies that a placebo test cannot be used to check whether Equation (2.3) in assumption 1, assumption 3 or assumption 4 fails.

Placebo tests can therefore be used to validate the underlying identifying assumptions for estimating population-level causal estimates. They may highlight the failure of several underlying assumptions but cannot delineate the bias from the failure of each individual

[4]If assumption 1 is violated and there is a constant difference between the potential outcomes in the target population and the RCT, then the PATT can still be identified by Theorem 1. See Section 2.6.1.

assumption. Also, the tests cannot exclude the possibility that each assumption is violated but the ensuing biases cancel one another out. Traditional, placebo tests have a null hypothesis, that there is no difference in the average outcome between groups, and the null hypothesis is rejected if the test statistic is significant. If the null hypothesis is not rejected then a standard conclusion is that there is evidence to support the identification strategy. However, the failure to reject the null hypothesis may be because of insufficient power to detect a true difference between the groups, particularly if treatment effects by subgroup are of interest, or if there are endpoints, such as cost, that have a high variance. CEA typically have both these features.

To address this concern, Hartman and Hidalgo (2011) introduce equivalence based placebo tests, whose null hypothesis is that "the data are *inconsistent* with a valid research design." In this context, the null hypothesis can be stated as: the adjusted endpoints for the treatment group in the RCT, are not equivalent to the treatment group in the target population. This null hypothesis of non-equivalence is only rejected if there is sufficient power.[5] Therefore a low p-value would indicate the two groups are statistically equivalent, whereas high p-values in traditional tests offer support for the identification strategy. The advantage of the proposed test is that it only supports the identification strategy when the test reports that the two groups are equivalent, *and* when the test has sufficient power. Specifying an alternative null hypothesis has implications for the test statistic and, just as in a sample size calculation, requires that the threshold for a meaningful difference in outcomes is pre-defined. Appendix A.3 and Hartman and Hidalgo (2011) give further details.

2.4 Estimating PATT

There are a growing number of estimation strategies for predicting population-level treatment effects from RCT data, which can be assigned into two broad classes. One class of strategies estimates the response surface using the RCT data and extrapolates this response surface to the target population, and includes methods such as Bayesian Additive Regression Trees (BART) (Chipman, George and McCulloch 2010; Green and Kern 2012), Classification and Regression Trees (CART) (Breiman 2001; Liaw and Wiener 2002; Stuart et al. 2011), and linear regression. The other prominent approach is to use weighting methods, such as Inverse Propensity Score Weighting (IPSW) (Stuart et al. 2011) and Maximum Entropy (MaxEnt) weighting (Jaynes 1957; Kullback 1997), which rely on ancillary information, for example from a NRS, to reweight the RCT data. The result in Theorem 1 is agnostic to the estimation strategy; the adjustment of the RCT data by W^T and W_T^C can either use predicted values from a model of the response surface, or

[5]This alleviates the issues of confounding the notion of statistical equivalence with a tests relationship to sample size discussed in Imai, King and Stuart (2008).

weights from the second class of estimators. Either way, in order to identify the population estimand of interest, the estimation strategy must pass the proposed placebo tests.

While Theorem 1 does not require a specific estimation strategy, we do provide a new research design that employs a weighting method. Our proposed strategy firstly matches treated and control units within the RCT to create matched pairs or strata (Diamond and Sekhon 2013; Sekhon 2011), from which we estimate the SATT overall and by subgroup. We then reweight the matched pairs according to the characteristics of the target population to report PATT, both overall and for pre-specified subgroups.

2.4.1 Matching Treated and Control Units within the RCT

We create matched pairs within the RCT data, by matching controls to treated units within the RCT using Genetic Matching (GenMatch) to maximise the balance between the randomized groups (Diamond and Sekhon 2013; Sekhon 2011). We recommend including in the matching algorithm, those covariates anticipated to influence not only the endpoints, but also the selection of patients into the RCT. Covariates related to the selection into the RCT are part of the conditioning sets W^T and W_T^C, and therefore care should be taken to ensure these covariates are balanced.

2.4.2 Weighting Methods

Stuart et al. (2011) consider IPSW, full matching and sub-classification for reweighting RCT estimates to the target population, and find that IPSW performs relatively well. Hence, we firstly consider IPSW, where in this context the propensity score estimates the predicted probability of a unit being in the RCT, conditional on baseline characteristics observed in the RCT and the NRS. IPSW then gives each individual in both the RCT and NRS settings a weight, calculated as the inverse of the probability of being in the RCT according to baseline characteristics. In general IPSW methods can be particularly sensitive to misspecification of the propensity score; the weights can be extreme, leading to unstable results (Porter et al. 2011; Kang and Schafer 2007; Kish 1992).

We consider an alternative approach to reweighting: MaxEnt. This approach goes back to at least Jaynes (1957); Kullback (1997); Ireland and Kullback (1968), and it has much in common with method of moments estimators (e.g. Hansen 1982; Hellerstein and Imbens 1999). In brief, this approach does not assume the propensity score is correctly specified, nor does it make additional assumptions about the distribution of weights. Under MaxEnt the cell weights, marginal distributions, or other population moments for the conditioning covariates, W^T, are used as constraints. MaxEnt ensures that the weights chosen for the matched pairs sum to one, but simultaneously satisfy the MaxEnt constraints given by the population characteristics. See Appendix A.4 for more details.

For estimating PATT, the requisite placebo tests can be conducted by comparing the average of the observed outcomes for the treated in the target population, to the weighted

outcomes for the treated in the RCT. Provided that the placebo test has sufficient power, failure to find equivalence between the above treated groups indicates that at least one assumption underlying Theorem 1 is not met, and that there is bias in the estimated PATT.

2.4.3 Response Surface Models

Response surface models can estimate the covariate endpoint relationships in the RCT, and use these estimates to predict population treatment effects in the target population. A common concern is that such estimation methods use the coefficients estimated from the RCT data in predicting population estimates for the observed covariate distribution in the population. For example, in the case of OLS regression, the response surface can be estimated from the RCT data, the βs held fixed and the population treatment effects predicted from the covariate distribution of the NRS treated. This approach assumes the response surfaces in the RCT and the target population are the same. This assumption may lead to efficiency gains relative to weighting approaches, especially if some, but not all, covariates included in the adjustment are strongly predictive of potential outcomes. The placebo tests proposed can be used with the response surface approach by comparing average predicted values from the model estimated from the NRS treated covariate distribution, to the average of the observed outcomes for the NRS treatment group. Again given sufficient power, a failure to find equivalence between the above outcomes indicates a failure of at least one assumption underlying Theorem 1 and bias in the estimated population effects.

2.5 Empirical Example: PAC

We illustrate our new strategy for extrapolating from an RCT to a target population using the PAC example, where a major concern is that baseline covariates anticipated to modify the treatment effect, differ between the RCT and those who receive PAC in the target population Table 2.1. Here, we estimate PATT overall and for pre-specified subgroups; patients' surgical status (elective, emergency, non-surgical) and type of admission hospital (teaching or not).

2.5.1 Matching and Weighting in the PAC Example

We used GenMatch to create matched pairs within the RCT data, by matching a control unit to each treated unit. The matching algorithm included those covariates anticipated to influence the selection of patients into the RCT and the endpoints (See Appendix A.6 Table A.1). The GenMatch loss function was specified to require that balance, according to t-tests and KS tests, was not made worse on those covariates

anticipated to be of high prognostic importance after matching. GenMatch matched 1-1 with replacement using a population size of 5,000. Matching was repeated within each subgroup to report SATT at the subgroup level.[6] Variance estimates were calculated conditional on the matched data (Imai, King and Stuart 2008). The matching identified a control for each treated observation, resulting in 507 matched pairs for the overall estimate. Each baseline covariate was well balanced after matching according to both t-tests and KS tests. The SATT results, both overall and at subgroup level, were similar to the SATE estimates from the RCT.

We use two weighting methods to derive weights for the matched pairs. These weights adjusted the distribution of observable baseline covariates in the matched RCT data to the distribution of the PAC patients in the NRS. We firstly estimated a propensity score to predict participation in the RCT, according to the baseline covariates listed in Appendix A.6, Table A.1. We recalculated the IPSW weights separately for each subgroup to enforce interactions of variables with the subgroup classifications. In recognition of the concern that the propensity score may be misspecified we used a machine learning algorithm for classification, random forests (Breiman 2001), implemented in the `randomForest` package with the default parameters (Liaw and Wiener 2002; Stuart et al. 2011).

We also used MaxEnt, and constructed the weights for the covariates listed in Appendix A.5, Table A.2 using marginal interacted covariate distributions from the population. The consistency constraints were taken as the covariate means from the PAC patients in the NRS. The weights were selected to satisfy these constraints while maximizing the entropy measure. As Figure 2.2 shows, population moments used in constructing the weights matched exactly.

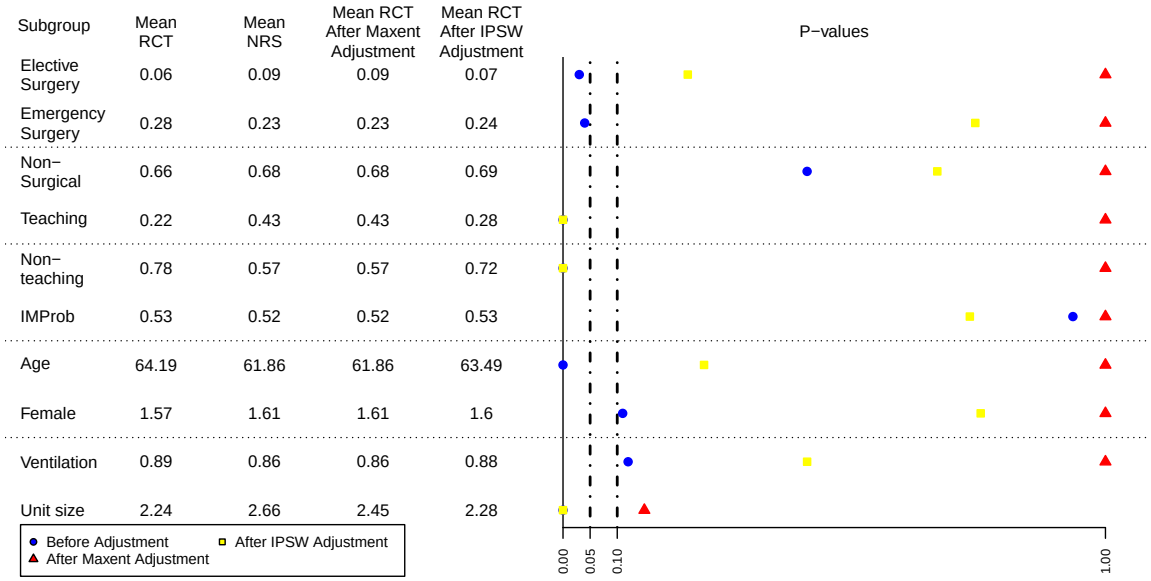
We applied both IPSW and MaxEnt to reweight the individual matched pairs from the RCT for the observable characteristics of the PAC patients in the NRS. To recognize the uncertainty in the estimation of the weights, standard errors for both SATT and PATT were estimated using subsampling (Politis and Romano 1994). Abadie and Imbens (2006) show that the bootstrap is not valid for estimating the standard error of a matching estimator, but suggest that subsampling (Politis and Romano 1994) is a valid estimating strategy. We used the algorithm described in Bickel and Sakov (2008) to select the subsampling size, m, and found that the optimal subsample was the sample size, n, in the RCT. We used 1000 bootstrap replicates.[7]

IPSW and MaxEnt provided weights that were reasonably stable (see Appendix A.7),

[6]The aggregated subgroup estimates may not be equivalent to the overall estimate because different matches are used for the overall and subgroup estimates.

[7]One of the arguments against using the bootstrap for matching estimators is that individual matches can be no better than in the full sample, and typically are worse. However, by the nature of the RCT design, where the true propensity of each individual to be assigned a particular treatment is random, there are many potential matches for each unit. Therefore, even within each iteration of the bootstrap, the probability of a close match for each unit is high. It may not be surprising then that the Bickel and Sakov (2008) algorithm selects $m = n$.

Figure 2.2: Balance on observable characteristics between the PAC patients in the NRS and the RCT, before and after adjustment of the RCT data with MaxEnt and IPSW



with mean weights of 1, and maxima of 4 (IPSW) and 8 (MaxEnt); no individual stratum was given an extreme weight. Figure 2.2 reports the covariate balance for the PAC patients in the NRS versus the RCT, after reweighting with either approach. While IPSW achieved good balance for some covariates, other prognostic factors such as attendance at teaching hospital, and ICU size, were still imbalanced. By contrast, MaxEnt balanced all covariate means between the PAC patients in the RCT and the NRS.

2.5.2 Results of the Placebo Tests

For each approach, we reported placebo tests that tested the underlying assumptions for estimating PATT by comparing the mean endpoints for the PAC patients in the NRS with the adjusted means for the PAC patients in the RCT. The results are reported in Figures 2.3 and 2.4, for all three endpoints (survival rates (black), cost (red) and INB (blue)). We present the equivalence based placebo test p-values (represented by squares) for the overall estimate and each subgroup, and allow for multiple comparisons, by presenting p-values with a false discovery rate correction using the Benjamini-Hochberg method (represented by triangles) (Benjamini and Hochberg 1995). After IPSW, there are still large differences in hospital survival rates between the PAC patients in the adjusted RCT treated group and the NRS. For the survival rate and net benefit endpoints, IPSW fails almost all the placebo tests. The placebo tests following MaxEnt report small survival

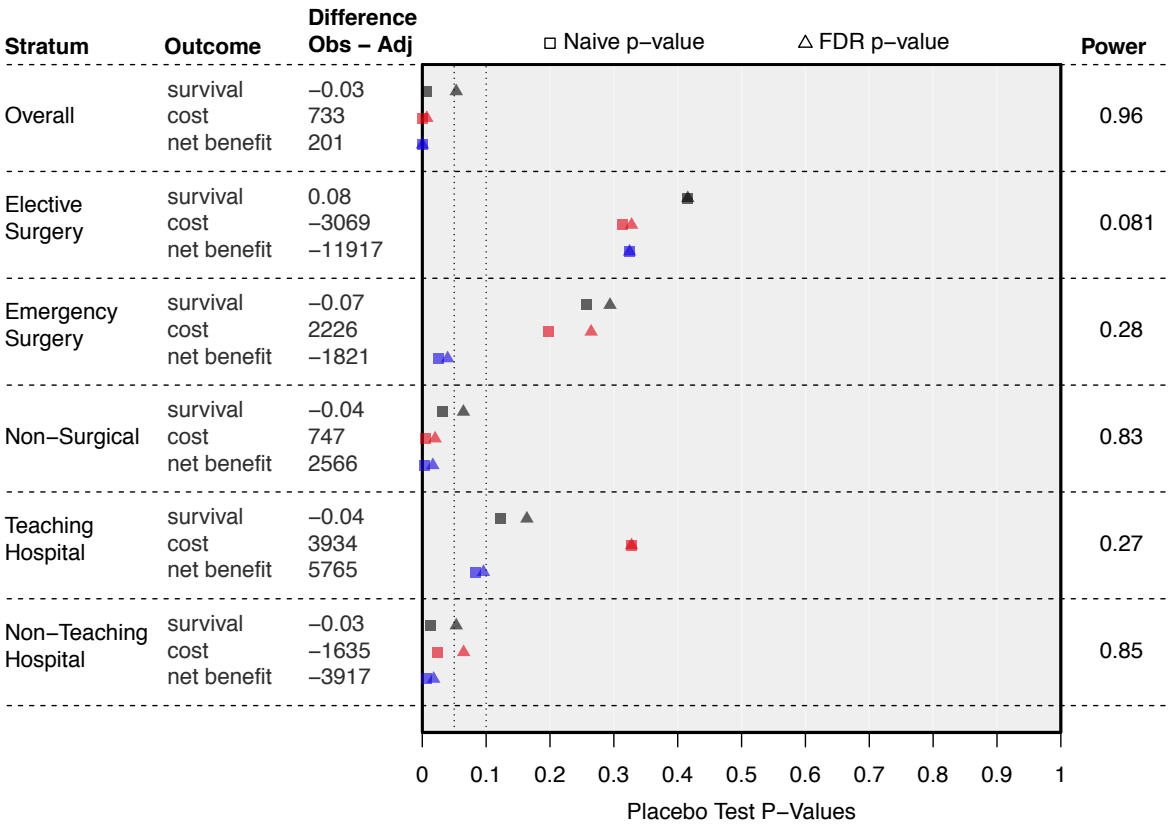


Figure 2.3: Maxent Placebo Tests
The column labelled "Difference" presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the RCT after reweighting. The column labelled "Power" presents the power of the equivalence t-test for each stratum.

differences between the PAC patients in the RCT (reweighted) versus the NRS settings, overall and for most subgroups. For the overall sample all the placebo tests are passed; mean differences between the settings are small, and there is sufficient power to assess whether such differences are statistically significant. For some subgroups (teaching hospitals, elective and emergency surgery) there is insufficient power to detect differences between the settings and the placebo test fails; for other subgroups (non-surgical and non-teaching hospital), the mean differences are small after weighting, and as there is also sufficient power, the placebo tests are passed.

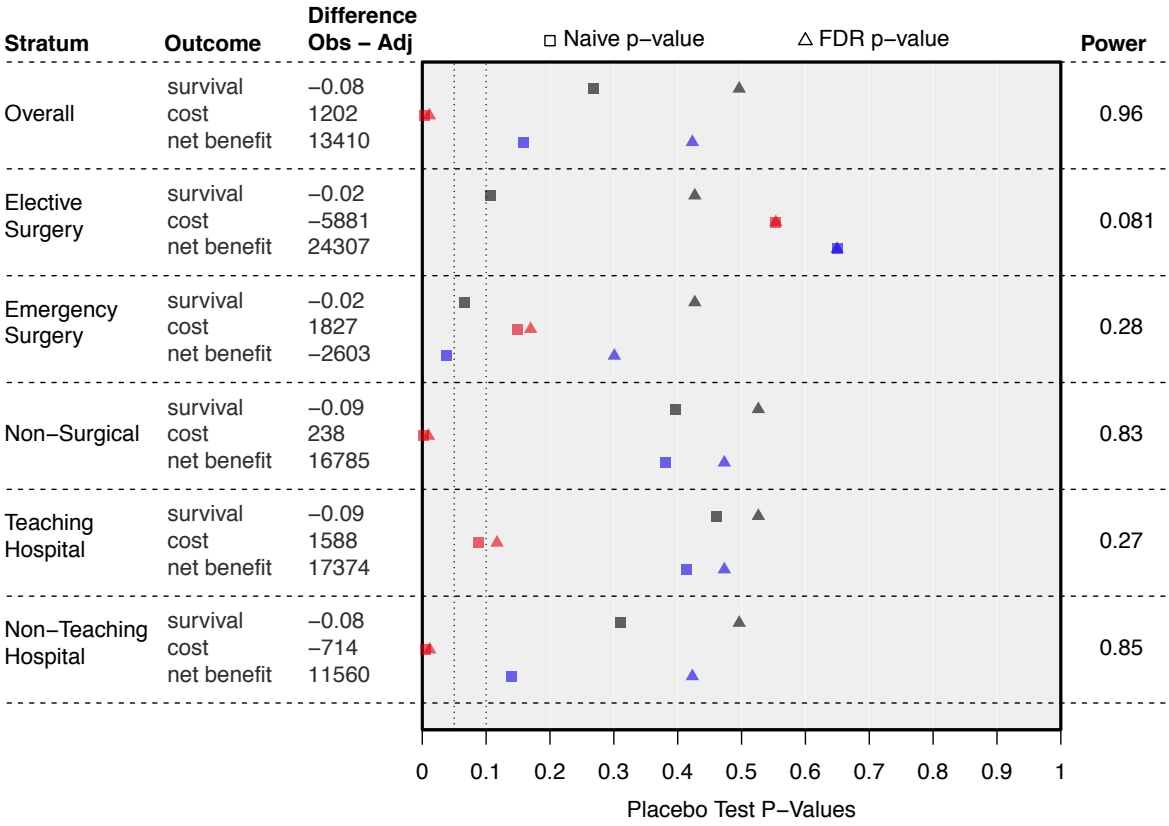
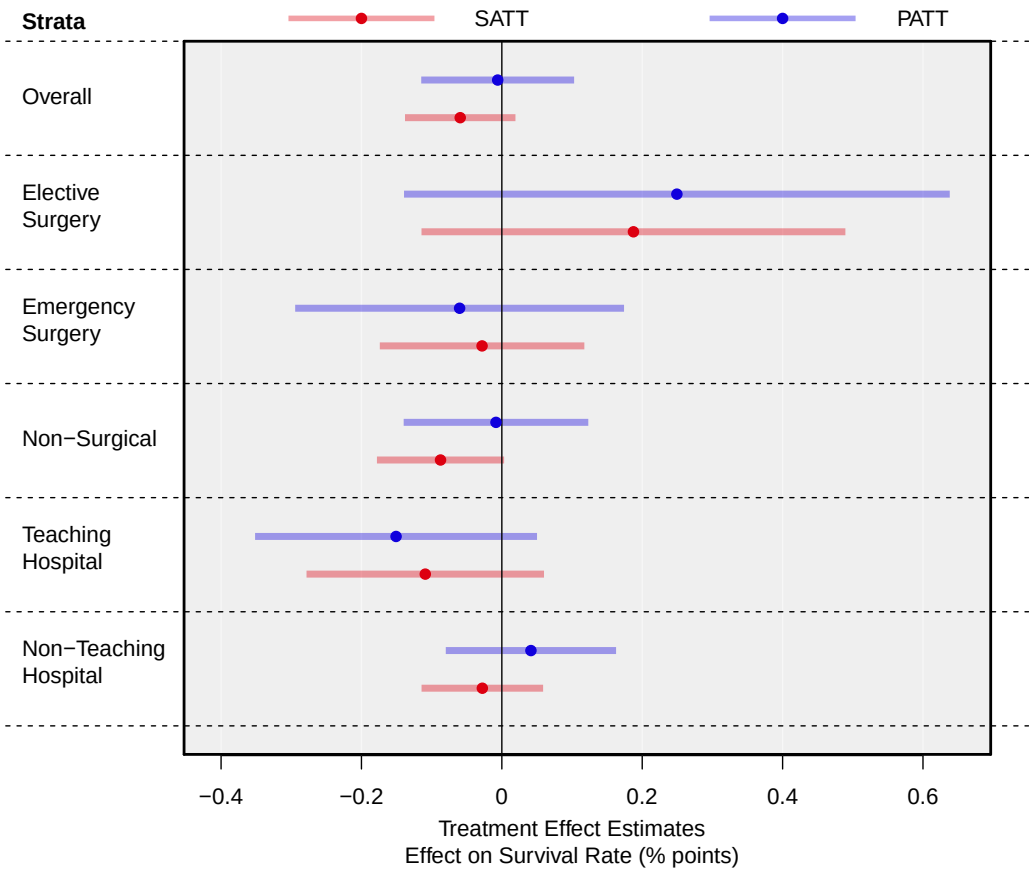


Figure 2.4: IPSW Placebo Tests
The column labelled "Difference" presents the difference between the observed outcomes for the PAC group in the NRS and the PAC group in the RCT after reweighting. The column labelled "Power" presents the power of the equivalence t-test for each stratum.

2.5.3 Population Estimates in the PAC Example

We report SATT estimated from the matched RCT data, and PATT after using the MaxEnt weights to adjust the SATT estimates. The 95% confidence intervals (CIs) reported use the standard errors obtained by subsampling (Figures 2.5–2.7). The IPSW did not pass the overall placebo tests, these results are not presented. For the overall group, the PATT and SATT estimates are similar for each endpoint. For the non-teaching hospital subgroup, which passed the placebo tests, the positive point estimate for PATT suggested a somewhat more beneficial effect for PAC on survival, than the corresponding SATT. The accompanying cost-effectiveness estimates reported a negative INB for the SATT, but for the PATT, the INB was positive. This finding suggests that for non-

Figure 2.5: Population Treatment Effects on Hospital Survival Rates



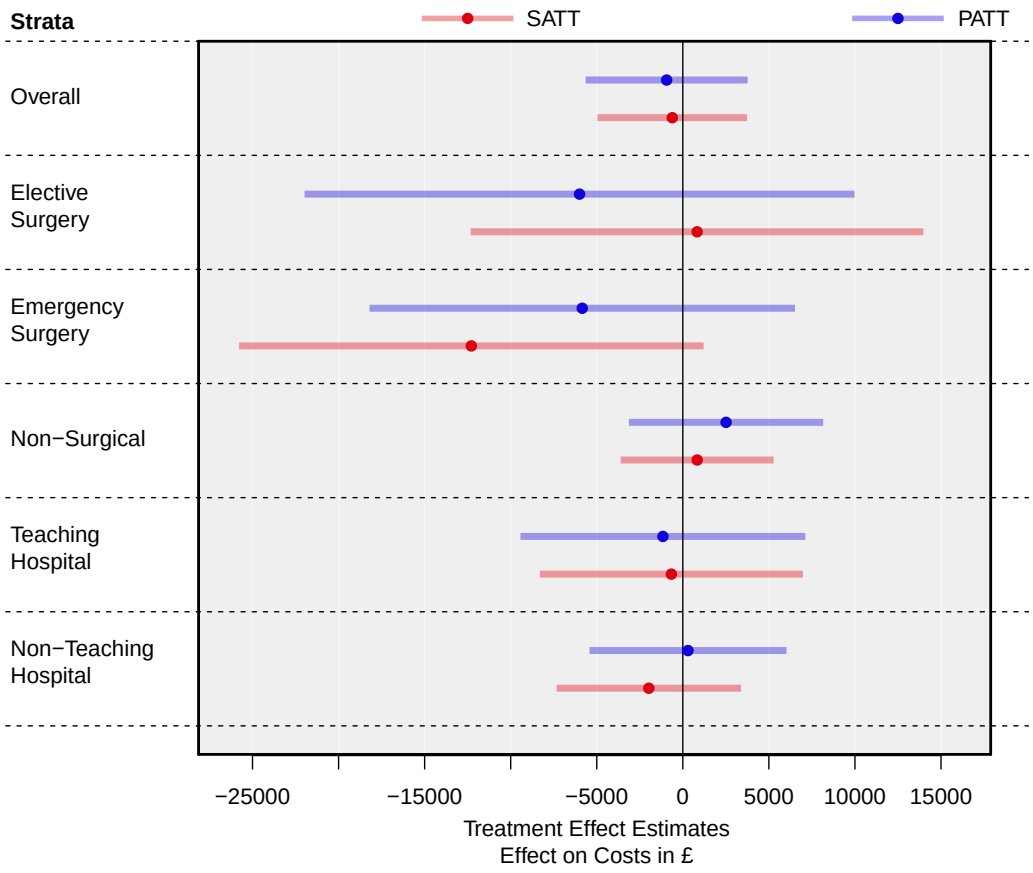
teaching hospitals in the target population, PAC was relatively cost-effective. However, the CIs for each estimate overlapped zero, and in general, the CIs for the PATT estimates were wider than those for the corresponding SATT estimates.

2.6 Alternative Designs Identified under Theorem 1

2.6.1 Using the Population Treated

A main assumption in the derivation of Theorem 1 is that selection on observables assumptions are sufficient to recognize the selection of the RCT participants. However if a placebo test rejects the null hypothesis given by Equation (2.7) then Equation (2.2) in assumption 1, assumption 2 or assumption 4 is violated. In such a case the results of

Figure 2.6: Population Treatment Effects on Costs (GBP £)



Theorem 1 are no longer valid. However, if assumption 4 is not violated and if assumption 3 and Equation (2.3) in assumption 1 are valid, PATT can still be identified by

$$\tau_{PATT} = \mathbb{E}(Y_i|S_i = 0, T_i = 1) - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T}, S_i = 1, T_i = 0)\}, \tag{2.8}$$

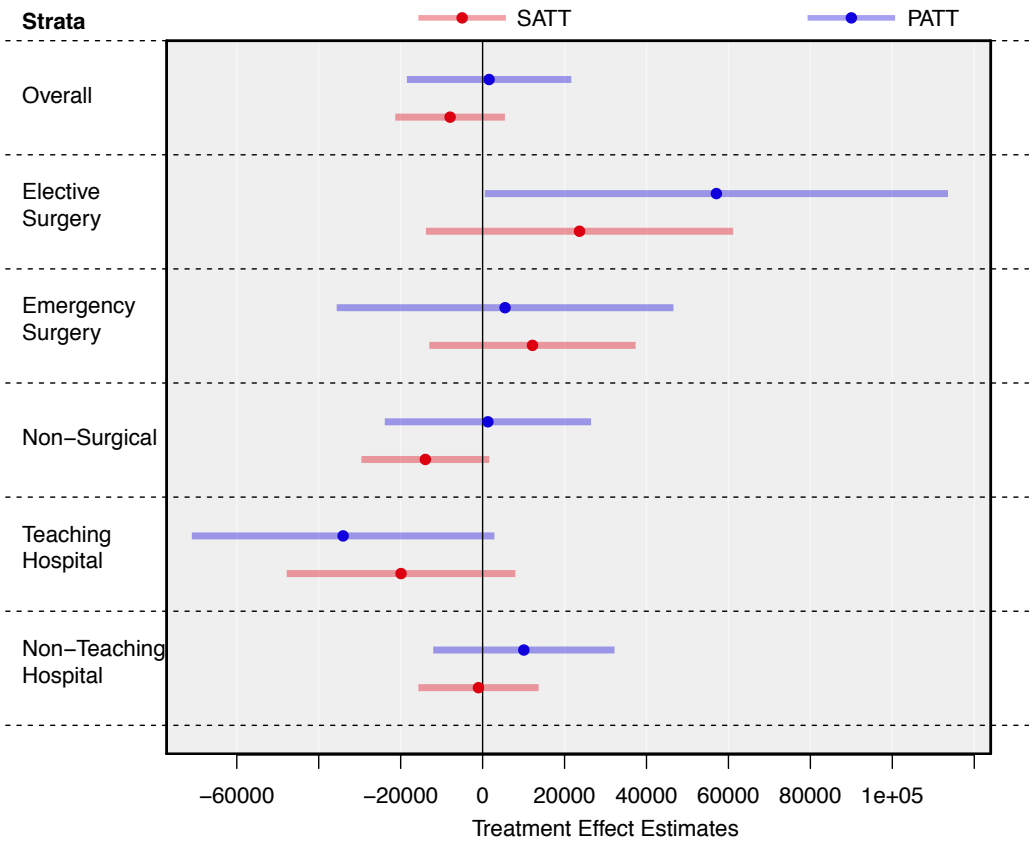
from (A.2) in Appendix A.1. This estimator makes direct use of the population treated, and it is valid if there is a constant difference in the potential outcomes between the population and the RCT. One can see this by rewriting (2.8) as:

$$\tau_{PATT_{DID}} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1) - \mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 0)\} \tag{2.9}$$

$$-[\mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}(Y_i|S_i = 0, T_i = 1)], \tag{2.10}$$

assuming $W_i^T = W_i^{C_T}$. The first difference (2.9) is the adjusted experimental estimand and is intuitively a measure of the adjusted average effect. The second difference (2.10) is

Figure 2.7: Population Treatment Effects on Incremental Net Monetary Benefits



Incremental Net Monetary Benefits calculated by valuing each quality-adjusted life year (QALY) gain at a £20,000 per QALY threshold

defined as the difference between the outcomes of the treatment groups in the RCT and the NRS.

The major concern with this estimator is that there is no longer a placebo test available to check if the identifying assumptions hold. Hence, while the main approach proposed makes a somewhat stronger identifying assumption, a key advantage is that this design allows the implications of the assumptions to be tested.

Appendix A.2 discusses alternative population estimands of interest to policy makers.

2.7 Discussion

This chapter derives conditions under which treatment effects can be identified from RCTs for the target population of policy relevance. We provide placebo tests that follow

directly from the conceptual framework and can assess whether the requisite assumptions are satisfied. These placebo tests contrast the reweighted RCTs endpoints, with those of the target population represented, for example by a NRS. The general framework is illustrated with estimation strategies that reweight the matched RCT data, according to IPSW or maximum entropy reweighting, but could also exploit alternative estimation strategies including double-robust estimators (Green and Kern 2012). Whichever estimation strategy is taken, the placebo tests presented can assess whether or not the assumptions required for identification are met. The chapter builds on previous approaches for considering external validity (Stuart et al. 2011; Green and Kern 2012; Imai, King and Stuart 2008; Kline and Tamer 2011), by defining the assumptions required for estimating population treatment effects, and providing a general strategy for assessing their plausibility.

We illustrate the framework for estimating population treatment effects in a context where the treatment, in this case a medical device, has been defused to the target population without adequate evaluation, and the parameter of interest is the PATT. The framework can be applied to other situations, for example in evaluations of new pharmaceuticals, where the only individuals who receive the treatment are those included in the phase III RCT. Here, the target population is best defined by those who would meet the criteria for treatment in routine practice but receive usual care, and the estimand of interest is the PATC. In these settings, the proposed framework, can assess the identification strategies with placebo tests that compare the weighted outcomes from the RCT control group, versus those receiving usual care in the target population. Failure of these placebo tests would indicate that either participants' unobserved characteristics, or "usual care," differs between the RCT and target population settings. Hence the underlying assumptions are violated leading to biased estimates of the effectiveness and cost-effectiveness of treatment in the target population.

In our illustration the sample average treatment effects were estimated from one RCT, and individual-level data from a single NRS were taken to represent the target population. More generally, the framework can assess the underlying assumptions with reweighting individual participant data from several RCTs to a target population. An advantage of having individual-level data for the target population is that the individual stratum from the RCT can be reweighted, not just according to the population means, but also to other moments of the distribution, such as the variance. While the proposed framework still applies to situations where only aggregated information is available for the target population, this implies further constraints to the reweighting approach. Firstly, the RCT data can only be reweighted according to aggregate statistics for the covariates reported. Secondly, while MaxEnt reweighting can be undertaken with summary data, IPSW requires individual-level data from the NRS. More generally, population treatment effects can be required for several target populations, for example according to geographical location or time period. The proposed framework can then assess whether or not the requisite

assumptions are met for each setting of policy interest.

This framework complements the move to RCTs with pragmatic designs, which requires that the participants and treatments included, represent those in the target population (Tunis, Stryer and Clancy 2003). As the case study illustrates, pragmatic RCTs can help ensure that the treatments delivered in RCTs are similar to routine practice, and that there is reasonable overlap in baseline characteristics between the settings. The PAC-Man RCT had broad inclusion criteria, many prognostic baseline covariates common to the RCT and NRS settings, good overlap in the distribution of the baseline covariates between the settings, and the RCT used the same treatment and usual care protocols as for routine practice. These design features were an important reason why the placebo test findings following MaxEnt reweighting, supported the underlying assumptions required for estimating the PATT, overall and for some subgroups. Where the placebo test results suggested that the underlying assumptions were violated, for some subgroups, for example teaching hospitals, the divergent point estimates suggested this was because of unobserved differences between the settings. For other subgroups, for example patients having elective surgery, the point estimates were similar, but the small sample size meant the placebo test had insufficient power. More generally RCTs apply restrictive exclusion criteria, or treat according to more rigid treatment protocols than would be applied in routine practice (Rothwell 2005). Here it would be anticipated that the assumptions pertaining to both the consistency of treatment, and unobserved differences in the populations would be violated; the placebo test would indicate the likely bias in the treatment effects estimated for the target population.

The proposed approach encourages future studies to fully recognize the uncertainty in estimating population treatment effects, which comprises not just the random error in the sample estimates, and the systematic differences between the RCT and the target population (Greenland 2005), but also the uncertainty in estimating the requisite weights. A recommended approach is to undertake subsampling (Politis and Romano 1994), and use the algorithm described in Bickel and Sakov (2008) to select the size of the subsample. It is anticipated that, as in the PAC example, when the treatment effects are estimated for the population rather than the sample, there will be increased uncertainty. An advantage of the proposed approach is that this additional uncertainty in estimating population treatment effects is made explicit. Future studies should anticipate the additional uncertainty at the design stage when developing the sampling strategy, for example in estimating the sample size required for estimating treatment effects for the population of interest.

The proposed framework warrants careful consideration in other settings for estimating population treatment effects, and the chapter raises the following areas for further research. First, the framework could be considered further in assessing the underlying assumptions, when estimating other parameters of interest (e.g. PATC, PATE), for alternative target populations, and with NRS data available at different levels of aggregation. Second, further research is required to consider how the proposed framework could be

useful in evidence synthesis and meta-analyses of individual participant data from several RCTs. Here, rather than weighting the data from each setting according to their relative sample size or variance, weights should partly reflect each study's relative relevance, according for example to elicited opinion (Turner et al. 2009). Our approach can be extended to recognize systematic differences in the populations and the treatments in each study versus those in the target population. Third, the current chapter illustrates an approach for reweighting evidence from head-to-head RCTs, but the framework can extend to those settings where indirect or mixed treatment comparisons are required, and there is a common comparator, for example usual care. Here the placebo tests can assess whether the underlying assumptions for estimating population treatment effects are met, by contrasting the reweighted endpoints from each RCT with those of the target population, for the common comparator (e.g. usual care). Fourth, there will be settings where the requisite assumptions for estimating population treatment effects are not met. Further research is required to examine how best to reduce the inevitable bias in the resultant estimates of population treatment effects.

Chapter 3

What's the Alternative?: An Equivalence Approach to Balance and Placebo Tests

Recent debates over the difficulties of causal inference in the social sciences have spurred a growing literature on how to judge the quality of observational research designs (Austin 2008; Dunning 2010*a*; Keele 2010). An important consequence of this debate is a growing expectation that scholars defend the merits of their research designs with tests of empirically refutable implications of the assumptions justifying their inferences (Sekhon 2010, p. 503), which we call "tests of design". The emerging norm in the literature is that only if the empirical implications of the scholar's causal identification assumptions are borne out in the data, can the causal effect estimates be treated as credible. In the "design-based inference" literature, the procedures used to check the assumptions justifying a design are just as important as those used to estimate causal effects (Rubin 2008).

The most common tests of design are balance and placebo tests. Balance tests, the most frequently used test of design, check if the distributions of pretreatment variables are approximately the same among treatment and control units. A common scenario where balance is tested in order to justify a design occurs in the analysis of natural experiments. The condition of covariate balance is implied by the "as if" randomization assumption invoked to justify analyzing the data as if it had come from a randomized experiment (Dunning 2010*a*). A related test is a placebo test, which examines the effect of the intervention on a post-treatment variable known to be unaffected by the cause of interest[1] (Rosenbaum 2002, p. 214). If the intervention were to show a correlation with the placebo outcome, then the validity of the research design is called into question. A common feature of these two standard tests is that it is incumbent upon the researcher to demonstrate that the difference between treated and control units on the pre-treatment covariate or the placebo outcome are substantively small and thus not indicative of a

[1]The definition of a placebo test is less well settled in the literature than the definition of a balance test. Some scholars appear to use balance and placebo tests interchangeably. We use Rosenbaum (2002)'s definition of a placebo outcome, which is a post-treatment outcome for which the effect is known, either by design or substantive theory. In almost all cases, this known effect is 0. Another type of placebo test, which we do not consider, is the use of an alternate treatment, related to the treatment of interest, but whose effect on the outcome is known. A classic example of such a placebo test is Di Nardo and Pischke (1997).

flawed design.

In this chapter, we argue that tests of design such as balance and placebo tests should be structured so that the burden of proof lies with researchers to positively demonstrate that the data is consistent with their identification assumptions. Current practice, however, is the opposite. The standard approach to validity checks is to use statistical tests which adopt the null hypothesis of no differences in pre-treatment characteristics and placebo outcomes; only rejecting this hypothesis if sufficient data exists to demonstrate otherwise. The implicit decision rule embedded in current practice is that treatment effects should only be estimated if one fails to reject the null hypothesis of no difference (at conventional levels of statistical significance) in tests of design. This approach could be loosely described as equating "a non-significant difference with significant homogeneity" (Wellek 2010, p. 3).

Instead of beginning with the assumption that the data is consistent with a valid design, we argue that one should begin with the initial hypothesis that the data is *inconsistent* with a valid research design. Only with sufficient data should one reject the null hypothesis of imbalance in pre-treatment covariates and post-treatment placebo outcomes and declare that the design is valid. The conceptual distinction between beginning with a null hypothesis of no difference, as is standard in current practice, versus beginning with a null hypothesis of a difference, as we advocate, may seem small, but the practical implications are substantial. Because one typically fixes ex-ante the probability of rejecting the null hypothesis when it is true (type I error), the consequences of increasing the sample size of a study are very different depending on the empirical content of the null hypothesis. The usual approach, as forcefully discussed by Imai, King and Stuart (2008, p. 494), has the unfortunate consequence that the fewer the units in the sample, holding all other factors constant, the more likely the validity tests will be passed. Reversing the null hypothesis, as we advocate, ensures that collecting more data does not reduce the likelihood of passing a validity test. Under our approach, when the research design is valid, additional data only improves one's ability to pass tests of design such as balance and placebo tests.

To implement our validity tests, we rely on the large literature in biostatistics on equivalence testing (Wellek 2010; Westlake 1976). The statistics of equivalence testing largely arose out of the problem in pharmokinetics of statistically demonstrating that two drugs had the same effect, an issue that arises frequently when considering regulatory approval of a generic versus brand-name drug. Drawing upon this line of research, we develop a general procedure for tests of design. First, we present the basic logic of equivalence testing and contrast it with the standard tests of difference. We show that equivalence tests are not susceptible to common critiques of balance tests. Second, we present the details of equivalence versions of frequently used parametric and nonparametric difference tests. Equivalence tests require the ex-ante selection of an "equivalence interval": the interval wherein any difference between treatment and control units is declared substan-

tively unimportant. We pay particular attention to the selection of an equivalence interval as it is a key distinction between equivalence and conventional hypothesis testing. We also argue for the importance of inverting equivalence tests to produce something akin to an equivalence confidence interval. Inversion of an equivalence test produces the smallest range, given the observed difference, in which a test would reject the null hypothesis of dissimilarity, at a pre-specified level. This inversion demonstrates the smallest difference supported by the data and can be an important metric in assessing the quality of a design.

To show that the use of equivalence tests over difference tests can alter the outcomes of design tests in practice, we apply our approach to Brady and McNulty (2011)'s natural experimental study of the costs of voting. We show that our approach improves over standard tests, by allowing the researcher to incorporate substantive considerations over what is "good" balance directly into the formulation of the null hypothesis. We then apply equivalence tests to ten prominent natural experiments in the political science and economics literature. In four of the ten cases, an equivalence test fails to reject the null hypothesis of difference: the opposite conclusion of standard tests.

3.1 Problems with Traditional Hypothesis Tests

To motivate the tests of design proposed in this essay, we consider two recent studies of voter turnout in political science. The first is a small randomized experiment where a newspaper advertisement encouraging turnout was published in four treatment cities, with four cities serving as controls, as discussed in Bowers (2011). The second example is Brady and McNulty (2011), a large natural experiment in Los Angeles in which millions of voters were assigned to new polling stations due to precinct consolidation, which resulted in many voters having to travel farther to vote, thereby increasing the "costs of voting". In both studies, the authors are quite careful in assessing their identification assumptions by conducting tests of design, namely tests of pre-treatment covariate balance.

In reporting balance across treatment and control cities in the newspaper study, Bowers (p. 11) finds a particularly large imbalance on share of the population that is black: a mean difference of -14.4 percentage points. The imbalance in this covariate may indicate that the control cities are poor counterfactuals for the treatment units, but Bowers argues against that conclusion by noting that the null hypothesis of no difference is not rejected at conventional levels with a p-value of 0.2. Despite passing the traditional balance test, a large observed difference is nevertheless troublesome in that it suggest the potential for biased effect estimates. In this case, the small size of the sample ensures that only extremely high levels of imbalance would lead to a rejection of the null of no difference.

A contrasting scenario is found in Brady and McNulty (2011), where after using a matching algorithm to match voters on a few important covariates, the authors report balance statistics on variables not used in the matching algorithm and report the mean differences at the precinct level. However, the authors do not provide t-test p-values

associated with those mean differences, providing the explanation,

> "For the rest of the results, it does not make a great deal of sense to present t-statistics because the large sample ensures that most of these differences are statistically significant. Rather, we focus on their size" (Brady and McNulty 2011, p. 9)

The authors then note that the magnitude of the differences are very small and unlikely to be indicative of hidden confounders, yet the size of their sample makes the traditional tests overly sensitive to these minute differences. Their caveat, we argue, reflects a conflict between the purpose for which the typical t-test was designed and the goal of tests of design, namely showing that differences on pre-treatment covariates are substantively unimportant. This chapter aims to outline the problems with how these tests are currently conducted and provide a set of statistical tests for equivalence, which are not subject to many of the concerns highlighted in these two examples.

3.2 Difference versus Equivalence

The current practice in observational studies in the social sciences is to conduct balance and placebo tests using difference-in-means tests, though test statistics which examine other aspects of covariate's distribution, such as Kolmogorov-Smirnov (K-S) tests, are becoming increasingly common. With these tests, a high p-value corresponds to evidence that the treatment and control groups are different, either in terms of the mean for the standard t-test or the empirical cumulative distribution function for the K-S test. In practice, most authors declare adequate balance or placebo equivalence, when these tests show no statistically significant difference between the two groups. A high p-value from such a test fails to provide statistical evidence that the two groups are different which is only indirectly related to showing that they are the same. The problem arises from the fact that the standard tests are designed to control for a type I false positive error of calling the two groups different when they are, in fact, the same. If the goal is to control the false positive error rate for calling the two groups the same when they are in fact different, then the standard test is controlling for the wrong type of error. We argue that a statistical test showing that two groups are equivalent should control for false positives in which the two groups are called equivalent when they are, in fact, different.

Although the practice of reversing the standard setup to make *difference* the null hypothesis and *sameness* the alternative hypothesis is absent from hypothesis testing in the social sciences, there exists a large statistical literature investigating the properties of precisely these types of tests, known as an "equivalence" or "bioequivalence" tests. Wellek (2010) and Berger and Hsu (1996) provide a review of the theory and main uses of equivalence testing and specific tests are discussed in Section 3.3. The main idea behind

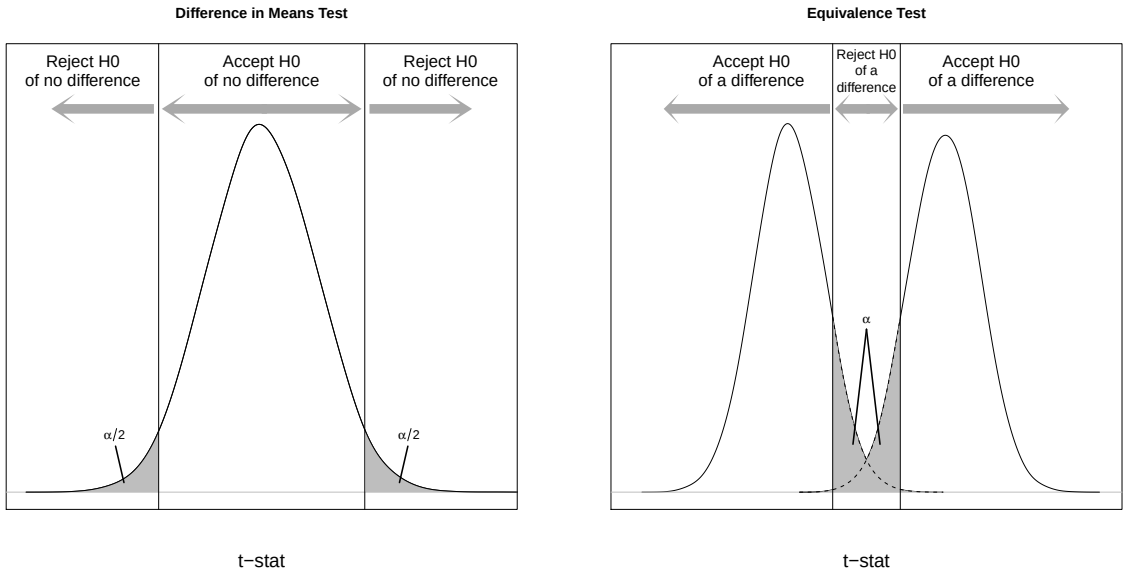


Figure 3.1: Tests of equivalence versus tests of difference. The left panel depicts the logic of tests of difference under the null hypothesis of no difference. The right panel depicts the logic of tests of equivalence under the null hypothesis of difference.

an equivalence test is to set up the test such that the groups are deemed equivalent if the data provides evidence that the two groups are the same, rather than failing to provide statistical evidence that they are different.

Operationally, the most important difference between equivalence testing and tests of difference is whether or not one needs to make an ex-ante decision over what range of values to define as "similar" versus "different". When using equivalence tests, the researcher must specify what is called an "equivalence range", the set of values within which the difference between the two variables are substantively equivalent. One example of a t-test for equivalence is set up as follows:

$H_0 : \mu_T - \mu_C \geq \epsilon_U$ or $\mu_T - \mu_C \leq \epsilon_L$ versus $H_1 : \epsilon_L < \mu_T - \mu_C < \epsilon_U$

where μ_T and μ_C refer to the mean of the treated and mean of the control groups, respectively, for a given variable. ϵ_U and ϵ_L refer to the upper and lower bounds for which two groups are considered equivalent. This test controls for the type I error of calling the two samples equivalent (as defined by the equivalence range) when, in fact, thImaiKingStuart2008ey are not. This is one illustrative example of an equivalence test. Alternative versions, which are designed for different types of data or sensitive to different departures of the null will be presented in subsequent sections.

Figure 3.1 depicts graphically how the traditional balance tests and equivalence tests differ. In traditional balance tests, the means of two groups are declared to be equivalent

if the observed t-statistic falls between the critical values. The shaded region corresponds to the region in which the two groups are classified as different when they are, in fact, the same, and the area corresponds to the level of the test. However, it is easy to see that this procedure is not controlling for the proper type I error as defined by the null of the equivalence test.

There are three factors that can result in the t-statistic lying in either the tails or the center of the t distribution. If the mean difference between the two populations is small, then the t-statistic will also be small, which is desirable for declaring the two groups equivalent. As the standard deviation grows, the two t-statistic will also move towards the center, which is also desirable behavior with respect to determining equivalence. More importantly, though, is the concern raised by Imai, King and Stuart (2008): holding the observed mean difference and standard deviation constant, the sample size can shift the t-statistic from the center to the tail. This is an undesirable property for balance and placebo tests because for any given mean difference, dropping observations will be beneficial in terms of "passing" the test. The converse problem is that when one has very large sample sizes, minute differences may be statistically significant even if substantively meaningless. A desirable property for a statistical test is that the power to detect the alternative increases in sample size, yet by conducting balance tests using tests of difference, the probability of rejecting the null of difference is inversely related to sample size. This problem is one of the reasons why Imai, King and Stuart (2008, p. 494) label the principle of balance testing as a "fallacy".

The right panel of Figure 3.1 illustrates why the t-test for equivalence is not subject to Imai, King and Stuart's critique of balance tests. In this example, the equivalence test is conducted by looking at the distribution of two non-central t-statistics. The lower curve is the distribution of the t-statistic around the hypothesized difference of ϵ_L and the upper curve is the distribution of the t-statistic around the hypothesized difference of ϵ_U. The two groups are considered equivalent if the observed t-statistics lie in the shaded region, i.e. the equivalence range. The area of the shaded region is equal to the level of the test, α. Therefore, this test controls for the correct type I error. There is an α probability that the two groups will be deemed equivalent when, in fact, they are different. Why are equivalence tests not subject to the same problems of sensitivity to sample size as the tests of difference? If the sample size is small, holding all else constant, the t-statistics will move towards the center of their respective distributions, thus making it less likely that we will call the two groups equivalent. Therefore, the power of the test behaves as we would expect with respect to sample size.

An example inspired by a simulation in Imai, King and Stuart (2008, p. 495) illustrates the effects of making the null a hypothesis about difference. Imai, King and Stuart show how sample size affects the t-statistic by taking a covariate from an imbalanced observational study and conducting a t-test after randomly dropping an increasingly large percentage of the controls. They are decreasing the sample size, but in expectation

they are not affecting the overall balance between the treated and control units. They then show that the t-statistic decreases, or moves towards insignificance, as more control observations are dropped, leading to the conclusion that "[t]he t-test can indicate that balance is becoming better whereas the actual balance is growing worse, staying the same, or improving". This simulation depicts the undesirable behavior of using a difference-in-means test. Austin (2008) also raises this point in justifying his claim that significance testing is inappropriate as a metric for post-matching balance because the post-matching p-values are confounded with sample size.

In Figure 3.2 we recreate this simulation, using data from Blattman and Annan's (2010) study on child soldiering. They examine the socioeconomic consequences of abduction by the Lord's Resistance Army, one of the main combatant groups in Uganda's civil war. In this simulation, we examine a balance test on age, which they point to as one of the most important covariates determining selection into treatment. Age is imbalanced, they argue, because the rebel army tended to target somewhat older children. The simulation study mimics Imai, King and Stuart's in that we randomly drop an increasingly large percentage of the controls (non-abductees). For each of the 5000 iterations, we conduct both a traditional and an equivalence based t-test. The figure shows the percentage of simulations that are declared non-equivalent. It is important to note that the two groups are imbalanced, and randomly dropping controls does not, on average, affect the level of imbalance. In the case of the difference of means t-test, the groups are declared non-equivalent if they are found to be statistically different at the 5% level. For the t-test for equivalence, the two groups are declared non-equivalent if they fail to reject the null hypothesis of non-equivalence at the 5% level. Our equivalence interval is 0.2 of a standard deviation in age. As was shown in the Imai, King and Stuart (2008) simulations, as the number of controls randomly dropped increases, the t-test for difference in means (red line) is increasingly likely to declare the two groups equivalent. However, the t-test for equivalence (blue line) is not subject to this problem. As the number of controls dropped increases, the t-test for equivalence still overwhelmingly declares the two groups non-equivalent. As the percentage of the controls drops approaches 85 to 90%, the t-test for equivalence does declare a few of the simulations equivalent. This may be due to the fact that a few of the random draws lead to control samples that were similar to the treated group, given the very small number of controls in these draws.

The main argument in defense of traditional hypothesis testing for validity tests is that although small sample sizes tend to make passing balance tests easier, small sample sizes also make finding significant treatment effects less likely, as articulated by Hansen (2008). Hansen points to the fact that the dependence on sample size, i.e. the $n^{1/2}$ factor in the standard error calculations, appears in both the balance and outcome tests. Therefore, if one artificially inflates the p-values of the balance tests with small sample sizes, then the p-values associated with the outcomes will also be large, leading to non-significant findings. While it may become easier to find treatment effects that achieve significance

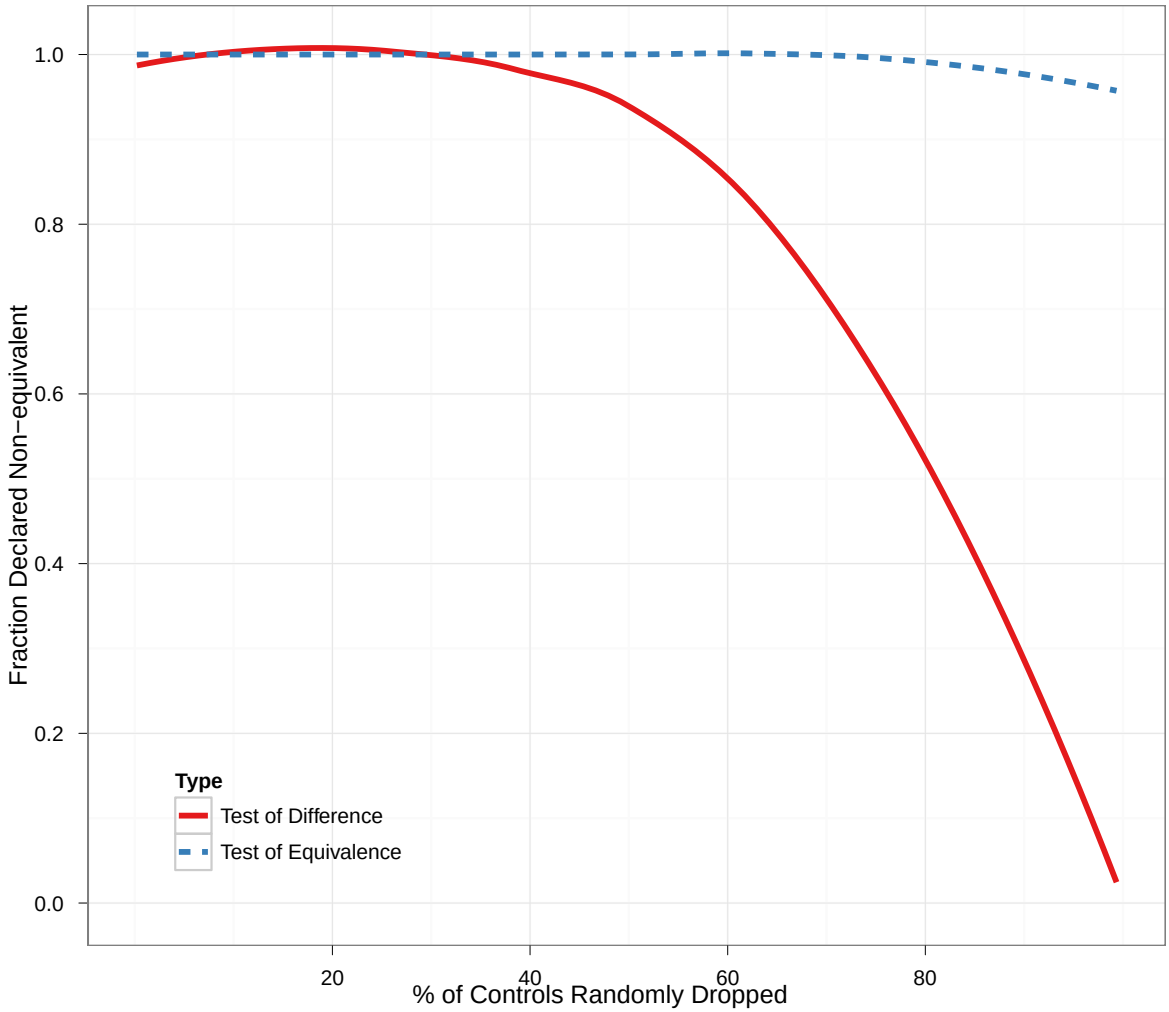


Figure 3.2: The behavior of tests of difference and equivalence when a varying percentage of the control units are dropped from the sample. The red line is the proportion of rejections of the null of no mean difference ($\alpha = .05$) using the difference in means t-test. The blue dashed line is the proportion of non-rejections of the null of difference using an equivalence t-test with an equivalence range of 0.2 of a standard deviation. For the difference test, as increasing numbers of control units are dropped, the share of tests falsely indicating increased balance increases. For the equivalence test, the share of tests falsely indicating increased balance are largely unaffected by sample size.

as sample sizes increase, the associated balance tests will also be harder to pass. This argument, however, hinges on an assumption that statistically significant findings on the outcome are the goal. This argument would not hold if a null finding was of interest. Using the equivalence tests discussed here, small sample sizes will tend to make it more difficult to call two groups equivalent, and will also lead to less significant results. Larger sample sizes will not, however, reduce the likelihood of passing the validating test. If one wants to put a higher burden on the validating tests as a test for design, then the equivalence range should be decreased as sample size increases. By decreasing the range of equivalence as sample size increases, increased burden is once again placed on the validating tests as the power to find significant differences in the outcome increases.

3.2.1 Selecting an Equivalence Range

To conduct an equivalence test, one must choose, prior to conducting the test, an equivalence range or $[\epsilon_L, \epsilon_U]$, i.e the range in which we can consider the parameter of interest in the two groups to be substantively equal. How should one select this interval? One is ultimately interested in bias, yet covariate balance (or the degree of placebo equivalence) is only a proxy for bias in our estimate of the treatment effect parameter. Consequently, what really matters is the unobservable mapping between covariate imbalance and bias. Because this mapping is fundamentally unobservable, our judgements about an adequate equivalence interval must ultimately depend on substantive considerations. Thus, when possible, one should specify an equivalence interval small enough to satisfy readers that imbalances within the interval are inconsequential.

It should be noted that the tradeoff to smaller intervals, however, is power to detect equivalence. If the intervals are very narrow, then a large amount of data will be required to obtain sufficient power to detect differences that small. As a result, we recommend that any results from an equivalence test be accompanied by the power of the test, under the assumption that the true difference is 0. In judging the results of a test of design, the power of the test will inform our expectations over the likelihood of rejecting the null of difference at a given equivalence interval.

Since there naturally will be disagreement over an appropriate equivalence interval, we also recommend inverting the equivalence test to produce an "inverted ϵ". The inverted ϵ is the largest difference at which the null hypothesis of difference would be rejected at a pre-specified α. Akin to a confidence interval, the inverted ϵ specifies the smallest difference supported by the data, given the observed difference between the treatment and control group. In other words, the difference between 0 and the inverted ϵ quantifies the degree of uncertainty we have over the true degree of imbalance. As long as the inverted ϵ is reported, readers can judge for themselves whether this difference constitutes equivalence on the pre-treatment covariate or placebo outcome.

Although we would argue that equivalence intervals are best chosen out of substantive

considerations, it is useful to specify default values. While this is an area in need of validation studies, we recommend using $\epsilon = \pm 0.2\sigma$, where σ is the pooled standard deviation of the variable being tested. The inspiration for this default value for ϵ are the simulation studies reported in Cochran and Rubin (1973), which showed that bias of this magnitude or less tended to produce only minor levels of bias when the relationship between imbalance and bias was linear, and outcome and covariates were normally distributed. Ho et al. (2006, p. 221) also make this recommendation for judging "adequate" balance. We stress, however, that default values of equivalence should still be given careful consideration for any particular application.

For most of the tests described in this chapter it is fairly simple to choose the equivalence range based on substantive knowledge of the data. For some tests, the range can be specified in terms of standardized differences, such as for the t-test for equivalence and the Mann-Whitney test. Other tests, which test directly for equivalence of the raw difference in means, can be specified on the scale of the variable. In the case of the TOST ratio test, the range of equivalence can be defined in terms of the ratio of the means. Table 3.1 provides either a standard range of equivalence used in the literature or an equation for translating a substantively defined ϵ on the scale of the variable into a standardized difference. When possible, researchers are better off defining the equivalence range based on substantive knowledge rather than using default values defined in the literature.

3.3 The Mechanics of Equivalence Tests

Just as their are a variety of tests for evaluating difference, there are many equivalence tests. The most appropriate test statistic depends on the type of variable, as well as desired sensitivity to different types of departures of the null. This section outlines some different types of equivalence tests that are available and the advantages of each type. Because most difference in means tests are conducted using t-tests, we discuss in detail the analogous t-test for equivalence. However, we will discuss other common tests for equivalence that are designed for different distributions and parameters of interest, and which may be more appropriate for small samples. The tests will be described briefly in this section and the formal mathematical outline of the tests can be found in the appendix.

3.3.1 t-test for Equivalence

As opposed to the t-test for difference, the t-test for equivalence is based on the standardized difference rather than the raw difference in means between the two groups. The standardized difference is a useful metric when testing for equivalence because, given some difference between the means of the two distributions, the two groups are increasingly indistinguishable as the variance of the distributions grows towards infinity, and increasingly disjoint as the variance of the distributions shrinks towards 0 (Wellek 2010).

For simplicity, assume that $X_{Ti} \sim N(\mu_T, \sigma)$ and $X_{Ci} \sim N(\mu_C, \sigma)$, then the equivalence t-test uses the following hypothesis test.

$$H_0 : \tfrac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \tfrac{\mu_T - \mu_C}{\sigma} \leq \epsilon_L$$
$$\text{versus}$$
$$H_1 : \epsilon_L < \tfrac{\mu_T - \mu_C}{\sigma} < \epsilon_U$$

We choose ϵ_L and ϵ_U appropriately, perhaps based on substantive knowledge. Section 3.2.1 discussed how to choose ϵ. Typically the range of equivalence is symmetric around zero. After defining an equivalence range, the realized test statistic is calculated. The test statistic is the typical two sample t-test statistic, where we define

$$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{ \sum_{i=1}^{m}(X_{Ti} - \bar{X}_T)^2 + \sum_{j=1}^{n}(X_{Cj} - \bar{X}_C)^2 \right\}^2}.$$

The only difference is now the test statistic is distributed as a non-central F distribution because it is the standardized difference instead of a raw difference. Assume that we define ϵ_U and ϵ_L to be symmetric around 0, i.e. $\epsilon = \epsilon_U = -\epsilon_L$, then the critical value can easily be obtained. The rejection rule now becomes:

$$|T| < C_{\alpha;m,n}(\epsilon)$$
$$\text{with}$$
$$C_{\alpha;m,n}(\epsilon) = F^{-1}(\alpha; df_1 = 1, df_2 = N-2, \lambda_{nc}^2 = mn\epsilon^2/N)^{\frac{1}{2}}$$

where $C_{\alpha;m,n}(\epsilon)$ is the square root of the inverse F distribution with level α, degrees of freedom $1, N-2$, and non-centrality parameter $\lambda_{nc}^2 = mn\epsilon^2/N$. If the ϵs were not symmetric, then we would have the rejection rule:

$$C_{\alpha;m,n}(\epsilon_L, \epsilon_U) < T < C_{\alpha;m,n}(\epsilon_L, \epsilon_U)$$

where the critical values must be determined so as to control for the level of the test. If $|T|$ is less than our critical value (or T lies within the critical values, in the case of asymmetric ϵs), then we reject the null hypothesis of a difference between the two groups, and declare the two groups equivalent. Otherwise we fail to reject the null of non-equivalence. In addition to the rejection decision, researchers should also analyze the inverted range, which provides more detail about the equivalence range that the data can support. In the case that the inverted range is small, then the researcher can be confident that the data provides strong evidence that the two groups are equivalent. If the inverted range is large, then the researcher may call in to question the equivalence of the two groups. In addition to the inverted range, the researcher should also look at the power of the test. If the test fails to reject the null of non-equivalence yet the sample size is small, the power of the test may provide insight in to whether it is a large difference between the two groups or a lack of statistical power which may lead to an increased p-value.

3.3.2 Alternative Tests for Equivalence

In many cases, researchers may be interested in testing for non-equivalence of different parameters of interest. This section outlines alternative tests for equivalence, some culled from the extant literature and others created for the problem at hand. The mathematical notation and steps for implementation for each test are described in detail in the appendix. Rather than studying the standardized difference, researchers may wish to conduct a test for equivalence of the raw mean difference. This can be accomplished using a Two-One-Sided-Test (TOST) (Berger and Hsu 1996). The TOST test is conducted using two one sided t-tests centered around the bounds of the equivalence range. One advantage of the TOST is that it allows for the researcher to define the equivalence range on the scale of the variable of interest as opposed to standardizing substantive ranges. The TOST test can also be adapted to test for equivalence of the ratio of the means of the two groups, instead of the raw difference between the means. The TOST ratio test has the advantage of having an absolute scale that is independent of the scale of the variable of interest. This test is used by the FDA for declaring generic drugs as equivalent to brand-name drugs. In that case, the two drugs are declared equivalent if the ratio of the mean effect of the two drugs falls within the range [0.8, 1.25].

The Fisher type exact test is well adapted to equivalence between two groups with binary outcomes. This test is based on the odds ratio as opposed to the mean difference between the two groups. Wellek (2010) discusses the advantages of choosing the odds ratio over the difference of p_T and p_C, however the basic point can be illustrated as follows. If the test statistic is defined as the difference in the probability of success between the two groups, i.e. $p_T - p_C$, then as p_T approaches 0 or 1, the range of values for which p_C could be called equivalent is diminished. If equivalence is defined as the two groups having a difference in probability of success of no more than 0.1, then if $p_T = 0$, p_C must be between 0 and 0.1. However, if $p_T = 0.5$, then p_C can be between 0.4 and 0.6. If the odds ratio is used as the test statistic, this shrinking of possibilities for p_C as p_T approaches 0 or 1, or vice versa, is not an issue. The Fisher type test for binary data tests whether the odds ratio is within a specified range, typically centered around 1. There are many other equivalence tests for binary data that focus on the raw difference in probabilities of success discussed in Barker et al. (2001).

Finally, the researcher may prefer to use a test sensitive to differences in distribution rather than differences in means, akin to the KS test (Sekhon 2007). The Mann-Whitney test for equivalence is an asymptotically distribution free test that is sensitive to divergences between two continuous distributions (Wellek 2010). If two distributions are equivalent then the probability that any treated observation is greater than any control observation should be approximately $1/2$, thus equivalence is defined as a range around this point. Therefore, the Mann-Whitney tests uses a rank-sum statistic to test whether or this probability is within a small range around $1/2$. If the two distributions are non-equivalent, then the bulk of the treated units should lie to one side of the median of the

Table 3.1: A summary of commonly used versions of equivalence tests.

Test Name	**Type of Data**	**Sample Specific**	**Test Statistic**	**Rejection Rule**	**Epsilon Range**
Equivalence t	Asympt. Normal sample mean	No	$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_T - \bar{X}_C)}{\left\{\sum_{i=1}^{m}(X_{Ti}-\bar{X}_T)^2 + \sum_{j=1}^{n}(X_{Cj}-\bar{X}_C)^2\right\}^2}$	$\lvert T\rvert < C_{\alpha;m,n}(\epsilon)$	$\epsilon_{def} = 0.2$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
Two-One Sided (TOST) t	Asympt. Normal sample mean	No	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{SE(\bar{X}_T - \bar{X}_C)}$ and $T_L = \frac{\bar{X}_T - \bar{X}_C - \epsilon_L}{SE(\bar{X}_T - \bar{X}_C)}$	$T_U < -t_{\alpha,m+n-2}$ and $T_L > t_{\alpha,m+n-2}$	$\epsilon_{def} = 0.2\sigma_{pooled}$ $\epsilon = \epsilon_{sub}$
TOST Ratio t	Asympt. Normal sample mean	No	$T_L = \frac{\bar{X}_T - \epsilon_L \bar{X}_C}{S\sqrt{1/m + \epsilon_L^2/n}}$ and $T_U = \frac{\bar{X}_T - \epsilon_U \bar{X}_C}{S\sqrt{1/m + \epsilon_U^2/n}}$	$T_U < -t_{\alpha,m+n-2}$ and $T_L > t_{\alpha,m+n-2}$	$\epsilon_{def} = [0.8, 1.25]$
Exact Fisher Binomial	Binary	Yes	$\rho = p_T(1-p_T)/p_C(1-p_C)$	$p_{m,n;\epsilon}(x\lvert s) < \alpha$	$\epsilon_{def} = 0.85$ $\epsilon = \frac{\log(1+2\epsilon_{sub})}{\log(1-2\epsilon_{sub})}$
Mann-Whitney	Any Continuous Distribution	No	$W_+ = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}(X_{Ti} - X_{Cj})$	$\frac{\left\lvert W_+ - 1/2 - \frac{\epsilon_1 - \epsilon_2}{2}\right\rvert}{\hat{\sigma}[W_+]} < C_{MW}(\alpha;\epsilon_1,\epsilon_2)$	$\epsilon_{def} = 0.2$ $\epsilon = \Phi\left(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}\right) - \frac{1}{2}$
Non-parametric Equivalence t	Any Continuous	No	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{\hat{\sigma}(\bar{X}_T - \bar{X}_C)}$ and	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = 0.2$ $\epsilon = \frac{\epsilon_{sub}}{\sigma_{pooled}}$
		Yes	$T_U = \frac{\bar{X}_T - \bar{X}_C - \epsilon_U}{\hat{\sigma}(\bar{X}_T - \bar{X}_C)}$		
Non-parametric TOST (npTOST)	Any Distribution	No	$T_U = \bar{X}_T - \bar{X}_C - \epsilon_U$ and $T_L = \bar{X}_T - \bar{X}_C - \epsilon_L$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = 0.2\sigma_{pooled}$ $\epsilon = \epsilon_{sub}$
		Yes			
Non-parametric Mann-Whitney	Any Continuous Distribution	No	$T_U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}(X_{Ti} - X_{Cj}) - (1/2 + \epsilon_U)$ and $T_L = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}(X_{Ti} - X_{Cj}) - (1/2 - \epsilon_L)$	Associated permutation p for both test statistics $< \alpha$	$\epsilon_{def} = 0.2$ $\epsilon = \Phi\left(\frac{\epsilon_{sub}}{\sqrt{2}\sigma_{pooled}}\right) - \frac{1}{2}$
		Yes			

ranked treated and control observations. This test is especially advantageous because it does not depend on the underlying distributions of the treated and control groups so long as they are both continuous. The rank-sum statistic is also robust to outliers in the data.

3.3.3 Sample Specific Equivalence Tests

Finally, a concern of many researchers is that validating tests, specifically balance tests conducted on pre-treatment covariates based on a hypothetical super-population, should not be conducted using hypothesis tests because they are contradictory to the non-random nature of the observed sample (Imai, King and Stuart 2008; Austin 2008). One solution to this issue is to conduct tests that are conditional on the realized sample using permutation based inference (Hansen and Bowers 2008). Each of the tests described above can be translated into a permutation based test. Using the Intersection-Union Test principle, each test statistic can be tested using a strict null hypothesis equal to the bounds of the equivalence range, and the overall null hypothesis of non-equivalence can be rejected if both corresponding permutation p-values are less than the level of the test, α. In addition to being conditional on the observed sample, the permutation tests are exact and do not rely on large sample approximations. Most importantly, the tests are designed to test for exchangeability of the two groups, a property that should be guaranteed by the random or quasi-random design of the study. Therefore, they are well suited for validating tests such as balance and placebo tests where we explicitly desire a test of exchangeability. They are also robust to outliers and sensitive to departures of the null above and beyond mean differences, such as differences in variability within the two groups.

Below, Table 3.1 summarizes the tests described above. The "Type of Data" column describes the type of data each test is appropriate for and the "Sample Specific" column describes whether the test is conditional on the observed sample, as opposed to referencing a super-population. The test statistic and rejection rule are also described for each test. Finally, the "Epsilon Range" column describes the typical epsilon range defined in the literature, denoted ϵ_{def}, and where available, the equation for translating substantively motivated ϵs, which are on the scale of the variable and denoted ϵ_{sub}, into the scale of the test. This table is intended to serve as a simple reference for practitioners. As mentioned above, the mathematical foundations and steps for implementation are described in detail in the appendix.

3.4 Example: Brady and McNulty (2011)

To illustrate the merits of equivalence tests over traditional tests, we return to the example of Brady and McNulty (2011), discussed in Section 3.1. Recall that Brady and McNulty (2011) argue that some polling stations in Los Angeles were consolidated close to "as if" random by the county registrar. Central to their argument about the quality of

their design is that prior to the consolidation, voters in treatment and control precincts had roughly equal "costs of voting", with distance between voters' residence and their polling station being their chief measure of cost. Balance on this variable is critical, yet the authors find that the pre-treatment difference is "highly significant", although "substantively rather small" (p. 5). If the conventional decision rule over adequate balance is followed, then one would reject the "as if" random identification assumption.

We replicate Brady and McNulty's balance check using the two sample t-test for equivalence. The observed average difference in distance between voters in treatment and control precincts is 0.033 miles or 58 yards. We use an equivalence interval ϵ of 0.2 standard deviations (amounting to about 0.05 miles or 88 yards). Note that is a case where the equivalence interval used to formulate the null hypothesis could be chosen on substantive grounds based on knowledge of factors affecting the decision to turnout. We also compute the inverted ϵ which is the smallest equivalence interval supported by the data ($\alpha = 0.05$) given the observed difference between treatment and control polling stations.

Can we reject the null hypothesis that the mean difference in the distance to polling stations in 2002 is greater than $\epsilon = 0.033$ miles? This null is rejected with a p-value that is essentially zero. Given our pre-specified equivalence interval, we consider the two samples to be well balanced on this variable. When we invert our test, we find that the smallest equivalence interval supported at the 0.05 level is 0.15 standardized units or 0.04 miles (70 yards). Whether or not 0.04 miles is of concern, worthy of further adjustment through matching or regression, should be debated by subject area experts.

3.5 Applying Equivalence Tests to Natural Experiments in the Social Sciences

Does the use of equivalence tests make a difference in practice? To show that it does, we apply the two sample t-test for equivalence to ten studies culled[2] from Dunning's (2010a) literature review of natural experiments in the social sciences. From each study, we selected one covariate that was tested for balance. Each study typically examined several covariates, so when possible we selected the pre-treatment outcome (the outcome variable as measured prior to the intervention) and, failing that, a variable that in our judgement, was closely related to the outcome of interest. The papers, which are on a diverse set of treatments in a variety of contexts, are listed in Table 3.2. For the equivalence interval, we chose 0.2 of a standard deviations, following Cochran and Rubin (1973, p. 422)'s discussion.

[2]In order to carry out the test, we required the mean difference, the standard error of the mean difference, and the sample size in each treatment condition. All natural experiments in Dunning's (2010a) list that reported this information were used.

Paper	Treat	Covariate	Mean Diff	N	P (diff)	P (equiv)	Power
Di Tella, Galiani and Schargrodsky (2007)	Property Rights	Years of Education	0.08	1080	0.75	0.00	0.88
Hyde (2008)	Observer Visit	ChallengerâĂŹs Vote Share	0.00	1763	0.78	0.00	0.82
Annan and Blattman (2010)	Abduction	Father's Years of Schooling	-0.05	741	0.86	0.00	0.68
Ferraz and Finan (2008)	Gov. Audit	Reelection rates (2004)	0.02	373	0.69	0.05	0.29
Chattopadhyay and Duflo (2004)	Reservations	Wells	-0.02	161	0.80	0.10	0.10
Krueger (2000)	Minimum Wage Increase	Employment - November 1992	-3.30	384	0.40	0.16	0.23
Dunning (2010*b*)	Reservations	Mean Scheduled Tribe Population	60.67	200	0.40	0.27	0.13
Lyall (2009)	Artillery Shelling	Rebel Presence	0.10	147	0.20	0.52	0.1
Ho and Imai (2008)	Ballot Order Position	Registered Democratic	-0.02	80	0.46	0.52	0.10
Lee (2008)	Democratic Victory	Democratic Win Prob	0.14	610	0.00	0.82	0.59

Table 3.2: Equivalence tests in ten natural experiments. Table shows the difference in means, the standard difference-in-means T-test p-value, the total number of units, the p-value from a two sample T-test of equivalence with an $\epsilon = .2$ of a standard deviation and the results of a power calculation. Studies are ordered by equivalence test p-value.

Paper	Covariate	Std. Inverted ϵ	Unstd. Inverted ϵ
Blattman (2010)	Father's Years of Schooling	0.11	0.59
Di Tella, Galiani and Schargrodsky (2007)	Years of Education	0.11	0.63
Hyde (2008)	ChallengerâĂŹs Vote Share	0.12	0.02
Ferraz and Finan (2008)	Reelection rates (2004)	0.20	0.13
Chattopadhyay and Duflo (2004)	Wells	0.28	0.20
Krueger (2000)	Employment - November 1992	0.32	17.27
Dunning (2010*b*)	Mean Scheduled Tribe Population	0.35	252.17
Lee (2008)	Democratic Win Prob	0.41	0.29
Lyall (2009)	Rebel Presence	0.48	0.33
Ho and Imai (2008)	Registered Democratic	0.77	0.13

Table 3.3: Inverted equivalence tests in ten natural experiments. Table shows upper boundary of the 95% confidence interval of the two sample T-test of equivalence in standardized and unstandardized units.

The results of the equivalence test on a pre-treatment covariate in the ten natural experiments are shown in Table 3.2, along with the conventional difference-in-means t-test p-value. Nine out of the ten natural experiments reported difference-in-means t-test p-values greater than 0.05, thus failing to reject the null hypothesis of no mean difference and consequently "passing" their balance test. If the equivalence test is used, however, only for five[3] of the ten studies can we reject the null hypothesis of a mean difference $|\epsilon| > 0.2\sigma$ with a 0.05 level of significance, where σ is the pooled standard deviation of the covariate. Four of the studies failed to reject the null hypothesis of a difference, but also failed to reject the null hypothesis of no mean difference. Consequently, in these four cases, the conventional decision rule would declare the natural experiments to be balanced, while our proposed test would not. Of course, failing to reject the null hypothesis of a difference by no means invalidates these studies' conclusions, but merely suggests that insufficient information exists to affirmatively declare that the treatment and control groups on these particular covariates are well balanced. At a minimum, our results suggest that these scholars could take special care to show that the design is valid using other design tests or robustness checks.

In Table 3.3, we present the maximum value of ϵ for which which we can reject the null hypothesis of non-equivalence, given the observed difference. We present both the standardized and unstandardized values of the inverted ϵ. This inverted ϵ is useful here because it can give the reader a sense of the smallest equivalence interval supported by the data at a given significance level. Because researchers' opinions may differ over how small an equivalence interval chosen ex-ante should be, reporting the inverted interval can allow readers to draw their own conclusion over the degree of balance evidenced in the data.

3.6 Conclusion

Increased skepticism about traditional research designs in economics, political science, and sociology has encouraged researchers to expend great efforts in defending their effect estimates from the critique that they suffer from omitted variable bias. In many areas of the social sciences, readers begin with the presumption that the observational design is flawed and must be convinced by empirical tests, direct or indirect, that this is not the case. The argument of this essay is that this skepticism should be directly embedded in the hypothesis tests that are used to convince readers over the validity of the design. By using equivalence tests, researchers begin with the assumption that the design is flawed, and this conclusion is only reversed if the data allows it. Furthermore, we believe that equivalence tests encourages researchers to directly address a substantive question about

[3]One study, Chattopadhyay and Duflo (2004) was borderline with a p-value of 0.1, but given the low power of the test for a study of that sample size, we would consider this covariate to be balanced.

their design: what is good balance? By requiring the researcher to specify an equivalence range *ex ante*, equivalence tests encourage a substantive discussion about imbalances that are small enough to be tolerated versus those that are not.

For sample sizes typically used in natural experiments and related designs in the social sciences, an equivalence approach will likely increase the difficulty of passing balance and placebo tests. As evidenced by our review of natural experiments in Section 3.5, covariates or placebo outcomes in many studies that currently pass tests of design when the null is sameness will not pass when the null is difference. Not passing an equivalence test does not by itself, of course, invalidate a design or indicate hopelessly biased estimates. Many other elements of a design should go into a evaluation of its quality, such as the degree to which the assignment to treatment is exogenous or "as if" random. For studies where the treatment assignment mechanism is well understood and the identifying assumptions seem quite plausible, our burden of proof should be lower. This might apply to cases where the treatment was actually randomized, although not necessarily by the researchers themselves, as in Chattopadhyay and Duflo (2004), Ferraz and Finan (2008), and Ho and Imai (2008). In other cases, such as for designs exploiting a discontinuity or those relying on a conditional independence assumption, more definitive evidence may be required to overcome doubt. For these cases, equivalence tests can improve on existing practice by ensuring that we encode our skepticism in the null hypothesis and require the researcher to marshall evidence against it.

Chapter 4

Methods to Alleviate Unit Non-Response Bias

The 2012 presidential election saw an explosion of publicly available polling conducted by a wide variety of organizations ranging from university survey centers to large news organizations. Advances in technology and the increasing use of voter files for list-based samples have made conducting surveys more accessible to organizations. Live and robo-call polls remain the most prominent modes of contact for political polls, though response rates to such polls have decreased over time, coinciding with the rise of public polling. More restrictive laws on predictive dialing of cell phones, the rise of a cell phone only population, and behavioral changes in voters' willingness to respond to such polls all increase the risk of unit non-response bias, which has the potential to severely bias the findings of these polls. This was seen in the 2013 Gubernatorial election in Virginia, where public polls were roughly five points more Democratic, on average, than the final election results. There was also large variance in the findings of individual polls conducted in the final weeks preceding the election.

As different types of individuals become likely to respond to surveys, the potential impact of unit non-response bias in political surveys grows. Traditionally, political surveys attempt to mitigate the impact of unit non-response at two stages: by using quota or stratified sampling in the design stage and by using *post hoc* weighting methods on the final data. This chapter first considers sources of bias caused by unit non-response, breaking down the bias into features that can be addressed in the design and *post hoc* adjustment stage, and unobservable behavioral bias that must be assumed away. Following the bias decomposition, current design phase approaches are considered. The chapter then considers a new sampling method, response rate sampling, that leverages past data on responsiveness to surveys to alleviate bias due to observable imbalance resulting from unit non-response. *Post hoc* weighting methods for correcting for observed imbalances are then considered, including a recommendation for a novel way of using traditional raking methods.

The chapter concludes with an application of response rate sampling and a new weighting method used in Obama for America's (OFA) analytics polling methodology, one of the largest and most accurate polling operations in democratic politics. OFA's internal analytics polling served as the data collection method for the campaign's successful predictive modeling operation. The dual nature of the data increased the necessity of representative data. Under-representation of key demographic groups, often those who do not respond

to surveys at high rates, can limit the power of models to detect key predictive variables.

The OFA analytics polling operation was also large, with tens of thousands of calls placed across a dozen key battleground states each week beginning in April. The campaign spoke with over one million voters across battleground states, collecting information on calls placed to over ten million voters in the context of random surveys. Given the large nature of the polling operation, and the associated costs, it was important to build a sustainable sampling methodology that leveraged past data to inform the best practices possible for collecting representative samples.

Given the close nature of the election in many of these states, and the limited budget of the campaign, it was also important to maximize the power of the surveys to accurately measure the opinions of key demographic subgroups so the campaign could maximize its return on investment for engagement methods with voters. Additionally, in the face of the changing dynamics of voter engagement, which impacts the likelihood a voter will respond to a survey, it was important that the campaign be able to discern true changes in voter opinion from changes in response bias, especially in the face of large campaign events such as the first presidential debate and the Republican vice presidential announcement.

The ability to accurately measure changes in opinion even in the face of differential voter engagement, allowed the campaign to most effectively allocate resources. The OFA analytics polling operation serves as a good case study of these new methods, showing how careful sampling designs, based on in-cycle data, and novel *post hoc* weighting methods create an accurate, scalable and cost-effective polling methodology.

4.1 Unit non-response and bias

Even in the case of meticulously planned survey designs, non-response can cause potential bias in survey estimates. Unit non-response, or vector non-response, occurs when the entire vector of response items is null. As discussed in Holt and Elliot (1991), "the fundamental problem arising from unit non-response is that the sample design breaks and, in particular, the selection probabilities are changed in an unknown way". Following their discussion, Holt and Elliot (1991) decompose the potential bias from unit non-response into two components, one related to the difference of observable characteristics between the return data and the target population of interest, and one related to the difference in the response curve between those who respond to the survey and those that don't.

Following the Holt and Elliot (1991) decomposition, define P_R as the proportion of the target population that responds to a survey, and P_{NR} as the proportion of the target population that does not respond to surveys. Define $\bar{Y}_R$ as the average outcome to a survey question among the responsive population, and $\bar{Y}_{NR}$ as the average outcome among the non-responsive portion of the population. The true population mean, then, is defined as

the convex combination of the two classes of respondents and their means:

$$\bar{Y} = P_R\bar{Y}_R + P_{NR}\bar{Y}_{NR}$$

In the case of a simple random sample[1], the observed mean outcome will be a function only of those who respond to the survey:

$$E[\bar{y}_R] = \bar{Y}_R$$

and the bias in this estimate is a function of the difference in the average response of respondents and non respondents, and the proportion of the population who does not respond.

$$bias(\bar{y}_R) = P_{NR}(\bar{Y}_R - \bar{Y}_{NR})$$

Assuming that the sample can be defined by a set of non-overlapping strata with proportions S_s, s.t. $\sum S_s = 1$, the population mean can be defined as:

$$\bar{Y} = \sum_s S_s\bar{Y}_s = \sum_i S_s(P_{iR}\bar{Y}_{sR} + P_{iNR}\bar{Y}_{iNR})$$

where P_{sR} and P_{sNR} are the proportions of respondents and non-respondents within each strata S_s, and $\bar{Y}_{sR}$ and $\bar{Y}_{sNR}$ are the respective population means.

The bias, then, can be restated as:

$$bias(\bar{y}_R) = \underbrace{\frac{\sum_s S_s\bar{Y}_{sR}(P_{sR} - \bar{P_R})}{\bar{P}_R}}_{\text{Observable Sample Imbalance}} \tag{4.1}$$

$$+ \underbrace{\sum_s S_s P_{sNR}(\bar{Y}_{sR} - \bar{Y}_{sNR})}_{\text{Unobservable Response Difference}} \tag{4.2}$$

where $\bar{P}_R$ is defined as overall proportion of respondents in the full population. Equation (5.1) refers to observable differences in the sample proportions of the strata as compared to the target population, based on the observable characteristics defined by the strata S_i. Equation (5.2) is due to differences in the average outcome of interest between respondents and non-respondents within strata.

The goal, then, of sampling design and weighting methods is to correct for the portion of bias related to sample representativeness. Design based methods correcting for non-response attempt to account for known differential rates of responsiveness in the sampling

[1]In the case of stratified sampling, the bias could be defined for each strata, and overall bias would be a function of the relative proportion of each strata in the population.

frame, whereas weighting methods attempt to account for representativeness on the back end. Assuming that a weighting method gives an estimate of the population proportion for each strata, $\hat{S}_s$, then assuming $E(\bar{y}_{sR}|\hat{S}_s) = \bar{Y}_{sR}$ and $E(\hat{S}_s) = S_s$, the bias in the first term goes to zero, and the remaining bias from the weighted estimate, $\tilde{y}$, is defined as:

$$bias(\tilde{y}) = \sum_s S_s P_{sNR}(\bar{Y}_{sR} - \bar{Y}_{sNR})$$

Holt and Elliot (1991) note that the bias in the weighted estimate will be lessened after weighting under two conditions, (1) if the bias due to the observable imbalance and the unobservable difference in response has the same sign, or (2) if the magnitude of the observable bias is more than twice the magnitude in the unobservable bias. Bias reduction only occurs if the weighting method returns unbiased estimates of the average response among respondents within strata, and the weights return unbiased estimates of population proportions for each strata.

This chapter discusses design phase approaches that emphasize representative final return data, thus minimizing the bias due to Equation (5.1). Following a discussion of sampling designs, the chapter discusses different weighting approaches, and addresses their ability to meet the criterion for reducing bias.

4.2 Design approaches for Addressing Unit Non-Response

Design based approaches to unit non-response attempt to use previous data on responsiveness in the population to account for known patterns of non-response in the sampling stage. As discussed in Groves et al. (2002), unit non-response can be accounted for in the framework of probability sampling. Non-response is the product of a probability of selection based on a selection mechanism, and if this selection mechanism is known or can be estimated, then the probability of selection can be accounted for in a probability sampling frame. Generally, the sampling weight is inversely proportional to the probability of selection. Pickery and Carton (2008) discuss how oversampling is related to substitution methods. Another design based approach is quota sampling. This section describes traditional designs that address unit non-response, followed by a new method, response rate oversampling.

4.2.1 Stratified Sampling

Stratified sampling is a proven methodology for improving the representativeness of samples. Lohr (2010) outlines the advantages of stratified samples. They protect against outlier samples, or "bad samples". They can be used to ensure a level of precision for a key

subgroup, they can be easier to administer and reduce costs, and they have been shown to reduce variance in population means and totals. Stratified samples were first studied by Jerzy Neyman in 1934, and the precision gains have been well studied since (Neyman 1934; Ericson 1965; Bickel and Freedman 1984; Imbens and Lancaster 1996).

Stratified sampling involves dividing the population into non-overlapping strata, and conducting simple random sampling within strata. Strata can be allocated proportionally to their population totals, or disproportionately. Optimal allocation, which considers the variance within strata, can reduce costs and increase precision of population estimates (Lohr 2010). Given known differential rates of response between strata, stratified sampling can be used in conjunction with quotas. If quotas are set at a strata level, the biases caused by quota sampling are minimized. Stratified sampling is highly related to the new method described in Section 4.2.3, response rate sampling. Response rate sampling uses stratified sampling, but alleviates the need for quotas by incorporating known rates of responsiveness.

4.2.2 Quota Sampling Methods

Probability sampling, such as stratified sampling, is the current standard for most surveys. Response rate oversampling, described below, is a new probability sampling method. However, many polling firms use quota-sampling methods. Online polling firms that use opt-in samples and polling firms that conduct Random Digit Dialing (RDD) often use quota-sampling methods. Quota-sampling methods are the non-probabilistic version of stratified sampling. As discussed in Berinsky (2006), quota sampling methods were the standard sampling design for public opinion firms until the 1950s.

In quota sampling, the population is divided in to non-overlapping strata, much like in stratified sampling; strata are created using variables relevant to the outcome of interest. Researchers then define target sizes for each of these strata, and interviews are conducted until each quota is met. Quota samples are useful when probability sampling cannot be conducted ahead of time, such as with random digit dialing, or when interviewers are responsible for recruiting subjects in the field as was the case with public opinion polls in the 1930s and 1940s. Quotas allow polling firms to exert direct control on the distribution of observable characteristics of their sample. As stated in Berinsky (2006), "Gallup and Roper did not trust that chance alone would ensure that their sample would accurately represent the sentiment of the nation".

While quota samples guarantee representativeness on observable characteristics, the non-probabilistic nature of the sample can lead to nonrandom selection within strata. In public opinion polls where there is interviewer discretion, this can be caused by interviewers choosing more approachable subjects (Berinsky 2006); in telephone polls, this can arise from self-selection of who answers the phone in a household, or generally leading to more responsive individuals having a higher likelihood of being included in the survey. This

leads to the bias described in Equation (5.2). The non-probabilistic nature of quota sampling leads to major concerns about non-random selection bias, and the representativeness of the opinions of survey respondents for non-respondents.

In the case that quotas are set at a strata level, representativeness will be achieved on observable variables. However, if quotas are only set on the margins of variables, i.e. setting a target number of men vs. women, and a target number of young vs. old, but not setting a target on $gender \times age$, then the representativeness of strata might not be met. If older women are more likely to answer the phone in a household, then quotas for older voters and women will be met first, meaning that the quotas for men and younger voters will have to be filled by young men. It is important, if using quotas, to set quotas at a deeply stratified level, not only on the margins, in order to achieve representative samples on observable variables. By including multiple interactions in the quotas, such as setting quotas on $age \times gender \times ethnicity$, then representativeness will be achieved on multiple interactions of key demographics variables. Note that response rate sampling can aid quota designs, ensuring that if certain strata are less likely to respond, then the impact of the selection in to the sample will be minimized. In the example above, more young men would be included in the sample, increasing the likelihood that that strata fills up the quota at a similar rate as older women. Additionally, when quotas begin to fill, it will minimize the need to call respondents in hard to fill quotas multiple times, but rather ensure that most survey respondents were called a similar number of times.

4.2.3 Response Rate Sampling

While stratified sampling helps ensure a balanced initial sample, there are known trends in the rates at which different demographic groups respond to surveys. For instance, age is a negative predictor of response, both due to data quality and behavioral factors, in political surveys. In surveys that use list-based methods from maintained voter files, phone numbers associated with youth tend to be out of date and associated with their parent's home. Additionally, conditional on reaching youth, they are less likely to complete a survey. Here we refer to the proportion of target respondents who are contacted and successfully complete a survey as the "response rate". By focusing on accounting for known differential response rates in the sampling stage, survey practitioners can improve the quality of their return data and minimize variance due to Equation (5.1).

The method described below, known as response rate sampling, provides a new approach to sampling that incorporates data on past response rates of individuals to increase the likelihood a sample yields a representative final data set. While past methods relied on stratified sampling, perhaps with quotas on strata, this method is novel in its approach of combining stratified sampling with past data to inform a non-uniform probability sample. By using past data to predict individual level responsiveness, there is no need to require quotas on strata, alleviating the known biases quotas can cause.

In order to account for unit non-response in the design phase of a survey, previously collected data must be used to estimate the response probability, or response rate, for a group or individual. Estimates of non-response can be obtained in many ways, ranging from estimates at a strata level to individual level estimates. Once estimates of non-response have been obtained, the estimates are used to create a probability sampling frame where each individual is included with probability inversely proportional to their estimated probability of response. This is referred to as response rate sampling. The goal of this method is that if the estimated probability of response is accurate, then the return data will be representative based on observed characteristics included in the response probability, minimizing the bias in $\bar{y}_R$ due to observable imbalance of the sample.

For example, say the only predictor of response rate is phone type: on average one in five cell phone owners will respond to a live phone survey, but only one in ten respondents contacted on a landline will respond to the live phone survey. Assume, also, that the target population is made up of half cell phone owners and half landline owners. Then, if a simple random sample of the target population is used, the final return data will be made up of two thirds cell phone respondents and one third cell phone respondents. If, however, the differential response rate is accounted for in the sampling design, and the sample is stratified, and one third of the final sample is cell phone owners and two thirds landline owners, then the final data will be representative, in expectation, on phone ownership.

In the context of political surveys, response rates vary systematically with demographics known to be predictive of many political behaviors of interest. These include age, gender, party registration, and measures of political engagement. Without accounting for the known differential response rates in the design phase, final survey return data will be demographically unrepresentative in predictable ways. There will be a lack of Democratic leaning groups and younger respondents, for instance.

The differential response rates are often due to both data quality issues and behavioral issues. Phone numbers on file are more likely to be out of date for more transient populations, such as youth. However, behaviorally, young people also tend to be less willing to take and complete surveys. While great care should be taken to confirm the quality of the underlying data and to encourage as many respondents as possible to participate in the survey, response rate corrections can be made conditional on the data and tailored to the survey design. While *post hoc* weighting methods can help adjust estimates, if the observable imbalances are substantial then the weights for some individuals will be extreme. The noise induced by these weights can increase the variance in estimates among key demographic groups, often those who are known to have heterogeneous political opinions, such as among younger respondents. Elliott and Little (2000) discuss the impact of truncating weights, which leads to slight increases in bias but can greatly decrease the variance. By increasing the representativeness of the survey, analysts will have greater stability of estimates and power to estimate opinion among key demographic subgroups.

Constructing response rate estimates requires a data set from previous surveys that contains demographic information and an outcome variable measuring whether or not a respondent begins or completes a survey. While the previous data does not need to be from surveys that are identical to the survey under consideration, the more similar the length, content, and topic of the survey, as well as the sampling frame, the more effective the response rate oversampling method will be. For instance, the attrition rate on a survey increases, non-linearly, with the length of a survey. If the goal is to model the likelihood to complete a five question survey, then response rate estimates based on a twenty question survey might lead to under estimates of likelihood to complete a five question survey. It is important to note that this will only matter if the drop off rates of longer surveys can be explained by variables that are predictive of response rates. If drop off rates are random conditional on starting a survey, then the relative ratios of the response rates will be the same, and response rate estimates based on longer surveys can be applied to shorter survey designs. In the absence of previous data, researchers will have to rely on *post hoc* weighting methods to correct for remaining imbalances due to unit non-response.

Response Rate Sampling: An Estimation Strategy

There are many ways to construct predictive models of responsiveness. One method would be to construct an individual level regression model to estimate of propensity to respond to or complete a survey. This can then be applied, as a predictive model, to the population of interest, and respondents can be selected inversely proportional to the predicted propensity score (Groves et al. 2002; Rosenbaum and Rubin 1983). An alternative approach, discussed here, is to construct response rate estimates at a strata level. This approach is compatible with stratified sampling methods since it provides oversampling rate estimates at the same level that the stratified sampling is conducted.[2]

When using a stratified sampling design, response rate oversampling is easiest to incorporate if the response rates are estimated at the strata level. While it is possible to divide the previous response data in to strata, and estimate a response rate within each strata, this can lead to over-fitting of the data and unstable estimates of responsiveness. Therefore, a method that estimates response rates within predefined strata without over-fitting the data is necessary. Classification trees, or CART methods, determine which interactions of qualitative variables are statistically significantly related to the outcome of interest, in this case whether or not someone completes a survey, while avoiding over-fitting the data.

Classification trees are a decision tree machine learning method which uses a predefined set of variables to predict an outcome of interest (Breiman et al. 1984). The tree finds the most predictive variable, divides the data set accordingly, and then finds the next

[2]Individual estimates can be aggregated up to the strata level.

most predictive variables within each division of the data. Divisions can group multiple categories of the data or fully stratify on a given variable. For instance, if non-white participants respond to a survey at similar rates, then the tree may divide respondents in two groups, white and non-white. However, if all races respond at significantly different rates, then the tree will fully stratify on race. The result of the tree is a set of interacted strata defined by the most to least statistically informative splits. Splits in the data are only made where variables explain remaining variance in a statistically significant manner. Where there is not a statistically informative split, the data is pooled.

Response rates are calculated as the percent of respondents who completed the survey within each terminal node of the tree. It is important to use cross-validation methods to determine important tuning parameters such as the complexity parameter, the minimum leaf size, etc.[3] An example of using CART to estimate response rates is shown in Figure 4.2. Further discussion of classification trees can be found in Breiman et al. (1984). Code can be implemented using the `rpart` package in `R`. Examples of the CART method are shown in Section 4.4.1. The outcome modeled is likelihood to complete a political survey, however there are situations in which survey designers may be interested in likelihood to begin or complete some part of the survey. For maximal flexibility in modeling responsiveness, when collecting meta data about a call, it is important to get clear and descriptive disposition codes from vendors.

Section 4.4.2 describes the CART method in the context of the OFA polling methodology. This section also describes further considerations that practitioners must take when implementing this method in the context of live calling data, including stratification variables and how to incorporate changes in election dynamics over time.

4.3 Weighting Methods to Correct for Unit Non-Response

Sample based design inference of surveys allows for analysis to be based on the known probability of inclusion in a sample based on the sampling framework (Lohr 2010). However, due to factors such as unit non-response, quota sampling methods, undercoverage in the sampling frame, and other factors, the true inclusion probability of an individual may not be known (Lohr 2010; Deville and Sarndal 1992; Lumley 2010; Holt and Smith 1979). In the face of unknown sampling probabilities, *post hoc* weighting methods can help practitioners overcome biases introduced by unrepresentative samples. As discussed in Section 4.1, weighting methods address the component of bias introduced by imbalance in observable characteristics related to the outcome of interest.

[3] One option is to use a cross-validation such as the Super learner to cross-validate the multidimensional space of parameters (van der Laan, Polley and Hubbard 2007).

Traditionally, survey weighting employs two main classes of weighting methods, post-stratification and raking, although other methods such as generalized regression and maximum entropy methods are also used. All of these classes of weighting methods are special cases of the general class of calibration estimators (Lumley 2010; Folsom and Singh 2000; Kott 2006; Schickler, Berinsky and Sekhon 2012). Calibration estimators find a set of weights that satisfy a set of population constraints while remaining as similar as possible to a set of reference weights (typically uniform weights). Following Deville and Sarndal (1992), let $\mathbf{T}_x$ represent a set of population totals across K auxiliary variables, let d represent a set of reference rates, then for a sample of size n, the final weights w must satisfy the following constraints:

$$T_{x_j} = \sum_i w_i x_{ij} \ \forall j \text{ in } K$$

minimizing the distance between w and d,

$$\sum_i D(w_i, d_i)$$

where $D(\cdot, \cdot)$ represents the choice of distance function. Many distance functions are described in Deville and Sarndal (1992). The following sections will describe common weighting methods and their associated distance functions. Another class of weighting methods, inverse propensity score weighting, is also discussed. The methodological and practical advantages and disadvantages of each weighting method are also discussed. Finally, a novel way of implementing raking is discussed.

4.3.1 Post-stratification and Linear Regression Weighting

Stratified sampling is one of the most common forms of survey design, with widely known properties that alleviate bias due to unrepresentativeness and increase precision. If the population is divided into S_s non-overlapping strata, s.t. $\sum_s S_s = 1$, then stratified sampling involves conducting a simple random sample within each strata. Population estimates are based on a weighted average of estimates within strata. Precision gains and implications for representativeness of the sample are well known (Neyman 1934; Ericson 1965; Bickel and Freedman 1984; Imbens and Lancaster 1996). Post-stratification is related to stratified sampling in that it divides the population in to non-overlapping strata for which population totals are known. Sample proportions are then adjusted relative to their known population proportions. Post-stratification weights are defined as:

$$w_p s, i = \frac{N_s/N}{n_s/n} \times d_i$$

where N_s and n_s are the number of units in a given strata s in the population and sample, respectively, and d_i is the reference weight in the calibration framework, typically defined as $d_i = 1 \ \forall \ i$. When used in conjunction with stratified sampling, post-stratification

is often used to correct for non-response bias occurring within pre-defined strata. However, it can also be used to allow for more flexible *post-hoc* weighting schemes that incorporate information not used in the sampling frame (Holt and Smith 1979). It can also be used to create weights in sampling designs that do not allow for pre-stratification, such as RDD or the increasingly common opt-in internet survey designs (McDonald, Mohebbi and Slatkin N.d.), or for incorporating additional ancillary information in to a quota sample design (Berinsky 2006; Schickler, Berinsky and Sekhon 2012). Assuming that the variables used to form the strata account for the selection mechanism for non-response, then population estimates. As discussed in Lumley (2010), p. 172, "If the non-response probability is homogenous within the group ... then post-stratification will give correct sampling weights, in the sense that the population total for any variable is correctly estimated."

An alternative approach to post-stratification is through generalized regression (GREG) estimators (Deville, Sarndal and Sautory 1993; Lehtonen and Veijanen 1998; Chen, Sitter and Wu 2002; Lumley 2010). Linear regression estimators are more flexible in that they can account for more flexible forms of variable interactions, continuous variables, in addition to categorical qualitative information. Linear regression estimators regress a target variable, Y, on a set of auxiliary variables. In the case of post-stratification, the regression equation is the full set of dummy variables for all cross-classifications of the auxiliary variables (Lumley 2010).

4.3.2 Inverse Propensity Score Weighting and Doubly Robust Estimation

When all auxiliary information is available for both the population and the sample, another approach that can be used is inverse propensity score weighting (IPSW) (Stuart et al. 2011; Kalton and Flores-Cervantes 2003; Freedman and Berk 2008). The idea behind inverse propensity score weighting is that the data can allow for the development of a model of the propensity for inclusion in the sample at a very granular level. This is related to the generalized linear regression framework, although the model for the propensity for inclusion does not have to be linearly additive in nature. Models are commonly built using a logistic functional form, however the theoretical underpinnings of the method do not depend on a given estimation strategy. Propensity scores can be built using machine algorithms (van der Laan, Polley and Hubbard 2007). Much like in the case of linear weighting, the effectiveness of the IPSW weighting method is sensitive to the propensity score specification.

The theory for IPSW is derived under the Horvitz-Thompson framework, proposed by Rosenbaum (1998) and others, is described further in Lunceford and Davidian (2004). Following the propensity score models of Rosenbaum (1998) and Rubin (1974) the IPSW framework views non-response as a missing data problem that can be overcome with a selection on observables assumption. The propensity score is defined as $e(\mathbf{X}) = P(Z =$

$1|\mathbf{X})$, s.t. $0 < e(\mathbf{X}) < 1$, where Z is an indicator for inclusion in the sample, and $\mathbf{X}$ is a set of covariates sufficient for selection on observables. Weights are then constructed as:

$$w_i = 1/\hat{e}(\mathbf{X})$$

where $\hat{e}(\mathbf{X})$ is the estimated propensity score. This class of weighting is related to the class of doubly robust estimators discussed in Robins, Rotnitzky and Zhao (1995). Doubly robust estimators rely on a model of both the response surface and the inclusion probability; as long as either the response surface or inclusion probability is modeled correctly, then doubly robust estimators are unbiased.

4.3.3 Raking and Maximum Entropy

Another common form of weighting is raking, or multiplicative weighting. Raking is a more flexible weighting method than post-stratification in that it allows for the incorporation of auxiliary information for which only population marginals are knowns (Berinsky 2006). The theory for raking, also known as iterative proportional fitting, was outlined by Little and Wu (1991), and was originally proposed by Deming and Stephan (1940). The raking algorithm is discussed further in Little and Wu (1991). It involves a series of post-stratification steps, in which post-stratification is done across the margin (or interactions of variables), repeating across variables until the weights converge within a specified tolerance.

In the calibration framework, raking uses the distance function:

$$D_{rake} = (w_i, d_i) = w_i \ln (w_i/d_i) - w_i + d_i$$

The last two terms cancel one another out, since $\sum_i w_i = \sum_i d_i$, or the final weights must sum to the same total as the original weights, meaning that raking converges to the Kullback-Leibler measure of cross-entropy, namely $\sum_i w_i \ln \frac{w_i}{d_i}$ (Ireland and Kullback 1968).

The idea behind maximum entropy is based on Shannon (1948) notion of information. These ideas are the foundations for what Jaynes (1957) later mathematically proved as the principle of maximum entropy. In essence, maximum entropy uses information to pick a probability distribution that both satisfies a set of constraints and maximizes entropy. As Kapur and Kesavan (1992) state, "out of all probability distributions satisfying given constraints, choose the distribution that is closest to the uniform distribution". The maximum entropy principle takes advantage of the information that is given, such as a set of linear constraints; by maximizing the entropy conditional on the given constraints, only the given information is used and no additional assumptions are made about the distribution of weights.

Raking, or MaxEnt weighting, allows for flexible definitions of target population margins. Where as full cross-classification of auxiliary variables must be used in post-stratification methods, raking allows for inclusion of fewer interactions of auxiliary variables, or simply for margins. It also allows for inclusion of auxiliary information for which there is a complete measure in the sample, but only marginal information in the target population.

There are drawbacks to raking algorithms. Given the iterative nature of estimation, it is possible that convergence cannot be achieved. Given there is no closed form solution for the weights, it is possible that the weights could be unstable and dependent on the order in which the algorithm proceeds (Bethlehem 2002). If the data is drawn from a simple random sample, and the survey does not suffer from non-response, linear weighting and raking algorithms will converge, however in the face of non-response, which method is preferable depends on the behavior of the weights in practice.

Raking On Cells

There is a clear tradeoff between traditional raking methods and post-stratification. Post-stratification accounts for more demographic factors and has well studied variance reduction properties. However, in practice, once an analyst accounts for more than a few variables, null strata, those in which there is no respondent in the sample, and small strata become an issue. Raking on margins provides more flexibility in the number of variables that can be included, however it requires the strong assumption of independence among those variables. Post-stratification on many variables often leads to extreme weights, whereas raking tends to have more stable weights.

In order to minimize the effect of extreme weights, rather than using post-stratification to weight on full cross-classification of variables, another novel approach is to rake on all combinations of lower-level interactions. For example, if one wishes to weight on age, gender, and ethnicity, then rather than weight the full cross classification of $age \times gender \times ethnicity$ in a post-stratification method, one can weight on all of the following margins: $age \times gender$, $age \times ethnicity$, $gender \times ethnicity$ (Schickler, Berinsky and Sekhon 2012). In cases such as political polling, where the population is not evenly distributed among the fullly classified strata, and smaller strata tend to also be more non-responsive, leading to an increased likelihood of strata being missing or sparsely populated, weighting on all combinations of lower-level interactions will minimize the chance of extreme weights on sparsely populated strata. While accounting for all lower-level interactions might not account for as much observable bias as a full cross-classification, there is evidence that extreme weights can cause instability in estimates that outweighs the overcome bias (Freedman and Berk 2008; Elliott and Little 2000) By accounting for all lower-level interactions, a significant portion of the bias is accounted for, leaving only the correlation of the fully interacted model unaccounted for. However, there is a risk with raking that if there is little relation between the additional dimension and the cell means, the variance

might increase rather than decrease (Lohr 2010).

This approach balances the trade off of post-stratification and raking methods, leading to more stable weights while still accounting for interactions among demographic variables. In practice, null strata are minimized, extreme weights are rare, and an analyst can easily account for many demographic variables, including four-way interactions among key variables. Additionally, analysts are afforded the flexibility of accounting for interactions among some variables, but only margins on variables that are either independent of other variables or for which only margins are known.

4.4 Application: Obama for America 2012 Analytics Polling

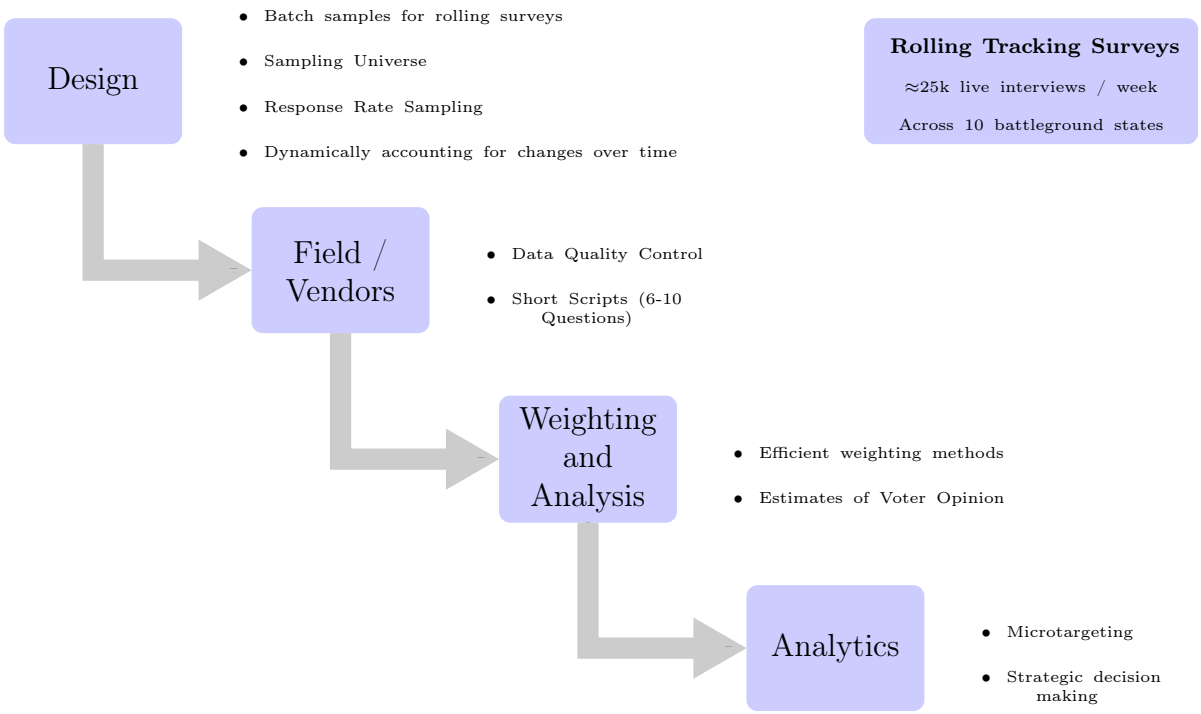
The Obama For America 2012 (OFA) campaign used a design based approach to unit non-response combined with a new flexible weighting scheme discussed in (Schickler, Berinsky and Sekhon 2012) for the large-scale polling operation fielding every day between the beginning of April 2012 and November 6, 2012. The campaign fielded live rolling, cross-sectional surveys during this time in key battleground states, ultimately contacting over one million voters through random surveys. Sampling, data collection, and weighting were conducted on a weekly basis, the process for which is outlined in Figure 4.1.

Working from an up-to-date voter file, OFA employed a list-based random sampling design of registered voters using state voter files. Random sampling of known registered voters alleviates the need for screener registration questions necessary in RDD designs, which can be subject to dishonesty and confusion of voters. Sampling was done using a stratified sampling design, including demographic variables available on the voter file. The target distribution was based on modeled estimates of the likely election day electorate.

Strata were created using seven key variables, including geographic, demographic, and behavioral measures. Stratification was done within polling regions within states. Polling regions were defined with the aim of determining areas of the state with roughly equal populations that exhibited similar patterns in voting behavior. The goal was for variance in support for the candidate to be greater across regions than within regions. Demographic measures included individual characteristics such as age, sex, race, and party registration where available. Finally, since the goal of stratification is to account both for variation in the primary outcome of interest, in this case support for the candidate, but also for behavioral measures related to self selection in to the survey, the sampling design used a proprietary proxy for political interest. This measure accounted for many factors, including previous vote history and other expressions of political engagement.

These stratification variables formed the basis of analysis for the response rate calculations, sampling design, and weighting methods. Each of the methodological steps in the polling process is outlined below.

Figure 4.1: Obama for America Analytics Polling Process



4.4.1 OFA: Response Rate Sampling

Call data from the campaign was collected on all attempted calls, including whether or not the target voter began and successfully completed a poll. Response rates were calculated to determine the likelihood that a given voter would complete a poll if contacted. Response rates were calculated on state-specific data sets where data allowed, and among data pooled for states in similar geographic regions and with a similar data structure where state-specific data was insufficient.

Given the changing dynamics of voters' interest and engagement in the campaign, this measure of responsiveness was updated frequently. For instance, as the election neared, given the large number of calls being placed by polling firms and campaigns, many voters began experiencing call fatigue, and response rates decreased across the full spectrum of voters. There were also also events during the political season that led voters to engage at different rates. Following the first debate, Democratic leaning voters became less likely to respond to polls in the face of a poor debate performance from the President. At the same time, Republican leaning voters became more engaged, evidenced through their increased rate of response to political polls. While there may have been a small change in the support levels for the candidates, much of the change seen in public polling was driven by this differential level of engagement and enthusiasm among partisans following the debate.[4] By updating baseline levels of responsiveness, the impact of changes in enthusiasm based on campaign events were mitigated, making it easier to discern true changes in public opinion.

Response rates were calculated using Classification and Regression Trees (CART), incorporating all of the stratification variables listed in Section 4.4. The complexity parameter was cross-validated, and the minimum bucket size was set such that overfitting was minimized. While modeling response rates at an individual level is an alternative, the campaign used CART methods in order to determine the strata for which response rates were statistically different. Calculating response rates at an aggregate level like this allowed for response rates to be different for groups whose responsiveness is statistically different, yet allowed for pooling of data for smaller or highly unresponsive groups. The advantage of pooling data for highly unresponsive groups is that, when sampling inversely proportional to response rate, groups with very low response rates will be oversampled at very high rates. Having a very precise measure for non-responsive groups, where noisy estimates can lead to very high, and unstable, weights, is important to minimize costs and increase the quality of the return data.

The tree in Figure 4.2 was constructed using previous data on whether or not respondents completed a survey, and the algorithm is given five qualitative variables: political

[4]Differential rates of enthusiasm in response to may also be a sign of a changing likely electorate. Changes in enthusiasm should be incorporated in to turnout expectations. Further research needs to be conducted on how measures of enthusiasm, such as responsiveness, are correlated with likelihood to turnout.

engagement, race, age, party, and gender. Political engagement is a propriety measure, dichotomized to create a measure of high vs. low political engagement levels. This is the most predictive variable of likelihood to complete a political survey, and therefore is the first split in the tree. Within the highly politically engaged group, race is the next most predictive variable, splitting among all three races in the data set. Among the low political engagement group, African Americans and Hispanics behave similarly, so the tree only splits out Whites separately. Further splits are conducted on age, although the dividing lines vary depending on the political engagement by race classification.[5] This process continues until either there are no remaining stratification variables or the variance explained by the remaining variables is not statistically significant.

Figure 4.2: Example of Classification Tree for Response Rates to Political Surveys

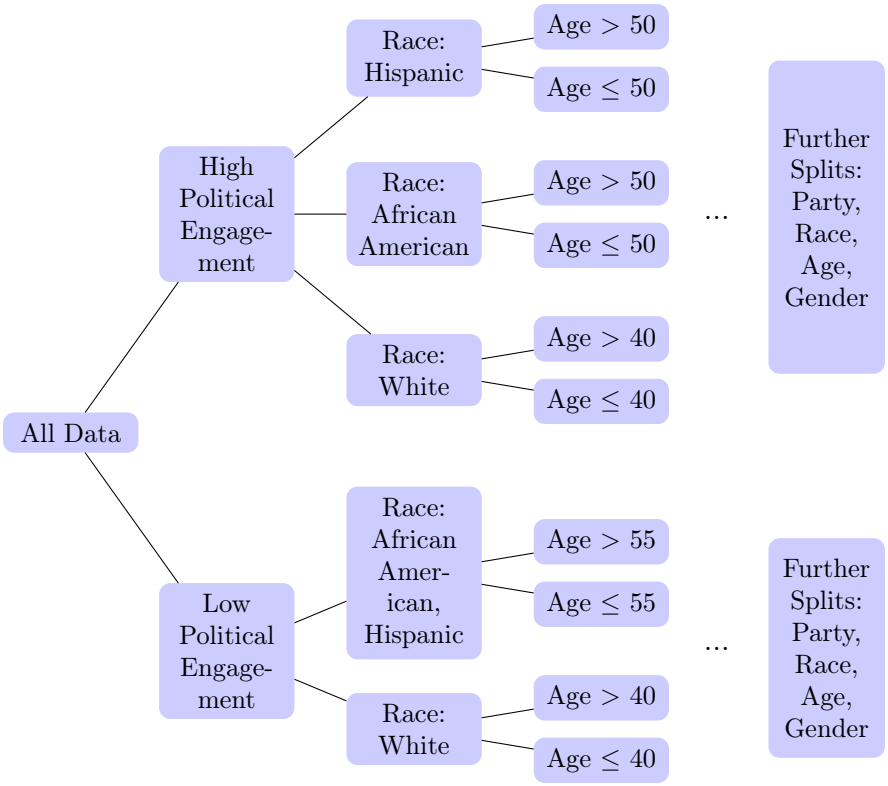


Table 4.1 shows an example of the most and least responsive groups to Obama For America's Analytics Tracking polls. There are many aspects of this table that exemplify

[5]Further splits can include additional splits on variables that have not already been fully stratified, such as additional splits on age within branches, or on strata groups that were combined earlier in the branch

Table 4.1: Examples of Response Rates from a Political Survey

Political Engagement*	Race	Age**	Party	Gender	Attempts	# Respondents	Response Rate†
Low	H	30 to 35	D, I, R	F, M, U	19,584	316	1.6%
...	...	...	...	...	...	...	...
High	W	55+	D	F	4,303	1,187	27.6 %

* Political engagement is a dichotomized variable based on a proprietary measure of engagement used by the campaign.
** Age bucket definitions have been changed slightly from the proprietary values used by the campaign
† Note that this is not conditional on contact, but rather includes the probability that the phone number is live and associated with the target respondent

typical trends in political surveys. First, the biggest predictor of likelihood to complete a political survey is captured by any measure of political engagement. Less political engaged respondents are less likely to complete political surveys. Age is also a big predictor of likelihood to complete a political survey, which is a function of both data quality and behavior. Younger respondents are more likely to be transient, and their phone numbers are more likely to be out of date. However, young respondents are also less willing to take a political survey conditional on contact. Data quality issues also give rise to differential response rates between races.

Another important thing to notice in Table 4.1 is that even though there is a lot more data in the least responsive group, the data is pooled among party and gender. Once splits have been made on political engagement, race, and age, the remaining variance is not significantly related to gender and party. The response rate is only about 1 in 62 respondents, which is very low.[6] Further splits on gender or party increase the instability in the estimates without providing any additional statistically significant information, so it is better to pool the data across strata. This is the advantage of CART methods for avoiding over-fitting the data.

The American electorate is not evenly distributed among cross classifications of these variables. For the most responsive groups, there is a lot of data among older, politically engaged, white respondents. It is important to note that response rate sampling conditions on responders, and does not address the bias due to Equation (5.2). However, in political data, there is evidence that the bias due to Equation (5.2) is small relative to the observable imbalance in Equation (5.1) (Keeter et al. 2000).

A balance must be struck between using more, older data to increase the power to detect differences in response rates and limiting to newer data that best reflects current election dynamics. In order to incorporate changes in responsiveness in response to changing election dynamics, flexibility in the measurements were allowed within terminal nodes

[6]Note that this includes disconnects and wrong numbers.

of the tree. Using the data from strata defined by terminal nodes of the CART tree, data within a node was divided into two groups based on a point of reference, such as a date three weeks prior. A t-test was conducted on data within the terminal node of the tree to determine if the response rate prior to the reference date was statistically significantly different than the response rate in the data collected after the reference date. If the response rates were significantly different, the response rate based on the most recent data was used. This allowed for flexibility in response to changing campaign dynamics, but prioritized demographic and behavioral attributes for determining response rates over the influence of the campaign cycle.

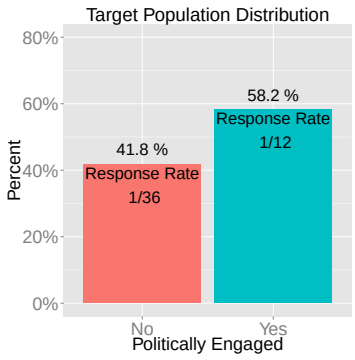
4.4.2 OFA: Checks of Design

Each week a new sample was constructed based on the refreshed response rates. OFA employed stratified sampling. Strata breakdowns were defined based on a target population, adjusted inversely proportional to estimated response rates. Since the CART method returns estimates at no smaller than a strata level, each strata defined by the stratification variables can be mapped to a terminal node in the CART tree. Within strata, a simple random sample of respondents was selected.

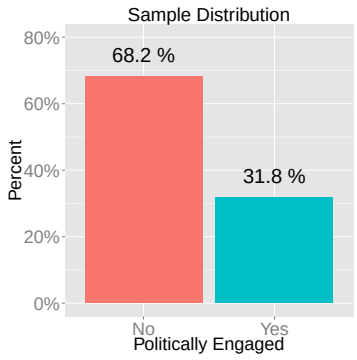
The final sample was randomized and calls were fielded through the week, spread equally throughout the week. Response rates varied systematically with day of the week, a factor that was not included in the campaign's sampling process but which could be included in samples for future studies. No quotas were placed on any of the stratification variables, only on the final sample size. Final data was monitored daily, and checks were conducted to ensure that the return data was representative of the target population.

Figure 4.3: Example Distributions of Sampling Process Adjusted by Response Rates Florida, 4/24/2012 - 4/30/2012

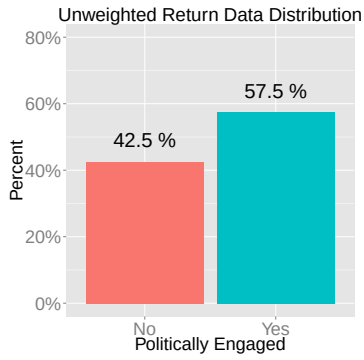
(a) Target Population Distribution with Response Rates



(b) Distribution of Sample Given Response Rates



(c) Distribution of Sample Given Response Rates

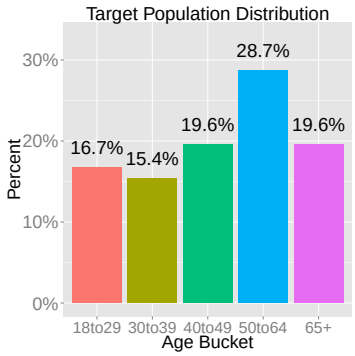


Response rate sampling requires that measures of non-response are accurate. Given a target population and an estimate of non-response, the final sample will not look representative of the target population, however the final, unweighted return data should be representative of the target population. Figure 4.3 shows an example of the sampling process with response rates. Figure 4.4(a) shows the target distribution of the political engagement in Florida, the state this example is drawn from. The political engagement variable is the most predictive variable in terms of explaining likelihood to complete a survey, and those who score highly on this measure are three times more likely to complete a survey than those who are not politically engaged. On average, one in twelve numbers dialed among those who classify as politically engaged will complete a survey, whereas only one in thirty six among the non-politically engaged will.[7]

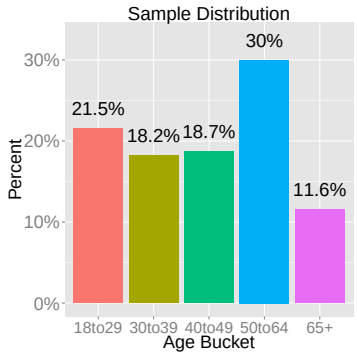
Figure 4.5(b) shows the distribution of the political engagement variable in the sample. Non-politically engaged voters are oversampled relative to their size in the target population. Figure 4.5(c) shows the distribution in the unweighted return data. Despite the oversample of non-politically engaged voters in the sample, the final data is representative of the target population, with the breakdown on the political engagement variable within statistical noise of the target distribution.

Figure 4.4: Example Distributions of Sampling Process Adjusted by Response Rates Michigan, January 2012

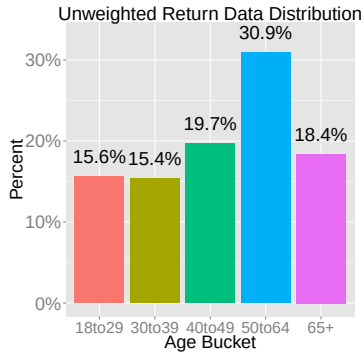
(a) Target Population Distribution with Response Rates



(b) Distribution of Sample Given Response Rates



(c) Distribution of Sample Given Response Rates



The political engagement measure is a proxy for a behavioral driver of non-response. Response rate sampling should incorporate measures related both to behavioral predictors of non-response and demographic predictors related to the outcome of interest. Some

[7]It is important to note that the response rate includes dials to numbers that are disconnected, wrong numbers, and other issues related to the quality of the data. The response rate is not a purely behavioral measure, and conditional on a target respondent being reached, the complete rate for surveys was, on average, between one in two to one in five, depending on the state and level of engagement during the campaign.

demographic variables, such as age, are highly predictive of non-response and many outcomes of interest. Figure 4.4 shows an example of how response rates relate to age. Young voters are more unresponsive than older voters, with those over the age of 65 significantly more likely to respond to a political poll than the younger cohorts.

As seen in Figure 4.3, Figure 4.4 shows that, despite a large differential in the likelihood of different age cohorts to respond to an OFA survey, adjusting for this known difference in the sample using response rates allows for the raw return data to be highly representative of the target population. Response rates account for both behavioral and quality of data issues that impact responsiveness to surveys. While political engagement factors are driven primarily by behavioral predictors of responsiveness, factors such as age are confounded by behavioral and data quality factors. Young voters are more likely to have old phone numbers associated with their names on the voter file, such as the phone number of their childhood home if they register at 18. More transient populations are more likely to have out-of-date contact information, and factors related to factors such as transience will confound behavioral and data quality factors.

4.4.3 OFA: Weighting

Sections 4.4.1 and 4.4.2 outlined how response rate sampling can increase the representativeness of survey return data. However, due to changes in voters' engagement during the election, changing dynamics of underlying responsiveness, or statistical noise, there are still small imbalances in the raw return data. Section 4.3 outlines various methods for *post hoc* weighting adjustments. Weighting return data can improve the accuracy of surveys, although extreme weights that occur in highly unrepresentative samples can induce noise in estimates.

OFA employed raking for *post hoc* weighting. Weighting was conducted using the same variables used in the the sampling frame. As shown in section 4.4.2, the return data was highly representative due to the response rates accounted for in sampling. This, combined with a flexible weighting method, allowed for an increased number of variables to be accounted for in the weighting step.

By using a raking method that accounted for all three-way interactions of all key variables, including demographics such as age, gender, race, political engagement, and region, the weighting method could account for remaining discrepancies between the raw return data and the target population. The representativeness of the raw return from week to week meant that the same factors could be accounted, minimizing the impact of an analyst's decisions on factors to control for on the estimates of public opinion.

Additionally, despite including numerous variables in the weighting adjustments, the flexible weighting method and representative samples meant that the final weights did not suffer from extreme weights. Typically, across states and time, 95% of weights were below 1.5, meaning the design effect of weighting on the survey was small and the noise

in the estimates induced by the weighting step was minimal.

Figure 4.5: Distribution of Weights in Final OFA Surveys

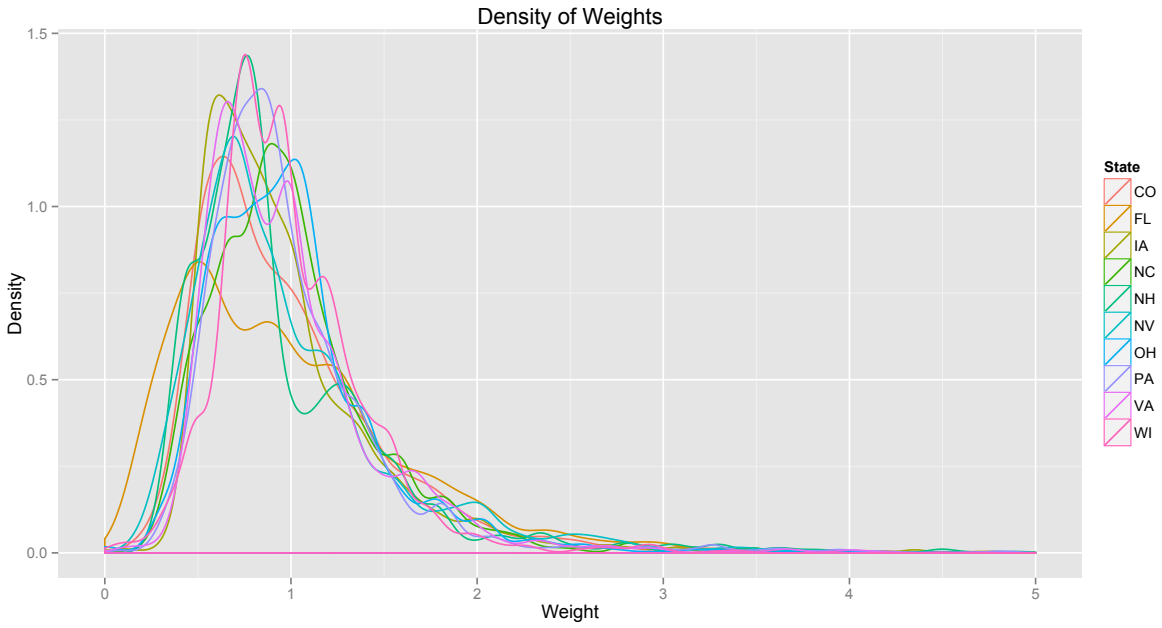


Figure 4.5 shows the density of weights by state for the final polls released by OFA. These polls were conducted between October 28, 2012 and November 4, 2012. What is important to note is that no weights exceed 5, and the bulk of the density is near 1. Most of the adjustments to the data were minor, despite accounting for up to four way interactions of key demographic variables.

The combination of response rate sampling and a flexible weighting method allowed for the campaign to have consistent, representative surveys that minimized the impact of non-response on the estimates. Estimates of public opinion remained steady over the course of the campaign. Changes that were observed were more easily attributed to changes in public opinion instead of changes in underlying demographics or differential voter engagement. This resulted in highly accurate estimates of public opinion, allowing the campaign to efficiently allocate resources. Ultimately, the polling proved to be not only consistent, but highly accurate.

4.5 Discussion and Conclusion

This chapter outlined multiple design based approaches to minimize bias due to unit non-response, as well as numerous weighting methods to account for observable imbalance

in final survey returns. A new design based approach, response rate sampling, is outlined. This method uses response data from previous surveys to estimate the the likelihood a strata or individual will complete a survey. Sampling can then be done inversely proportional to these known response rates, increasing the likelihood the survey response data will be representative. This minimizes the impact of *post hoc* weighting methods, as well as reduces the number of impactful substantive decisions a practitioner must make on how to conduct *post hoc* weighting. Additionally, in the face of observable imbalance, a new approach to raking methods is discussed.

While this chapter addressed the advantages response rate sampling serves in reducing bias due to unit non-response, especially with respect to the bias outlined in Equation (5.1), there is further work to be done on the impact response rate sampling has on variance reduction. Response rate sampling increases the representativeness of the final data, decreasing the emphasis placed on *post hoc* weighting adjustments. This reduces variance, and minimizes the likelihood of extreme weights. However, response rate sampling may also lead to increased confidence in the assumption that data is missing at random (MAR). The MAR assumption states that nonresponse depends only on observed variables (Lohr 2010), since hard to reach groups are overrepresented in the sample, increasing the likelihood that these groups are not represented purely by unrepresentative early responders who are given high weights. Further work will explore the impact of response rate sampling on this assumption, and its impact on variance reduction of estimates of opinion.

In addition to introducing response rate sampling, this chapter applied this method in the context of a large scale presidential campaign. Obama for America's (OFA) 2012 internal polling employed response rate sampling in the design of their internal analytics polling operation. The campaign spoke with over one million voters across battleground states, collecting information on calls placed to over ten million voters in the context of random surveys. This massive polling operation allowed for frequent refreshes of the CART models estimating the probability of response among demographic strata. By incorporating up-to-date estimates of responsiveness in the sampling design, OFA was able to minimize the impact large campaign events, such as the first presidential debate, had on estimates of public opinion. Large events such as these impact not only voter opinion, but voter's enthusiasm, which translates into differential rates of response among different partisan and demographic groups. Using accurate baseline measures of responsiveness, updated in the face of voter fatigue towards surveys that increased over the course of the campaign, OFA was able to mitigate the changes due to unit non-response in internal surveys, and keep a consistent and accurate measure of public opinion.

Response rate sampling also allowed the campaign to collect representative polling data, minimizing the impact of *post hoc* weighting methods. Weighting methods often assign high, unstable weights, in political polls, to groups that are heterogenous in their opinions, such as younger voters and hispanics. By focusing on collecting observably

representative survey data, the weighting methods were both able to account for more demographic characteristics while also keeping the weights stable.

By incorporating previous data, response rate sampling, and a novel approach to raking weighting methods, the OFA analytics polls allowed the campaign to have stable estimates of voter opinion, even in times that public polls showed wild swings. Combined with careful analytical approaches to a consistent and accurate target election day electorate, these methods allowed the campaign to have highly accurate measures of public opinion.

Chapter 5

Conclusion

Social scientists and policy makers continue to put increased emphasis on identifying causal effects in their research, employing the myriad of novel approaches that have been developed in recent years. With this rise in the use of causal analysis tools, the importance of understanding the assumptions underlying such methods has increased. Methods and estimates are occasionally stretched beyond the limits of the assumptions. This has increased the need for a thorough understanding of what assumptions are necessary to identify causal estimates outside of a purely experimental framework. Additionally, it has lead to a need for validation tests that allow researchers to provide sound evidence that estimates are unbiased.

This dissertation has discussed one area, methods for increasing the external validity of experimental estimates, of interest to researchers and policy makers alike. This problem has been addressed from multiple angles, both theoretical, practical, and from a design perspective. While each chapter addresses a piece of this methodological problem, the use cases for each method extend well beyond the scope of the particular problem outlined here. This chapter will outline the findings of each chapter and discuss further research.

5.1 Overview of Findings and Further Research

This dissertation has focused primarily on identification methods for extrapolating experimental results to target populations. This has been addressed from multiple avenues. Chapter 1 outlined the problem, showing the inherent biases that can impact both sample and population average treatment effects. Following Heckman et al. (1998), we decomposed bias into an internal and external component. This bias decomposition allows us to outline what parts of potential bias can be remedied by choice of estimand, which can be be biased with estimating strategies such as matching and weighting. Additionally, Chapter 1 outlines components of bias that can be alleviated with smart design of experiments and data collection, as well as potential biases that can be assumed away. The presence of the biases can, in many cases, be tested with validation checks.

5.1.1 Identifying Externally Valid Experimental Estimates

Following a discussion of the potential biases that can impact causal estimates, Chapter 2 provides the basis of the argument for extrapolating experimental estimates to larger population of estimates. Past research has focused on using non-random studies to investigate population effects. Chapter 2 adds to this literature by deriving the assumptions required to identify population average treatment effects from RCTs aided by NRS data. Additionally, an observable implication of the assumptions allows for a set of placebo tests. Finally, the chapter outlined a new research design for estimating population effects that use non-randomized studies to adjust the RCT data. This design employs a matching approach to create matched strata within the RCT, and then considers two weighting methods for adjusting the experimental data using population information from the NRS.

The main findings of this chapter are described in Theorem 1, replicated below,

Theorem 1. *Assuming consistency and SUTVA hold, if*

$$\begin{aligned}&\mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 0, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 0, T_i = 1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{is1}|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_{is0}|W_i^{C_T}, S_i = 1, T_i = 1)\},\end{aligned}$$

or sample assignment for treated units is strongly ignorable given W_i^T *and sample assignment for controls is strongly ignorable given* $W_i^{C_T}$ *then*

$$\tau_{PATT} = \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T, S_i = 1, T_i = 1)\} - \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T}, S_i = 1, T_i = 0)\},$$

where $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$ *denotes* $\mathbb{E}_{W_i^T|S_i=0,T_i=1}\{\mathbb{E}(\cdot|W_i^T, \ldots)\}$
and $\mathbb{E}_{01}\{\mathbb{E}(\cdot|W_i^{C_T}, \ldots)\}$ *denotes* $\mathbb{E}_{W_i^{C_T}|S_i=0,T_i=1}\{\mathbb{E}(\cdot|W_i^{C_T}, \ldots)\}$.

Additionally, an observable implication of this derivation is a validation test. If the estimand of interest is PATT, then, regardless of adjustment method, adjusted experimental treated units should be equivalent to true population treated. This can be tested using an equivalence placebo test, discussed in Chapter 3. If all assumptions are met, and there is sufficient power in the data, then researchers can use this equivalence test to provide evidence in favor of the extrapolation method.

There are strong policy implications for this study. Governments with universal healthcare systems rely on cost-effectiveness measures to determine their standards of care and coverage. Experiments, given their non-representative samples and non-pragmatic treatment regimes, do not always provide valid estimates of cost-effectiveness of medical interventions among target populations. This method provides policy-makers the ability to identify the impact of these medical interventions in use among target populations.

The theoretical method derived in Chapter 2 is applied to an important, practical example. The chapter discusses the example of Pulmonary Artery Catheterization (PAC). The literature on PAC is, thus far, mixed in the findings of its effectiveness. A commonly

used procedure, experimental findings indicated no overall benefits on measures of survival, whereas NRS methods showed possible negative effects of the intervention. Using a new estimation strategy, we recover a null effect in the population on survival, cost, and cost-effectiveness measures. However, we find subgroups for which the intervention proves to be cost-effective, such as for patients in non-teaching hospitals, or those undergoing elective surgery. More importantly, we show that while one weighting method, MaxEnt, is able to pass a set of placebo tests for most subgroups, inverse propensity score weighting is unable to pass such validity checks. Additionally, using the tests derived in Chapter 3, we are able to discern groups for which the placebo tests mail fail due to lack of power from those where unobservable differences persist.

In addition to further research on the substantive findings of the study, there is work to be done on extending this method. The assumptions necessary for proving Theorem 1 can be adjusted to allow for indirect comparisons of treatments. If treatment A is compared to treatment B in study 1, and to treatment C in study 2, then this method allows for the comparison of treatment B to treatment C without the need for an additional experiment. The framework provides a straightforward path for identifying other population treatment effects beyond PATT. Finally, it provides a path forward for conducting more sophisticated meta-analyses.

Theorem 1 is agnostic to estimating strategy or substantive field. Beyond the substantive field of health economics research, there are many applied examples that can benefit from this method. In Political Campaign and Political Behavior research, there is a wealth of experimental data on Get-Out-The-Vote, persuasion, and other experiments. These experiments have been conducted under a myriad of electoral conditions and with numerous iterations of treatments. Methods such as the one outlined in Chapter 2 provide a path forward for comparing these treatments in the face of differential selection criterion and environments.

5.1.2 An Equivalence Approach to Balance and Placebo Tests

Following the bias decompositions in Chapter 1, and the assumptions outlined in Chapter 2, Chapter 3 addresses an estimating strategy for validity checks of the exchangeability assumption key to most causal identification methods. This assumption is met by design in experiments for sample treatment effect estimates, however for population treatment effect estimates, or in observational methods, this assumption should be validated. Recent debates over the difficulty of causal inference in the social sciences has led to the expectation that scholars justify their research designs by testing the plausibility of their causal identification assumptions, often through balance and placebo tests.

This chapter outlines a new approach to validation tests. The current practice in observational studies in the social sciences is to conduct balance and placebo tests using difference-in-means tests, though test statistics which examine other aspects of covari-

ate's distribution, such as Kolmogorov-Smirnov (K-S) tests, are becoming increasingly common. With these tests, a high p-value corresponds to evidence that the treatment and control groups are different, either in terms of the mean for the standard t-test or the empirical cumulative distribution function for the K-S test. In practice, most authors declare adequate balance or placebo equivalence, when these tests show no statistically significant difference between the two groups. A high p-value from such a test fails to provide statistical evidence that the two groups are different which is only indirectly related to showing that they are the same. The problem arises from the fact that the standard tests are designed to control for a type I false positive error of declaring the two groups different when they are, in fact, the same. If the goal is to control the false positive error rate of declaring the two groups the same when they are in fact different, then the standard test is controlling for the wrong type of error.

We argue that a statistical test showing that two groups are equivalent should control for false positives in which the two groups are called equivalent when they are, in fact, different. When the hypothesis test is correctly specified so that *difference* is the null and *equivalence* is the alternative, the problems afflicting traditional tests are alleviated. Leveraging the statistical literature on tests of equivalence to provide an alternative framework for testing covariate and placebo balance, this chapter provides a new approach to validity checks. Traditional parametric tests are outlined, as well as a set of permutation based tests that leverage the randomized designs of experiments.

This chapter outlines an example, using Brady and McNulty (2011), who argue that some polling stations in Los Angeles were consolidated close to "as if" random by the county registrar. Central to their argument about the quality of their design is that prior to the consolidation, voters in treatment and control precincts had roughly equal "costs of voting", with distance between voters' residence and their polling station being their chief measure of cost. Balance on this variable is critical, yet the authors find that the pre-treatment difference is "highly significant", although "substantively rather small" (p. 5). If the conventional decision rule over adequate balance is followed, then one would reject the "as if" random identification assumption. This case study provides a good example of when tests of difference are inappropriate, such as in this large n, high powered design. We find, using equivalence tests, that their data does provide statistically significant evidence in favor of their design, and that the treatment and control group in this natural experiment are statistically equivalent.

There are many additional applications that would benefit from equivalence based validation tests. This form of balance and placebo test should be come standard practice when providing evidence for the strength of a design in observational studies. Additionally, though, extensions of this method could show how these tests could be used in surveys or data collection methods to justify the representativeness of the sample. Further work will be done on creating and releasing an `R` library to make these types of validity tests readily available to researchers. Finally, additional theoretical work will show how these

tests relate to multiple testing problems, especially given practitioners are evaluating balance on numerous characteristics. Traditional balance tests cannot fit in the multiple testing framework, since they control for the wrong type of error. However, equivalence tests allow practitioners to leverage the multiple testing literature to adjust for tests on multiple covariates.

The equivalence tests outlined in Chapter 3 are also used in the placebo tests outlined in Chapter 2. The use cases for equivalence based placebo and balance tests are highly varied. Practitioners who conduct matching methods would benefit from equivalence based balance tests as a metric for automated algorithms' loss functions, such as GenMatch (Diamond and Sekhon 2013). Additionally, experimentalists would benefit from conducting equivalence placebo tests as further evidence of successful randomizations. While this method aids in the use case of extrapolation methods from experiments to larger populations of interest, it also has far reaching implications for testing the key identifying assumption in most causal analysis methods.

5.1.3 Increasing Representativeness of Survey Data for Externally Valid Estimates

The final chapter of this study turns to the design step of studies. Many studies rely on surveys for collecting outcome data. In addition to understanding the underlying assumptions necessary to provide valid population estimates, Chapter 4 explores methods for collecting representative data that increase the validity of population estimates, both in experimental and observational methods. Chapter 4 explores the impact of unit nonresponse on estimates, following the Holt and Elliot (1991) decomposition of bias. Bias in survey estimates can be decomposed as follows:

$$bias(\bar{y}_R) = \underbrace{\frac{\sum_s S_s \bar{Y}_{sR}(P_{sR} - \bar{P_R})}{\bar{P}_R}}_{\text{Observable Sample Imbalance}} \tag{5.1}$$

$$+ \underbrace{\sum_s S_s P_{sNR}(\bar{Y}_{sR} - \bar{Y}_{sNR})}_{\text{Unobservable Response Difference}} \tag{5.2}$$

where Equation (5.1) refers to observable differences in the sample proportions of the strata as compared to the target population, based on the observable characteristics defined by the strata S_i. Equation (5.2) is due to differences in the average outcome of interest between respondents and non-respondents within strata.

Chapter 4 discusses design phase approaches that emphasize representative final return data, thus minimizing the bias due to Equation (5.1). Following sampling design

approaches, the paper discusses different weighting methods, and addresses their ability to meet the criterion for reducing bias. There are two main innovations in Chapter 4, response rate sampling and a new use of raking weighting methods.

Response rate sampling, discussed in Chapter 4, is a new design stage approach to counteracting the influence of unit non-response. In order to account for unit non-response in the design phase of a survey, previously collected data must be used to estimate the response probability, or response rate, for a group. Estimates of non-response can be obtained in many ways, ranging from estimates at a strata level to individual level estimates. Once estimates of non-response have been obtained, the estimates are used to create a probability sampling frame where each individual is included with probability inversely proportional to their estimated probability of response. This is referred to as response rate sampling. If the estimated probability of response is accurate, then the return data will be representative based on observed characteristics included in the response probability, minimizing the bias in $\bar{y}_R$ due to observable imbalance of the sample.

In addition to outlining the theoretical advantages of response rate sampling, Chapter 4 provides an estimation strategy for calculating response rates in previously collection survey response data, the machine learning algorithm Classification and Regression Trees (CART). CART is used to estimate response rates at a strata level, and incorporating these estimates in a stratified sampling framework. The CART method provides safeguards against over fitting while simultaneously easily providing oversampling rates for defined strata.

In addition to a design based approach for addressing unit non-response, Chapter 4 outlines a new approach to raking weighting methods, raking on stratification cells. This approach balances the trade offs between post-stratification and raking methods, leading to more stable weights while still accounting for a large number of interactions among demographic variables. In practice, null strata are minimized, extreme weights are rare, and an analyst can easily account for many demographic variables, including four-way interactions among key variables. Additionally, analysts are afforded the flexibility of accounting for interactions among some variables, but only margins on variables that are either independent of other variables or for which only margins are known.

The analytical strategies outlined in Chapter 4 are applied to an example, the Obama for America (OFA) 2012 Analytics Polling operation. The OFA analytics polling operation was large, with tens of thousands of calls placed in a dozen battleground states each week beginning in April. The campaign spoke with over 1 million voters across battleground states, collecting information on calls placed to over 10 million voters in the context of random surveys. Given the large nature of the polling operation, and the associated costs, it was important to build a sustainable sampling methodology that leveraged past data to inform the best practices possible for collecting representative samples.

Given the close nature of the election in many of these states, and the limited budget of the campaign, it was also important to maximize the power of the surveys to accurately

measure the opinions of key demographic subgroups so the campaign could maximize its return on investment for engagement methods with voters. Additionally, in the face of the changing dynamics of voter engagement, which impacts the likelihood a voter will respond to a survey, it was important that the campaign be able to discern true changes in voter opinion from changes in response bias, especially in the face of large campaign events such as the first presidential debate and the Republican vice presidential announcement.

The ability to accurately measure changes in opinion, even in the face of differential voter engagement, allowed the campaign to most effectively allocate resources. The OFA analytics polling operation was a good case study of these new methods, showing how careful sampling designs, based on in-cycle data, and novel *post hoc* weighting methods can create an accurate, scalable and cost-effective polling methodology. The campaign was able to accurately predict outcomes, within the margin-of-error, in all of the battleground states in which it was regularly polling.

While this paper addresses the advantages response rate sampling serves in reducing bias due to unit non-response, especially with respect to the bias outlined in Equation (5.1), there is further work to be done on the impact response rate sampling has on variance reduction. Response rate sampling increases the representativeness of the final data, decreasing the emphasis placed on *post hoc* weighting adjustments. This reduces variance, and minimizes the likelihood of extreme weights. However, response rate sampling may also lead to increased confidence in the assumption that data is missing at random (MAR). The MAR assumption says that nonresponse depends only on observed variables (Lohr 2010). Hard to reach groups are overrepresented in the sample, increasing the likelihood that these groups are not represented purely by early, unrepresentative respondents who are given high weights. Further work will explore the impact of response rate sampling on this assumption, and its impact on variance reduction of estimates of opinion. Additionally, further study needs to be done on the relationship between increases in enthusiasm measures and estimates of the target population from which samples are being drawn.

5.2 Concluding Remarks

Incredible progress has been made in the field of causal analysis. From Fisher (1925), who outlined the advantages of experiments and the estimands they identify to the work on the potential outcomes framework (Holland 1986; Rosenbaum 2002; Rubin 1974, 2006) to recent work on understanding causal mediation analysis (Pearl 2010), the field has made leaps and bounds in its ability to allow researchers to identify important and interesting quantities of interest. This dissertation has addressed a growing question: how to move more fluidly between experimental and population estimates. This question has been addressed from a theoretical standpoint by outlining the necessary identifying assumptions. A new estimating strategy for testing the key identifying assumption central to most causal methods, the notion of exchangeability, is outlined. Finally, a new approach to

improving the underlying data from which we explore questions and estimate quantities of interest is provided. This multifaceted approach will allow researchers to improve their ability to estimate causal quantities of interest with more confidence.

Bibliography

Abadie, Alberto and Guido Imbens. 2006. "On the Failure of the Bootstrap for Matching Estimators." Working Paper.

Annan, Jeannie and Christopher Blattman. 2010. "The Consequences of Child Soldiering." *The Review of Economics and Statistics* 92(4):882–898.

Austin, Peter C. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Statistics in Medicine* 27(12):2037–2049.

Banerjee, Abhijit V. and Esther Duflo. 2008. "The Experimental Approach to Development Economics." Department of Economics, MIT.

Barker, Lawrence, Henry Rolka, Deborah Rolka and Cedric Brown. 2001. "Equivalence Testing for Binomial Random Variables: Which Test to Use?" *The American Statistician* 55(4):279–287.
URL: *http://www.jstor.org/stable/2685688*

Benjamini, Yoav and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57(1):289–300.

Berger, Roger and Jason Hsu. 1996. "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets." *Statistical Science* 11(4):283–302.
URL: *http://www.jstor.org/stable/2246021*

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s The Analysis of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70(4):499–529.

Bethlehem, Jelke G. 2002. Weighting nonresponse adjustments based on auxiliary information. In *Survey nonresponse*, ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little. Wiley.

Bickel, Peter J. and Anat Sakov. 2008. "On the Choice of m in the m out of n Bootstrap and Confidence Bounds for Extrema." *Statistica Sinica* 18:967–985.

Bickel, Peter J. and David A Freedman. 1984. "Asymptotic Normality and the Bootstrap in Stratified Sampling." *Annals of Statistics* 12(2):470–482.

Bowers, Jake. 2011. Making Effects Manifest in Randomized Experiments. In *Cambridge Handbook of Experimental Political Science*. Cambridge University Press.

Brady, Henry and John McNulty. 2011. "Turning Out to Vote: The Costs of Finding and Getting to the Polling Place." *American Political Science Review* 105(1):1–20.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone and R.A. Olshen. 1984. *Classification and Regression Trees*. New York: Chapman & Hall.

Chattopadhyay, Raghabendra and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72(5):1409–1443.

Chen, J., R. R. Sitter and C. Wu. 2002. "Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys." *Biometrika* 89(1):230–237.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1):266–298.

Chittock, Dean R, Vinay K Dhingra, Juan J Ronco, James A Russell, Dave M Forrest, Martin Tweeddale and John C Fenwick. 2004. "Severity of illness and risk of death associated with pulmonary artery catheter use." *Critical Care Medicine* 32:911–915.

Cochran, William and Donald Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya: The Indian Journal of Statistics* 35(4):417–446.

Cole, Stephen R and Constantine E Frangakis. 2009. "The consistency statement in causal inference: a definition or an assumption?" *Epidemiology* 20(1):3–5.

Cole, Stephen R and Elizabeth A Stuart. 2010. "Generalizing Evidence From Randomized Clinical Trials to Target Populations The ACTG 320 Trial." *American journal of epidemiology* 172(1):107–115.

Connors, Alfred F, Theodore S Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, William J Fulkerson, Humberto Vidaillet, Steven Broste, Paul Bellamy, Joanne Lynn and William A Knaus. 1996. "The effectiveness of right heart catheterization in the initial care of critically ill patients." *Journal of the American Medical Association* 276:889–897.

Dalen, James E. 2001. "The Pulmonary Artery Catheter—Friend, Foe, or Accomplice?" *Journal of the American Medical Association* 286:348–350.

Deaton, Angus. 2009. Instruments of Development: Randomization in the tropics, and the search for the elusive keys to economic development. NBER Working Paper 14690 National Bureau of Economic Research.

Deming, W. Edwards and Frederick F. Stephan. 1940. "On a Least Squares Adjustment

of a Sampled Frequency Table When the Expected Marginal Totals are Known." *The Annals of Mathematical Statistics* 11(4):427–444.

Deville, Jean-Claude and Carl-Erik Sarndal. 1992. "Calibration estimators in survey sampling." *Journal of the American Statistical Association* 87(418):376–382.

Deville, Jean-Claude, Carl-Erik Sarndal and Olivier Sautory. 1993. "Generalized raking procedures in survey sampling." *Journal of the American Statistical Association* 88(423):1013–1020.

Di Nardo, J.E. and J.S. Pischke. 1997. "The Returns to Computer Use Revisted: Have Pencils Changed the Wage Structure Too?" *The Quarterly Journal of Economics* .

Di Tella, Rafael, Sebastian Galiani and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* pp. 209–241.

Diamond, Alexis and Jasjeet S. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95(3):932–945.

Dunning, Thad. 2010*a*. Design-Based Inference: Beyond the Pitfalls of Regression Analysis? In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. David Collier and Henry Brady. 2nd ed. Rowman & Littlefield.

Dunning, Thad. 2010*b*. "Do Quotas Promote Ethnic Solidarity? Field and Natural Experimental Evidence from India.".

Elliott, Michael R. and Roderick J.A. Little. 2000. "Model-Based Alternatives to Trimming Survey Weights." *Journal of Official Statistics* 16(3):191–209.

Ericson, W. A. 1965. "Optimum Stratified Sampling Using Prior Information." *Journal of the American Statistical Association* 60(311):750–771.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Releaed Audits on Electoral Outcomes." *Quarterly Journal of Economics* pp. 703–745.

Finfer, Simon and Anthony Delaney. 2006. "Pulmonary artery catheters as currently used, do not benefit patients." *British Medical Journal* 333:930–1.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Folsom, Ralph E. and Avi C. Singh. 2000. "he generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification." *Proceedings of the American Statistical Association, Survey Research Methods Section* .

Freedman, David A. 2006. "Statistical Models for Causation: What Inferential leverage

Do They Provide?" *Evaluation Review* 30:691–713.

Freedman, David A and Richard A Berk. 2008. "Weighting Regressions by Propensity Scores." *Evaluation Review* 32(4):392–409.

Gheorghe, Adrian, Tracy E. Roberts, Jonathan C. Ives, Benjamin R. Fletcher and Melanie Calvert. 2013. "Centre Selection for Clinical Trials and the Generalisability of Results: A Mixed Methods Study." *PLOS ONE* 8(2).

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Large-Scale Experiments Using Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3):491–511.

Greenhouse, Joel B, Eloise E Kaizar, Kelly Kelleher, Howard Seltman and William Gardner. 2008. "Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users." *Statistics in medicine* 27(11):1801–1813.

Greenland, Sander. 2005. "Multiple-bias modelling for analysis of observational data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(2):267–306.

Groves, Robert M., Don A Dillman, John L Eltinge and Roderick JA Little, eds. 2002. *Survey Nonresponse*. Wiley.

Hansen, Ben B. 2008. "The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin,Statistics in Medicine." *Statistics in Medicine* 27(12):2050–2054.

Hansen, Ben B. and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23(2):219–236.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50(4):pp. 1029–1054.
URL: *http://www.jstor.org/stable/1912775*

Harrison, David A, Anthony R Brady and Kathy Rowan. 2004. "Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database." *Critical Care* 8:R99–111.

Hartman, Erin K. and F. Daniel Hidalgo. 2011. "What's the Alternative?: An equivalence approach to Placebo and Balance Tests." Working Paper.

Harvey, SE, CA Welch, DA Harrison and MA Singer. 2008. "Post hoc insights from PAC-Man—the UK Pulmonary artery catheter trial." *Critical Care* 35(6):1714–1721.

Harvey, Sheila, David A Harrison, Mervyn Singer, Joanne Ashcroft, Carys M Jones, Di-

ana Elbourne, William Brampton, Dewi Williams, Duncan Young and Kathryn Rowan. 2005. "An assessment of the clinical effectiveness of pulmonary artery catherters in patient management in intensive care (PAC-Man): a randomized controlled trial." *Lancet* 366:472–77.

Heckman, James J. 1992. Randomization and Social Policy Evaluation. In *Evaluating Welfare and Training Programs*, ed. Charles Manski and I. Garfinkel. Cambridge, MA: Harvard University Press.

Heckman, James J. 1996. "Randomization as an Instrumental Variable." *The Review of Economics and Statistics* 78(2):336–341.

Heckman, James J and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Heckman, James J, Hidehiko Ichimura, Jeffery Smith and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5):1017–1098.

Heckman, James J and Sergio Urzua. 2009. Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. IZA Discussion Papers 3980 Institute for the Study of Labor (IZA).

Hellerstein, Judith K and Guido W Imbens. 1999. "Imposing moment restrictions from auxiliary data by weighting." *Review of Economics and Statistics* 81(1):1–14.

Ho, D E, K Imai, G King and E A Stuart. 2006. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Ho, Daniel and Kosuke Imai. 2008. "Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: The California Alphabet Lottery, 1978-2002." *Public Opinion Quarterly* 72(2):216–240.

Hoch, Jeff S, Andrew H Briggs and Andy R Willan. 2002. "Something old, something new, something borrowed, something BLUE: A framework for the marriage of econometrics and cost-effectiveness analysis." *Health Economics* 11:415–430.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Holt, D and D Elliot. 1991. "Methods of Weighting for Unit Non-Response." *Journal of the Royal Statistical Society. Series D (The Statistician)* 40(3):333–342.

Holt, D and TMF Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society. Series A (General)* 142(1):33–46.

Hyde, Susan D. 2008. "The Observer Effect in International Politics: Evidence from a Natural Experiment." *World Politics* 60(1):37–63.

Imai, Kosuke, Dustin Tingley and Teppei Yamamoto. 2013. "Experimental Design for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society, Series A* 176(1):5–51.

Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2008. "Misunderstandings among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171(2):481–502.

Imbens, Guido. 2009. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). NBER Working Paper 14896 National Bureau of Economic Research.

Imbens, Guido and Tony Lancaster. 1996. "Efficient estimation and stratified sampling." *Journal of Econometrics* 74(2):289–318.

Ireland, C. T. and S. Kullback. 1968. "Contingency Tables with Given Marginals." *Biometrika* 55(1):pp. 179–188.
URL: *http://www.jstor.org/stable/2334462*

Jaynes, Edwin T. 1957. "Information Theory and Statistical Mechanics." *Physical Reviews* 106(4):620–630.

Kalton, Graham and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81–97.

Kang, Joseph D. Y. and Joseph L. Schafer. 2007. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from imcomplete data (with discussion)." *Statistical Science* 22:523–539.

Kapur, J. N. and H. K. Kesavan. 1992. *Entropy Optimization Principles with Applications*. Academic Press, Inc.

Keele, L. 2010. "How Much is Minnesota Like Wisconsin? States as Counterfactuals." *Unpublished paper The Ohio State . . .* .

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves and Stanley Presser. 2000. "Consequences of reducing nonresponse in a national telephone survey." *Public Opinion Quarterly* 65(2):125–148.

Kish, Leslie. 1992. "Weighting for unequal P_i." *Journal of Official Statistics* 8:183–200.

Kline, Brendan and Elie Tamer. 2011. "Using Observational vs. Randomized Controlled Trial Data to Learn About Treatment Effects." Available at SSRN: http://ssrn.com/abstract=1810114 or http://dx.doi.org/10.2139/ssrn.1810114.

Kott, Phillip S. 2006. "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology* 32(2):133.

Krueger, AB. 2000. "JSTOR: The American Economic Review, Vol. 90, No. 5 (Dec.,

2000), pp. 1397-1420." *American Economic Review* .

Kullback, Solomon. 1997. *Information theory and statistics*. New York: John Wiley.

Lee, DS. 2008. "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* .

Lehmann, Erich L. 1975. *Nonparametrics*. Springer.

Lehtonen, R. I. S. T. O. and Ari Veijanen. 1998. "Logistic generalized regression estimators." *Survey Methodology* 24:51–56.

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

Little, Roderick JA and Mei-Miau Wu. 1991. "Models for Contingency Tables with Known Margins when Target and Sampled Populations Differ." *Journal of the American Statistical Association* 86(413):87–95.

Lohr, Sharon L. 2010. *Sampling: Design and Analysis*. Second ed. Brooks/Cole.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology Wiley.

Lunceford, Jared K. and Marie Davidian. 2004. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in medicine* 23(19):2937–2960.

Lyall, J. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks?: Evidence from Chechnya." *Journal of Conflict Resolution* 53(3):331–362.

Mattos, Rogerio and Alvaro Viega. 2004. "Entropy Optimization: Computer implementation of the maxent and minxent principles." Working Paper.

McDonald, Paul, Matt Mohebbi and Brett Slatkin. N.d. "Comparing Google Consumer Surveys to Existing Probability and Non-Probability Based Internet Surveys." Google, Inc White Paper.

Miratrix, Luke W., Jasjeet S. Sekhon and Bin Yu. 2013. "Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments." *Journal of the Royal Statistical Society, Series B* 75(2):369–396.

Mitra, Nandita and Alka Indurkhya. 2005. "A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data." *Health Economics* 14:805–815.

Mojtabai, Ramin and Joshua G Zivin. 2003. "Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis." *Health Services Research* 38:233–59.

Neyman, Jerzy. 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.

Nixon, Richard M and Simon G Thompson. 2005. "Incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations." *Health Economics* 14:1217—1229.

Pearl, Judea. 2010. "Causal Analysis in Theory and Practice: On Mediation, counterfactuals and manipulations." Online reply to Robins and Thomas. `http://www.mii.ucla.edu/causality/wp-content/uploads/2010/05/mediation-part1.pdf`.

Pickery, Jan and Ann Carton. 2008. "Oversampling in relation to differential regional response rates." *Survey Research Methods* 2(2):83–92.

Politis, Dimitrius N. and Joseph P. Romano. 1994. "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions." *The Annals of Statistics* 22:2031–2050.

Porter, Kristin E., Susan Gruber, Mark J. van der Laan and Jasjeet S. Sekhon. 2011. "The Relative Performance of Targeted Maximum Likelihood Estimators." *The International Journal of Biostatistics* 7(1).
URL: *http://www.bepress.com/ijb/vol7/iss1/31/*

Robins, J. M., A. Rotnitzky and L. P. Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90(429):106–121.

Rosenbaum, Paul R. 1998. Propensity Score. In *Encyclopedia of Biostatistics*. Wiley.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York: Springer-Verlag.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.

Rothwell, Peter M. 2005. "External validity of randomised controlled trials:"to whom do the results of this trial apply?"." *The Lancet* 365(9453):82–93.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(6):688–701.

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, Donald B. 2008. "For objective causal inference, design trumps analysis." *The Annals of Applied Statistics* 2(3):808–840.

Sakr, Yasser, Jean-Louis Vincent, Konrad Reinhart, Didier Payen, Christian J Wiedermann, Durk F Zandstra and Charles L Sprung. 2005. "Sepsis Occurrence in Acutely

Ill Patients Investigators. Use of the pulmonary artery catheter is not associated with worse outcome in the ICU." *Chest* 128(4):2722–31.

Schickler, Eric, Adam Berinsky and Jasjeet S. Sekhon. 2012. "Collaborative Research: The American Mass Public in the Early Cold War Years." National Science Foundation, Political Science Program Grant SES - 1155143.

Sekhon, Jasjeet S. 2007. "Alternative balance metrics for bias reduction in matching methods for causal inference." *Working Paper* .

Sekhon, Jasjeet S. 2011. "Matching: Multivariate and Propensity Score Matching with Automated Balance Search." *Journal of Statistical Software* 42(7):1–52. Computer program available at *http://sekhon.berkeley.edu/matching/*.

Sekhon, Jasjeet S. and Richard Grieve. 2012. "A Nonparametric Matching Method for Covariate Adjustment with Application to Economic Evaluation (Genetic Matching)." *Health Economics* 21(6):695–714.

Sekhon, Jasjeet Singh. 2010. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12:487–508.

Shadish, William R, Thomas D Cook and Donald Thomas Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *The Bell Systems Technical Journal* 27:379–423, 623–656.

Stevens, K, C McCabe, C Jones, J Ashcroft, S Harvey and K Rowan. 2005. "The incremental cost effectiveness of withdrawing pulmonary artery catheters from routine use in critical care." *Appl Health Econ Health Policy* 4(4):257–264. On behalf the PAC-Man Study Collaboration.

Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw and Philip J. Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society, Series A* 174(2):369–386.

Tunis, Sean R, Daniel B Stryer and Carolyn M Clancy. 2003. "Practical clinical trials." *JAMA: the journal of the American Medical Association* 290(12):1624–1632.

Turner, Rebecca M, David J Spiegelhalter, Gordon Smith and Simon G Thompson. 2009. "Bias modelling in evidence synthesis." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172(1):21–47.

van der Laan, Mark J., Eric C. Polley and Alan E. Hubbard. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6(1):1–21.

Wellek, S. 1996. "A New Approach to Equivalence Assessment in Standard Comparative

Bioavailability Trials by Means of the Mann-Whitney Statistic." *Biometrical Journal* 38(6):695–710.
URL: *http://dx.doi.org/10.1002/bimj.4710380608*

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press.

Westlake, WJ. 1976. "Symmetrical confidence intervals for bioequivalence trials." *Biometrics* .

Willan, Andrew R and Andrew H Briggs. 2006. *Statistical Analysis of Cost-Effectiveness Data*. Wiley.

Willan, Andrew R, Andrew H Briggs and Jeffrey S Hoch. 2004. "Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data." *Health Economics* 13:461–475.

Willan, Andrew R. and D. Y. Lin. 2001. "Incremental net benefit in randomized clinical trials." *Statistics in Medicine* 20(11):1563–1574.

Willan, Andrew R., Eric Bingshu Chen, Richard J. Cook and D. Y. Lin. 2003. "Incremental net benefit in randomized clinical trials with quality-adjusted survival." *Statistics in Medicine* 22(3):353–362.

Chapter A

Chapter 2 Appendix

A.1 Proof of Theorem 1

Proof. From (2.2) and (2.4)

$$\begin{aligned}\mathbb{E}(Y_{i01}|S_i=0,T_i=1) &= \mathbb{E}(Y_{i11}|S_i=0,T_i=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=1,T_i=1)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^T,S_i=1,T_i=1)\}.\end{aligned} \tag{A.1}$$

From (2.3) and (2.5)

$$\begin{aligned}\mathbb{E}(Y_{i00}|S_i=0,T_i=1) &= \mathbb{E}(Y_{i10}|S_i=0,T_i=1)\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=1,T_i=0)\}\\ &= \mathbb{E}_{01}\{\mathbb{E}(Y_i|W_i^{C_T},S_i=1,T_i=0)\}.\end{aligned} \tag{A.2}$$

The result follows by substituting Eqs. (A.1) and (A.2) in the quantity of interest τ_{PATT} in Equation (2.1). Without strong ignorability of sample assignment, from (2.6),

$$\begin{aligned}&\mathbb{E}(Y_{i01}|S_i=0,T_i=1)-\mathbb{E}(Y_{i00}|S_i=0,T_i=1)\\ &=\mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=0,T_i=1)\}-\mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=0,T_i=1)\}\\ &=\mathbb{E}_{01}\{\mathbb{E}(Y_{i11}|W_i^T,S_i=1,T_i=1)\}-\mathbb{E}_{01}\{\mathbb{E}(Y_{i10}|W_i^{C_T},S_i=1,T_i=1)\},\end{aligned}$$

and the result follows from randomization. □

A.2 Identifiability of Alternative Causal Quantities

The main population treatment effect considered in this paper is PATT, however there are numerous population treatment effects that policy makers might be interested in. If PATC is of interest then assumptions 2 and 3 can be replaced by

$$(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i^T, T_i = 0) \qquad \text{and} \qquad (Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i^{C_T}, T_i = 0), \tag{A.3}$$

respectively. Additionally, if Equation (2.3) in assumption 1, $(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | (W_i^{C_T}, T_i = 0)$ and assumption 4 hold then $\mathbb{E}(Y_i | S_i = 0, T_i = 0) = \mathbb{E}_{00}\{\mathbb{E}(Y_i | W_i^{C_T}, S_i = 1, T_i = 0)\}$. Therefore the mean outcomes would be the same for the control group in the target population and the adjusted RCT, adjusted such that $W_i^{C_T}$ follows its distribution in the target control group. A placebo test can then be used to check the validity of the required assumptions. However, this is not necessary to apply Theorem 1 because (A.3) is not assumed in the current analysis.

In circumstances where the estimand of interest is the PATE then the estimand of interest is the effect in the entire target population, where $\tau_{PATE} = \mathbb{E}(Y_{i01} - Y_{i00} | S_i = 0)$. In such a case, assumptions 1–4, as well as $(Y_{i01}, Y_{i11}) \perp\!\!\!\perp S_i | (W_i^{C_T}, T_i = 0)$ and $(Y_{i00}, Y_{i10}) \perp\!\!\!\perp S_i | W_i^{C_T}, T_i = 0$, are sufficient for identification. Assuming $W_i^T = W_i^{C_T}$, these assumptions and randomization imply that $Y_{ist} \perp\!\!\!\perp (S_i, T_i) | W_i^T$, which means that the potential outcomes for units with the same W_i^T are exchangeable, regardless of whether they are assigned to treatment or control and whether they are in the target population or RCT. Under these assumptions and randomization, it can be shown that

$$\mathbb{E}(Y_{ist} | S_i = 0) = \mathbb{E}_{W_i^{C_T} | S_i = 0}\{\mathbb{E}(Y_i | W_i^{C_T}, S_i = 1, T_i = t)\} \tag{A.4}$$

$$\tau_{PATE} = \mathbb{E}_{W_i^{C_T} | S_i = 0}\{\mathbb{E}(Y_i | W_i^{C_T}, S_i = 1, T_i = 1) - \mathbb{E}(Y_i | W_i^{C_T}, S_i = 1, T_i = 0)\}, \tag{A.5}$$

for $t = 0, 1$. Equation A.4 implies that the mean outcome in the target population is the same as in the adjusted RCT $T_i = t$ group, adjusted such that $W_i^{C_T}$ follows its distribution in the target population. Equation A.5 implies that the results from an adjusted RCT can be used to identify PATE for a target population. The analysis of Stuart et al. (2011) makes the stronger assumptions required to justify Equations A.4 and A.5. Stuart et al. (2011) verify the assumptions by confirming the validity of Equation A.4, for $t = 0$, which then justifies the generalizability of the RCT results, from Equation A.5. It is possible for Equation (A.4) to be violated and Equation (A.5) to still hold. This occurs if treatment consistency is violated but the potential outcomes in the target population and RCT differ by some constant.

The assumptions required for Equation A.4 can be checked by a placebo test of the mean outcome in the target population and the adjusted RCT $T_i = t$ group, for $t = 0, 1$. However, since the assumptions used in the analysis here are weaker and do not imply Equation A.4, such a test is not done.

A.3 Equivalence Tests

Equivalence tests begin with the null hypothesis:

$$H_0 : \frac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} \geq \epsilon_U \qquad \text{or} \qquad \frac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} \leq \epsilon_L$$
$$\text{versus}$$
$$H_1 : \epsilon_L < \frac{\mu_{\text{adj samp}} - \mu_{\text{pop}}}{\sigma} < \epsilon_U$$

where $\mu_{\text{adj samp}}$ is the true mean of the reweighted sample treated and μ_{pop} is the true mean of the populated treated, and σ is the pooled standard deviation of the two groups. We define $\epsilon_L = 0.2$ and $\epsilon_U = 0.2$, as discussed above. The test uses the test statistic

$$T = \frac{\sqrt{mn(N-2)/N}(\bar{X}_{\text{adj samp}} - \bar{X}_{\text{pop}})}{\left\{ \sum_{i=1}^{m}(X_{\text{adj samp}i} - \bar{X}_{\text{adj samp}})^2 + \sum_{j=1}^{n}(X_{\text{pop}j} - \bar{X}_{\text{pop}})^2 \right\}^2}$$

where $\bar{X}_{\text{adj samp}}$ is the observed mean of the reweighted sample treated, $\bar{X}_{\text{pop}}$ is the observed mean of the population treated, standardized by the observed standard deviation. m refers to the number of observations in the reweighted sample, and n to the number of observations in the population treated, and $N = m + n$. The test rejects the null of non-equivalence if:

$$|T| < C_{\alpha;m,n}(\epsilon)$$
$$\text{with}$$
$$C_{\alpha;m,n}(\epsilon) = F^{-1}(\alpha; df_1 = 1, df_2 = N - 2, \lambda_{nc}^2 = mn\epsilon^2/N)^{\frac{1}{2}}$$

where $C_{\alpha;m,n}(\epsilon)$ is the square root of the inverse F distribution with level α, degrees of freedom $1, N-2$, and non-centrality parameter $\lambda_{nc}^2 = mn\epsilon^2/N$. One important aspect of equivalence testing is that it requires the definition of a range over which observed differences are considered substantively inconsequential. We follow the recommendations of Hartman and Hidalgo (2011), and define equivalence as a mean difference between the reweighted sample treated and the true population treated of no more than 0.2 standardized differences and use the t-test for equivalence defined in Wellek (2010).

A.4 Maximum Entropy Weighting

The principle of maximum entropy is defined as:

$$\max_{\mathbf{p}} S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i \tag{A.6}$$

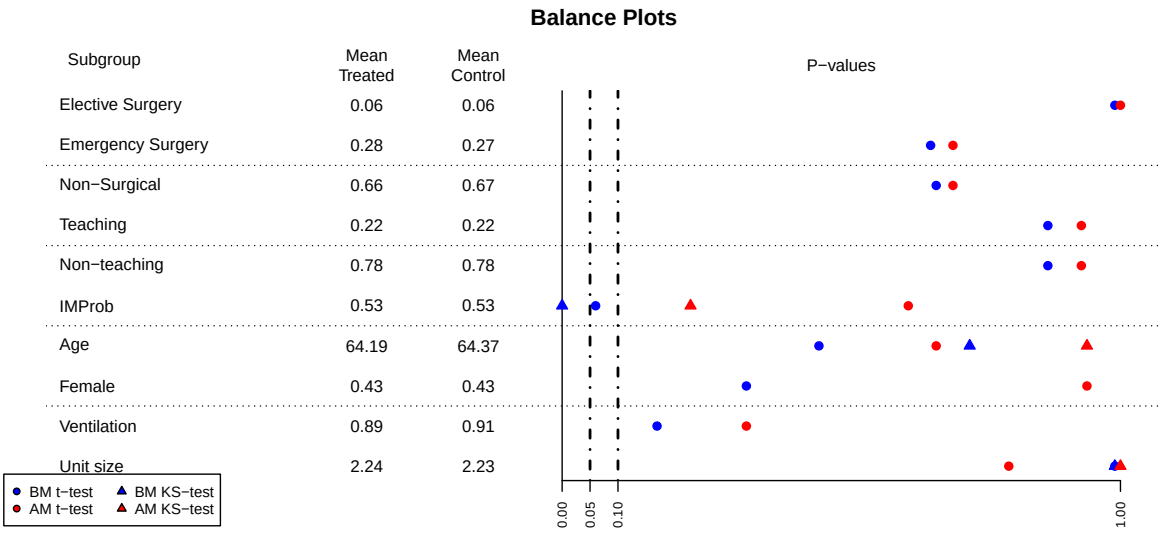
$$s.t. \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_r(x_i) = \sum_{i=1}^{n} p_i g_{ri} = a_r \qquad r = 1, \ldots, m \\ p_i \geq 0 \qquad i = 1, 2, \ldots, n \end{cases} \tag{A.7}$$

where equation (A.6) maximises Shanon's measure of entropy, which is a form of probabilistic uncertainty. The first constraint in equation (A.7) is referred to as the natural constraint, and it simply states that all the probabilities must sum to one. The m moment constraints are referred to as the consistency constraints. Each a_r represents an r-th order moment, or characteristic moment, of the probability distribution (i.e. $g_{ri} = (x_i - \mu)^r$ where μ is the distribution mean). The distribution chosen for $\mathbf{p}$ is that most similar to the uniform that still satisfies the constraints .[1] In this context this ensures that individuals in the treatment group in the RCT who have identical values for all of the covariates used in the constraints are given equal weights. Here, the matched pairs from the RCT are reweighted using constraints from the target population as, for example represented by the NRS. The consistency constraints are constructed using moments such as the covariate means from a NRS. Typically this is done using covariate data contained in both the NRS and the RCT, however information from several external sources (e.g. disease registries) about the target population can also be incorporated into the constraints. Once the consistency constraints have been created, a set of weights that simultaneously satisfies the constraints while maximizing the entropy measure is calculated. PATT can then be reported by weighting the SATT for each of the individual matched pairs.

[1] Due to the fact that there are $m + n$ equations and $m + n$ unknowns, corresponding to m Lagrange multipliers and n probabilities, it is not possible to derive an analytical solution for p_i and λ_r simultaneously using only the known moments. A solution must be found using an iterative search algorithm (Mattos and Viega 2004).

A.5 GenMatch Balance in RCT

Figure A.1: Covariate Balance in the RCT



A.6 Covariates used in Matching and Reweighting

Table A.1: Covariates used in GenMatch

GenMatch Covariates	
Priority (balance enforced to be no worse than initial balance)	Age, baseline probability of death, elective surgery indicator, emergency surgery indicator, size of ICU, teaching hospital indicator, mechanical ventilator at admission, base excess
Additional Covariates used for Matching	physiology score, admission diagnosis, gender, past medical history variables on cardiac, respiratory, liver, and immune measures, heart rate and blood pressure physiology measures, temperature measures, respiratory measures
Additional covariates used for Balance	admission diagnosis, blood gas rate, Pf rate, Ph, Creatinine, Sodium, Urine output, white blood cell counts, Glasgow coma, cardiac and respiratory measures of organ failure, indicator for sedation or paralyzation, baseline PAC rate in unit, geographical region, APACHE II probability of death, indicators for missing values

Table A.2: Covariates used in MaxEnt

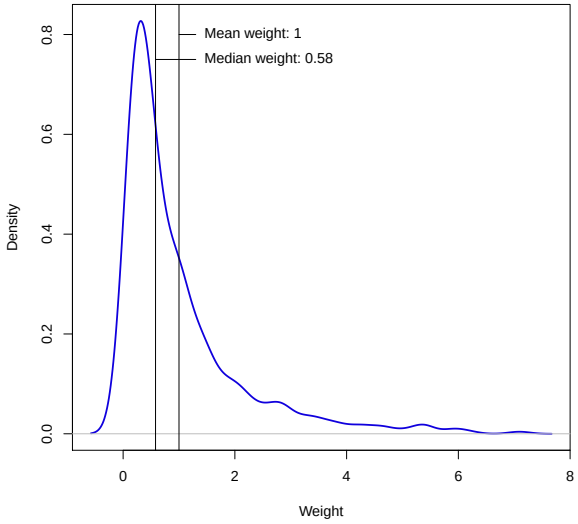
MaxEnt Margins
age, elective surgery indicator, emergency surgery indicator, teaching hospital indicator, gender, baseline probability of death, mechanical ventilator at admission, chemical measure of decline, past medical history variables on cardiac, respiratory, liver, and immune measures, categorical variables on blood pressure rates, categorical measures on temperature, geographical region, categorical variables on age (0-56, 57-66, 67+), categorical classification of diagnostic variable, categorical classification of base excess, base excess categories × age categories, unit size × teaching hospital indicator, teaching hospital indicator × base excess categories, mechanical ventilation × base excess categories, teaching hospital indicator × mechanical ventilator at admission, unit size × mechanical ventilator at admission, gender × teaching hospital, teaching hospital × age categories, gender × age categories, emergency surgery indicator × gender, elective surgery indicator × gender, teaching hospital indicator × past medical history variables on cardiac, respiratory, liver, and immune measures, age categories × base excess × gender, gender × past medical history variables on cardiac and respiratory measures, mechanical ventilation at admission × pasty medical history variables on cardiac, respiratory, and renal measures

Table A.3: Covariates used in IPSW Estimation

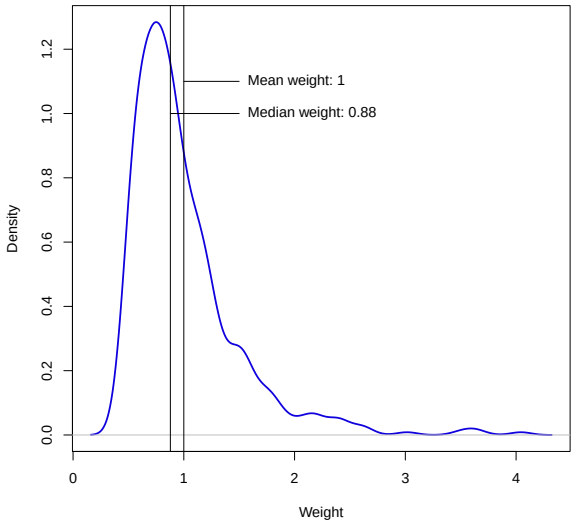
Propensity Score Covariates
gender, age, categorical age variables, elective surgery indicator, emergency surgery indicator, past medical history variables on cardiac, respiratory, liver, and immune measures, categorical diagnostic variable, chemical decline variable, base excess categorical variables, heart rate categorical variables, blood pressure categorical variables, temperature categorical variables, blood gas rate categorical variables, Pf rate categorical variables, Ph categorical variables, Creatinine categorical variables, Sodium categorical variables, Urine output categorical variables, white blood cell counts categorical variables, Glasgow coma categorical variables, cardiac and respiratory measures of organ failure, mechanical ventilation at admission, unit size categorical variable, teaching hospital indicator

A.7 Weights

Figure A.2: Weight Distributions



(a) Maxent Weights



(b) IPSW Weights

Chapter B

Chapter 3 Appendix

B.1 Two-One-Sided Test and Intersection Union Tests

The Two-One-Sided Test (TOST), a type of intersection union test, is the most commonly used equivalence test for studying bioequivalence. It is the test recommended by the FDA, for instance. Intersection union tests are a way of testing multiple hypotheses at once. They are set up in the following manner:

$$H_0 : \theta \in \cup_{i=1}^{k} \Theta_i \qquad \text{versus} \qquad H_1 : \theta \in \cap_{i=1}^{k} \Theta_i^c \tag{B.1}$$

where θ is the parameter if interest and Θ is the parameter space. The overall null hypothesis, H_0 is rejected at the α level if all of the individual null hypotheses, H_{0i}, are rejected and the α level. Note that this can be a conservative test, depending on how the rejection region for the combined test is determined (Berger and Hsu 1996). The typical TOST t-test is a type of intersection union test in which the hypotheses are set up as:

$$H_0 : \mu_T - \mu_C \geq \epsilon_U \cup \mu_T - \mu_C \leq \epsilon_L \qquad \text{versus} \qquad H_1 : \epsilon_L < \mu_T - \mu_C < \epsilon_U \tag{B.2}$$

A t-test is conducted for both of the null hypotheses, i.e. a test one sided test for $\mu_T - \mu_C \geq \epsilon_U$ and a one sided test for $\mu_T - \mu_C \leq \epsilon_L$. The overall null hypothesis is rejected at level α if the associated p-value for each of the individual hypotheses is less than α. Commonly, the null hypothesis is defined in terms of the ratio of μ_T and μ_C, thus making the hypotheses of the form:

$$H_0 : \frac{\mu_T}{\mu_C} \geq \epsilon_U \cup \frac{\mu_T}{\mu_C} \leq \epsilon_L \qquad \text{versus} \qquad H_1 : \epsilon_L < \frac{\mu_T}{\mu_C} < \epsilon_U \tag{B.3}$$

This test, using the ratios, is used frequently to test the bioequivalence of generic drugs versus non-generic drugs in medicine. In that case, the ϵs are chosen as $\epsilon_U = 1.25$ and $\epsilon_L = 0.8$, the current standard of the FDA. Setting up the hypotheses as a ratio has advantages such as putting the metric of difference on an absolute scale instead of on the scale of the variable. Berger and Hsu (1996) show that the ratio test is also conducted

using a t-test, however the test statistic is adjusted as such:

$$T_L = \frac{\bar{X_T} - \epsilon_L \bar{X_C}}{S\sqrt{1/m + \epsilon_L^2/n}} \qquad T_U = \frac{\bar{X_T} - \epsilon_U \bar{X_C}}{S\sqrt{1/m + \epsilon_U^2/n}} \tag{B.4}$$

The overall null hypothesis is rejected $T_L \geq t_{\alpha,m+n-2}$ and $T_L \leq -t_{\alpha,m+n-2}$.

B.2 Exact Fisher Binomial Test for Equivalence

In the case of a binary treatment with a binary response, an equivalence can take the form of an Exact Fisher type test. In this case, the data can be described by the percentage of units taking on value 1 in either the treatment or control condition. We will call the rate of units with a response value of 1 in the treatment condition p_T and the rate of units with a response value of 1 in the control condition p_C. The test statistic is the odds ratio of the two groups, $\rho = p_T(1-p_T)/p_C(1-p_C)$, the advantages of which are discussed in Section 3.3. The hypothesis using the odds ratio as the test statistic is then set up as:

$$H_0 : 0 < \rho \leq \epsilon_L \text{ or } \epsilon_U \leq \rho < \infty \qquad \text{versus} \qquad H_1 : \epsilon_L < \rho < \epsilon_U \tag{B.5}$$

with $\epsilon_L < 1 < \epsilon_U$. The optimal solution to this test is based on R.A. Fisher's exact test for the homogeneity of two binomial distributions, based on the conditional distribution of the odds ratio sum of the number of successes in the treated and control groups. The distribution of this test statistic follows an extended hypergeometric distribution (Wellek 2010). For simplicity, assume that the sample sizes are the same and that the ϵs are chosen symmetric around 1. The test rejects the null hypothesis of non-equivalence if the associated p-value of the test statistic is less than the α level of the test, where the p-value is calculated as:

$$p_{n,\epsilon}(x|s) = \sum_{j=s-\max(x,s-x)}^{\max(x,s-x)} h_s^{n,n}(j;\epsilon) \tag{B.6}$$

with

$$h_s^{n,n}(x;\epsilon) = \frac{\binom{n}{x}\binom{n}{s-x}\epsilon^x}{\sum_{j=\max(0,s-n)}^{\min(s,n)}\binom{n}{j}\binom{n}{s-j}\epsilon^j} \qquad , \max(0, s-n) \leq x \leq \min(s,n) \tag{B.7}$$

Wellek (2010) outlines the rejection rule in the case of unequal sample size and/or a non-symmetric equivalence range. We have implemented these scenarios in our accompanied `R` package, but the intuition behind the test is the same. In the case of binary data, multiple tests for testing the equivalence of the probabilities of success of the two groups instead of the odds ratio are also discussed in Barker et al. (2001). Most of these tests are based on the $100(1-\alpha)$ confidence interval of the t-test, which corresponds to a TOST t-test.

B.3 Mann-Whitney Test for Equivalence

Sometimes a test for equivalence of the means of the distribution is not sufficient to prove equivalence between the two groups. Instead, we may wish to test for equivalence of the distributions of two groups. A non-parametric equivalence test for comparing two continuous distributions, F and G, is based on the Mann-Whitney statistic. This test is asymptotically distribution free and robust to outliers in the data (Wellek 2010). Failure to reject the null of nonequivalence in this test implies not simply that the two groups differ in their means, but is designed to test for departures in other parts of the distribution as well. Lehmann (1975) originally outlined the properties of the U-statistic, and Wellek (2010) further discusses the implementation for equivalence testing. The basic outline of the test is as follows. Let $X_{Ti} \sim F \ \forall i = 1, \ldots, m$ and $X_{Cj} \sim G \ \forall j = 1, \ldots, n$, then they equivalence hypothesis for the non-parametric test can be set up as:

$$H_0 : \pi_+ \leq 1/2 - \epsilon_1 \text{ or } \pi_+ \geq 1/2 + \epsilon_2 \qquad \text{versus} \qquad H_1 : 1/2 - \epsilon_1 < \pi_+ < 1/2 + \epsilon_2 \quad \text{(B.8)}$$

where $\pi_+ = P[X_{Ti} > X_{Cj}]$. Here, π_+ is estimated using the Mann-Whitney statistic, W_+ defined as:

$$W_+ = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}(X_{Ti} - X_{Cj}) \tag{B.9}$$

Intuitively, if the two samples are equivalent, then the chance that any given treated unit's value of X_{Ti} lies above any given control unit's value X_{Cj} is about one half. The epsilons, then define the tolerance around one half for which the two groups would still be equivalent. The hypothesis test is thus set up with a null hypothesis that $P[X_{Ti} > X_{Cj}]$ is either smaller or larger than the range of equivalence, and the alternative is that $P[X_{Ti} > X_{Cj}]$ lies within the range of equivalence. The statistical test is carried out with the following rejection rule:

$$\text{Reject nonequivalence iff} \qquad \frac{\left|W_+ - 1/2 - \frac{\epsilon_1 - \epsilon_2}{2}\right|}{\hat{\sigma}[W_+]} < C_{MW}(\alpha; \epsilon_1, \epsilon_2) \tag{B.10}$$

where

$$C_{MW}(\alpha; \epsilon_1, \epsilon_2) = \chi^{2^{-1}}(\alpha; df = 1, \lambda^2_{nc} = \tfrac{(\epsilon_1 + \epsilon_2)^2}{4\hat{\sigma}^2[W_+]})$$

The Mann-Whitney statistic is asymptotically distributed, thus allowing for the approximation of the critical value[1]. The properties of the Mann-Whitney test for equivalence are studied further in Wellek (1996).

[1]The variance of W_+, regardless of the underlying distributions F and G is always defined as $\text{Var}[W_+] = \frac{1}{mn}\left(\pi_+ - (m+n-1)\pi_+^2 + (m-1)\Pi_{X_T X_T X_C} + (n-1)\Pi_{X_T X_C X_C}\right)$ where $\Pi_{X_T X_T X_C} = P[X_{Ti_1} > X_{Cj}, X_{Ti_2} > X_{Cj}]$ and $\Pi_{X_T X_C X_C} = P[X_{Ti} > X_{Cj_1}, X_{Ti} > X_{Cj_2}]$ (Wellek 2010)

B.4 Sample specific versions of parametric tests

If the data is drawn from an experiment, or quasi-experiment, where the assignment mechanism is known and random, we can conduct our tests using the permutation distribution of the data, also known as randomization inference. Here we discuss generally how to conduct the permutation tests and specifically how to conduct non-parametric versions of the tests described above. Permutation tests are tests designed to test for the exchangeability of two groups and are well suited to the problem at hand of validating quasi-experimental designs. In theory, these observational designs should guarantee exchangeability between the two groups. The non-parametric versions of the above tests all use an IUT approach where the the bounds of the equivalence range are used as the strict nulls, and TOST tests are conducted based on the permutation distribution of the test statistics. If the p-value for both associated tests is less than α, then the test rejects the null of non-equivalence.

The non-parametric TOST t-test (npTOST) is set up using the same hypotheses as in (B.2). To test the null that $\mu_T - \mu_C \geq \epsilon_U$ the permutation distribution given the assignment mechanism and the strict null hypothesis that $\mu_T - \mu_C = \epsilon_U$ is calculated, or approximated if the number of permutations is large, using a one-sided test with the strict null of a treatment effect of ϵ_U (Rosenbaum 2002). Then, an exact p-value corresponding to the null $\mu_T - \mu_C = \epsilon_U$ is calculated. The p-value for the analogous test given the assignment mechanism and the strict null hypothesis that $\mu_T - \mu_C = \epsilon_L$ is also calculated. If both p-values are less than the level of the test, α, then the two groups are statistically equivalent, with the overall p-value corresponding to the maximum of the two individual one sided test p-values.

The non-parametric equivalence t-test is set up similar to the npTOST, using the same hypotheses as for the equivalence t-test. To test that $\frac{\mu_T - \mu_C}{\sigma} \geq \epsilon_U$, the permutation distribution of the the standardized difference is computed using the strict null hypothesis that $\frac{\mu_T - \mu_C}{\sigma} = \epsilon_U$. The permutation distribution is also constructed for the strict null of $\frac{\mu_T - \mu_C}{\sigma} = \epsilon_L$, and the appropriate one-tailed permutation p-value is calculated for each distribution. The test rejects the null of non-equivalence if both p-values are less than α. The non-parametric Mann-Whitney test is constructed analogously. However, the test statistic there is the W_+, as defined in Table 3.1, and it is tested around the strict null of $W_+ = 1/2 - \epsilon_L$ and $W_+ = 1/2 + \epsilon_U$. As before, the two one-sided permutation p-values are calculated, and the test rejects the null of non-equivalence if both p-values lie below α. For further discussion of the permutation based one sided test, see Lehmann (1975).