DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Essays on Political Methodology and Data Science

Citation
Kashin, Konstantin Daniel. 2015. Essays on Political Methodology and Data Science. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:17464583

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Essays on Political Methodology and Data Science

a dissertation presented
by
Konstantin Daniel Kashin
to
The Department of Government

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Political Science

Harvard University
Cambridge, Massachusetts
May 2015

©2015 – Konstantin Daniel Kashin
all rights reserved.

Dissertation Advisor: Professor Gary King Konstantin Daniel Kashin

Essays on Political Methodology and Data Science

Abstract

This collection of six essays makes novel methodological contributions to causal inference, time-series cross-sectional forecasting, and supervised text analysis. The first three essays start from the premise that while randomized experiments are the gold standard for causal claims, randomization is not feasible or ethical for many questions in the social sciences. Researchers have thus devised methods that approximate experiments using nonexperimental control units to estimate counterfactuals. However, control units may be costly to obtain, incomparable to the treated units, or completely unavailable when all units are treated. We challenge the commonplace intuition that control units are necessary for causal inference. We propose conditions under which one can use post-treatment variables to estimate causal effects. At its core, we show when one can obtain identification of causal effects by comparing treated units to other treated units, without recourse to control units.

The next two essays demonstrate that the U.S. Social Security Administration's (SSA) forecasting errors were approximately unbiased until about 2000, but then began to grow quickly, with increasingly overconfident uncertainty intervals. Moreover, the errors all turn out to be in the same potentially dangerous direction, each making the Social Security Trust Funds look healthier than they actually are. We also discover the cause of these findings with evidence from a large number of interviews we conducted with participants at every level of the forecasting and policy processes.

Finally, the last essay develops a new dataset for studying the influence of business on public policy decisions across the American states. Compiling and digitizing nearly 1,000 leaked legislative proposals made by a leading business lobbying group in the states, along with digitized versions of all state legislation introduced or enacted between 1995 and 2013, we use a two-stage supervised classifier to categorize state bills as either sharing the same underlying concepts or specific language as business-drafted model bills. We find these business-backed bills were more likely to be introduced and enacted by legislatures lacking policy resources, such as those without full-time members and with few staffers.

Contents

Listing of figures

My dissertation is dedicated to my family, especially to my parents, Vsevolod and Julia Kashin. I owe the majority of my accomplishments to their love, unyielding support, and selfless sacrifice.

Acknowledgments

I would like to acknowlege my committee members for their support and guidance throughout my graduate school career. Professor Gary King, my dissertation chair, for aiding in my transformation into a data scientist. Thank you for always being accessible, even when skiing, 35,000 feet in the sky, or riding camels in Morocco. Your dedication to your students never ceases to impress me. To Professor Adam Glynn, for instilling me with a passion for causal inference and a healthy skepticism for most scientific claims. I have learned more from our Skype discussions than I could have from most courses. To Professor Arthur Spirling, for always having my best interests at heart, asking rigorous questions, and your dry British wit. Finally, to Professor Daniel Ziblatt, for your endless enthusiasm for my research. I am truly thankful for all that you have taught me about politics.

Beyond my advisors, I have learned a tremendous amount from my interaction with graduate students and faculty at the Institute for Quantitative Social Science (IQSS), the Department of Government, and the Department of Statistics at Harvard. I could not have asked for a better environment to pursue my graduate studies. I am indebted to the staff of IQSS for patiently addressing my technical queries and to Thom Wall for his administrative support. Furthermore, I am grateful to all the scholars, of whom there are too many to name, who have lent their thoughts and comments to one or more of these chapters. I owe a special debt of gratitude to Samir Soneji, without whom several chapters of this dissertation would not have come to fruition. Thank you for sharing your wealth of knowledge about demography with me.

I would like to thank my family and friends for their goodnatured humor and keeping me grounded. This dissertation is dedicated to my parents, who have made an incalculable sacrifice for their children. To Alex, you were the best friend, roommate, and colleague one could ask for. I will forever cherish our memories in graduate school and look forward to many more to come. To Erin, Jeff, Leslie, Mikey, Noam, and Volha, thank you for sustaining me throughout graduate school and providing many laughs.

Finally, to Karl, my partner, I am extremely thankful for your constant presence and for keeping me going in the most difficult of times with your love and enthusiastic support. As I have come to understand that the life of a researcher can be peculiar, to say the least, especially to someone who inhabits a world removed from academia, thank you for your perseverance through my moments of stress and occasional aloofness.

1

Front-door versus Back-door Adjustment with Unmeasured Confounding

1.1 Introduction

1.1.1 Job Training Program Evaluation

There exists a rich literature on the effects of job training programs, much of it assessing the ability of nonexperimental estimators to reproduce the experimental estimates of program impact. Influential studies that cast doubt on the utility of nonexperimental estimates (e.g. see LaLonde (1986); Fraker and Maynard (1987)) have in turn sparked fruitful research on how to ensure the comparability of a nonexperimental control group to the treated group on observed and unobserved characteristics. Researchers have had success approximating experimental impact estimates in nonrandomized program evaluation with methods such as matching on the propensity score (see e.g. Dehejia and Wahba (1999)) and semi-parametric models for selection bias (see e.g. Heckman et al. (1998)).

However, the success of these and other such observational studies depends on the measurement of two types of variables that can be difficult/expensive to obtain. First, observational program evaluations of this type must collect outcome information on the control units. For job training programs, this means that earnings information must be collected for individuals that did not sign up for the program. As evidenced by Hotz (1992), this is often an expensive and laborious task. Second, for these observational evaluations to be successful, Heckman et al. (1998) demonstrates that detailed (and expensive) covariate information will often need to be collected.

In this paper, we explore the conditions under which the need for such expensive (and potentially missing) outcome and covariate information can be obviated by the collection of post-treatment compliance information and the use of a front-door adjustment. Specifically, we evaluate the effects of signing up for the Job Training Partnership Act program on 18-month earnings by leveraging post-treatment compliance information via front-door adjustment. Following the literature comparing observational estimates to experimental benchmarks in the case of job training programs, we take signing up for the program to be the treatment of interest. However, unlike nearly all previous work, we do not use the nonexperimental control group for estimation of this causal effect, instead using a subset of treated units to construct the desired counterfactual. This makes possible estimation of causal effects even when the nonexperimental control group is fundamentally incomparable to the treated group, or more remarkably, where control units are unavailable.

The JTPA Title II-A program is a prototypical government job training program where eligible economically disadvantaged individuals receive a variety of services such as classroom training, job search assistance, and on-the-job training (Heckman et al., 1999). Broadly created to boost employment and earnings, the JTPA program annually served approximately one million participants in the 1980s and 1990s at a cost of roughly $1.6 billion per year. Given the scale of the program, the Department of Labor commissioned an experimental evaluation of the program in 1986. The resultant National JTPA Study thus randomized one-third of eligible JTPA applicants into an experimental control group that was not allowed to access program services at a small subset of program sites (Orr et al., 1994; Bloom et al., 1993, 1997). Crucially, the Study collected information on service uptake of treated units, allowing us to measure compliance with the treatment. Moreover, non-compliance was approximately one-sided, meaning that individuals barred from receiving JTPA services were unlikely to actually receive any services.

In addition to experimental estimates for the impact of the JTPA program, the National JTPA Study set out to create a nonexperimental comparison group, termed the eligible nonparticipant (ENP) sample, to assess the viability of nonexperimental estimators. While the initial JTPA Study design proposed ENP samples at ten of the sixteen participating study sites, high data collection costs forced ENP sampling to be conducted at only four sites (Hotz, 1992).[1] The ENP sample was composed of individuals who were eligible for, but did not apply to, the JTPA program. They were selected using random sampling of economically disadvantaged households (Smith, 1994). The availability of a nonexperimental comparison group allows us to estimate program impact using standard covariate adjusment. Ultimately, we use data from the National JTPA Study to demonstrate that front-door adjustment provides estimates that are closer to the experimental benchmark than standard covariate adjustments.

[1] All of our analysis is thus restricted to the four sites, or service delivery areas, at which an ENP sample is available.

1.1.2 Review of the Front-Door Criterion

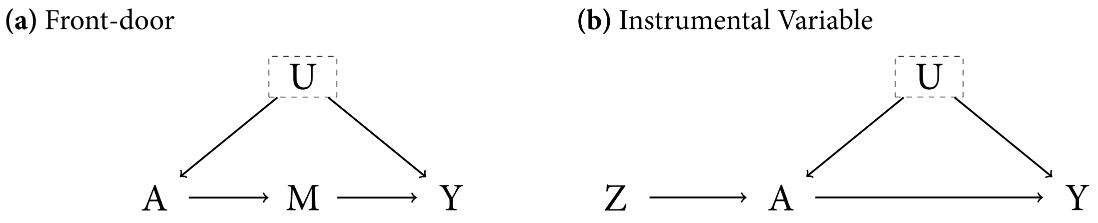
The front-door criterion and adjustment formula (Pearl, 1995) and its extensions provide a means for nonparametric identification of treatment effects via the pathways or mechanisms by which treatment affects the outcome.[2] Importantly, the front-door criterion can hold even in the presence of unmeasured common causes of the treatment and the outcome. Despite these results, the front-door approach has seen relatively little use (VanderWeele, 2009) due to concerns that the assumptions required for point identification are "exceptional" (Cox and Wermuth, 1995; Imbens and Rubin, 1995).[3]

In order to provide intuition and avoid confusion due to the limited use of the front-door approach, we present an informal summary of the results from (Pearl, 1995) below, as well as a comparison between the front-door approach and the more well known (and somewhat related) instrumental variables approach (e.g. Angrist et al. (1996); Balke and Pearl (1997)). The following sections provide a more formal presentation. Figure 1.1 (a) presents a directed acyclic graph (DAG)– equivalent to Figure 3 of (Pearl, 1995)– representing a causal structure amenable to front-door adjustment. Informally, the goal of analysis here is to estimate the effect of the treatment/action A on the outcome Y. Standard covariate adjustments (also known as back-door adjustments) will be insufficient for identification here because U is an unmeasured common cause of A and Y. Pearl (1995) shows that when 1) an exclusion restriction holds so that the effect of A on Y is entirely mediated by a set of post-treatment variables

[2] Extensions of the front-door criterion have highlighted more complicated graph structures under which it is possible to obtain point identification of total effects (Kuroki and Miyakawa, 1999; Tian and Pearl, 2002a,b; Shpitser and Pearl, 2006; Chalak and Halbert, 2011).

[3] One exception is Winship and Harding (2008) which outlines how the front-door criterion can aid in the identification of age-period-cohort models. Additionally, there are several papers that use post-treatment variables to gain some type of information about total effects. Cox (1960) and Ramsahai (2012) examine when post-treatment variables can improve the efficiency of total effects estimates. VanderWeele (2008) and VanderWeele and Robins (2009) show that post-treatment variables can help identify the direction of bias in point estimates of total effects. Joffe (2001) and Glynn and Quinn (2011) both use post-treatment variables to calculate bounds for total effects while Kaufman et al. (2009) provides a variety of bounds, some of which involve measuring post-treatment variables, using linear programming via the OPTIMIZE program (Balke, 1995).

M, and 2) there are no unmeasured common causes of *A* and *M* or of *M* and *Y* (that cannot be blocked by *A*), then a front-door adjustment can identify the effect of *A* on *Y*. As a stylized representation of the job training context of Heckman et al. (1998), the treatment *A* indicates whether an individual has signed up for the program, the mediator *M* indicates whether an individual enrolled in the program, and the outcome *Y* represents post-program earnings. As we discuss in detail below, the assumptions implicit in this graph will not hold for job training programs, but this presentation clarifies the inferential approach.



(a) Front-door

(b) Instrumental Variable

Figure 1.1: Comparison of Front-door and Instrumental Variable Directed Acyclic Graphs.

Instrumental variables approaches are tangentially related to front-door approaches but rely on similar assumptions and vocabulary, especially within the job training context. It is helpful to contrast the two approaches. Figure 1.1 (b) presents a directed acyclic graph (DAG)– equivalent to Figure 1 of (Balke and Pearl, 1997)– representing a causal structure amenable to instrumental variables analysis. Informally, the goal of analysis here is to estimate the effect of the treatment/action *A* on the outcome *Y*, where again standard covariate adjustments will be insufficient for identification because *U* is an unmeasured common cause of *A* and *Y*. An instrumental variable, *Z*, can facilitate the estimation of the effect of *A* on *Y* when *Z* is randomly assigned and an exclusion restriction holds such that the effect of *Z* on *Y* is completely mediated by *A*. This presentation omits many important details but these are not pertinent to the present discussion (see Angrist et al. (1996) and subsequent discussion for details).

For the analysis of job training programs, instrumental variables approaches are typically utilized when *Z* represents randomized sign-up, *A* represents enrollment, and *Y* again represents post-program earnings (Orr et al., 1994; Bloom et al., 1997; Abadie et al., 2002). Hence, the inferential target of an instrumental variables analysis in the job training context is the effect of enrollment on earnings. In contrast, the front-door analysis we explore in this paper uses enrollment M to facilitate estimation of the effect of sign-up A on earnings Y for contexts where sign-up has not been randomized. Not only is the effect of sign-up on earnings the parameter of interest in the econometrics literature utilizing JTPA data (Heckman et al., 1998, 1997; Heckman and Smith, 1999), but it is a relevant parameter for policymakers faced with the decision of whether or not to offer the opportunity for job training to a group of individuals.[4] Ultimately, "[the effect of sign-up] is informative on how the availability of a program affects participant outcomes" (Heckman et al., 1999). Since policymakers cannot compel participation, non-compliance and attrition are a natural part of most programs, and as a result, the effect of the opportunity for training is often a more pragmatic estimate than the effect of actual participation.

1.1.3 Plan for the Paper

In this paper, we consider the applicability of the front-door adjustment in situations where the front-door criterion does not hold by providing formulas for the asymptotic bias of front-door adjustments for both Average Treatment Effects on the Treated (ATT) and Average Treatment Effects (ATE). These formulas are derived in a manner analogous to the asymptotic bias formulas of VanderWeele and Arah (2011) for standard covariate adjustments, and similarly allow for sensitivity analysis. Additionally, because they are derived without using potential outcomes beyond those used for the definition of ATT and ATE, our formulas do not require causal effects to be well defined for variables other than the treatment. Specifically, we do not require that the stable unit treatment value assumption (SUTVA) (Rubin, 1980) hold for variables other than the treatment. This fact along with the manner of derivation allows for direct comparisons to the bias formulas of VanderWeele and Arah (2011). Therefore, we are able to

[4] For interest in the effect of sign-up outside the JTPA program, see for example Lee (2009) and Zhang et al. (2009).

conduct comparative sensitivity analyses to assess whether a front-door approach will be preferred (in bias terms) to a standard covariate adjustment (sometimes known as a back-door adjustment).

To provide intuition, we also present these bias comparisons in two special cases: the estimation of ATT for nonrandomized program evaluations with one-sided noncompliance, and the estimation of ATE using linear structural equation models. These comparisons demonstrate that there are broad classes of applications for which the front-door or hybrid adjustments will be preferred to the back-door adjustments.

The paper is organized as follows. Section 1.2 presents the bias formulas for the front-door approach to ATT and compares this to the bias from covariate adjustments, both within the general framework and for nonrandomized program evaluations with one-sided noncompliance. Section 1.3 presents an application of these methods to the National JTPA (Job Training Partnership Act) Study. Section 1.4 concludes. Because the presentation for ATE is somewhat parallel and redundant to the presentation for ATT, we provide the bias formulas and comparisons, both within the general framework and for linear structural equation models, in the supplementary material.

1.2 The Front-Door Approach for ATT

For an outcome Y and a treatment/action A, we define the potential outcome under a generic treatment as $Y(a_1)$ and the potential outcome under control as $Y(a_0)$. While the presentation of the front-door approach in Pearl (1995, 2000, 2009) focuses on ATE ($E[Y(a_1)] - E[Y(a_0)]$), in many applications ATT ($E[Y(a_1)|a_1] - E[Y(a_0)|a_1]$) is the question of interest. See Supplement A for an extended discussion of the front-door adjustment for ATE.

We assume consistency which implies that the expectation over the treatment potential outcomes conditional on treatment, $\mu_{1|a_1} = E[Y(a_1)|a_1]$, equals the expectation over the observed outcomes conditional on treatment, $E[Y|a_1]$. As in VanderWeele and Arah (2011), we also assume that $E[Y(a_0)|a_1]$

can be equated to expectations over observed outcomes by conditioning on observed covariates X and unobserved covariates U. For simplicity in presentation we assume that X and U are discrete, such that

$$\mu_{0|a_1} = E[Y(a_0)|a_1] = \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x, a_1) \cdot P(x|a_1), \tag{1.1}$$

but continuous variables can be easily accommodated.[5] Finally, for (1.1), we also need probabilistic assignment to hold such that there is a positive probability of both a_1 and a_0 for all values of U and X among the a_1 units (Rubin, 2010).

The front-door adjustment for a set of measured post-treatment variables M can be written as the following:

$$\mu^{fd}_{0|a_1} = \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1), \tag{1.2}$$

where these sums are taken over values of x and m with positive probability. The asymptotic bias for $E[Y(a_0)|a_1]$ is the following (see equation (A.1) in the appendix for a proof):

$$\begin{aligned} B^{fd}_{0|a_1} = &\sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x) \cdot E[Y|a_1, m, x, u] \cdot P(u|a_1, m, x) \\ &- \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|a_1, x) \end{aligned} \tag{1.3}$$

Throughout the rest of this paper, we will use the term bias to mean asymptotic bias. It is straightforward to show that this bias will be zero when the following two conditions hold:

Assumption 1 (Y is mean independent of A conditional on M, X, and U)

$E[Y|a_1, m, x, u] = E[Y|a_0, m, x, u]$ *for all* m, x, u.

[5] We also note that for formulas of this type throughout the paper, when any of the densities take the value zero (e.g., $P(u|x, a_1) = 0$ or $P(x|a_1) = 0$) we mean the entire term to be zero.

Assumption 2 (*U* **is independent of** *M* **conditional on** *X* **and** *a_0* **or** *a_1*)

$P(m|a_0, x) = P(m|a_0, x, u)$ and $P(u|a_1, m, x) = P(u|a_1, x)$.

These are analogous to the assumptions of Pearl (1995) for the identification of ATE. Hence, it is possible for the front-door approach to provide an approach to estimation even when there is an unmeasured confounder.

The back-door adjustment for $E[Y(a_0)|a_1]$ and the associated bias can be written as the following (see equation (A.2) in the appendix for a proof):

$$\mu^{bd}_{0|a_1} = \sum_x E[Y|a_0, x] \cdot P(x|a_1) \tag{1.4}$$

$$B^{bd}_{0|a_1} = \sum_x P(x|a_1) \sum_u E[Y|a_0, x, u][P(u|a_0, x) - P(u|a_1, x)] \tag{1.5}$$

This is nearly identical to the formula from page 3 of VanderWeele and Arah (2011), except we have not included a contrast with a reference value for the unmeasured confounder. Since consistency implies that $E[Y(a_1)|a_1] = E[Y|a_1]$, the front-door ATT bias is $B^{fd}_{ATT} = -B^{fd}_{0|a_1}$ and the back-door ATT bias is $B^{bd}_{ATT} = -B^{bd}_{0|a_1}$. Hence, the front-door ATT bias can be smaller than the back-door ATT bias even when the aforementioned front-door independence conditions do not hold exactly. It is also possible to form hybrid estimators that utilize the front-door approach for some values of X and the back-door approach for other values of X. Finally, we note that these are direct comparisons in the sense that we did not define additional potential outcomes in order to derive the front-door result (i.e., we are agnostic as to whether SUTVA holds with M as a treatment variable).

1.2.1 Special Case: Nonrandomized Program Evaluations with One-Sided Noncompliance

In order to develop some intuition about the front-door approach for ATT, we next consider the special case of nonrandomized program evaluations with one-sided noncompliance. Following a robust literature in econometrics on social program evaluation (e.g., Heckman et al. (1998)), we define the program impact as the ATT where the active treatment (a_1) is assignment into a program (perhaps self selected assignment), (a_0) is non-assignment to the program, (m_1) is participation in the program (perhaps self selected participation), (m_0) is non-participation in the program.[6]

Assumption 3.A (One-sided Noncompliance)

$P(m_0|a_0, x) = P(m_0|a_0, x, u) = 1$ *for all* x, u.

Assumption 3.B (Probabilistic Noncompliance for Treated Units)

$0 < P(m_0|a_1, x, u) < 1$ *for all* x, u.

Part A of Assumption 3 prevents units who were not assigned to (or did not select into) the program from participating in the program, whereas Part B requires that each unit who was assigned to (or selected into) the program has a possibility of both participating and not participating in the program. We also note that the variance of this approach can be large unless there are a sufficient number of noncompliers.

[6] There is some ambiguity regarding the use of the term ATT. Some authors continue to refer to the assigned treatment as "the treatment", and $\mu_{1|a_1} - \mu_{0|a_1}$ as ATT, while other authors would refer to the received treatment as "the treatment", and $\mu_{1|a_1} - \mu_{0|a_1}$ would be more properly characterized as the Intent to Treat Effect on the Intended (ITI). For continuity, we will continue to refer to $\mu_{1|a_1} - \mu_{0|a_1}$ as the ATT. This is consistent with the parameter of interest in the econometrics literature utilizing JTPA data (Heckman et al., 1998, 1997; Heckman and Smith, 1999), and as mentioned above, is a relevant parameter from the point of view of policymakers (Heckman et al., 1999). For interest in the Intent to Treat Effect outside of the JTPA program, see for example Lee (2009) and Zhang et al. (2009).

Consider the bias in the front-door approach for ATT when M indicates whether the active treatment (a_1) was actually received and there is one-sided noncompliance such that Assumption 3 holds. In this case, the front-door formula reduces to the following:

$$\begin{aligned}
\mu^{fd}_{ATT} &= \mu_{1|a_1} - \mu^{fd}_{0|a_1} \\
&= E[Y|a_1] - \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1) \\
&= E[Y|a_1] - \sum_x \underbrace{E[Y|a_1, M = 0, x]}_{\text{treated non-compliers}} \cdot P(x|a_1)
\end{aligned} \tag{1.6}$$

Compare this to the standard back-door formula for ATT:

$$\begin{aligned}
\mu^{bd}_{ATT} &= \mu_{1|a_1} - \mu^{bd}_{0|a_1} \\
&= E[Y|a_1] - \sum_x \underbrace{E[Y|a_0, x]}_{\text{controls}} \cdot P(x|a_1)
\end{aligned} \tag{1.7}$$

Equations (1.6) and (1.7) demonstrate the principal difference between front-door and standard back-door approaches in this context. Standard back-door estimates compare units that were assigned treatment to similar units that were assigned control. Front-door estimates compare units that were assigned treatment to similar units that were assigned treatment but did not receive treatment. More specifically, those that were assigned treatment and received treatment are implicitly compared to similar units that were assigned treatment but did not receive treatment. Those that were assigned treatment and did not receive treatment (i.e., non-compliers) are implicitly compared to themselves. The front-door and the back-door ATT bias under one-sided noncompliance (Assumption 3) can be written as the following (see Appendix A.1.3):

$$B^{fd}_{ATT} = \sum_x P(x|a_1)P(m_1|a_1,x) \sum_u E[Y|a_0,m_0,x,u] \cdot [P(u|a_1,m_1,x) - P(u|a_1,m_0,x)] \tag{1.8}$$

$$+ \sum_x P(x|a_1) \sum_u \Bigg\{ E[Y(a_0)|a_1,m_1,x,u] \cdot P(m_1|a_1,x,u)$$

$$- E[Y(a_0)|a_1,m_0,x,u] \cdot \left[\frac{E[Y|a_1,m_0,x,u]}{E[Y(a_0)|a_1,m_0,x,u]} - P(m_0|a_1,x,u)\right] \Bigg\} \cdot P(u|a_1,m_0,x) \tag{1.9}$$

$$B^{bd}_{ATT} = \sum_x P(x|a_1) \sum_u E[Y|a_0,M=0,x,u][P(u|a_1,x) - P(u|a_0,x)] \tag{1.10}$$

The idea of leveraging non-compliers in this manner was briefly explored in Heckman et al. (1997), although it was not mentioned in the abstract or conclusion, and it was not discussed in connection to the front-door approach.

Given the non-standard nature of these comparisons, it is helpful to consider an intuitive justification for the technique along with the formal presentation. The question is under what conditions we might expect the front-door bias to be approximately zero, or when we would expect the front-door bias to be preferable to the back-door bias.

In a simplified framework, with additional assumptions, a principal stratification approach can be helpful to develop intuition. Under one-sided noncompliance there are two principal strata:

Compliers: $M(1) = 1$ and $M(0) = 0$

Never-takers: $M(1) = 0$ and $M(0) = 0$

If we further assume that the treatment can only have an effect on the outcome through the mediator (i.e., that an exclusion restriction holds such that $E[Y|A = 1, M(1) = 0, M(0) = 0] = E[Y(0)|A =$

$1, M(1) = 0, M(0) = 0]$ for never takers, then randomization of M conditional on $A = 1$ is sufficient for identification:

$$\begin{aligned} E[Y|A = 1, M = 0] &= E[Y|A = 1, M(1) = 0, M(0) = 0] \\ &= E[Y(0)|A = 1, M(1) = 0, M(0) = 0] \\ &= E[Y(0)|A = 1, M = 0] \\ &= E[Y(0)|A = 1] \end{aligned}$$

It is straightforward to extend this to situations where M is randomly assigned conditional on pre-treatment covariates X as well.

Returning to our more general presentation under one-sided non-compliance, the exclusion restriction can be written as the following:

Assumption 4 (Exclusion Restriction)

$E[Y|a_1, m_0, x, u] = E[Y(a_1)|a_1, m_0, x, u] = E[Y(a_0)|a_1, m_0, x, u]$ *for all* x, u.

Note however that this condition would not necessarily be sufficient as an exclusion restriction if Assumption 3 did not hold. If Assumptions 3–4 hold, the front-door bias simplifies to the following:

$$B^{fd}_{ATT} = \sum_x P(x|a_1)P(m_1|a_1, x) \sum_u E[Y|a_0, m_0, x, u] \cdot [P(u|a_1, m_1, x) - P(u|a_1, m_0, x)] \tag{1.11}$$

$$+ \sum_x P(x|a_1) \sum_u P(m_1|a_1, x, u) \Big[E[Y(a_0)|a_1, m_1, x, u] - E[Y(a_0)|a_1, m_0, x, u]\Big] P(u|a_1, m_0, x) \tag{1.12}$$

and the randomization assumption can be written as the following:

Assumption 5 (Conditional Randomization of M)

M is randomly assigned for a_1 units within levels of x, such that M is independent of all variables that are not affected by M conditional on $A = a_1$ and $X = x$ for all x.

Assumptions 3 - 5 are now sufficient for zero large sample bias because $P(u|a_1, m_1, x) = P(u|a_1, m_0, x)$ and $E[Y(a_0)|a_1, m_1, x, u] = E[Y(a_0)|a_1, m_0, x, u]$ when M is conditionally independent of the unmeasured covariates and the potential outcomes. However, as explained below, we do not expect Assumption 5 to hold in practice due to unmeasured common causes of compliance and the outcome. Moreover, there will be many applications where Assumption 4 does not hold as well. The key question, then, is whether the bias will be smaller from the front-door or back-door approaches.[7]

It is helpful to separately consider the two components of front-door bias represented by (1.8) and (1.9). If we assume that the second component, (1.9), is zero for illustrative purposes, then the bias comparison between these two approaches is a comparison between (1.8) and (1.10). In this scenario, when the treated compliers and non-compliers are better matched on the unobserved covariates than the treated and control units, then the front-door approach will produce less bias. In fact, because the front-door imbalance in U due to compliance types is scaled by the proportion of compliers among the treated (an estimable quantity), the front-door imbalance in U can be greater than the back-door imbalance in U, and the front-door approach will still produce less bias. This means that in cases where the (1.9) component is zero and the treatment has low compliance, we will often prefer the front-door approach.

As we consider in the next section, (1.8) will sometimes be smaller than (1.10), so the key question will often be the magnitude of (1.9). If Assumption 4 holds, then the front-door bias simplifies and

[7] A subtle point that deserves clarification is that while the exclusion restriction and one-sided noncompliance (in combination with randomization of A) would allow identification of certain more fine grained counterfactual expectations (e.g., $E[Y(a_0)|a_1, m_1]$), these assumptions do not alter the standard back-door formula for ATT (Equation 1.7).

(1.9) becomes (1.12). The simplified bias term (1.12) might not be zero because an unmeasured variable $v \notin U$ can be a common cause of both M and Y. This would imply that $E[Y(a_0)|a_1, m_1, x, u] - E[Y(a_0)|a_1, m_0, x, u] \neq 0$. Furthermore, non-zero (1.12) can occur with unmeasured variables $v, w \notin U$ such that v is a common cause of M and U and w is a common cause of U and Y (similarly if v is a common cause of M and X and w is a common cause of X and Y.)

Despite this complication, it will be possible in some circumstances to glean information from these equations that will be useful in a comparison of front-door and back-door results. For example, if we are willing to assume that (1.8), (1.9), and (1.10) all have the same sign (or (1.11), (1.12), and (1.10) with an exclusion restriction), then the large sample bias of both the front-door and back-door approaches will have the same sign, and we should have a preference between the two approaches in large samples upon observing whether front-door or back-door estimate is larger. An example of this will be provided in the next section.

Simplified Comparative Sensitivity Analysis

Equations (1.8) and (1.9) (or (1.11) and (1.12) under an exclusion restriction) form the basis for a sensitivity analysis of the front-door approach. Moreover, this sensitivity analysis can be fruitfully compared with a sensitivity analysis for the back-door approach based on (1.10). In order to illustrate this, we use the simplifying assumptions used in VanderWeele and Arah (2011), although as discussed in that article it is straightforward to relax these assumptions at the cost of complicating the analysis. This simple comparative sensitivity analysis is sufficient to establish that the front-door approach might be preferred to the back-door approach when these assumptions may hold approximately. We leave an illustration of more complicated comparative analysis for future work.

VanderWeele and Arah (2011) shows that when U is binary and when

$$\gamma = E[Y|U=1, a_0, M=0, x] - E[Y|U=0, a_0, M=0, x] \tag{1.13}$$

and

$$\delta = P(U=1|a_1, x) - P(U=1|a_0, x) \tag{1.14}$$

do not depend on x, then the back-door ATT bias under one-sided noncompliance in (1.10) can be written as $\gamma \cdot \delta$. In this case, δ can be interpreted as the imbalance on U across the treatment and control groups, and γ can be interpreted as a "controlled direct effect" (VanderWeele and Arah, 2011) of U on Y while forcing the treatment variable to be in the control condition. Note, however, that this "effect" need not be a causal effect with SUTVA holding for U as a treatment variable. If we additionally assume that Assumption 4 holds, as well as the following three assumptions,

Assumption 6 (Front-door imbalance on U does not depend on x)

$\varepsilon = P(U=1|a_1, m_1, x) - P(U=1|a_1, m_0, x) = P(U=1|a_1, m_1) - P(U=1|a_1, m_0)$,

Assumption 7 (Front-door bias due to confounding of the M and Y relationship does not depend on x or u[8])

$\eta = E[Y(a_0)|a_1, m_1, x, u] - E[Y(a_0)|a_1, m_0, x, u] = E[Y(a_0)|a_1, m_1] - E[Y(a_0)|a_1, m_0]$,

Assumption 8 (The compliance rate does not depend on x or u)

$P(m_1|a_1) = P(m_1|a_1, x) = P(m_1|a_1, x, u)$,

[8] The dependence on x is testable.

then the front-door bias under one-sided noncompliance can be written as $P(m_1|a_1)[\gamma \cdot \varepsilon + \eta]$ (see Appendix A.1.5 for proof). The compliance rate $P(m_1|a_1)$ can be estimated, and we can think of $\gamma \cdot \varepsilon + \eta$ as the bias for the estimation of the "effect of the mediator" (with $\gamma \cdot \varepsilon$ corresponding to the bias due to U and η corresponding to the bias due to other confounding variables). Although, note again that we do not require a formal definition of this "effect." Therefore, comparing the bias from the front-door and back-door approaches, the key question is whether the bias for the "effect of the mediator," scaled by the compliance rate, is larger than the bias for the back-door approach. This will be illustrated in the next section.

1.3 Application: National JTPA Study

In this section, we compare the performance of the front-door approach derived in the previous section to the performance of the back-door approach in the context of the National JTPA Study, a job training evaluation for which we have both experimental data and a nonexperimental comparison group.[9] We measure program impact as the ATT on 18-month earnings in the period post-randomization or post-eligibility. The National JTPA Study is amenable to the use of the front-door approach because of the presence of nearly one-sided noncompliance, satisfying Assumption 3.[10]

The National JTPA Study was commissioned by the Department of Labor to gauge the efficacy of the Job Training Partnership Act (JTPA) of 1982. Implemented between November 1987 and September 1989, the National JTPA Study randomized participants at 16 study sites (technically called *service delivery areas*, or SDAs) across the United States into a treatment and control group. Active treatment consisted of being allowed to receive JTPA services following application for the program, while the

[9] See Appendix A.3 for a more thorough description of the data.

[10] There were a very small number of individuals that received the training program without being assigned to it, however these do not affect the results.

control group was barred from receiving program services for a period of 18 months following random assignment (Bloom et al., 1993; Orr et al., 1994). The key feature of this study for our analysis is that there was sizable noncompliance among the treated units. In our sample, roughly 57% of adult men and 55% of adult women who were allowed to receive JTPA services actually utilized JTPA services (see Table A.1).

The Study also collected a sample of eligible nonparticipants (ENPs) at 4 service delivery areas as a nonexperimental comparison group. The sample was selected following a screening interview.[11] To match the ENP sample, we restrict the experimental sample to only the 4 sites. Furthermore, we focus on two of five *target groups* defined in the initial study: (1) male adults and (2) female adults.[12] Participants were considered adults if they were at least 22 years old at random assignment. We conduct our analysis separately for the two target groups.

We establish the experimental benchmarks for adult males and adult females by comparing the mean earnings in the 18 months after random assignment of the experimental active treatment group sample to the experimental control group sample. We utilize kernel-based regularized least squares (KRLS) to present covariate-adjusted experimental estimates of program impact across a variety of simple conditioning sets (Hainmueller and Hazlett, 2014). For the null conditioning set, the program impact for adult males was, on average, an increase of $699.60 in the 18-month earnings. For adult females, the impact was $702.17.[13]

Using rich historic data on labor market participation for both the experimental control group and the nonexperimental control group, Heckman et al. (1998) were able to characterize selection bias and thus apply their semiparametric sample selection estimator. As the authors explain, "detailed informa-

[11] See Appendix A.3 for additional information regarding the ENP sample. See Smith (1994) for details of ENP screening process.

[12] The other 3 target groups were female youths, non-arrestee male youths, and male youth arrestees.

[13] See discussion of how we created our sample and the earnings data in Appendix A.3.

tion on recent labor force status histories and recent earning are essential in constructing comparison groups that have outcomes close to those of an experimental control group" (1020). However, what if such rich data is not available or it is too costly? Are we then unable to create a comparison group that resembles an experimental control group?

Our results from the front-door approach suggest that even with extremely limited covariates we have recourse to the treated noncompliers in the creation of a comparison group. This was confirmed in Heckman et al. (1997) which shows that no-shows in the National JTPA Study are similar to the experimental control group by calculating their respective conditional probabilities of being enrollees. We expand on this result here. Figure 1.2 presents the comparative performance of the front-door and back-door approaches across a variety of simple conditioning sets for adult males and adult females, respectively. As for the experimental benchmarks, we do not assume linearity or additivity in the conditional expectation function $E[Y|a, m, x]$ and thus use kernel-based regularized least squares (KRLS) to obtain three conditional expectations: $E[Y|a_0, M = 0, x]$, $E[Y|a_1, M = 0, x]$, and $E[Y|a_1, M = 1, x]$.[14] Due to significant rates of non-compliance, we are able to rather precisely estimate the conditional expectation function for non-compliers.

The result is striking in that for adult males, the front-door estimates exhibit uniformly less estimation error than the back-door estimates across all the specifications we examine. The error using the null conditioning set from the back-door estimate is -6745.98. This negative error in the back-door estimate persists even when we condition on age, race, or site. The error in the back-door estimates becomes positive whenever conditioning on the total monthly earnings in the month of random assignment / eligibility screening.[15] The stable performance of the front-door estimates is noteworthy. Without recourse to more detailed data on labor force participation and historic earnings, we find that

[14] We report results from KRLS here due to our reluctance to make strong parametric assumptions, but we obtain similar results when using other methods, such as OLS.

[15] $t = 0$ is the month of random assignment for the experimental samples and the month of eligibility screening for the nonexperimental control sample.

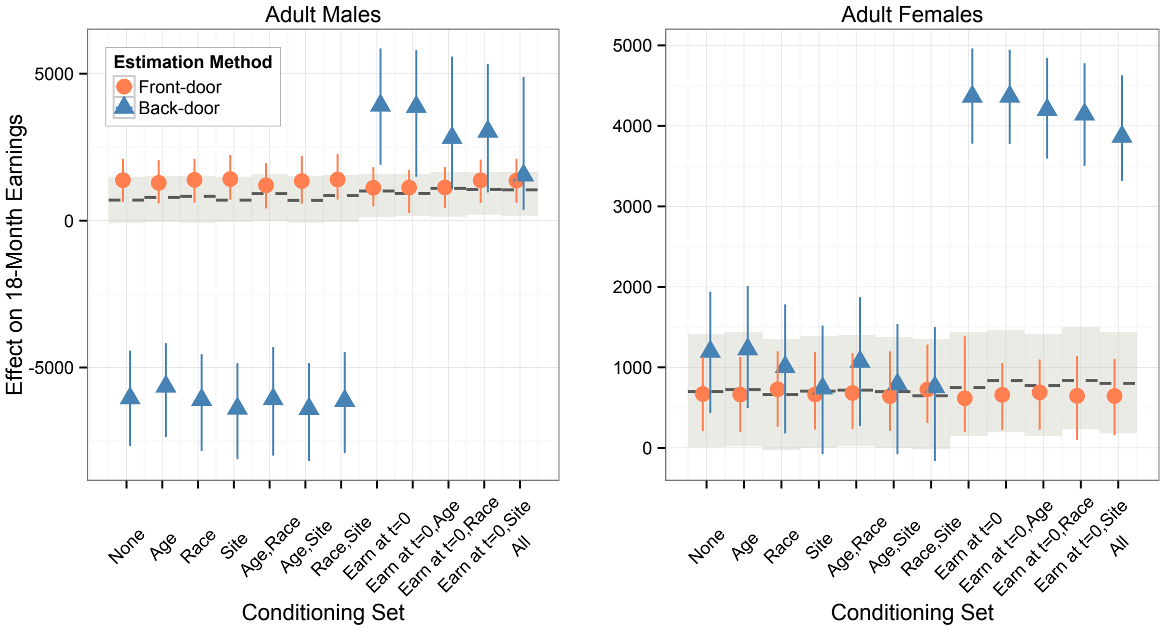


Figure 1.2: Comparison of Back-door and Front-door Adjustment on JTPA Dataset by Target Group using KRLS. The conditioning sets include permutations of the following variables: age; race dummies for white, black, and other; site dummies; and total earnings in month of random assignment/eligibility screening (RA/ES). The experimental estimates are denoted as dashed dark grey lines, with the shaded grey regions representing the 95% confidence intervals. 95% bootstrapped percentile confidence intervals for both adjustment methods and the experimental benchmark are based on 10,000 replicates.

front-door estimates are preferable to back-door adjustment. The front-door estimates for adult females are similarly stable across specifications, and we would also prefer the front-door estimates compared to the back-door estimates in all but one specification if considering only the point estimates (and even in that specification, the back-door estimate has less absolute error only by $8 relative to the front-door estimate).

In sum, we find that for all but one covariate set for adult females, the front-door adjustment has less error than typical back-door adjustment. Moreover, the improvement due to the front-door adjustment is often dramatic, and there is no covariate set where the front-door adjustment has large error. In fact the strong performance of the front-door adjustment relative to the back-door adjustment meant that we were unable to find a hybrid approach that improved on the front-door approach for this application. Rephrasing our result, we find that using treated units that did not comply and receive JTPA services as proxies for experimental control units yields better estimates than using the nonexperimental control group as the counterfactual for what would have happened to the treated units had they not received treatment. To emphasize this point, we note that it was not actually necessary to collect information on any control units (experimental or nonexperimental) in order to get front-door estimates that are quite close to the experimental benchmark.

1.3.1 Comparative Sensitivity Analysis for the JTPA

In most applications, we will not have the experimental benchmark presented above. However, using the simplified comparative sensitivity analysis discussed in Section 1.2.1, we show that we would likely prefer the front-door estimates to the back-door estimates for adult males in this application, even if we did not know the experimental benchmark.

It is helpful to consider how we would react to a simple sensitivity analysis on the back-door estimates for adult males. Suppose we did not have the experimental benchmark or the front-door estimates; we only had available the back-door estimates in Figure 1.2 (a). Suppose further that we only consider the conditioning sets that include baseline earnings, as these are seen as more credible. If we are willing to assume that a back-door approach would be approximately unbiased if we could measure earning potential as a binary variable U, and if we assume that effects do not depend greatly on the values of the measured covariates, then we can use the simple sensitivity parameters from VanderWeele and Arah (2011). If we think of $U = 1$ individuals as having high earning potential and $U = 0$ individuals as having low earning potential, then the difference in expected post-program earnings represented by γ (Equation 1.13) is clearly positive across all conditioning sets. Additionally, due to the well-established pre-program earnings dip – those that select into treatment have temporarily low average earnings immediately preceding the start of training programs (Heckman and Smith, 1999) –

treated individuals within a given strata of baseline earnings have a greater probability of having high earnings potential than control units with the same value of baseline earnings. Therefore, by including baseline earnings earnings in the covariate set, it is likely that $\delta > 0$ and that the back-door bias $\gamma \cdot \delta$ is positive.

Now suppose we have the front-door estimates for these sets. We immediately notice that the front-door estimates are all smaller than the back-door estimates. Given that we assume the back-door approach to have positive bias, we would prefer the front-door estimates if we believed they exhibited positive bias as well. Furthermore, we would also prefer the front-door estimates if we believed they exhibited negative bias whose absolute value was smaller than the magnitude of the back-door bias. To rephrase this, we would only prefer back-door estimates if two conditions held: 1) front-door approach had negative bias and 2) absolute value of front-door bias was greater than back-door bias. An examination of the front-door sensitivity parameters makes it seem unlikely that both of these conditions would hold.

We can compare back-door bias to the simplified front-door bias formula seeing as Assumption 4 is likely to hold approximately for this application (signing up without showing up should have little effect). Using notation defined in Section 1.2.1, we would only prefer back-door estimates to front-door estimates if the following two conditions both held: 1) $\gamma \cdot \varepsilon + \eta < 0$ and 2) $|\gamma \cdot \varepsilon + \eta| > \frac{\gamma \cdot \delta}{P(m_1|a_1)}$. As mentioned above, we estimate $P(m_1|a_1)$ to be roughly 57% for men, so the bias for the "effect of the mediator" on the outcome would have to not only be negative, but also have at least $175\% = (1/.57)\%$ the magnitude of the bias for the effect of the treatment in order for the back-door approach to be preferred.

Let's examine the sign of the front-door bias. We have already assumed that $\gamma > 0$. Moreover, it is quite likely that ε is positive as well considering that high earning potential individuals are bound to be diligent and participate in the program upon signing up. That is, the probability that an individual who signs up for and participates in the job training program has high earning potential is higher than for an individual who signs up but does not show up. What about η? It is very difficult to come up with an unmeasured common cause of sign up and earnings that would lead to a negative enough η so as to overwhelm the positive $\gamma \cdot \varepsilon$ term. As a result, it seems very likely that the front-door approach exhibits positive bias and is thus preferable over the back-door approach.

As with all sensitivity analyses, this analysis is speculative. Yet, simple reasoning about the likely signs of the bias from both the back-door and front-door approaches, combined with observing both sets of estimates, should lead one to prefer the estimates from the front-door approach. At the very least, front-door estimates should be presented along with back-door estimates when the conditions discussed above are reasonable.

1.4 Conclusion

In this paper, we have provided formulas for the asymptotic bias of front-door adjustments for both ATT and ATE (in the supplementary material). These formulas only utilize potential outcomes in terms of the treatment, and they provide a means for sensitivity analysis with front-door adjustment. We have further demonstrated that these bias formulas can be compared directly to the bias formulas of VanderWeele and Arah (2011) for standard back-door covariate adjustments. This allows the consideration of when the front-door approach will be preferred to the back-door approach.

In order to provide intuition, we have also presented these bias comparisons in the special case of nonrandomized program evaluations with one-sided noncompliance. These comparisons demonstrated that there are broad classes of applications for which the front-door or hybrid adjustments will be preferred to the back-door adjustments. In particular, we illustrated the case of nonrandomized program evaluations with one-sided noncompliance with an application to the National JTPA (Job Training Partnership Act) Study. We show that the front-door adjustment performs remarkably better than the back-door adjustment over a wide variety of sets of covariates. We also develop a comparative sensitivity analysis that demonstrates the front-door approach likely should have been preferred to the back-door approach even prior to seeing the experimental benchmark. The sensitivity analysis makes clear that in certain applications it is sufficient to evaluate the direction of bias, and not the magnitude, in order to prefer the front-door estimates to the back-door estimates.

The results in this paper have implications for research design and analysis. First, the JTPA example demonstrates the importance of collecting post-treatment variables that represent compliance with, or uptake of, the treatment. This is true even for the analysis of total effects. In this application, the enrollment information was more useful than all other pre-treatment covariates we examined. If such compliance information can be collected, we recommend that researchers use the formulas presented in this paper, combined with substantive knowledge of the application of interest, to reason about the likely direction and magnitude of bias under both the front-door and back-door approaches. While one approach may be preferable to the other, as in this application, it is possible that the two approaches taken together bracket or provide informative bounds on the causal effect of interest in other applications. Even in applications where the exclusion restriction is not expected to hold, the simplified sensitivity analysis in this paper can be augmented with well-substantiated assumptions regarding the direction of the direct effect of treatment on the outcome. Generally, researchers should present front-door estimates alongside back-door estimates, at the very least as a robustness check for results derived using standard adjustment techniques. However, in the presence of strong prior beliefs that front-door bias will be smaller than back-door bias, a researcher may choose to eschew the back-door approach, in which case it may be unnecessary to collect any information on control units. This could be extremely helpful in cases where it is costly to collect pretreatment covariates, or to follow up with the control units to measure outcomes.

Finally, we note that this approach provides a method for analysis when it is unethical or otherwise not feasible to withhold treatment from individuals in a study. One need not look past the JTPA example to see that it can be politically impossible to mandate participation and extremely difficult to secure voluntary participation in an experiment. The Manpower Demonstration Research Corporation — the group hired by the Department of Labor to implement the experiment — was able to secure only a nonrandom sample of 16 JTPA training centers out of the 228 contacted about participation. Moreover, this effort required "more than a year of searching and ... the help of hundreds of thousands of dollars in side payments" (Heckman and Smith, 1995, p. 104). A survey of these 228 training centers revealed that 61.8% and 54.5% of training centers were concerned about the ethical and public relations implications of random assignment and denial of services, respectively (Doolittle and Traeger, 1990). Given that these concerns extend to randomization in other social programs and in other domains, the ability to estimate causal effects using only treated units is a promising alternative. We are currently working to further validate, and expand upon, the techniques proposed in this paper using additional applications ranging across diverse domains.

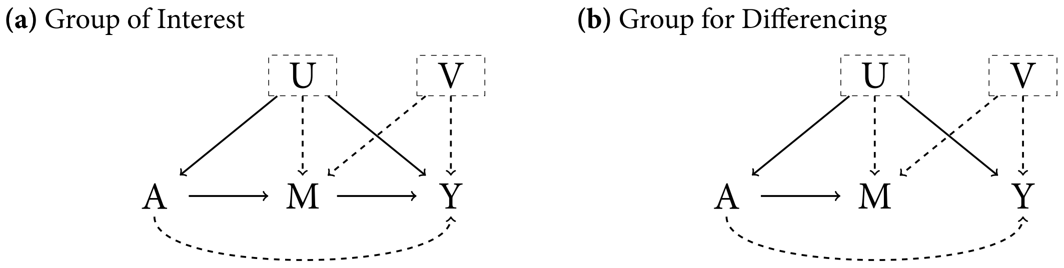
2

Front-door Difference-in-Differences Estimators

2.1 Introduction

One of the main tenets of observational studies is that post-treatment variables should not be included in an analysis because naively conditioning on these variables can block some of the effect of interest, leading to post-treatment bias (King et al., 1994). While this is usually sound advice, it seems to contradict recommendations from the process tracing literature that information about mechanisms can be used to assess the plausibility of an effect (Collier and Brady, 2004; George and Bennett, 2005; Brady et al., 2006).

The front-door criterion (Pearl, 1995) and its extensions (Kuroki and Miyakawa, 1999; Tian and Pearl, 2002a,b; Shpitser and Pearl, 2006) resolve this apparent contradiction, providing a means for nonparametric identification of treatment effects using post-treatment variables. Importantly, the front-door approach can identify causal effects even when there are unmeasured common causes of the treatment and the outcome (i.e., the total effect is confounded). Figure 2.1a presents the directed acyclic graph associated with the front-door criterion. The formal definition of this graph can be found in Pearl (1995), but for our purposes, it will suffice to note the following: A represents the treatment/action variable, M represents a set of mediating variables (possibly singleton), Y represents the outcome, U and V represent unobserved variables, and arrows represent the possible existence of an effect from one variable to another. Solid arrows are allowed for the front-door criterion to hold. Note the existence of solid arrows from U to both A and Y. Hence, unmeasured common causes of the treatment and outcome are allowed.


(a) Group of Interest

(b) Group for Differencing

Figure 2.1: Front-door Directed Acyclic Graphs (DAGs). A represents the treatment/action variable, M represents a set of mediating variables, Y represents the outcome, U and V represent unobserved variables. Solid arrows are allowed for the front-door criterion to hold. Dashed arrows are not allowed for the front-door criterion to hold.

However, the front-door adjustment has been used infrequently (VanderWeele, 2009) due to concerns that the assumptions required for point identification will rarely hold (Cox and Wermuth, 1995;

Imbens and Rubin, 1995). These assumptions are represented by the dashed arrows in Figure 2.1a. Hence, while common causes of A and Y are allowed for the front-door criterion to hold, common causes of M and Y (not mediated by A) are not allowed. Additionally, the front-door criterion will not hold when A has a direct effect on Y.

A number of papers have proposed weaker and more plausible sets of assumptions (Joffe, 2001; Kaufman et al., 2009; Glynn and Quinn, 2011) that tend to correspond to conceptions of process tracing. However, these approaches rely on binary or bounded outcomes, and even in large samples these methods only provide bounds on causal effects (i.e., partial instead of point identification). In this paper, we develop bias formulas for the front-door approach and demonstrate that we can remove or ameliorate this bias via a difference-in-differences approach when there is one-sided noncompliance. We illustrate that under one-sided noncompliance, the front-door estimator implies substituting treated noncompliers for controls.

In this paper, we take a difference-in-differences (DD) approach to removing the bias from the front-door estimator. At the most basic level, a DD estimator tries to correct the bias coming from a standard estimator by estimating this bias from a set of observations for which for which there should be no effect. If an apparent effect is found for these observations, then this is taken to be bias. If one assumes this bias to be equal to the bias from the standard estimator, then it can be subtracted from the standard estimate.

The front-door difference-in-differences approach extends the front-door approach in a similar manner. First, we identify the treated units of interest, which we will refer to as the group of interest. The graph in Figure 2.1a corresponds to this group. Second, if we can identify a group of treated units distinct from our group of interest for which we believe the treatment should have no effect through the mediator, then a non-zero front-door estimate for this group can be attributed to bias. The graph in Figure 2.1b corresponds to this group. Note the missing arrow from M to Y. We will refer to this group as the differencing group. For example, in the context of the early voting application to follow, we consider the effects of an early in-person (EIP) voting program on turnout for elections in 2008 and 2012. One differencing group we consider is potential voters that used an absentee ballot in the previous election. EIP was unlikely to have much of an effect on these voters, as they had already demonstrated their ability to vote by another means of early voting. Therefore, we consider non-zero front-door estimates of the turnout effect for this group to be evidence of bias.

If we further assume that the bias for the differencing group is equal to the bias for our group of interest, then by subtracting the front-door estimator for this group from the front-door estimator for the group of interest, we can remove the bias from our front-door estimate for the group of interest. Note that if all effects and bias are positive, then when the bias from the differencing group is larger than the bias for the group of interest and/or the treatment has an effect for the differencing group, then this differencing approach can provide a lower bound on the effect of the program. We demonstrate this within the context of a job training study. However, we also demonstrate that the bias for each group is related to the proportion of compliers in the group, and therefore, an equal bias assumption is untenable without an additional adjustment. This will be described in detail below.

One question that arises regarding this approach is why use a differencing group instead of the more traditional differencing over time (for standard difference-in-difference estimators)? We note that over time differencing is still possible in many applications. For example, we use an over time analysis as a robustness check for the job training study. That said, for many applications, such as the early voting study, conditioning on past outcome values may be more appropriate than over time differencing. Our focus on differencing groups in this paper is more akin to the third differencing in a traditional difference-in-difference-in-differences (DDD) strategy (see pages 242–243 of Angrist and Pischke (2009) for a description of this "higher order contrast" approach).

The paper is organized as follows. Section 2.2 presents the bias formulas for the front-door approach to estimating average treatment effects on the treated (ATT), both for the general case and the simplification for nonrandomized program evaluations with one-sided noncompliance. Section 2.3

presents the difference-in-differences approach for front-door estimators for the simplified case and discusses the required assumptions. Section 2.4 presents an application of the front-door difference-in-differences estimator to the National JTPA (Job Training Partnership Act) Study. Section 2.5 presents an application of the front-door difference-in-differences estimator to election law: assessing the effects of early in-person voting on turnout in Florida. Section 2.6 concludes.

2.2 Bias for the Front-Door Approach for ATT

In this section, we present large-sample bias formulas for the front-door approach for estimating the average treatment effect on the treated (ATT). This is often the parameter of interest when assessing the effects of a program or a law. For an outcome variable Y and a binary treatment/action A, we define the potential outcome under active treatment as $Y(a_1)$ and the potential outcome under control as $Y(a_0)$.[1] Our parameter of interest is the ATT, defined as $\tau_{att} = E[Y(a_1)|a_1] - E[Y(a_0)|a_1] = \mu_{1|a_1} - \mu_{0|a_1}$. We assume consistency, $E[Y(a_1)|a_1] = E[Y|a_1]$, so that the mean potential outcome under active treatment for the treated units is equal to the observed outcome for the treated units such that $\tau_{att} = E[Y|a_1] - E[Y(a_0)|a_1]$. The ATT is therefore the difference between the mean outcome for the treated units and mean counterfactual outcome for these units, had they not received the treatment.

We also assume that $\mu_{0|a_1}$ is potentially identifiable by conditioning on a set of observed covariates X and unobserved covariates U. To clarify, we assume that if the unobserved covariates were actually observed, the ATT could be estimated by standard approaches (e.g., matching). For simplicity in presentation we assume that X and U are discrete, such that

$$\mu_{0|a_1} = \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|a_1, x) \cdot P(x|a_1),$$

[1] Note that we must assume that these potential outcomes are well defined for each individual, and therefore we are making the stable unit treatment value assumption (SUTVA).

but continuous variables can be handled analogously. However, even with only discrete variables we have assumed that the conditional expectations in this equation are well-defined, such that for all levels of X and U amongst the treated units, all units had a positive probability of receiving either treatment or control (i.e., positivity holds).

The front-door adjustment for a set of measured post-treatment variables M can be written as the following:

$$\mu_{0|a_1}^{fd} = \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1).$$

We can thus define the large-sample front-door estimator of ATT as:

$$\tau_{att}^{fd} = \mu_{1|a_1} - \mu_{0|a_1}^{fd}.$$

The large-sample bias in the front-door estimate of ATT, which is entirely attributable to the bias in the front-door estimate of $\mu_{0|a_1}$, is the following (see Appendix A.1.1 for a proof):

$$\begin{aligned} B_{att}^{fd} = &\sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|a_1, x) \\ &- \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x) \cdot E[Y|a_1, m, x, u] \cdot P(u|a_1, m, x). \end{aligned}$$

As the bias formula shows, it is possible for the front-door approach to provide a large-sample unbiased estimator for the ATT even in the presence of an unmeasured confounder that would bias traditional covariate adjustment techniques such as matching and regression. Specifically, the front-door bias will be zero when three conditions hold: (1) $E[Y|a_1, m, x, u] = E[Y|a_0, m, x, u]$, (2) $P(m|a_0, x) = P(m|a_0, x, u)$, and (3) $P(u|a_1, m, x) = P(u|a_1, x)$. The first will hold when Y is mean independent of A conditional on U, M, and X, while the latter two will hold if U is independent of M conditional on

X and a_0 or a_1. This is an alternative derivation of the result from (Pearl, 1995), although we focus on ATT instead of ATE and do not require the definition of potential outcomes beyond those required for the definition of ATT.

For the difference-in-differences estimators we consider in this paper, we use the special case of non-randomized program evaluations with one-sided noncompliance. Following the literature in econometrics on program evaluation, we define the program impact as the ATT where the active treatment (a_1) is assignment into a program (Heckman et al., 1999), and when M indicates whether the active treatment (a_1) was actually received. We use the short-hand notation m_1 to denote that active treatment was received and m_0 if it was not.

Assumption 3.A from Chapter 1 implies that only those assigned to treatment can receive treatment, and the front-door large-sample estimator reduces to the following under this assumption:

$$\begin{aligned}
\tau_{att}^{fd} &= \mu_{1|a_1} - \mu_{0|a_1}^{fd} \\
&= E[Y|a_1] - \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1) \\
&= E[Y|a_1] - \sum_x \underbrace{E[Y|a_1, m_0, x]}_{\text{treated non-compliers}} \cdot P(x|a_1) \qquad (2.1) \\
&= \sum_x P(x|a_1) \cdot P(m_1|x, a_1) \cdot \left\{ \underbrace{E[Y|a_1, m_1, x] - E[Y|a_1, m_0, x]}_{\text{“effect” of receiving treatment}} \right\} \qquad (2.2)
\end{aligned}$$

The formulas in (2.1) and (2.2) are interesting because they do not rely upon outcomes of control units in the construction of proxies for the potential outcomes under control for treated units (see Appendix A.1.2 for the derivation of (2.2)). This is a noteworthy point that has implications for research design that we will revisit subsequently. The formula in (2.1) can be compared to the standard large-sample covariate adjustment for ATT:

$$\begin{aligned}\tau_{att}^{std} &= \mu_{1|a_1} - \mu_{0|a_1}^{std} \\ &= E[Y|a_1] - \sum_x \underbrace{E[Y|a_0,x]}_{\text{controls}} \cdot P(x|a_1). \end{aligned} \tag{2.3}$$

Roughly speaking, standard covariate adjustment matches units that were assigned treatment to similar units that were assigned control. On the other hand, front-door estimates match units that were assigned treatment to similar units that were assigned treatment but did not receive treatment. This sort of comparison is not typical, so it is helpful to consider the informal logic of the procedure before presenting the formal statements of bias. The fundamental question is whether the treated noncompliers provide reasonable proxies for the missing counterfactuals: the outcomes that would have occurred if the treated units had not been assigned treatment. Therefore, in order for the front-door approach to be unbiased in large samples, we are effectively assuming that 1) assignment to treatment has no effect if treatment is not received and 2) those that are assigned but don't receive treatment are comparable in some sense to those that receive treatment. This will be made more precise below.

The front-door formula in (2.2), with the observable proportions $P(x|a_1)$ and $P(m_1|a_1,x)$ multiplying the estimated effect of receiving the treatment, is helpful when considering the simplified front-door ATT bias, which can be written in terms of the same observable proportions (see Appendices A.1.3 and A.1.6 for proofs):

$$\begin{aligned} B_{att}^{fd} = \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u \Bigg[& E[Y|a_0,m_0,x,u]\cdot[P(u|a_1,m_1,x) - P(u|a_1,m_0,x)] \\ & + \Bigg\{ E[Y(a_0)|a_1,m_1,x,u]\frac{P(m_1|a_1,x,u)}{P(m_1|a_1,x)} \\ & - E[Y(a_0)|a_1,m_0,x,u]\cdot\frac{\frac{E[Y|a_1,m_0,x,u]}{E[Y(a_0)|a_1,m_0,x,u]} - P(m_0|a_1,x,u)}{P(m_1|a_1,x)}\Bigg\}\cdot P(u|a_1,m_0,x)\Bigg] \end{aligned}$$

The unobservable portion of this bias formula (i.e., everything after the $\sum_u$), can be difficult to interpret, but there are a number of assumptions that allow us to simplify the formula. For example, we might assume that treatment does not have an effect on the outcome for noncompliers, as we did in the exclusion restriction Chapter 1. When combined with the consistency assumption, Assumption 4 can also be written as $E[Y(a_1)|a_1, m_0, x, u] = E[Y(a_0)|a_1, m_0, x, u]$. If this exclusion restriction holds, then the bias simplifies to the following:

$$\begin{aligned} B^{fd}_{att} = \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u \Bigg[& E[Y|a_0, m_0, x, u] \cdot [P(u|a_1, m_1, x) - P(u|a_1, m_0, x)] \\ & + \left\{ E[Y(a_0)|a_1, m_1, x, u] \frac{P(m_1|a_1, x, u)}{P(m_1|a_1, x)} - E[Y(a_0)|a_1, m_0, x, u] \cdot \frac{P(m_1|a_1, x, u)}{P(m_1|a_1, x)} \right\} \cdot P(u|a_1, m_0, x) \Bigg] \end{aligned}$$

If instead we assume that compliance rates are constant across levels of u within levels of x,

Assumption 9 (Constant compliance rates across values of u within levels of x)

$P(m_1|a_1, x, u) = P(m_1|a_1, x)$ for all x and u,

then due to the binary measure of treatment received, we know that $P(u|a_1, m_1, x) = P(u|a_1, m_0, x)$ (see Appendix A.1.7), and the bias simplifies to the following:

$$\begin{aligned} B^{fd}_{att} = \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u \Bigg[& \Big\{ E[Y(a_0)|a_1, m_1, x, u] \\ & - E[Y(a_0)|a_1, m_0, x, u] \cdot \frac{\frac{E[Y|a_1, m_0, x, u]}{E[Y(a_0)|a_1, m_0, x, u]} - P(m_0|a_1, x, u)}{P(m_1|a_1, x)} \Big\} \cdot P(u|a_1, m_0, x) \Bigg] \end{aligned}$$

Assumption 9 can be strengthened and the bias simplified further in some cases of clustered treatment assignment. Because the front-door estimator uses only treated units under Assumption 3.A, it is possible that all units within levels of x were assigned in clusters such that U is actually measured at the cluster level. We present an example of this in the application, where treatment (the availability of early in-person voting) is assigned at the state level, and therefore all units within a state (e.g., Florida) have the same value of u. Formally, the assumption can be stated as the following:

Assumption 10 (u is constant among treated units within levels of x)

For any two units with a_1 and covariate values (x, u) and (x', u'), $x = x' \Rightarrow u = u'$.

When Assumption 10 holds, the u notation is redundant, and can be removed from the bias formula which simplifies as the following:

$$B^{fd}_{att} = \sum_{x} P(x|a_1)P(m_1|a_1,x)\Big\{E[Y(a_0)|a_1,m_1,x] - E[Y(a_0)|a_1,m_0,x] \cdot \frac{\frac{E[Y|a_1,m_0,x]}{E[Y(a_0)|a_1,m_0,x]} - P(m_0|a_1,x)}{P(m_1|a_1,x)}\Big\} \tag{2.4}$$

Finally, it can be instructive to consider the formula when both Assumption 4 and Assumption 10 hold. In this scenario, the remaining bias is due to an unmeasured common cause of the mediator and the outcome.

$$B^{fd}_{att} = \sum_{x} P(x|a_1)P(m_1|a_1,x)\{E[Y(a_0)|a_1,m_1,x] - E[Y(a_0)|a_1,m_0,x]\}$$

In some applications, the bias B^{fd}_{att} may be small enough for the front-door estimator to provide a viable approach. For others, we may want to remove the bias. In the next section, we discuss a difference-in-differences approach to removing the bias.

2.3 Front-door Difference-in-Differences Estimators

If we define the front-door estimator within levels of a covariate x as $\tau^{fd}_{att,x}$, then the front-door estimator can be written as a weighted average of strata-specific front-door estimators where the weights are relative strata sizes for treated units:

$$\tau_{att}^{fd} = \sum_x P(x|a_1)\tau_{att,x}^{fd}.$$

If we further define the group of interest as the stratum g_1 and the differencing group as the stratum g_2, then the front-door estimators within levels of x for these groups can be written as:

$$\begin{aligned}
\tau_{att,x,g_1}^{fd} &= P(m_1|x, a_1, g_1)\{E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1]\},\\
\tau_{att,x,g_2}^{fd} &= P(m_1|x, a_1, g_2)\{E[Y|a_1, m_1, x, g_2] - E[Y|a_1, m_0, x, g_2]\}.
\end{aligned}$$

Using these components, the front-door difference-in-differences estimator can be written as

$$\begin{aligned}
\tau_{att}^{fd-did} &= \sum_x P(x|a_1, g_1)\big[\tau_{att,x,g_1}^{fd} - \frac{P(m_1|a_1, x, g_1)}{P(m_1|a_1, x, g_2)}\tau_{att,x,g_2}^{fd}\big] && (2.5)\\
&= \sum_x P(x|a_1, g_1)P(m_1|x, a_1, g_1)\big[\{E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1]\}\\
&- \{E[Y|a_1, m_1, x, g_2] - E[Y|a_1, m_0, x, g_2]\}\big]. && (2.6)
\end{aligned}$$

Hence, (2.5) shows that within levels of x, the front-door difference-in-differences estimator is the difference between the front-door estimator from the group of interest and a scaled front-door estimator from the differencing group, where the scaling factor is the ratio of the compliance rates in the two groups. Then, the overall front-door difference-in-differences estimator is a weighted average of the estimators within levels of x, where the weights are determined by the group of interest proportions of x for treated units. Intuitively, the scaling factor is necessary because it places the front-door estimate for the differencing group on the same compliance scale as the front-door estimate for the group of interest. The necessity of this adjustment can be most easily seen in (2.6), where we see that the main goal is to remove the bias from the $\{E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1]\}$ component of group 1 with the $\{E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1]\}$ component of group 2.

In order for the front-door difference-in-differences estimator to give us an unbiased estimate of the ATT for the group of interest in large samples, we need the following two assumptions to hold. If we further define bias within levels of x for a generic group g as

$$\begin{aligned}
B^{fd}_{att,x,g} = P(m_1|a_1, x, g) \sum_u \Bigg[& E[Y|a_0, m_0, x, u, g] \cdot [P(u|a_1, m_1, x, g) - P(u|a_1, m_0, x, g)] \\
& + \{E[Y(a_0)|a_1, m_1, x, u, g] \frac{P(m_1|a_1, x, u, g)}{P(m_1|a_1, x, g)} \\
& - E[Y(a_0)|a_1, m_0, x, u, g] \cdot \frac{\frac{E[Y|a_1, m_0, x, u, g]}{E[Y(a_0)|a_1, m_0, x, u, g]} - P(m_0|a_1, x, u, g)}{P(m_1|a_1, x, g)} \} P(u|a_1, m_0, x, g) \Bigg],
\end{aligned}$$

then the assumption we need for the differencing group is the following:

Assumption 11 (No effect for the differencing group)

$\tau^{fd}_{att,x,g_2} = B^{fd}_{att,x,g_2}$ for all x.

Note that Assumption 11 can often be weakened. If we believe there are effects for the differencing group, but these have the same sign for the group of interest, then subtracting the scaled estimated effect from the differencing group will remove too much from the estimated effect in the group of interest. For example, if we believe that effects for the group of interest and the differencing group would be positive, then the front-door difference-in-differences estimator would tend to be understated. If we additionally believe that the bias in the front-door estimator is positive prior to the differencing, then the front-door and front-door difference-in-differences estimator will bracket the truth in large samples.

We also need to assume that the bias in the group of interest (g_1) can be removed using the bias from the differencing group (g_2):

Assumption 12 (Bias for g_1 equal to scaled bias for g_2 within levels of x)

$B^{fd}_{att,x,g_1} = \frac{P(m_1|a_1,x,g_1)}{P(m_1|a_1,x,g_2)} B^{fd}_{att,x,g_2}$ *for all x.*

If Assumptions 3.A, 11, and 12 hold, then τ^{fd-did}_{att} has no large-sample bias for τ_{att} (see Appendix A.2.1 for a proof). However, the interpretation of Assumption 12 will often be simplified when Assumptions 4, 9, or 10 hold. This will be discussed in the context of the applications, but one special case is useful to consider for illustrative purposes. When Assumptions 3.A, 4, 9, 10, and 11 hold, then Assumption 12 is equivalent to the following:

$$\{E[Y(a_0)|a_1, m_1, x, g_1] - E[Y(a_0)|a_1, m_0, x, g_1]\} = \{E[Y(a_0)|a_1, m_1, x, g_2] - E[Y(a_0)|a_1, m_0, x, g_2]\}$$

Note that this equality is analogous to the parallel trends assumption for standard difference-in-differences estimators.

2.4 Illustrative Application: National JTPA Study

We now illustrate how front-door and front-door difference-in-differences estimates for the average treatment effect on the treated (ATT) can bracket the experimental truth in the context of the National JTPA Study, a job training evaluation with both experimental data and a nonexperimental comparison group. This section builds upon Chapter 1, which demonstrates the superior performance of front-door adjustment compared to standard covariate adjustments like regression and matching when estimating the ATT for nonrandomized program evaluations with one-sided noncompliance. Specifically for the National JTPA Study, matching adjustments using the nonexperimental comparison group can come close to the experimental estimates only when one has "detailed retrospective questions on labor force participation, job spells, earnings" (Heckman et al., 1998). However, in the absence of detailed labor force histories, Chapter 1 show that it is possible to create a comparison group that more closely resembles an experimental control group using the front-door approach. Nonetheless, while the front-door approach is preferable to standard covariate adjustments for the National JTPA Study, the front-door estimates exhibit positive bias. We thus attempt to address this bias using the front-door-difference-in-differences approach developed in this paper.

In order to implement the difference-in-differences approach, we focus on currently or once married adult men as the group of interest (henceforth referred to as simply married men).[2] We measure program impact as the ATT on 18-month earnings in the post-randomization or post-eligibility period. For married males, the experimental benchmark is $703.27 .[3] Since front-door adjustment requires compliance information, we cannot use the typical over time difference-in-differences approach to remove possible bias. However, the focus on married men enables us to use single adult men as the differencing group in a front-door difference-in-difference approach. Ideally, we would select a differencing group where Assumption 11 holds; that is, a group where the treatment would have no effect. In many circumstances, though, it may not be possible to find a differencing group where treatment will have no effect. We were unable to find such a group for the JTPA Study, and indeed it is likely that Assumption 11 is violated for our differencing group because the job training program should have effects for single men. However, we anticipate the effects of the training program will be positive, yet smaller, for single men than for married men due to evidence that marriage improves men's productivity (e.g., see Korenman and Neumark (1991)). In this case, the front-door difference-in-differences approach should provide a lower bound on the job training effect for married men.

The Department of Labor implemented the National JTPA Study between November 1987 and September 1989 in order to gauge the efficacy of the Job Training Parternship Act (JTPA) of 1982. The Study randomized JTPA applicants into treatment and control groups at 16 study sites (referred to as

[2] Age for adult men ranges from 22 to 54 at random assignment / eligibility screening. Once married men comprises individuals who report that they are widowed, divorced, or separated.

[3] See discussion of how we created our sample and the earnings data in Appendix A.3.

service delivery areas, or SDAs) across the United States. Participants randomized into the treatment group were allowed to receive JTPA services, whereas those in the control group were prevented from receiving program services for an 18-month period following random assignment (Bloom et al., 1993; Orr et al., 1994). Crucially for our analysis, 61.4% of married men allowed to receive JTPA services actually utilized at least one of those services. Moreover, the Study also collected a nonexperimental comparison group of individuals who met JTPA eligibility criteria but chose not to apply to the program in the first place.[4] Since this sample of eligible nonparticipants (ENPs) was limited to 4 service delivery areas, we restrict our entire analysis to only these 4 sites.

2.4.1 Results

The front-door and front-door difference-in-differences estimates for the effect of the JTPA program on married males - our group of interest - are presented in Figure 2.2 across a range of covariate sets. Additionally, we present the standard covariate adjusted estimates for comparison. We use OLS separately within experimental treated and observational control groups (the ENPs) for the standard estimates. For front-door estimates, we use OLS separately within the "experimental treated and received treatment" and "experimental treated and didn't receive treatment" groups. Therefore, these estimates assume linearity and additivity within these comparison groups when conditioning on covariates, albeit we note that we obtain similar results when using more flexible methods that relax these parametric assumptions. The experimental benchmark (dashed line), is the only estimate that uses the experimental control units.

First, the front-door estimates exhibit uniformly less estimation error than estimates from standard covariate adjustment across all conditioning sets in Figure 2.2. The error in the standard estimates for the null conditioning set and conditioning sets that are combinations of age, race, and site are nega-

[4] See Appendix A.3 for additional information regarding the ENP sample. See Smith (1994) for details of ENP screening process.

tive. The error becomes becomes positive when we include baseline earnings in the conditioning set. In sharp contrast, the stability of front-door estimates is remarkable. We thus find that front-door estimates are preferable to standard covariate adjustment when more detailed information on labor force participation and historic earnings is not available.

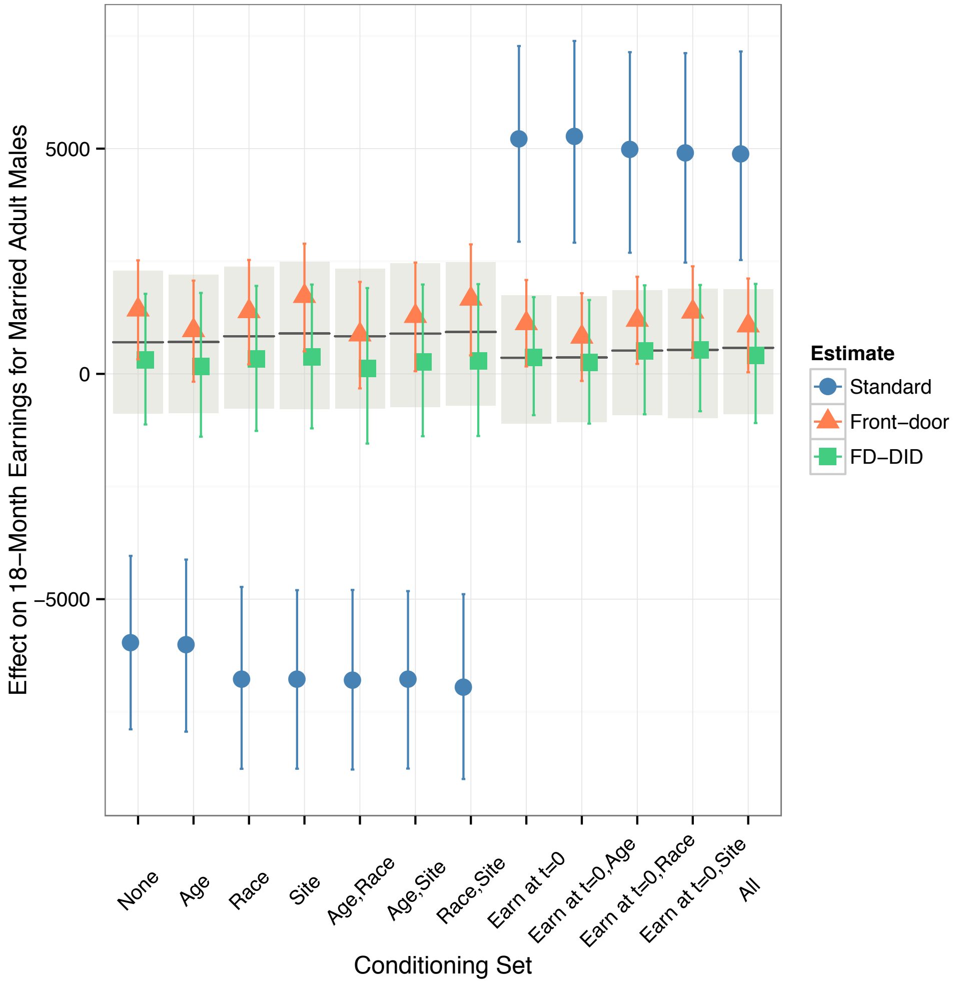


Figure 2.2: Comparison of standard covariate adjusted estimates, front-door, and front-door difference-in-differences estimates for the JTPA effect for married adult males. The dashed line is the experimental benchmark. 95% bootstrapped confidence intervals are based on 10,000 replicates.

In spite of the superior performance of front-door estimates compared to standard covariate adjustment, the front-door estimates are slightly above the experimental benchmark across all covariate sets. Even without seeing the experimental benchmark, these estimates are likely affected by positive bias because those that fail to show up to the job training program are likely to be less diligent individuals than those that show up. Given the anticipated positive bias in the front-door estimates, we use the front-door difference-in-differences estimator to either recover an unbiased point estimate or obtain a lower bound, depending on our assumptions as to the effect of the program in the differencing group. If we believe that the JTPA program had no effect for single males (i.e., Assumption 11 holds), and we also believe that Assumptions 3.A and 12 hold, then the difference-in-differences estimator will return an unbiased estimate of the effect for the group of interest in large samples. If, on the other hand, we believe there might be a non-negative effect for single males, then we would obtain a lower bound for the effect for the group of interest. In this application, it is more likely that there was positive effect of the JTPA program for single males, albeit one smaller than for married males. Hence, the front-door difference-in-differences estimator will likely give us a lower bound for the effect of the JTPA program for married males. In fact, in many applications we may be unable to find a differencing group with no effect, yet still be able to use front-door and front-door difference-in-differences approaches to bound the causal effect of interest given our beliefs about the sign and relative scale of effects in the group of interest and the differencing group.

When examining the empty conditioning set, the front-door estimate that we obtain for single males is $946.09. In order to construct the front-door difference-in-differences estimator, we have to scale this estimate by the ratio of compliance for married males to compliance for single males, which is equal to $0.614/0.524 \approx 1.172$. Subtracting the scaled front-door estimate for single males from the front-door estimate for married males as shown in (2.5), we obtain an estimate of $315.41. This is slightly below the experimental benchmark and thus indeed functions as lower bound. In sharp contrast to the front-door and front-door difference-in-differences estimates that rather tightly bound the truth, the bias in the standard estimate is -$6661.90. It is noteworthy that the front-door estimate acts as an upper bound and the front-door difference-in-differences estimate acts as a lower bound across all conditioning sets presented in Figure 2.2.

2.5 Illustrative Application: Early Voting

In this section, we present front-door difference-in-differences estimates for the average treatment effect on the treated (ATT) of an early in-person voting program in Florida. We want to evaluate the impact that the presence of early voting had upon turnout for some groups in the 2008 and 2012 presidential elections in Florida. In traditional regression or matching approaches (either cross sectional or difference-in-differences), data from Florida would be compared to data from states that did not implement early in-person voting. These approaches are potentially problematic because there may be unmeasured differences between the states, and these differences may change across elections. One observable manifestation of this is that the candidates on the ballot will be different for different states in the same election year and for different election years in the same state. The front-door and front-door difference-in-differences approaches allows us to solve this problem by confining analysis to comparisons made amongst modes of voting within a single presidential election in Florida.

Additionally, by restricting our analysis to Florida, we are able to use individual-level data from the Florida Voter Registration Statewide database, maintained since January 2006 by the Florida Department of State's Division of Elections. This allows us to avoid the use of self-reported turnout, provides a very large sample size, and makes it possible to implement all of the estimators discussed in earlier sections because we observe the mode of voting for each individual. The data contains two types of records by county: registration records of voters contained within *voter extract files* and voter history records contained in *voter history files*. The former contains demographic information – including, crucially for this paper, race – while the latter details the voting mode used by voters in a given election.

The two records can be merged using a unique voter ID available in both file types. However, voter extract files are snapshots of voter registration records, meaning that a given voter extract file will not contain all individuals appearing in corresponding voter history file because individuals move in and out of the voter registration database. We therefore use voter registration files from four time periods to match our elections of interest: 2006, 2008, and 2010 book closing records, and the 2012 post-election registration record. Our total population, based on the total unique voter IDs that appear in any of the voter registration files, is 16.4 million individuals. Appendix A.4 provides additional information regarding the pre-processing of the Florida data.

Information on mode of voting in the voter history files allows us to define compliance with the program for the front-door estimator (i.e., those that utilize EIP voting in the election for which we are calculating the effect are defined as compliers). Additionally, we use information on previous mode of voting to partition the population into a group of interest and differencing groups. In order to maximize data reliability, we define our group of interest as individuals that used EIP in a previous election (e.g., 2008 EIP voters are the group of interest when analyzing the turnout effect for the 2012 election). In other words, we are assessing what would have happened to these 2008 EIP voters in 2012 if the EIP program had not been available in 2012. To calculate the EIP effect on turnout for the 2012 election, we separately consider 2008 and 2010 EIP voters as our groups of interest. For the 2008 EIP effect on turnout, we rely upon 2006 EIP voters as our group of interest. An attempt to define the group of interest more broadly (e.g., including non-voters) or in terms of earlier elections (e.g., the 2004 election) would involve the use of less reliable data, and would therefore introduce methodological complications that are not pertinent to the illustration presented here.[5] Therefore, the estimates presented in this

[5] Following Gronke and Stewart (2013), we restrict our analysis to data starting in 2006 due to its greater reliability than data from 2004. We also might like to extend the group of interest to those that did not vote in a previous election, but we avoid assessing either 2008 or 2012 EIP effects for these voters because it is difficult to calculate the eligible electorate and consequently the population of non-voters. In their analysis of the prevalence of early voting, Gronke and Stewart (2013) use all voters registered for at least one general election between 2006 and 2012, inclusive, as the total eligible voter pool. However, using registration records as a proxy for the

application are confined only to those individuals that utilized EIP in a previous election and hence we cannot comment on the overall turnout effect.

We consider two differencing groups for each analysis: those who voted absentee and those that voted on election day in a previous election. When considering the 2012 EIP effect for 2008 EIP voters, for example, we use 2008 absentee and election day voters as our differencing groups. It is likely that the 2012 EIP program had little or no effect for 2008 absentee voters and perhaps only a minimal effect for 2008 election day voters, as these groups had already demonstrated an ability to vote by other means. Therefore, any apparent effects estimated for these groups will be primarily due to bias, and this bias can then be removed from the estimates for the group of interest. As discussed in earlier sections, the estimates from the differencing groups must be scaled according to the level of compliance for the group of interest. Finally, the existence of two differencing groups allows us to conduct a placebo test by using election day voters as the group of interest and the absentee voters as the differencing group in each case. This analysis is explored below.

Despite the limited scope of the estimates presented here, these results have some bearing on the recent debates regarding the effects of early voting on turnout. There have been a number of papers using cross state comparisons that find null results for the effects of early voting on turnout (Gronke et al., 2007, 2008; Fitzgerald, 2005; Primo et al., 2007; Wolfinger et al., 2005), and Burden et al. (2014) finds a surprising negative effect of early voting on turnout in 2008.[6] However, identification of turnout effects from observational data using traditional statistical approaches such as regression or matching rely on the absence of unobserved confounders that affect both election laws and turnout (Hanmer, 2009). If these unobserved confounders vary across elections, then traditional difference-in-differences estimators will also be biased. See Keele and Minozzi (2013) for a discussion within the context of election

eligible electorate may be problematic (McDonald and Popkin, 2001). By focusing on the 2008 voting behavior of individuals who voted early in 2006, we avoid the need to define the eligible electorate and the population of non-voters.

[6] Burden et al. (2014) examine a broader definition of early voting that includes no excuse absentee voting.

laws and turnout. Additionally, a reduction in Florida's early voting program between 2008 and 2012 provided evidence that early voting may encourage voter turnout (Herron and Smith, 2014).

The front-door estimators presented here provide an alternative approach to estimating turnout effects with useful properties. First, front-door adjustment can identify the effect of EIP on turnout in spite of the endogeneity of election laws that can lead to bias when using standard approaches. Second, unlike traditional regression, matching, or difference-in-differences based estimates, the front-door estimators considered here only require data from Florida within a given year. This means that we can effectively include a Florida/year fixed effect in the analysis, and we do not have to worry about cross-state or cross-time differences skewing turnout numbers across elections. We also include county fixed effects in the analysis in order to control for within-Florida differences.

However, in addition to the limited scope of our analysis, it is important to note that the exclusion restriction is likely violated for this application. Since early in-person voting decreases waiting times on election day, it is possible that it actually increases turnout among those that only consider voting on election day. This would mean that front-door estimates would understate the effect if all other assumptions held because the front-door estimator would be ignoring a positive component of the effect. Alternatively, Burden et al. (2014) suggest that campaign mobilization for election day may be inhibited, such that early voting hurts election day turnout. This would mean that front-door estimates would overstate the effect because the front-door estimator would be ignoring a negative component of the effect. This can also be seen by examining the bias formula (2.4) (because the EIP treatment is assigned at the state level, Assumptions 3.A and 10 will hold).

Taken together, the overall effect of these exclusion restrictions is unclear and would depend on the strength of the two violations. The predictions also become less clear once we consider the front-door difference-in-differences approach, where additional bias in front-door estimates might cancel with bias in the estimates for the differencing group. For the remainder of this analysis, we will assume that all such violations of the exclusion restriction cancel out in the front-door difference-in-differences estimator. This is implicit in Assumption 12.

2.5.1 Results

In order to construct the front-door estimate of the 2008 EIP effect for our group of interest, we calculate the turnout rate in 2008 for all individuals who voted early in 2006. We also calculate the non-complier turnout rate in 2008 by excluding all individuals who voted early in 2008 from the previous calculation. The front-door estimate of the 2008 EIP effect for 2006 early voters is thus the difference between the former and latter turnout rates. Quite intuitively, the counterfactual turnout rate without EIP for the group of interest is the observed turnout rate of non-compliers in that group. We do not devote much attention to the front-door estimates seeing as they are implausibly large.[7] The positive bias stems from the fact that 2006 EIP voters would be more likely to vote in 2008, even in the absence of EIP, than the 2006 non-EIP group (this group includes individuals that did not vote in 2006). In terms of the bias formula in (2.4), this is equivalent to saying that $E[Y(a_0)|a_1, m_1, x] > E[Y(a_0)|a_1, m_0, x]$.

In order to address this bias, we present front-door difference-in-differences estimates for the 2008 EIP program in Figure 2.3. The estimates all utilize county fixed effects and are calculated separately across the racial categories.[8] The front-door difference-in-differences estimates for the group of interest (2006 EIP voters) are in green, with 2008 absentee voters (triangles) and 2008 election day voters (squares) as the differencing groups. The former, for example, is constructed as the difference between front-door estimates for 2006 early voters and the front-door estimates for 2006 absentee voters, with the front-door estimates for the differencing group scaled by the ratio of early voter compliance to ab-

[7] Front-door estimates are available in Appendix A.5.

[8] We calculate the FD-DID estimates within each county and then average using the population of the group of interest as the county weight. Due to very small sample sizes in a few counties, we are occasionally unable to calculate front-door estimates. In these cases, we omit the counties from the weighted average when calculating the front-door estimates with fixed effects. We note that due to their small size, these counties are unlikely to exert any significant impact upon the estimates regardless.

sentee voter compliance as shown in (2.5). The purple estimates in Figure 2.3 represent the placebo test, with 2006 election day voters standing in as the group of interest and the absentee voters as the differencing group. In general, we note that if there exists more than one plausible differencing group, then one should conduct the analysis using each differencing group separately, as well as a placebo test to verify the plausibility of Assumption 11.

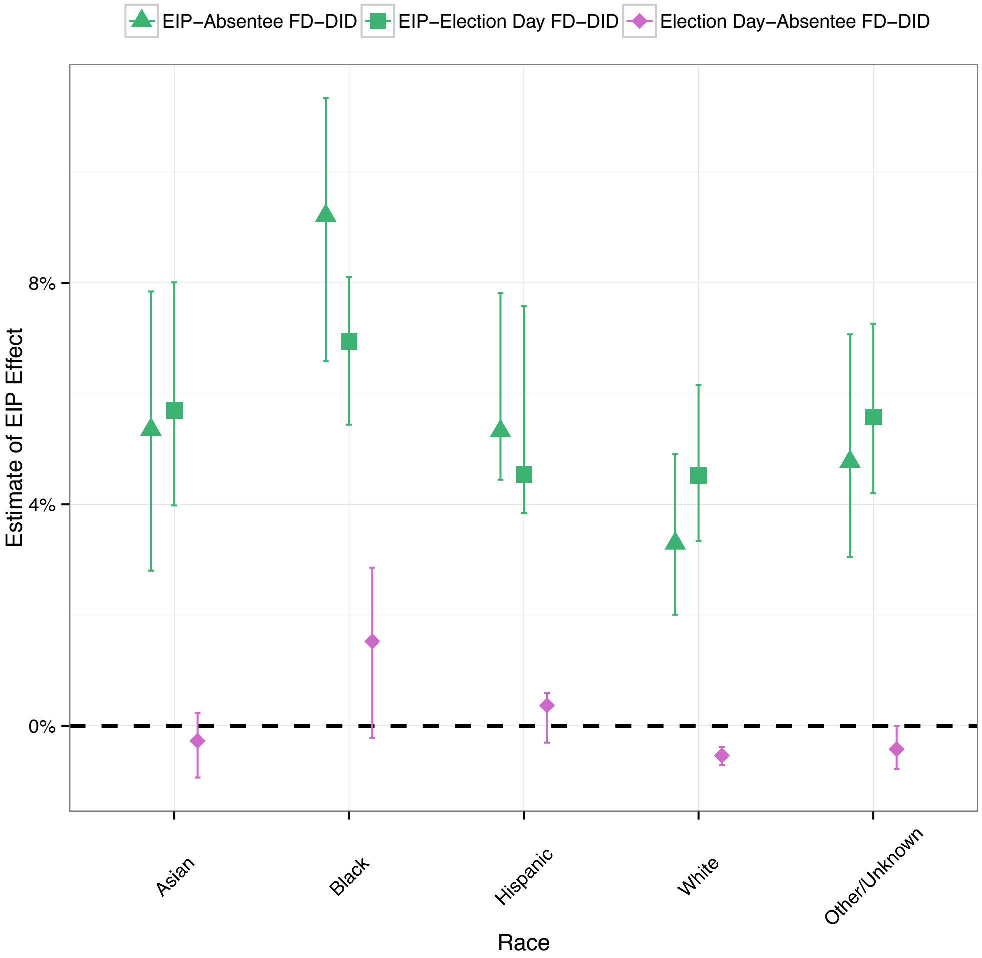


Figure 2.3: Front-door difference-in-differences estimates for the turnout effect in 2008 for voters who voted early in 2006 (by race). All estimates include county fixed effects. 99% block bootstrapped confidence intervals are based on 10,000 replicates.

The EIP program estimates are positive and significant at the 99% level. All placebo tests, with the exception of the white estimate, are indistinguishable from zero, giving us confidence in the estimated EIP effects. Even if the slightly negative placebo estimate for whites indicates a true negative effect of the 2008 EIP program, and not bias, the weighted average of the green and the purple effects (i.e., the 2008 EIP effect for the 2006 EIP and election day voters together), again produces a slightly positive estimate. Therefore, we generally find evidence that early voting increased turnout for the subset of individuals who voted early in 2006. Moreover, comparing the point estimates across races, we find some evidence that the program had a disproportionate benefit for African-Americans.

Our methodology uses voting behavior in 2006 only to define groups and does not compare turnout of voters across elections. Thus any differences between presidential election and midterm election voters (see e.g. Gronke and Toffey (2008)) does not pose a prima facie problem for the analysis. Moreover, using early voters in a midterm election as the group of interest for calculating the EIP effect in a presidential election does not require additional assumptions beyond what one would need if using early voters in a presidential election. Nonetheless, a potential downside of the preceeding results is that the estimated 2008 EIP is limited to those individuals who voted early in the 2006 midterm election, whereas we might want to extend the group of interest to early voters in a presidential election. Unfortunately, we cannot present the estimates with the group of interest and the differencing groups defined in terms of 2004 behavior because the data from 2004 are not reliable (as mentioned above). As a robustness check, we also estimate the effect of the early voting program in the 2012 election, for which we can define the group of interest and the differencing groups using 2008 voting behavior. However, as discussed above, Florida's early voting program was reduced between 2008 and 2012, so we should not expect the results to be equivalent.

The results of this analysis are presented in Figure 2.4. For the 2008 EIP voters, the 2012 EIP front-door difference-in-differences estimates (green) are positive and significant at the 99% level (based on 10,000 block bootstraps at the county level). There is some evidence of differences between the racial categories, but these differences change depending on which differencing group is used. The purple estimates are for the most part indistinguishable from zero, indicating that the placebo tests have mostly been passed. The slightly negative purple estimate for whites again indicates either bias, or perhaps a negative effect of the 2012 EIP program for white 2008 election day voters. Note that even if we believe this estimate, the weighted average of the green and the purple effects for whites (i.e., the 2012 EIP effect for the 2008 EIP and election day voters together) produces a slightly positive estimate, albeit this estimate is indistinguishable from zero. In sum, the evidence points to a slightly positive turnout effect of the 2012 EIP program on the 2008 EIP users.

It is also notable that the size of the estimated EIP effect for 2012 is less than half the estimated EIP effect for 2008 when looking at EIP voters as the group of interest across all races. There are two potential reasons for this. First, our estimates for the 2008 EIP program are obtained using groups defined by 2006 midterm election behavior and as already mentioned, midterm election early voters are likely different than presidential election early voters. Second, the nature of the early voting program changed between the 2008 and 2012 elections, notably removing the option of voting early on the Sunday prior to the election and all-together near-halving of the early voting period from 14 days to 8 days (Gronke and Stewart, 2013; Herron and Smith, 2014). This change might possibly reduce the effect of the EIP program in 2012 when compared to 2008 - a finding consistent with the conclusion made by Herron and Smith (2014) that individuals who voted in 2008 on the Sunday prior to the election were disproportionately less likely to vote in 2012.

In order to isolate the consequences of the change in the early voting program from changes in the construction of the group of interest and differencing groups, we re-estimate the effects of the 2012 EIP program using 2010 EIP voters as the group of interest (green), and using 2010 absentee (triangles) and election day voters (squares) as the differencing groups. Placebo tests are reported using 2010 election day voters as the group of interest and 2010 absentee voters as the differencing group (purple). These results are presented in Figure 2.5, and they are quite similar to the results in Figure 2.3. This provides

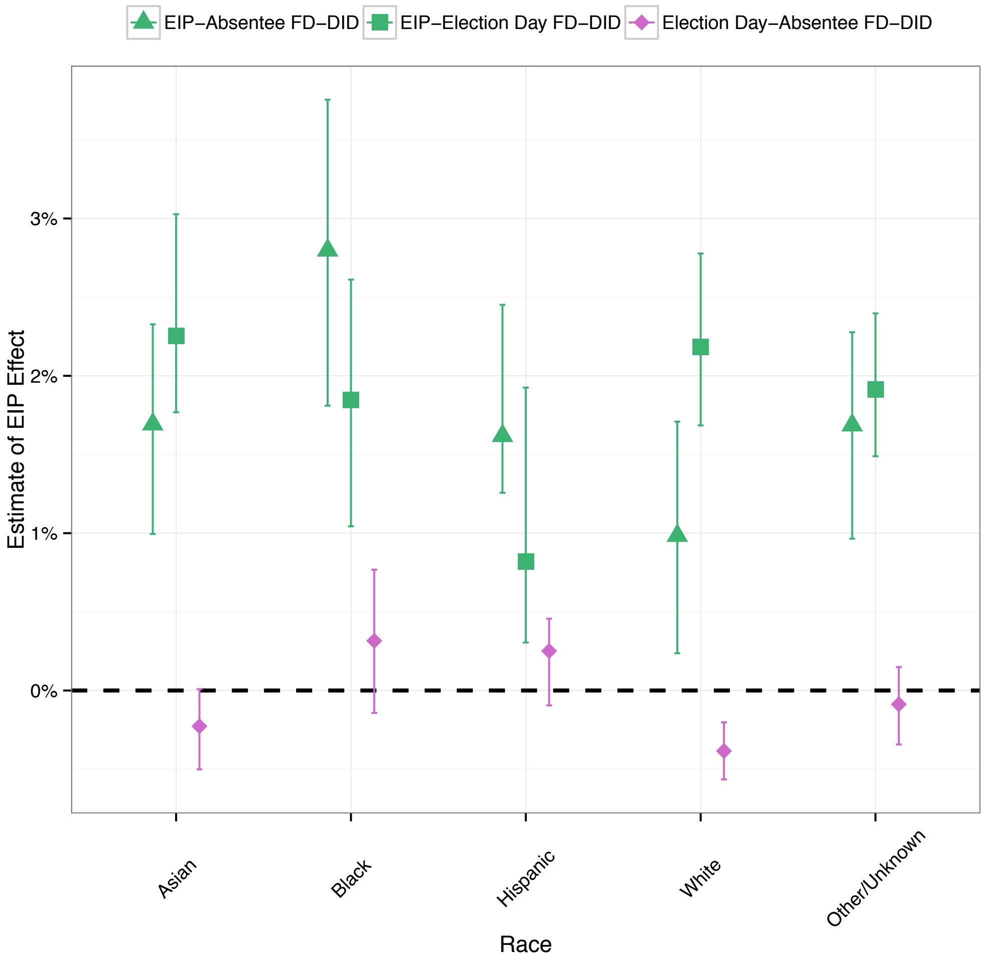


Figure 2.4: Front-door difference-in-differences estimates for the turnout effect in 2012 for voters who voted early in 2008 (by race). All estimates include county fixed effects. 99% block bootstrapped confidence intervals are based on 10,000 replicates.

some evidence that if we were able to obtain reliable data from the 2004 election, our estimates for the 2008 EIP program would likely have produced something similar to Figure 2.4 when using 2004 EIP voters as the group of interest, and 2004 absentee (triangles) and election day voters (squares) as the differencing groups. However, the estimates in Figure 2.3 are slightly larger than estimates in Figure 2.5. This is consistent with the reduction in the early voting window for the 2012 election.

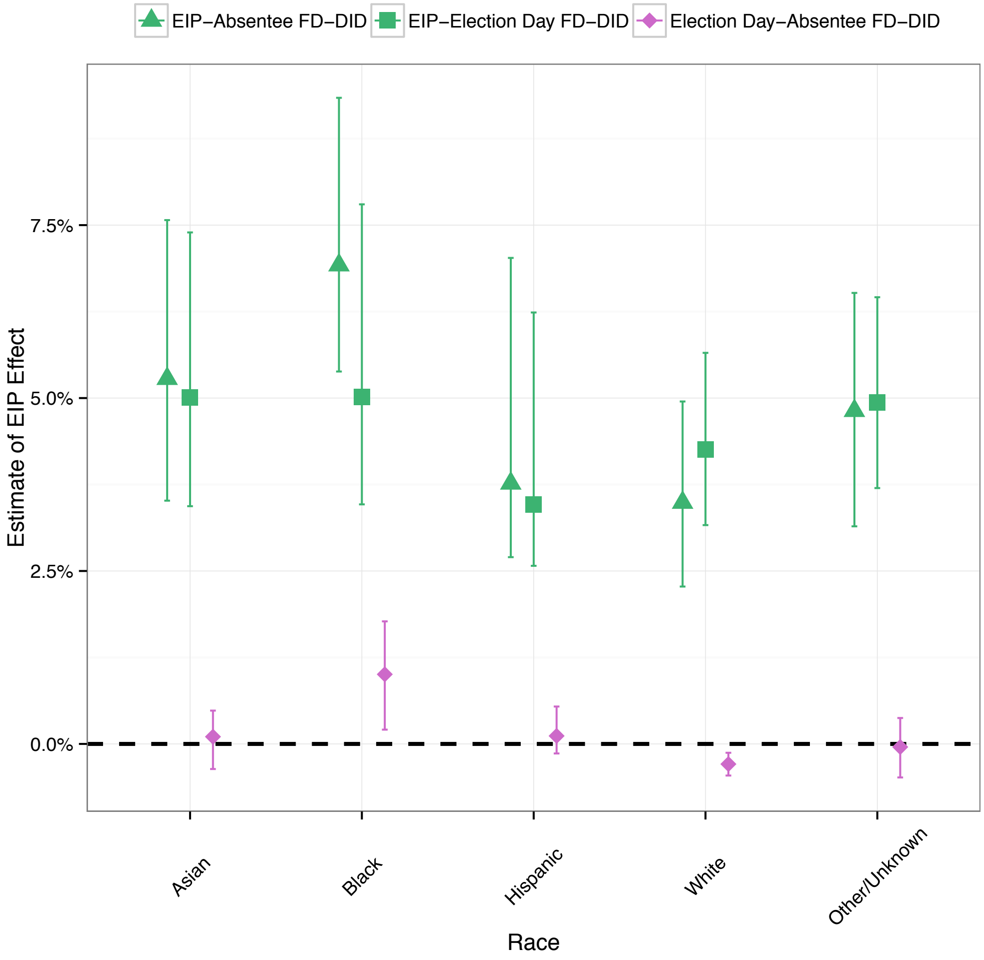


Figure 2.5: Front-door difference-in-differences estimates for the turnout effect in 2012 for voters who voted early in 2010 (by race). All estimates include county fixed effects. 99% block bootstrapped confidence intervals are based on 10,000 replicates.

2.6 Conclusion

In this paper, we have developed front-door difference-in-differences estimators for nonrandomized program evaluations with one-sided noncompliance and an exclusion restriction. These estimators allow for asymptotically unbiased estimation via front-door techniques, even when front-door estimators have significant bias. Furthermore, this allows for program evaluation when all of the relevant units have been assigned to treatment.

We illustrated this technique with an application to the National JTPA (Job Training Partnership Act) Study and with an application to the effects of Florida's early in-person voting program on turnout. For the job training application, we showed that front-door and front-door difference-in-differences could be used to bracket the experimental benchmark. For the application to the effects of an early in-person (EIP) voting program on turnout in Florida in 2008 and 2012, we found that for two separate differencing groups, the program had small but significant positive effects. While the scope of the analysis is limited, this result provides some evidence to counter previous results in the literature that early voting programs had either no effect or negative effects.

Finally, the results in this paper have implications for research design and analysis. First, the examples demonstrate the importance of collecting post-treatment variables that represent compliance with, or uptake of, the treatment. Such information allows front-door and front-door difference-in-differences analyses to be carried out as a robustness check on standard approaches. Second, the bracketing of the experimental benchmark in the JTPA application show that control units are not always necessary for credible causal inference. This is a remarkable finding that should make a number of previously infeasible studies possible (e.g., when it is unethical or impossible to withhold treatment from individuals).

3

Causal Inference without Control Units

3.1 Introduction

The necessity of control units is taken as a fundamental tenet of research design for causal inference regarding treatment effects. However, outside of the research community, causal inference is routinely attempted without control units. In some cases, the lack of control units merely reflects a lack of planning (or understanding of research principles); for others, there are defensible reasons. Sometimes ethical concerns rule out withholding treatment from anyone. Other times, logistical hurdles prevent withholding treatment. Still other times, withholding treatment is ruled out due to concerns about disrupting business practices, or the costs assumed to be associated with foregoing treatments that are "known" to work.

All of these impediments to the use of control units can also occur when attempting to replicate a small study that used control units (and perhaps randomized treatment assignment). For example, it is quite common for studies to be halted once treatment has been shown to be beneficial for some units. For multi-site studies, it is possible for some sites to allow experimental control, while others do not (e.g., the Job Training Partnership Study which is discussed in detail below). Once a treatment has been demonstrated to have an effect in a small study or a single site, it can be difficult/imprudent to conduct follow up controlled studies to replicate the finding. Yet there may be concerns about the continuing efficacy of treatments, or about heterogeneity across sites.

Furthermore, causal inferences are routinely attempted by the non-research community, even in the absence of control. In some cross-sectional cases, these inferences are made via process tracing. For example, if a program is started to assist individuals in the completion of a task, then it is common (although misguided) to take the number of individuals partaking in the program as direct evidence of the effect of the program on completion of the task. This approach is particularly routine when the person doing the analysis has an incentive to justify the money spent on the program. In longitudinal cases where all units receive treatment at the same time, causal inferences are made through the use of before/after studies or some close variant (e.g., interrupted time series studies).

In this paper, we show that for either the cross-sectional or longitudinal case, compliance information can be leveraged via front-door techniques to improve causal inferences. In order to establish the efficacy of the approach, we attempt to estimate or bound experimental benchmarks from a number of experimental studies, using only the treated units from these studies. In Section 3.2, we consider the use of the front-door approach with one-sided noncompliance. We separately consider this approach for cross sectional studies on voter turnout and health insurance, and a longitudinal study of a job training program. In Section 3.3, we discuss the use of the front-door difference-in-difference approach in three scenarios. The first involved the use of a prototypical differencing group and serves as a proof of concept. The second is a scenario where previous studies have established the differencing group. The third is a scenario where randomized control has been allowed at one study site, but not at others. The differencing group is chosen here by studying the heterogeneity of effects in the first site. We conclude the paper by discussing the possibilities and drawbacks of causal inference without control units in future studies.

3.2 Front-door

In this section, we consider the use of the front-door technique to establish the primary parameter of interest in program evaluations: the average treatment effect on the treated. In intuitive terms, this parameter expresses the difference between the average observed outcome and the average outcome for the same individuals if the program had not been implemented. In formal terms, we define active treatment (being assigned or self-selected into the program) as a_1, while we define control as a_0. We define the observed outcome as Y and the potential outcome under control as $Y(a_0)$ (the outcome that would have been observed if the program had not been implemented). The average treatment effect on the treated can be defined as the following:

$$\tau_{ATT} = E[Y - Y(a_0)] \tag{3.1}$$

Throughout this paper, we additionally assume that if not assigned to the program one cannot receive the program. For example, if you don't receive such a call, it is not possible to not comply and answer the call. In formal terms, we define compliance with treatment as m_1 and noncompliance as m_0. Hence we assume one-sided noncompliance as in Assumption 3.A in Chapter 1.

Assumption 3.A states that it is not possible to receive the active treatment if one is assigned to control. Under this assumption, Chapter 1 shows that the front-door estimator can be written as the following (where x indicates control variables and these are assumed to be discrete for simplicity in presentation):

$$\tau_{att}^{fd} = E[Y|a_1] - \sum_x \underbrace{E[Y|a_1, m_0, x]}_{\text{treated non-compliers}} \cdot P(x|a_1) \tag{3.2}$$

$$= \sum_x P(x|a_1) \cdot P(m_1|x, a_1) \cdot \left\{ \underbrace{E[Y|a_1, m_1, x] - E[Y|a_1, m_0, x]}_{\text{"effect" of receiving treatment}} \right\} \tag{3.3}$$

The first thing to note about this estimator is that it does not involve the use of control units (i.e., a_0 does not appear). The second thing to note, is that we are being a bit informal with this notation. Some versions of this estimator don't allow two compliers to have the same value of x (e.g., nearest neighbor matching without replacement). Finally, both (3.2) and (3.3) allow for useful interpretations of the front-door approach vis-a-vis other approaches that do not use control units, in both cross sectional and longitudinal settings.

3.2.1 Cross-sectional

In the cross-sectional setting, the simplest possible "no controls" estimator is $E[Y|a_1]$, that is, the average outcome among the treated. For example, one might estimate the effect of a phone GOTV program as the proportion that voted among those that received a call. Of course, this estimator is fundamentally flawed because it does not consider the counterfactual of what might have happened to these individuals in the control condition. For example, would the phone call recipients have voted even if they did not receive the call? The front-door estimator (3.2) improves on this naive estimator by using non-compliers as proxies for the counterfactuals (and weighting according to covariates).

Note that the front-door analysis is premised on the idea that the treatment could only have an effect through compliance. This is known as an exclusion restriction and is stated formally as Assumption 4 in Chapter 1. However, once we make this assumption, there is an alternative to the simple "no controls" estimator: average the outcomes for the compliers and zeros for the noncompliers ($\sum_x P(x|a_1) \cdot P(m_1|x, a_1) \cdot E[Y|a_1, m_1, x]$). This is a process tracing approach. For example, instead of estimating the effect of a phone GOTV program as the proportion that voted among those that received a call, one might estimate it as the proportion that both answered the phone and voted among those that received a call. The front-door (3.3) clearly improves on this by examining the proportion that received a call, didn't answer, but voted anyway, and using this group to estimate what the voting rate would have been among those that did answer the phone, if instead they had not been called.

Of course, even when Assumptions 3.A and 4 hold, the front-door estimator can still be biased. Informally, this can be seen in (3.3), where one-sided noncompliance and the exclusion restriction are sufficient to transfer the problem of estimating the treatment effect into the problem of estimating the compliance effect. Hence remaining bias will be due to bias in the estimation of this effect (see Chapter 1 for a formal presentation).

However, even though the front-door approach is likely to be biased for many applications, for many of those applications we have expectations about the sign of that bias. For example, those that answer the GOTV phone call are at home and likely to be older. Older people vote more regularly and hence are more likely to vote in the absence of the phone call. This means that the front door estimator (as conceptualized in (3.3)) is likely to have positive bias. Because we have already established that the estimator will produce smaller estimates then the naive approaches for this application, the front door approach produces a tighter upper bound.

To demonstrate the potential efficacy of the front-door approach in this scenario, we replicated the experimental estimates from 19 GOTV phone studies (from 8 papers). Details of the studies, as well as our criteria for inclusion of these studies, is available in Appendix A.6. See the black horizontal line segments in Figure 3.1. We also present front-door estimates that only use the treated units from these studies. These are represented by the orange triangles on the same plot. Note that the front-door estimates are, as expected, larger than the experimental estimates. Therefore, even if these experiments had not been run, we could have estimated upper bounds on the effect sizes, which may be useful when doing a cost-benefit analysis.

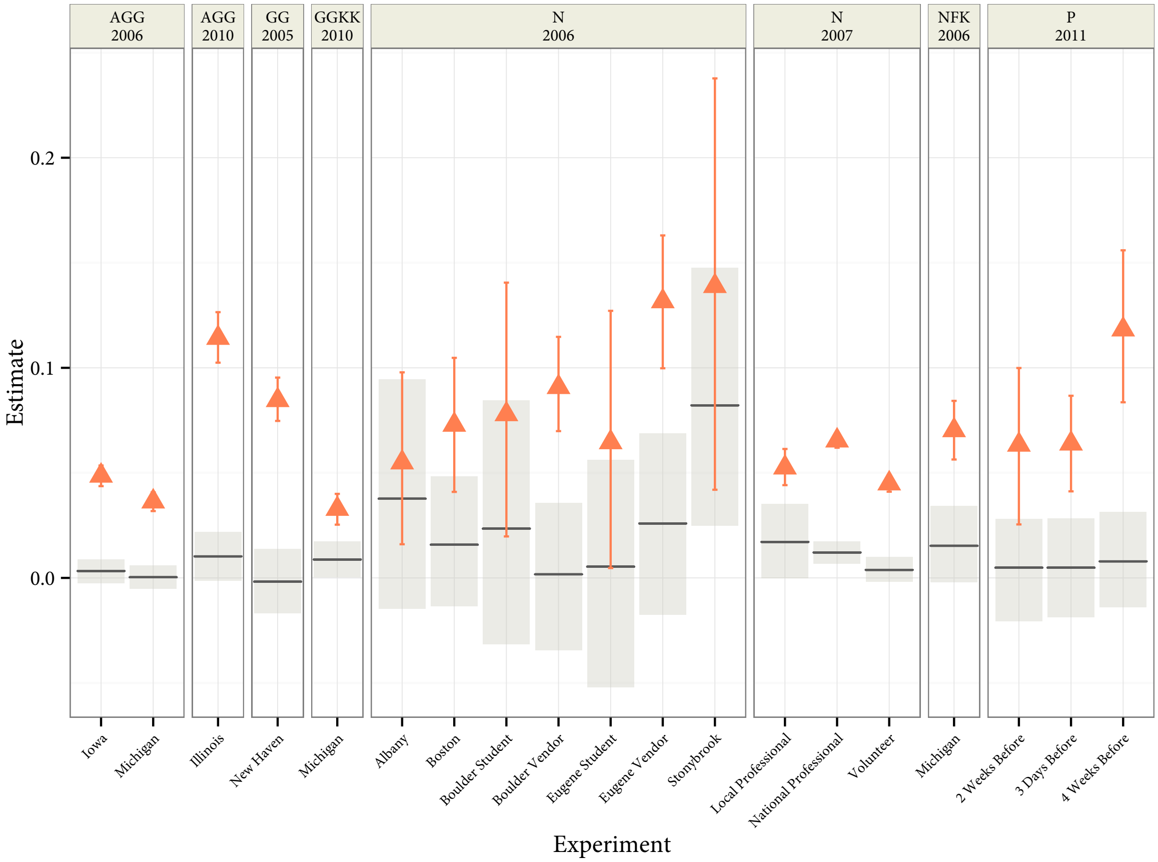


Figure 3.1: Front-door Estimates of Get-Out-The-Vote Phone Call Treatment.

The GOTV phone example demonstrated that the front-door approach with noncompliers can establish useful upper bounds. However, it is also possible that the front-door approach will establish useful lower bounds. The Oregon Health Insurance Experiment provides a good example of this.

The Oregon Health Insurance Experiment (OHIE) randomized low-income, uninsured individuals into a lottery for Medicaid in Oregon in 2008. If treated, the individual could apply for the Oregon Health Plan. We replicate Tables 5-6 and 8-10 of Finkelstein et al. (2012), encompassing the effect of Medicaid access upon 5 different outcome categories, as measured by survey responses. This represents the entirety of tables replicable from the public use file. Outcome data is based on a mail survey conducted roughly 12 months after randomization. The sample we use are those that answered the mail survey (analysis is robust to including survey weights, but we leave them out for simplicity). We also control for eight covariates available from the form used to register for lottery. Details of the study and the covariates are available in Appendix A.7.

The replicated experimental effects are presented as the black horizontal line segments in Figure 3.2. These experiments also collected information on compliance with the program: whether or not an individual, once accepted, actually sent in a completed application within 45 days. The compliance rate in our sample is 71%. In constructing the front-door estimates note that because less healthy individuals are more likely to comply with the treatment, we expect the front-door estimates to have negative bias. The front-door results, reported as orange triangles on Figure 3.2, demonstrate this downward bias. However, we note that even by only using the treated units from these experiments, we were able to establish lower bounds on the size of effects for the program. In some cases these lower bounds are significantly larger than than zero.

3.2.2 Longitudinal

The previous subsection presented analysis and applications that did not use longitudinal data, however, there are many applications where longitudinal data is used to make causal inferences without control units. For example, in assessing a job training program where all individuals have been invited to join at the same time, we might estimate the effect of the program with average gain scores (i.e. the average of the difference between after program earnings and before program earnings). If more than

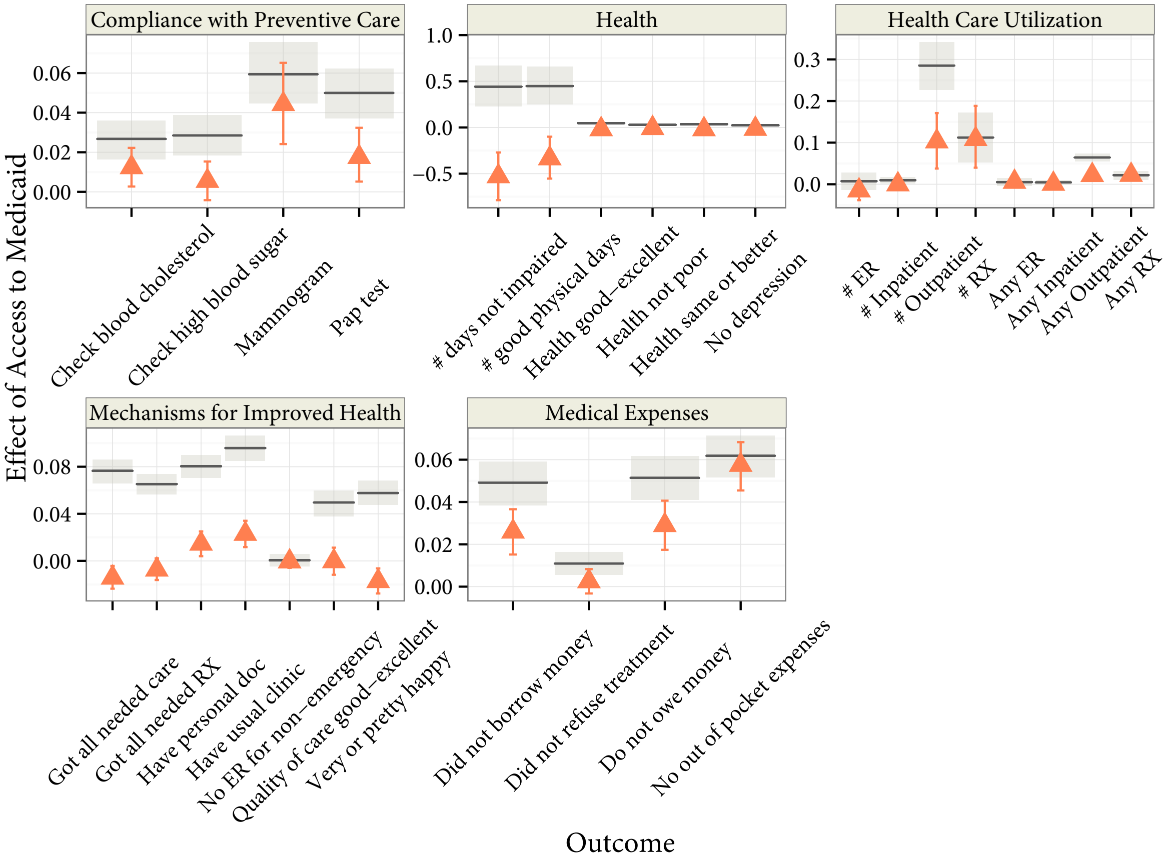


Figure 3.2: Front-door Estimates for Oregon Health Insurance Experiment (OHIE). Treatment is effect of access to Medicaid.

two time periods are available, we might average the results from an interrupted time-series analysis for each individual. Both average gain score and average interrupted time-series analysis can fit neatly into the notation above if we define Y to be either the gain score for an individual or the result from an interrupted time-series analysis for an individual. If compliance information is available (e.g., whether an individual actually enrolled in the job training program after signing up), then we might also scale these simple "no control" estimators with compliance as above. Specifically, we could take average the gain scores for those that comply and zeros for those that do not comply.

For front-door analysis in the longitudinal setting, there are three options. First, one could simply ignore the longitudinal nature of the data and base the analysis on the compliance comparison as shown above. This is not recommended generally, but it is important to note that this is possible, while there is no analogous possibility without the compliance information. Second, one could run the front-door analysis with Y defined to be the gain score or the result from an interrupted time-series analysis for an individual. Third, because the front-door approach accommodates the use of covariates, one could incorporate the pre-treatment outcome values as covariates in the analysis.

In order to compare these options, we use data from the National JTPA Study. The Department of Labor implemented the National JTPA Study between November 1987 and September 1989 in order to gauge the efficacy of the Job Training Parternship Act (JTPA) of 1982. Detailed information on the data is available in Appendix A.3. The three front-door options along with the compliance scaled gain score analysis are reported for the JTPA study data in Figure 3.3. The experimental benchmark is also reported as the dashed black line in the figure. Note that while the simple scaled gain-score analysis results in an estimate that is much too large, all three front-door approaches minimally overestimate the benchmark and their confidence intervals cover the benchmark.

3.3 Front-door Difference-in-Differences

Although we have shown that the front-door approach may produce useful bounds, there may applications where we would like to remove additional bias from the front-door estimator, or where we would like to produce an additional bound to bracket the parameter. For example, when the front-door estimator produces an upper bound, it would sometime be useful to also have a lower bound. The front-door difference-in-differences estimator Chapter 2 will allow us to accomplish one or the other of these goals.

The front-door difference-in-differences estimator functions similarly to standard difference-in-

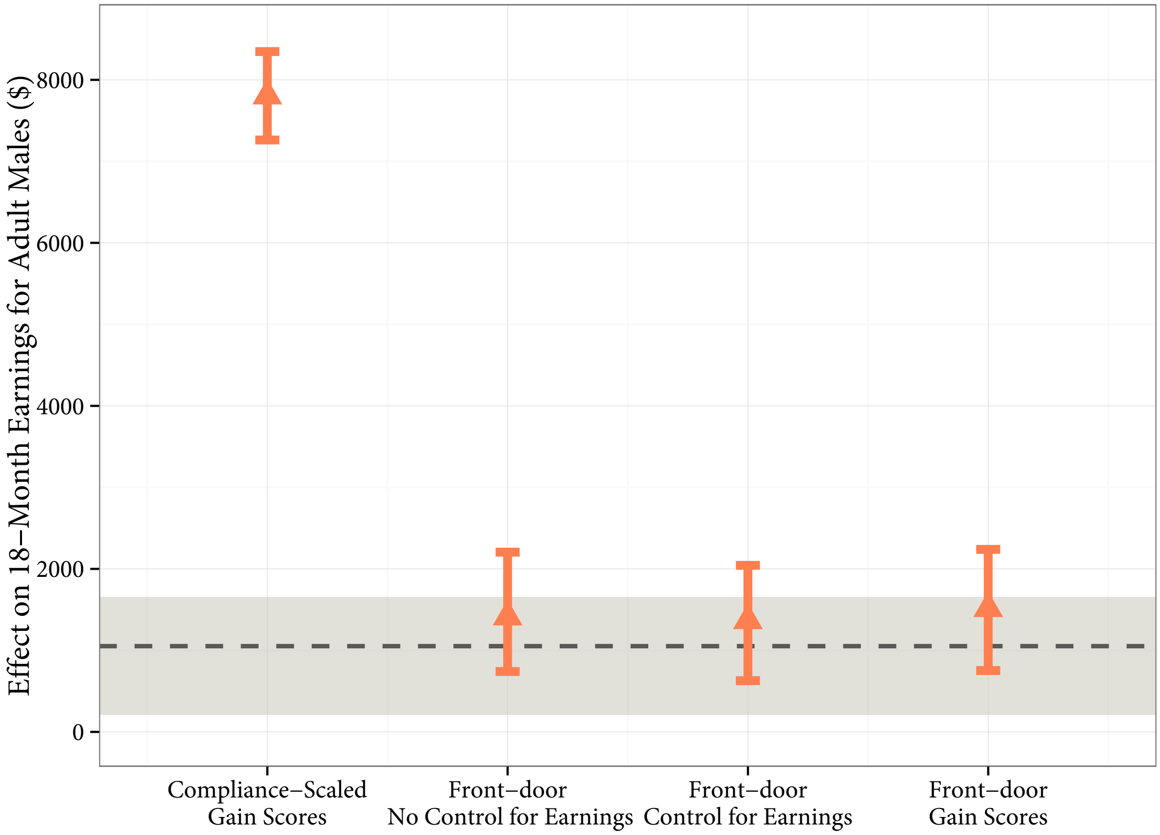


Figure 3.3: Front-door Estimates of the JTPA Program.

differences estimators in that it uses a group of observations for which there should be no effect to estimate a proxy for the bias in the group of interest. In the standard approach, the pre-treatment comparison between the treatment and the control units functions as the observations for which there should be no effect. The front-door version of this estimator works, as above, by shifting the treatment effect problem to a compliance effect problem. Hence, in order for this estimator to work well, one must find units for which the compliance effect is zero (or has known sign if bounding is the goal). We call this group of units a differencing group.

The formal presentation of the front-door difference-in-differences estimator is below in (3.4), where we further define the group of interest as the stratum g_1 and the differencing group as the stratum g_2:

$$\tau_{att}^{fd-did} = \sum_x P(x|a_1, g_1)P(m_1|x, a_1, g_1)\left[\{E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1]\}\right.$$
$$\left. - \{E[Y|a_1, m_1, x, g_2] - E[Y|a_1, m_0, x, g_2]\}\right]. \tag{3.4}$$

Note that $E[Y|a_1, m_1, x, g_1] - E[Y|a_1, m_0, x, g_1] - \{E[Y|a_1, m_1, x, g_2] - E[Y|a_1, m_0, x, g_2]\}$ looks like a standard difference-in-differences estimator, but for the compliance effect. The front-door version of this estimator converts back to the treatment effect by using compliance rate for the group of interest ($P(x|a_1, g_1)$). See Chapter 2 for details. The key question is how does one arrive at a differencing group. In the following subsections we discuss the prototypical differencing group and groups chosen on the basis of previous studies. The first of these is meant only to be illustrative. It presents an idealized scenario that is not useful for research purposes, but provides intuition about how the estimator should work. The second shows how a difference-in-differences approach can be used fruitfully in practice.

3.3.1 The Prototypical Case

When should we expect the front-door difference-in-differences estimator to work? First, we would need to find a differencing group for which the effect of compliance should be zero. For two of the GOTV phone studies (Arceneaux et al. (2010) and Gerber et al. (2010)) we have such a group of units, because one of the treatment arms in each of these studies received a placebo treatment. In Arceneaux et al. (2010), the placebo treatment was a "buckle up" phone message. In Gerber et al. (2010), the placebo treatment was a recycling phone message. Clearly, compliance with the placebo treatment (i.e., answering the phone for the "buckle up" phone message or the recycling phone message) should not affect the decision to turnout.

Furthermore, while compliance (answering the phone) was not randomized in either the treatment group (GOTV message) or the placebo group ("buckle up" or recycling message), the treatment and placebo groups were exchangeable due to randomization. This property, combined with the fact that the content of the phone message was unlikely to affect compliance, means that the bias in the estimate of the compliance effect should be the same for both the treatment group and the placebo group. Hence, the assumptions of the front-door difference-in-differences estimator should be satisfied (see Chapter 2 for details.)

In fact, this is exactly what we see when we calculate the front-door difference-in-differences estimates for the Arceneaux et al. (2010) and Gerber et al. (2010) studies. The experimental results, the front-door estimates, and the front-door differnece-in-differences estimates are presented in Figure 3.4. Note that while the front-door estimates are too high, likely for the aforementioned reason that older people are more likely to answer the phone, the front-door difference-in-differences estimates hit the experimental benchmark.

3.3.2 Differencing Groups Chosen on the Basis of Previous Studies

When we only have treated units, we will need to choose a differencing group from among those units. Often, this will be done on the basis of previous studies. For example, although the correspondence between treatments is not perfect,[1] Imai and Strauss (2011) use tree-based methods to identify heterogeneity in causal effects for a GOTV experiment from Dale and Strauss (2009). They found that age explains much of heterogeneity in treatment effects (voters ages 20-24 are very responsive to treatment, their model predicts zero or even negative treatment effects for ages 18-19).

Figure 3.5 presents the results of a front-door difference-in-differences analysis for the GOTV data, and following Imai and Strauss (2011) we use 18 and 19 year olds as the differencing group because

[1] The experiment randomly assigned newly registered voters to receive a text message urging them to vote in the 2006 midterm election.

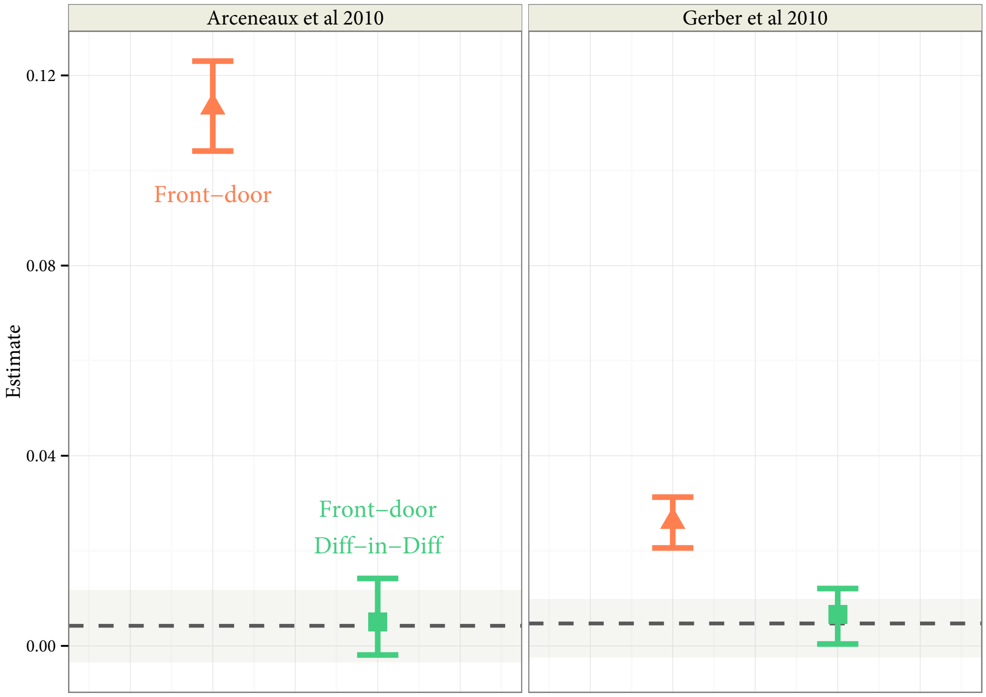


Figure 3.4: Prototypical GOTV Front-door Difference-in-Differences.

we expect the treatment to have approximately zero effect for this group. The front-door estimates are identical to those in Figure 3.1 and reprinted here for comparison. Note that the differencing approach appears to provide a better estimate of the experimental benchmark in all but two of the studies, although the standard errors are quite large due to the size of the differencing group.

We can similarly develop front-door difference-in-differences analysis for the JTPA data by consulting previous studies. There is evidence from economics literature that marriage increases male productivity (Korenman and Neumark, 1991; Hellerstein et al., 1999). As explained in Hellerstein et al. (1999), " The equality of relative marginal productivity and wages of married workers shows that the marriage wage premium reflects an underlying productivity differential and is not attributable to discrimination in favor of married workers. However, the result does not help sort out whether marriage increases the

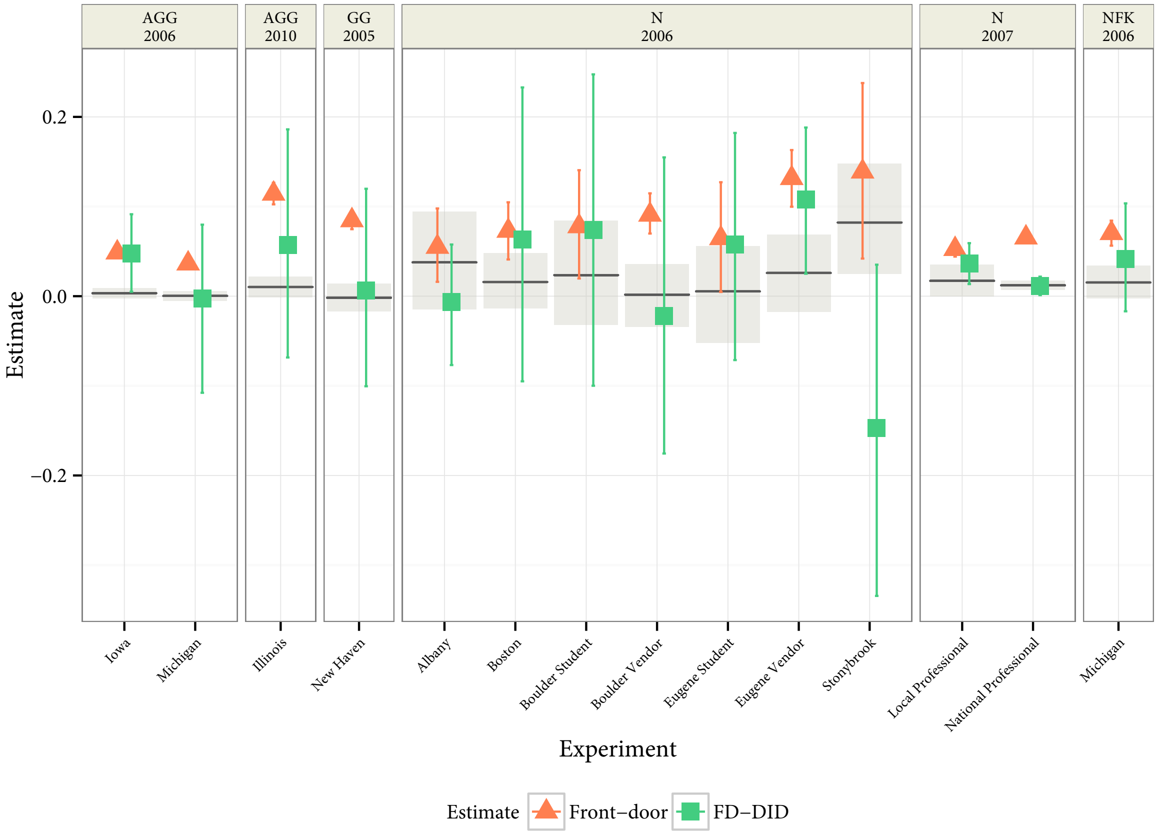


Figure 3.5: Front-door and Front-door Difference-in-Differences Estimates for Get-Out-the-Vote (GOTV) Phone Treatment. Differencing group is 18 and 19 year olds.

productivity of men or whether high productivity men are selected into marriage.

These studies might lead us to use single adult males as a differencing group, but note that unlike the previous two analyses, we only have reason to believe that the JTPA effect is smaller for single men. Therefore, a front-door difference-in-differences analysis will have negative bias (subtract off too much). However, in combination with the positive bias from the front-door analysis, we should now be able to bracket the experimental benchmark.

In fact, this is what we see in Figure 3.6, which presents the experimental results for adult males along with the previously presented front-door estimates and the front-door difference-in-differences analysis using single adult males as the differencing group. Note that the front-door and front-door difference-in-differences analyses bracket the experimental benchmark as expected.

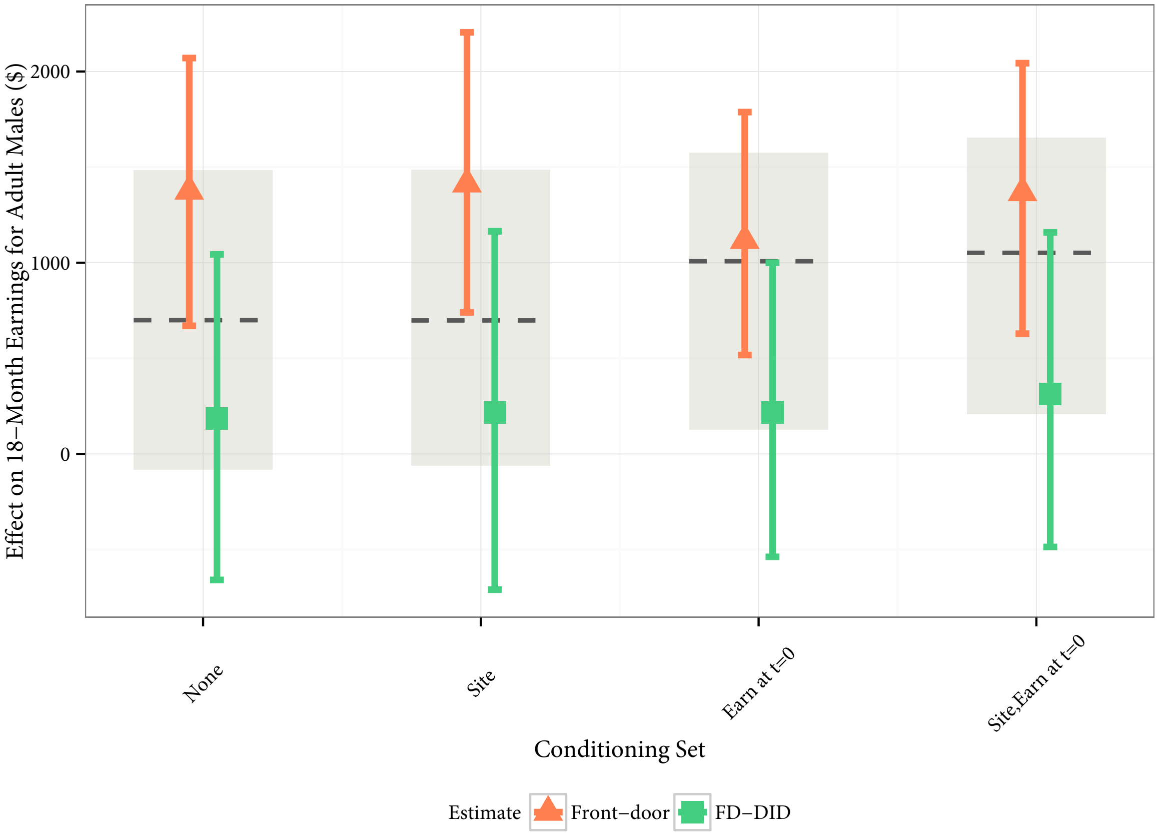


Figure 3.6: JTPA Front-door and Front-door Difference-in-Differences Estimates. Differencing group is single adult males.

3.4 Conclusion

In this paper, we have explored the possibility of making causal inferences without the use of control units, and shown that front-door and front-door difference-in-differences approaches can improve on traditional process tracing or gain-score type approaches to this task. Additionally, we have replicated the experimental benchmarks from a number of studies and shown that front-door approaches with treated units can provide useful information about these benchmarks (even when control units are unavailable).

Given some theoretical knowledge about the compliance mechanisms, we have shown the front-door approach can provide reliable bounds on the causal effect. In some cases (e.g., phone GOTV), this can provide an upper bound on a positive effect that might be useful for cost-benefit analysis. In other cases (e.g., OHIE), this can provide a lower bound on a positive effect that might be useful in establishing the efficacy of a program.

Furthermore, if previous studies suggest a suitable differencing group, a front-door difference-in-differences approach can be used to remove bias from the front-door estimates. We have demonstrated that in applications where the effect of compliance on the outcome for the differencing group is plausible zero (e.g., GOTV), then front-door difference-in-differences will come close to the experimental benchmark. In other cases where the effect for the differencing group is smaller than for the group of interest but non-zero, such as the analysis for the JTPA, the front-door and front-door difference-in-differences estimates will bracket the experimental benchmark.

While we have established the promise of the technique, future work is needed to fine tune problems in its implementation. First of all, it should be noted that front-door difference-in-differences estimates may exhibit high variance due to small differencing groups. Future work should explicitly address the bias-variance tradeoff inherent in choosing differencing groups. Additionally, more work needs to be done investigating research designs, such as multi-site studies, that would allow researchers to automate selection of differencing groups.

4

Systematic Bias and Nontransparency in US Social Security Administration Forecasts

Since the passage of the Social Security Act of 1935, a central concern of the Board of Trustees has been the demographic and financial forecasts neces- sary to assess the long-term solvency of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. These forecasts are used in a variety of ways. For example, they affect decisions about whether the rate of payroll taxes or the amount of benefits should be raised or lowered for the 210 million workers and 58 million beneficiaries in 2013, respectively. The methodology rooted in the forecasts is used by the Social Security Administration to evaluate policy proposals put forward by Congress to modify the program. The forecasts are also used as essential inputs in assessing the finances of Medicare and Medicaid and are central to research in demography, economics, political science, public health, public policy, and sociology. Although the Social Security Administration has performed these forecasts since 1942, no systematic and comprehensive evaluation of their accuracy has ever been published.

Each year, the Office of the Chief Actuary of the Social Security Administration carries out the mandate for producing forecasts in the *Annual Report Of The Board Of Trustees Of The Federal Old-Age And Survivors Insurance And Federal Disability Insurance Trust Funds* ("Trustees Report"). Actuaries at the Social Security Administration separately forecast demographic variables (e.g., mortality rates) and economic variables (e.g., labor force participation rates) that ultimately combine to produce solvency forecasts. In this paper, we offer the first evaluation of Social Security forecasts that compares the SSA forecasts with observed truth; for example, we look at forecasts made in the 1980s, 1990s, and 2000s with outcomes that are now available. We do this for demographic forecasts and for financial forecasts in the next section.

Forecasts, of course, should not be expected to be precisely accurate. However, our analysis reveals several problems. First, Social Security Administration forecasting errors — as evaluated by how accurate the forecasts turned out to be — were approximately unbiased until 2000 and then became systematically biased afterward, and increasingly so over time. Second, most of the forecasting errors since 2000 are in the same direction, consistently misleading users of the forecasts to conclude that the Social Security Trust Funds are in better financial shape than turned out to be the case. And finally, the Social Security Administration's informal uncertainty intervals are increasingly inaccurate since 2000. Although the Social Security Administration has recently begun to follow the recommendations of its outside advisers on including certain types of more formal uncertainty estimates, a step that should be part of all government reporting (Manski, 2013), these estimates have also not been systematically evaluated.

At present, the Office of the Chief Actuary does not reveal in full how its forecasts are made and, as a result, no other person, party, or organization, in or out of government, has been able to make fully independent quantitative evaluations of policy proposals about Social Security. Even the Congressional Budget Office, which produces Social Security Trust Fund solvency forecasts, relies on the demographic forecasts produced by the Office of the Chief Actuary as inputs for its models. Thus, the Office of the Chief Actuary holds an unusual position within American politics of being the sole supplier of Social Security forecasts, as well as heading the only organization producing fully independent quantitative evaluations of policy proposals to alter Social Security. For each evaluation of a proposed policy, the Office of the Chief Actuary estimates the effect on key financial outcomes that assess the solvency of the Trust Funds. For the vast majority of policy proposals evaluated by the Office of the Chief Actuary, the estimated financial impact is smaller than most of the systematic forecasting errors since 2000. Social Security Administration forecasts of current law and its counterfactual evaluation of policy proposals share the same growing bias because both are based on the same forecasting method- ology. Additionally, the uncertainty surrounding the estimated effects of proposed policies, which would likely be larger than the uncertainty in the forecasts under current law, usually dominate the estimated effect of the policy.

In the conclusion of the chapter we argue that the Social Security Administration and the Office of the Chief Actuary should follow best practices in academia and many other parts of government and make their forecasting procedures public and replicable, and calculate and report calibrated uncertainty intervals for all forecasts. The subsequent chapter explains the possible origin of the biases reported here and proposes simple structural ways of changing the system to fix the problems going forward.

4.1 Demographic Forecasts

Demographic variables important to solvency forecasts from the Social Security Administration include fertility, migration, and mortality. Higher levels of fertility and migration increase the number of workers who contribute payroll taxes and increase long-term solvency. Lower levels of mortality, especially among those age 65 years and older, increase the number of retirees who receive benefits and decreases long-term solvency. Moreover, if Americans live longer than the forecasts predict, they will draw benefits for more years than expected and the Trust Funds will become exhausted sooner than anticipated. As Diamond and Orszag (2005, p.63) explain, the increase in benefit payments from longer lives is not counterbalanced by the increase in payroll tax receipts because the system is designed to be approximately fair on average from an actuarial standpoint — and thus higher life expectancies are not taken into account in what is paid into the system earlier in working life. Adults with longer working careers also receive higher benefits compared to those starting their careers at later ages.

Observed Demographic Data As a baseline, we present four observed time series in Figure 4.1. Life expectancy for males and females, both at birth and at age 65, are relatively smooth over almost the entire time period, as can be seen in all four graphs. Indeed, three of the four are approximately linear; the fourth, female life expectancy at 65, is not far from linear. The highly regular nature of these data suggests that relatively accurate forecasts should be possible.

There is a standard conceptual difficulty in measuring current life expectancy: How can the analyst describe life expectancy when people are still alive? We follow common practice here by using the concept of "period life expectancy." This approach calculates life expectancy in a given year as the average number of years a person would expect to live if that person experienced the mortality rates in that given year over the course of a lifetime. Thus, life expectancy is a function of age-specific mortality rates and the average number of person-years contributed by those who die in each age. The mortality

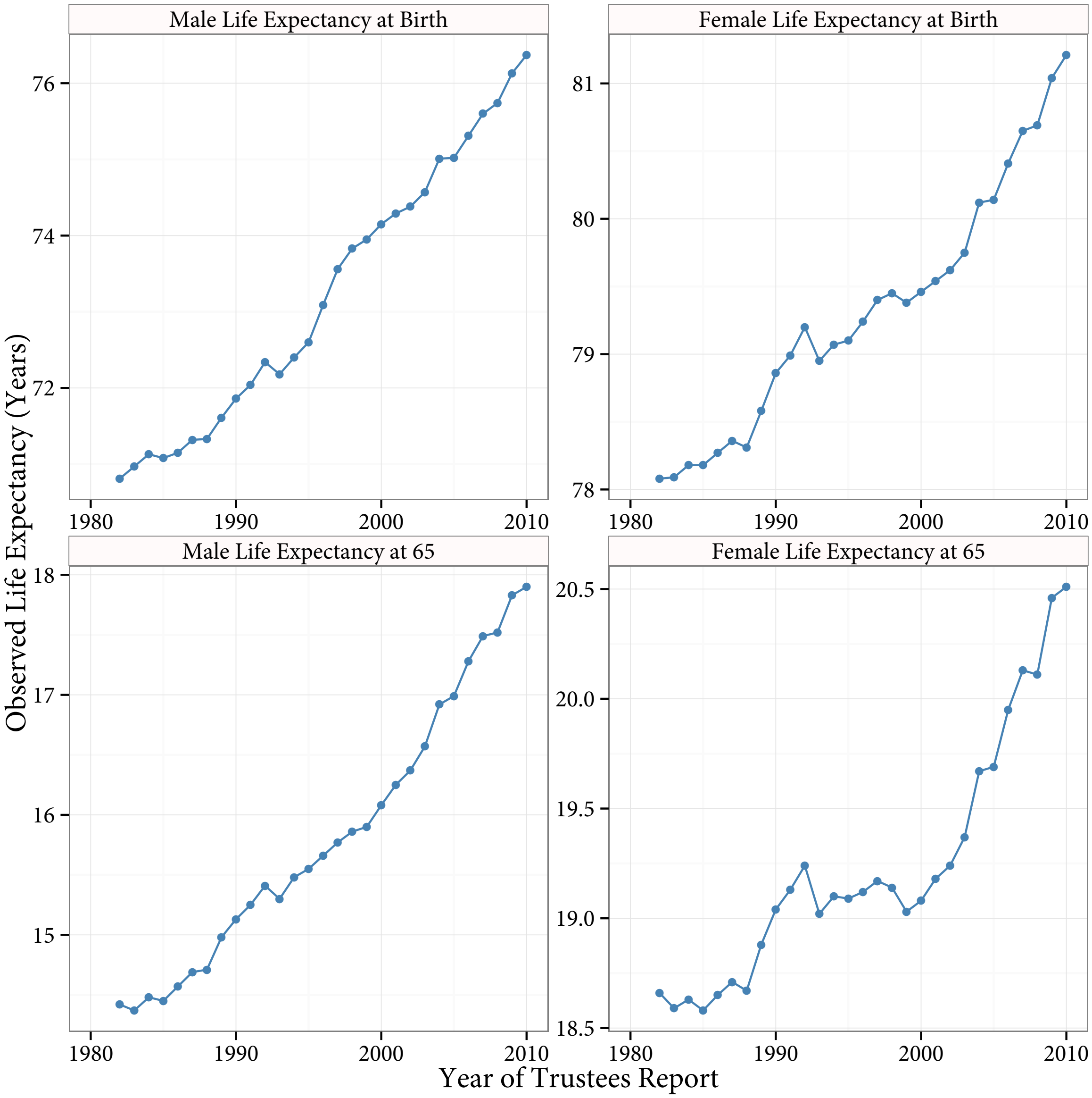


Figure 4.1: Observed Period Life Expectancy. As described in the text, "period life expectancy" for a year is a single-number summary of all the age-specific mortality rates for that same year and is interpreted as the average number of years a person could expect to live if he or she experienced the mortality rates of a given year over the course of their life.

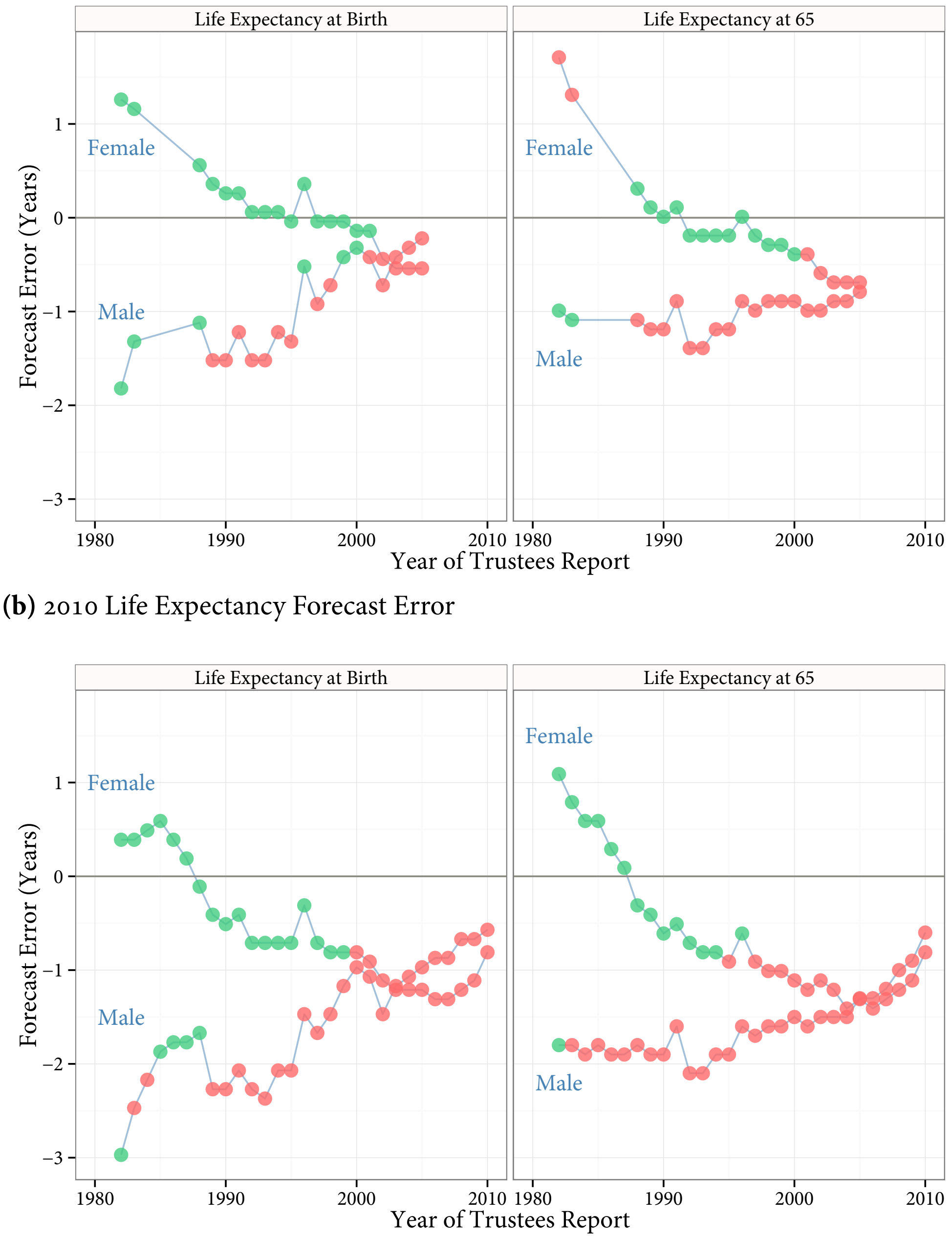
rate for people of a given age equals the number of deaths in that age divided by the number of person-years lived in that age (the exposure).[1]

The Office of the Chief Actuary forecasts male and female life expectancy separately. The male and female population counts are then combined with sex-specific economic factors like estimates of the labor force participation rates, and sex-specific beneficiary rates like disability incidence rates to project the population of workers and beneficiaries. In turn, the number of male and female workers and beneficiaries serve as inputs for predictions of Trust Funds' operations and actuarial status.

We begin in 1982, the earliest year with regular life expectancy forecasts from the Social Security Administration, and continue until 2010, the last year for which observed actual data have been released. For the years before 2001, the Social Security Administration only reveals information about its demographic forecast for years divisible by five. However, the observed time series are quite smooth and, hence, interpolations to other years should be accurate.

Forecasts We present the results in stages beginning in Figure 4.2 with an evaluation of forecasts for 2005 and 2010, the two years forecast by the largest number of Trustees' Reports (with details for all years in Appendix B). We compute forecast error (here and throughout) as the "intermediate scenario" forecast minus the observed value, so that positive values represent overestimates and negative values represent underestimates. The vertical axis is the forecast error for each of the four demographic variables and the horizontal axis is the year of the Trustees Report when the forecast was made.

[1] We use observed life expectancy based on Human Mortality Database life tables rather than the life tables from the Social Security Administration. The small differences in estimated life expectancy between the two sources do not account for the much larger error rates and patterns reported in this paper. Both sources seek to estimate the conditional probability of death (and life expectancy, its single number summary), but the Human Mortality Database is the standard in the scientific literature for its emphasis on "comparability, flexibility, accessibility, [and] reproducibility," for subjecting US Census counts and National Center for Health Statistics death counts to international quality standards, and for including all potential beneficiaries in the analyses, as discussed in an overview to the database (http:// www.mortality.org/Public/Overview.php; accessed March 20, 2015). Appendix B repeats all demographic analyses in this section with Social Security Administration data; no important differences arise compared with the results presented here.


(a) 2005 Life Expectancy Forecast Error

(b) 2010 Life Expectancy Forecast Error

Figure 4.2: Forecast Error of Life Expectancy in 2005 (panel a) and 2010 (panel b) by Year of Trustees Report. Circles (females) and triangles (males) colored green when truth falls within SSA uncertainty intervals and colored red when the truth falls outside SSA uncertainty intervals.

First, despite the strong resemblance and very high correlation between male and female life expectancy in Figure 4.1, the forecast errors are substantially worse for males than females over most of the range of the forecasts. In some of the forecasts of the mid-1980s, the overestimate of female life expectancy is more-or-less offset by the underestimate of the male life expectancy, but in later years, both are underestimated.

Second, the patterns of error persist. For example, Figure 4.2 shows that every single Trustees Report for 23 years (1982–2005) underestimated male life expectancy in 2005. Similarly, every forecast for 28 years (1982–2010) underestimated male life expectancy in 2010.

Third, the forecast for males do pass an obvious test by more closely approximating the truth as the year being forecast approaches. For females, errors have been smaller than males until recently, but in the years from 2000–2005, when projecting female life expectancy at 65, forecast errors of female life expectancy actually increased as the year of the Trustees Report approached the year being forecast.

Fourth, a large number of the forecasts fall outside the uncertainty intervals offered by the Social Security Administration. In the forecasts, these uncertainty interval are categorized as "high cost", "intermediate cost", and "low cost" scenarios. The high and low cost scenarios form the Social Security Administration uncertainty interval. In Figure 4.2, we color points green if the truth falls within these "uncertainty intervals" and red if the truth falls outside of these intervals. Although any forecast is of course uncertain and errors are to be expected, uncertainty intervals should still capture the truth with some known frequency. We find that only 1 of the 29 uncertainty intervals for male life expectancy at age 65 for 2010 actually captured the truth. In the years after 2000, every single forecast for year 2010 male and female life expectancy at birth and at age 65 was underestimated, and the truth fell outside the uncertainty intervals.

The uncertainty intervals reported by the Social Security Administration are given no formal statistical basis in published materials. Therefore, we assessed how these intervals were qualitatively presented. In Trustees Reports from the earlier part of our period, the early and mid-1980s, the Social

Security Administration wrote about the intervals as one would discuss wide confidence intervals, perhaps at a 90 percent confidence interval, and readers were warned that the confidence intervals might not necessarily cover the truth. In recent years, especially after 2000, the Trustees Reports became more confident in these intervals. Since 2003, the Trustees Reports have included an appendix referring to a stochastic model that attempts to formalize the uncertainty of their forecasts. The model itself is not publicly available, so outside analysts cannot evaluate how it has been calibrated or evaluated for model dependence. But the Trustees describe the uncertainty intervals in qualitative terms that one would typically use to discuss something stronger than a 95 percent confidence levels. For example, in the 2011, 2012, and 2013 Trustees Reports, the Trustees repeated the same definition: "In the future, the costs of OASI, DI, and the combined OASDI programs as a percentage of taxable payroll are unlikely to fall outside the range encompassed by alternatives I [low-cost] and III [high-cost] because alternatives I and III define a wide range of demographic and economic conditions."

In short, the post-2000 forecasts all indicate that both men and women would have lived shorter lives than they did, and also offered uncertainty ranges implying that the Trust Funds were on firmer financial ground than turned out to be warranted. We reach an identical conclusion when we examine the forecast error over all Trustees Reports for all observed years, as shown in Appendix B.

Uncertainty Intervals Finally, we analyze the set of all the Social Security Administration life expectancy forecasts with respect to uncertainty interval coverage. Figure 4.3 plots the year of the Trustees Report (horizontally) by the year of the forecast (vertically), with one square for each forecast colored green when the truth fell within the uncertainty interval and red when the truth fell outside the interval.

The results in Figure 4.3 demonstrate systematic problems with the uncertainty intervals used by the Social Security Administration. The uncertainty intervals failed to capture the truth for every forecast made since 2000 for all four demographic variables. For the graphs on male life expectancy at birth and

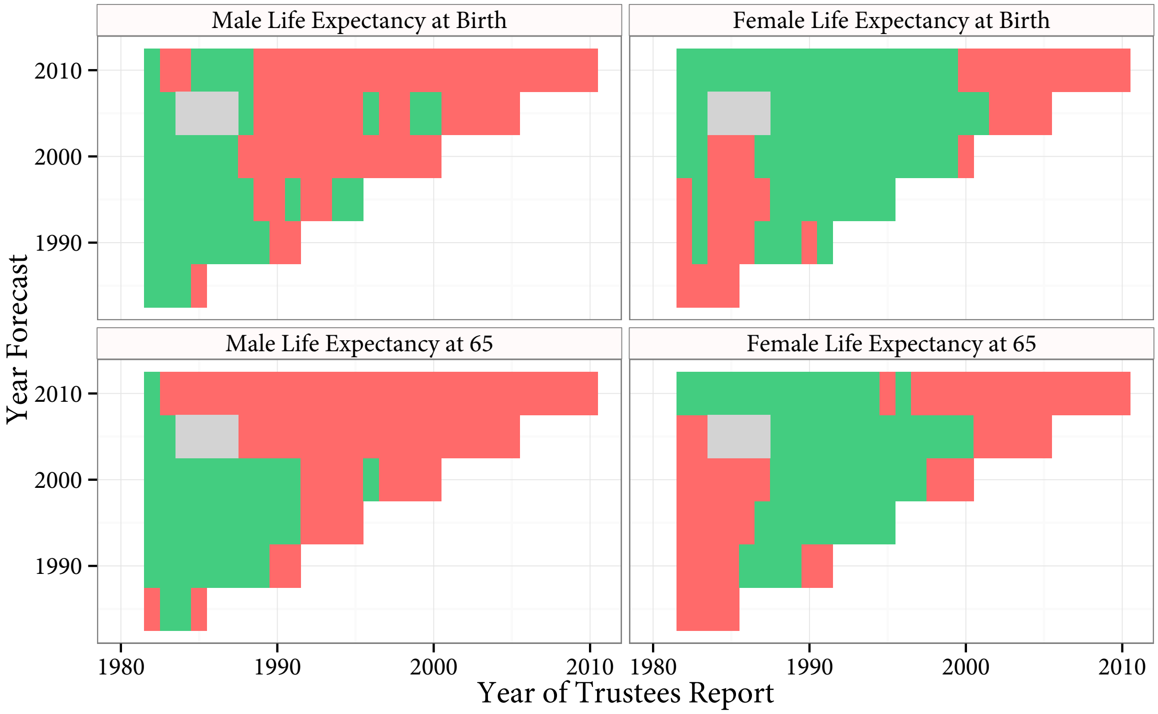


Figure 4.3: Uncertainty interval coverage by year of Trustees Report and year of forecast. Green indicates uncertainty interval covered the truth, red indicates that it did not, and gray indicates that SSA did not provide an uncertainty interval. Contemporaneous forecast error is possible because of the time lag (typically three to four years) in finalizing mortality data.

at age 65 (the two graphs on the left), the problem began approximately in 1990.

We might expect that some uncertainty intervals fail to capture the eventually observed truth, especially when the forecast was made many years earlier than the year forecast. Yet, since about 2000, the uncertainty intervals consistently failed to capture the truth for male and female life expectancy at birth and age 65. And seemingly the Social Security Administration did not perform any correction if and when these systematic errors became known.

Forecast Biases Are Not Explained by the Great Recession Could the systematic forecasting biases documented in this section be caused to some extent by the Great Recession, which lasted from

December 2007 to June 2009? Historically, increases in unemployment have led to lower mortality primarily because of fewer accidental deaths (like road traffic fatalities) (Granados, 2005; Stuckler et al., 2011) not counterbalanced by a small increase in the comparatively fewer number of suicides. Thus, a lengthy recession could potentially explain life expectancies longer than predicted. But although the recession may explain some of the forecasting error, it cannot explain most of it.

First, the Great Recession began in December 2007, when the Social Security Administration had already been underestimating Americans' life expectancy for several years prior. Second, the mortality data and errors in forecasting mortality from one year to the next are relatively smooth functions of time — that is, the errors do not increase when the recession arrived. Finally, the Great Recession cannot account for the 0.6-year forecast error in male life expectancy and 0.8-year forecast error in female life expectancy made by the Social Security Administration in 2010. During the 18-month recession, unemployment increased 4.6 percentage points from a trough of 4.9 percent in February 2008 to a peak of 9.5 percent in June 2009. Previous US- and European-based studies estimate mortality rates decline approximately 0.5 percent for every 1 percent in unemployment (Ruhm, 2000). So the 4.6 percentage point increase in employment during the Great Recession would approximately correspond to a 2.3 percent decline in mortality rates. For comparison, the inaccuracies in projected male and female life expectancies correspond to a 5.2 and 7.6 percent decline in mortality rates, respectively.

A Note About Fertility & Immigration We also evaluated the performance of Social Security Administration forecasts of fertility and migration (see Appendix B). Recent forecasts of the total fertility rate exhibited persistent and growing error, and the forecasts were overly confident. For example, the error in forecasts of the total fertility rate in 2010 grew—rather than shrank—across successive Trustees Reports. The forecast error of the 2010 total fertility rate in the 2010 Trustees was 0.15, which translated to approximately 315,000 more births forecasted than actually occurred (8% of total births in 2010).[2] As with mortality, forecast biases in fertility are not explained by the Great Recession. The rise in unemployment during the Great Recession led to a fertility decline of approximately 5%. Yet, the inaccuracy in the total fertility rate forecasted in 2010 corresponded to an approximately 8% difference in fertility rates. Overall, the forecast error in fertility makes the US population seemingly younger than it really is and, consequently, the Social Security Trust Funds healthier than they may be.

In contrast to mortality and fertility, Social Security Administration forecasts of legal immigration (the largest component of overall immigration) were far less biased and were appropriately confident. For example, the error in forecast of the net legal immigration in 2010 shrank across successive Trustees Reports. By the 2010 Trustees Report, the forecast error was less than 1% of the observed number of net legal immigrants in 2010.

The results of mortality, fertility and immigration forecasts may illuminate some of the reasons why the Social Security Administration varies in its performance of forecasting these three demographic components. As we describe in detail in Chapter 5, a constellation of factors may interact to produce biased mortality (and fertility) forecasts. First, the forecasting method itself allows for the introduction of unintentional bias because they involve a very large number (previously 210, now 150) interrelated subjective decisions about rates of mortality decline. Second, as Social Security reform became increasingly charged politically, the Social Security Administration became more insular and disregarded the continued advice of outside advisers to assume a more rapid increase in life expectancy. Third, mortality rates decreased at an ever faster pace after about 2000, but the Social Security Administration mortality forecasts did not keep pace with this change in input data.

Some of the same factors that possibly produce biased mortality forecasts may occur for fertility, too. The Social Security Administration forecasting method for fertility also involves subjective decisions about future levels of fertility rates. In contrast to mortality and fertility, the level of legal immigra-

[2] The 2010 Trustees Report included historical fertility up to 2006 because of the time lag in reporting final birth data.

tion is annually set by Congress. The Social Security Administration forecast of net legal immigration largely follows this Congressionally set level and applies an empirically based percentage of emigration.

4.2 Financial Forecasts

We next consider Social Security Administration forecasts of Trust Fund solvency, for which demographic forecasts serve as a key input. In particular, we examine forecasts and observed outcomes of the three most commonly cited financial indicators when discussing the health of Social Security: the *cost rate*, the *trust fund balance*, and the *trust fund ratio*. The cost rate equals the overall cost of the Social Security program in a given year dived by the taxable payroll for that year. The trust fund balance equals the difference between projected annual income and projected annual cost, as a percentage of the taxable payroll. The trust fund ratio equals the assets of the Social Security Trust Funds at the beginning of a calendar year divided by the expected expenditure for that year.

We collect all forecasts for each measure published in the annual Trustees Reports from 1978, when the reports began consistent reporting of financial indicators, until 2013. The reports usually include yearly forecasts between the year of the report and 10 years in the future and then every fifth subsequent year. After 2000, single-year supplemental tables are available online.

These three financial indicators directly relate to economic and public policy debates that have occurred over nearly the entire lifetime of Social Security. After the Social Security Amendments of 1983, for example, the trust fund balance increased primarily because of higher payroll tax rates, although benefit levels increased, too. Numerous economic studies find Social Security affects personal savings through reduction of disposable income because of payroll taxes and anticipated benefits during retirement (Harris, 1941; Feldstein, 1974; Diamond and Hausman, 1984). Gramlich (1996) argues that proposed Social Security reform faces competing challenges in political economy: ensuring long-run actuarial balance while not lowering the ratio of discounted benefits to discounted taxes paid (the "money's worth" ratio). The long-run actuarial balance, a function of the trust fund balance and cost rate, can be maintained by raising payroll tax rates or lowering benefit level, although such changes would reduce the money's worth ratio. Other proposed reforms, such as individual accounts and personal savings accounts, extends the solvency of Social Security and maintain the money's worth ratio, but face intense public scrutiny (Samwick, 1999).

The Cost of Mortality Forecasting Errors Before turning to the three financial indicators, we begin by comparing the forecast errors in cost specifically due to forecast errors in mortality (dashed line) versus overall forecast errors in cost (solid line, Figure 4.4). In theory, either of these forecast errors in cost could be larger because of forecast errors in cost from many other inputs.

For each Trustee Report and forecast year, we estimate the number of additional retirees that the Social Security Administration did not expect because of errors in predicting life expectancy. For example, the 2005 Trustee Report under-forecasted male life expectancy at age 65 in the year 2010 by 1.3 years (forecast 16.6 years, truth 17.9 years). The 1.3 year under-forecast of life expectancy corresponds to approximately 151,000 male beneficiaries. We estimate the forecast errors in costs due to forecast error in mortality as the product of the total number of additional beneficiaries and the average benefit amount per year. For this figure, we plot the forecast year on the horizontal axis and the forecast error in cost (in billions of 2010 dollars) on the vertical axis. Each panel presents forecasts from different Trustees Reports. The years of the Great Recession are denoted by the grey shaded region. To put the figure into perspective, the total cost of the Social Security program in 2010 was $712.5 billion.

Figure 4.4 emphasizes four points. First, mortality is a highly predictable part of the overall forecast error in cost, as evidenced by the highly smooth and almost linear dashed lines in each panel. Second, for many years, forecast errors in cost specifically due to forecast errors in mortality were a large fraction of the overall forecast error in cost. Third, the forecast errors in cost due to forecast errors in mortality are neither random nor constant. The errors increase secularly and thus strongly suggest the

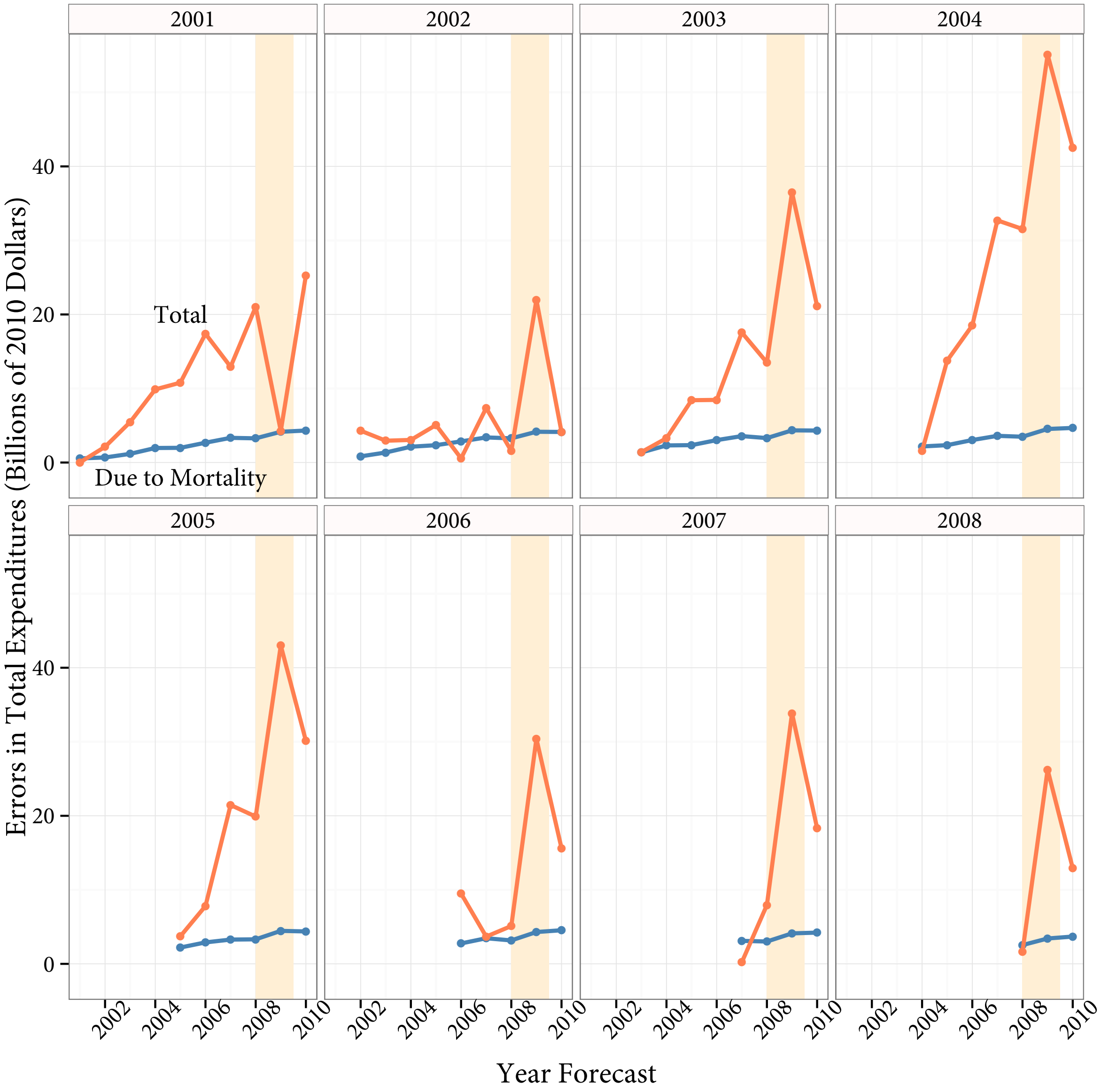


Figure 4.4: Cost of Mortality Forecasting Errors (in billions of 2010 dollars). Each panel of the figure corresponds to a Trustees Report. Within each panel, we plot the forecast error in total Social Security expenditures (orange lines) and the forecast error in total Social Security expenditures due to mortality forecasting errors (blue lines). Finally, we represent the Great Recession as a vertical yellow shaded region.

existence of information that can be used to improve forecasting performance. Finally, the overall forecast error in costs are highly variable relative to errors due to mortality. They are much larger during the Great Recession, shown by the vertical shaded area, but these overall forecast errors in costs were also large at times well before the onset of the Great Recession.

Cost Rate Forecasting Errors Figure 4.5 reports the forecast error in the cost rate (the vertical axis in each panel) made in a Trustees Report in the given year (the horizontal axis in each panel) for a time a number of years out into the future (in the title of each panel). For example, the upper left panel shows the forecast error in the cost rate for forecasts made one year in advance of the year forecast. A value of zero indicates that the forecast was perfectly accurate. Green points fall within SSA's forecast uncertainty interval and red points fall outside. To enhance readability, we superimpose on each panel a smoothed line showing the path of the errors.[3]

The pattern of forecast errors in Figure 4.5 is striking: Forecasts from Trustees Reports until about 2000 were approximately unbiased, which can be seen by the roughly random scatter of points vertically around the horizontal line at zero forecast error. However, forecasts from Trustees Reports after roughly the year 2000 were increasingly biased over time, and all in the same direction. Congress and other users of these forecasts would have been misled into thinking that the cost of the Social Security program was less than it actually turned out to be. The biases here are as true for forecasts one year into the future (top left) and for forecasts 10 years into the future (bottom right). As expected, the errors are larger for forecasts farther into the future.

Finally, Figure 4.5 shows that the largest errors are also most likely to be outside the uncertainty intervals. The purpose of uncertainty estimates is to protect oneself from drawing overconfident conclusions from the data, and if estimates are consistently falling outside those uncertainty intervals, then

[3] The smoothed line is estimated with a locally weighted scatterplot smoothing (LOESS) procedure, in which the predicted error for Trustees Report *t* is calculated based on a local polynomial of degree 2 fit to neighboring observations. These observations are weighted by their tricubic distance from the Trustees Report in question.

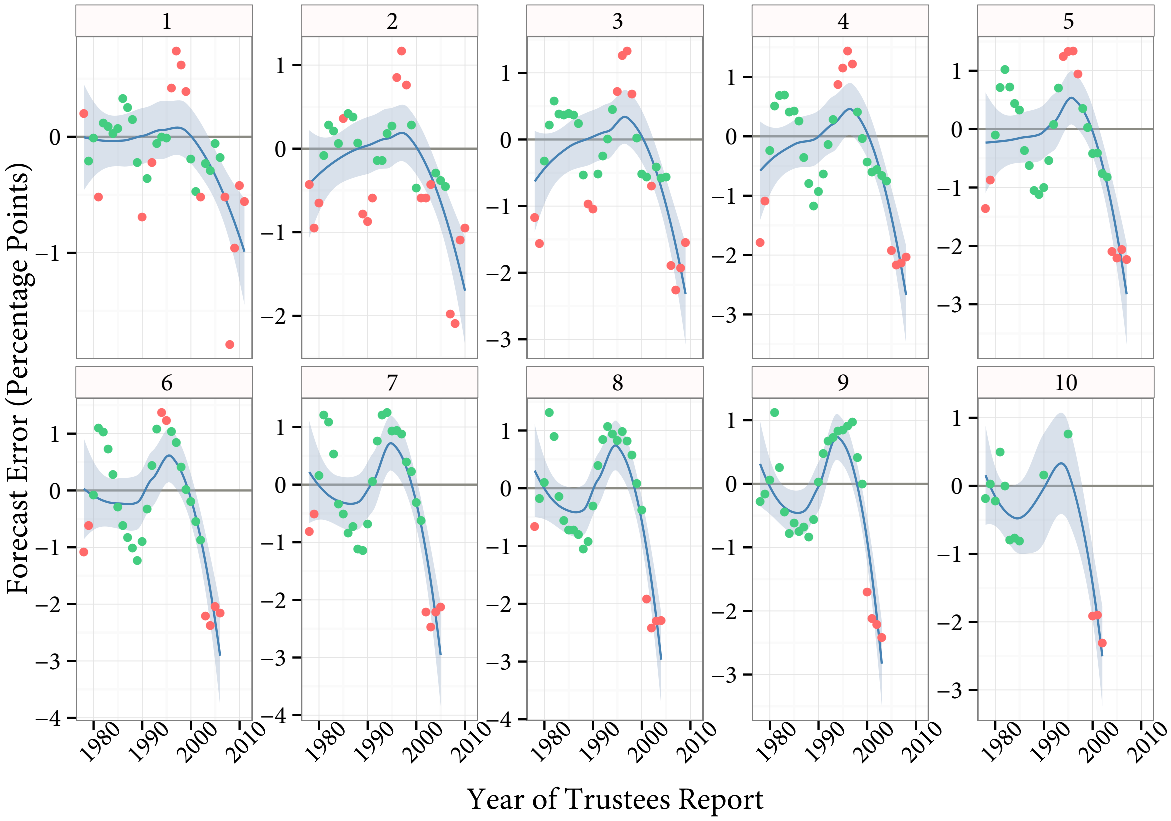


Figure 4.5: Cost Rate Forecasting Errors. Forecast errors in the cost rate (vertically) by the year of the forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Cost rate forecasting errors are overestimates if positive and underestimates if negative. Points are green if the error is within SSA's uncertainty interval and red otherwise.

improvement in the forecasting process should follow.

Trust Fund Balance Forecasting Errors A positive annual trust fund balance indicates the program has a surplus for the year and a negative trust fund balance translates to a deficit. We present Figure 4.6 in the same format as Figure 4.5. The evaluation of forecasting errors in the trust fund balance leads us to the same conclusions as forecasting errors in the cost rate. The Social Security Administration forecasts of trust fund balances were approximately unbiased until about 2000, after which they become substantially biased. Moreover, the direction of the biases are all in the same direction, making the Social Security trust funds look healthier than they turned out to be. The reported uncertainty intervals are again overconfident.

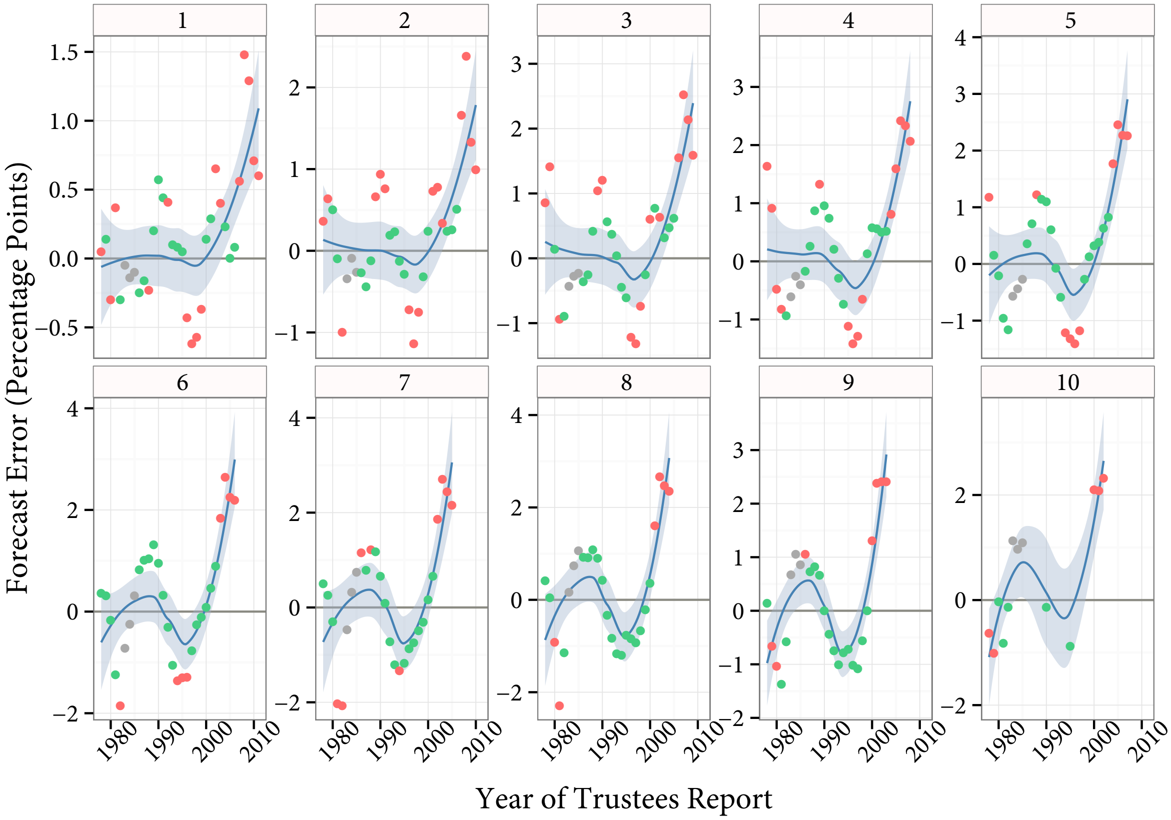


Figure 4.6: Balance Forecasting Errors. Forecast errors in balance (vertically) by the year of the forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Positive errors overestimate Trust Fund assets; negative errors underestimate them. Points are green if the error is within SSA's uncertainty interval and red otherwise.

Trust Fund Ratio Forecasting Errors When the trust fund ratio equals zero percent or becomes negative, the Social Security Trust Funds are insolvent. The Trust Funds are deemed financially adequate in the short term if the ratio stays above 100 percent for the first 10 forecasted years. Insolvency does not release the federal government from its obligation to pay benefits to qualified individuals (Meyerson, 2014). The Social Security Act stipulates that every fully insured individual is entitled to receive benefits. Yet, the Antideficiency Act prohibits the federal government from paying Social Security benefits beyond the balance of the Trust Funds. Once insolvency occurs, beneficiaries would either receive delayed or lower benefit payments.

In Figure 4.7, we present results in a form parallel to Figure 4.5. While SSA's uncertainty intervals appear to have better coverage when compared to the cost rate and trust fund balance metrics, the results in this figure confirm the main results from our analysis of the cost rate and trust fund balance. First, trust fund ratio forecast errors are approximately unbiased from 1978 through about the year 2000, as indicated by the dots scattered randomly above and below the vertical line drawn at zero. After 2000, forecast errors became increasingly biased, and in the same direction. Trustees Reports after 2000 all overestimated the assets in the program and overestimated solvency of the Trust Funds. The size of this bias has increased over time, with the more recent Trustee Reports being less and less reliable. Finally, the coverage of uncertainty estimates did not improve over time and were strongly and positively correlated with the size of the absolute error.

Implication of Financial Forecasting Errors for Proposal Scoring In addition to producing the annual Trustees Report, the Office of the Chief Actuary also scores policy proposals to alter Social Security submitted by members of Congress, the administration, and select professional organizations. For each of the policy proposals it scores, the Office of the Chief Actuary makes point estimate predictions about what would happen to one or more financial metrics, such as those we study above, if the proposal became law. Although the Office of the Chief Actuary includes no uncertainty measures with its predictions, we can estimate their uncertainty on the basis of our evaluation of their forecasts.

Uncertainty in these counterfactual predictions have two components. The first is the inherent uncertainty of the effect of the intervention if the law changed as proposed. The second is the uncertainty

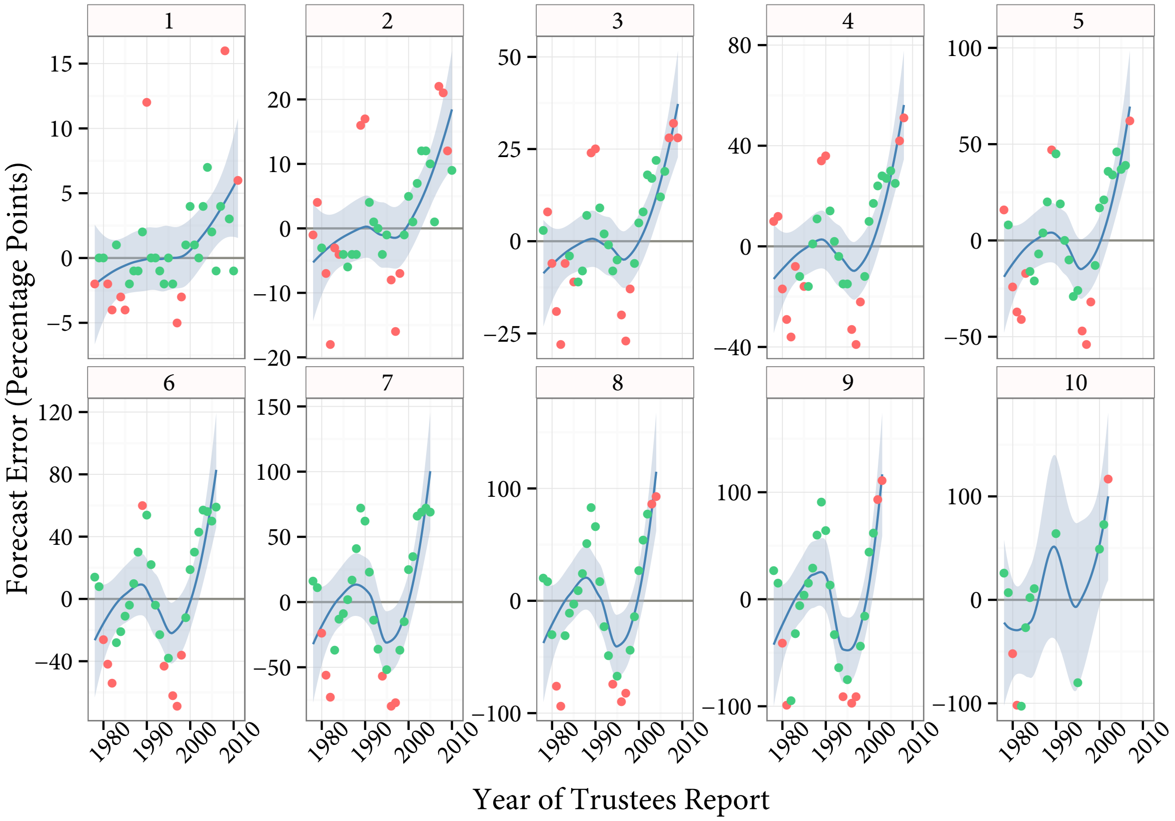


Figure 4.7: Trust Fund Ratio Forecasting Errors. Forecast errors in the trust fund ratio (vertically) by the year of the Trustees Report forecast (horizontally) by how many years into the future the forecast is made (in the title of each panel). Positive errors overestimate Trust Fund assets; negative errors underestimate them. Points are green if the error is within SSA's uncertainty interval and red otherwise.

in forecasting the same financial indicators under current law, as we do earlier. We use our evaluation of the second component as a lower bound for the uncertainty of the Office of the Chief Actuary's policy scoring.

The Social Security Administration evaluated 63 proposals since 2003, which resulted in 104 assessments of financial indicators for which we can evaluate forecasting performance (ssa.gov/oact/solvency). For example, in 2015, the Social Security Administration evaluated the effect of President Obama's Executive Actions for immigration on Social Security solvency. The Chief Actuary concluded immigration reform would increase the cost rate by 0.04%, which is considerably smaller than most cost rate forecasting errors made since 2000. Overall, we found 43% of policy assessments by the Office of the Chief Actuary predicted changes in Social Security finances that were smaller than the average forecasting error made since 2000. And 95% of the assessments concluded changes in Social Security finances that were smaller than the maximum forecasting error made since 2000.

Members of Congress and the public devote considerable energy debating policy proposals on the basis of these evaluations. And presidents and their opponents tout the merits of policy proposals to engender public support. Yet, if this lower bound on the magnitude of forecasting errors exceeds the estimated effect of the reforms, then these discussions and debates will not be grounded in the best information available and may lead to biased policy conclusions.

4.3 Conclusions and Recommendations

In recent years, especially after about 2000, the Social Security Administration began issuing systematically biased forecasts with overconfident assessments of uncertainty. Reliance on such forecasts led policymakers and other users of the forecasts to conclude that the Social Security Trust Funds were on firmer financial ground than actually turned out to be the case. We focus on four crucial steps SSA should take to ensure this problem is addressed, with extensions and explanations for these patterns in Chapter 5.

First, forecasting mistakes are no embarrassment unless the forecaster fails to learn from them. Thus, we recommend that the Social Security Administration publish annually a systematic and comprehensive evaluation of its forecasting performance for both demographic factors and financial solvency. This best practice of forecasting self-evaluation is routine among academic researchers (j.mp/LeeMiller01) and professional actuaries (Lu and Wong, 2011), for other countries' social security programs (Shaw, 2007), and in other parts of the U.S. Government, such as the Congressional

Budget Office (j.mp/CBOeval), the Census Bureau (j.mp/CensusPopEval), the Bureau of Labor Statistics (j.mp/BLSeval), and even other parts of the Social Security Administration itself (j.mp/SSAevals, j.mp/AFR13). Every future Trustees Report, without exception, should include a routine evaluation of all prior forecasts, and a discussion of what forecasting mistakes were made, what was learned from the mistakes, and what actions might be taken to improve forecasts going forward.

Second, the Social Security Administration withholds from public view much of the data and procedures it uses to make many of its forecasts. The Office of the Chief Actuary, which produces the demographic and economic forecasts, does not share much of its data and procedures even with other parts of the Social Security Administration. Currently, the best anyone can do to understand how the Social Security Administration forecasts work is to attempt to reverse engineer their results (as done by many involved in the policy process and authors of simulation programs such as SSASIM by the Policy Simulation Group); see also King and Soneji (2011); Soneji and King (2012). The "replication standard" for data sharing is the widely understood and accepted best practice throughout the scientific community (King, 1995, P.444) and echoed in the Obama Administration's executive orders requiring "a presumption in favor of openness," and that data produced by government be "accessible, discoverable, and usable by the public" (j.mp/ObamaOpenData).

Finally, it appears to us and to other outside observers that the forecasting procedures used by the Social Security Administration are informal and qualitative. These approaches also fail to take advantage of many of the dramatic improvements in statistical modeling over the last several decades (e.g., Girosi and King, 2008; King and Soneji, 2011). Even some explicitly quantitative parts of the methods seem idiosyncratic or unnecessarily model dependent.

Our study reveals systematic errors in both demographic and Trust Fund solvency forecasts. Forecasting errors in economic variables, such as the labor force participation rate and growth in average wages, may also contribute to systematic errors in Trust Fund solvency forecasts. For the disability program of Social Security, forecasting errors in the disability incidence rate may be an especially important source of solvency forecast error.

This list of "best practices" is neither new nor controversial. There is a Social Security Advisory Board Technical Panels on Assumptions and Methods made up of outside experts. The Social Security Administration's own outside advisors have repeatedly and emphatically recommended that the Office of the Chief Actuary make its data and procedures widely available, and allow its work to be replicated by outside groups. The collective efforts of the scientific community could easily be marshaled to improve the difficult forecasting task that confronts the Social Security Administration, all essentially without cost to the taxpayer. The creation of transparent forecasting procedures will also enable members of Congress and partisans on all sides to consider alternative assumptions explicitly when they debate proposals to ensure the solvency of Social Security. Forecasts of Social Security solvency also shape debates on immigration, public health, taxation, and income redistribution from working age adults to retirees. Accurate forecasts would help ensure these debates are based on the best information available.

5

Explaining Systematic Bias and Nontransparency in US Social Security Administration Forecasts

Social Security is the single largest program in the U.S. government, currently providing benefits to over 58 million retirees and disabled and levying payroll taxes on another 210 million workers. It is also one of the most popular programs, ending retirement-generated impoverishment for a vast segment of the population. The program functions via an intergenerational transfer of wealth. The government levies payroll taxes on today's workers and deposits the revenue into interest-earning Social Security Trust Funds; retirees, disabled workers, and their families receive benefits paid from the Trust Funds. The Trust Funds function as a bank account that enables the Social Security Administration (SSA) to smooth out temporary imbalances between workers and beneficiaries. For example, while the baby boom generation remains in the workforce, the Trust Funds collect more money than SSA needs to pay for current beneficiaries. When this generation retires, it withdraws more benefits from the Trust Funds than future workers' payroll taxes will generate.

The viability of Social Security depends fundamentally on the solvency of the Trust Funds, which, in turn, depends upon accurate demographic and financial forecasts that enable members of Congress to make sound decisions on Social Security policy. For example, without adjustments to payroll taxes or benefit levels, the Trust Funds could become exhausted sooner than anticipated if medical advancements extend the life expectancy of retirees or an economic recession increases unemployment and decreases payroll tax revenue.

The goal of SSA's demographic and financial forecasts is to provide sufficiently accurate information to policymakers to address excess cash outflows or low cash inflows via one or more of the available policy levers. Among many others, these include gradual increases in payroll tax rates or changes in the retirement age. The earlier an accurate forecast becomes available, the more options Congress has to ensure the solvency of Social Security through incremental, politically palatable, and economically achievable changes. In contrast, inaccurate demographic and financial forecasts narrow the range of feasible policy levers, often to proposals that are fiscally disruptive, politically challenging, or otherwise infeasible.

The errors in SSA forecasts are large and mostly in the same direction, making the Trust Fund look healthier than it actually is. With these forecast evaluations, we also provide the first honest uncertainty estimates of SSA's policy scores, which constitute the only real evaluation of every major proposal to change the system in the last two decades; we find that more than 90% of SSA's numbers are overwhelmed by forecast uncertainty.

Of course, even the most skilled forecasters will sometimes make inaccurate predictions due to unforeseeable events. Such surprises do not reflect negatively on the forecasters. However, continuing to make systematic forecasting errors is evidence of bias or flawed methodology. In this case, it suggests a failure to meet the "best practices" in scientific evaluation procedures commonplace in academia, industry, and other government agencies. In fact, we find that the SSA lags behind best practices benchmarks in at least three critical respects.

First, SSA has never published systematic and comprehensive evaluations of its forecasts. Second, the SSA Office of the Chief Actuary (OCACT), which produces the forecasts, withholds many aspects of its data and forecasting procedures from the public, the scientific community, and even other parts of SSA. This makes independent replication impossible. Third, critical aspects of OCACT's forecasting procedures are informal or qualitative, even though far better systematic quantitative techniques exist and continually improve. Consequently, SSA is not set up to learn optimally from its pattern of forecasting errors, or to meet the "replication standard" widely supported in academia (King, 1995) and even in the U.S. government via President Obama's executive orders (j.mp/ObamaOpenData). As a result, important parts of OCACT procedures are difficult or impossible to replicate, understand, or implement, which means that errors introduced are unlikely to be corrected. Indeed, some steps in SSA's forecasting involve committees or individuals making large numbers of interrelated qualitative judgments that are extremely difficult to do well without computer assistance and, at the same time, involve high levels of discretion for decision makers hidden from public view.

As we show here, SSA's forecasting procedures turn out to meet all the major conditions for unintended political, social, and psychological biases to be introduced, even when those involved try diligently to produce forecasts free from external influence. Our research seems to indicate, consistent with the social-psychological literature, that these biases do not occur because of individuals making avoidable mistakes or not trying hard enough. Greater effort for the same task performed in the same manner would likely not help. Rather, the biases occur because of the lack of formal procedures at SSA designed to avoid these biases in the first place.

Indeed, OCACT's own scientific panel of distinguished outside advisers, the Social Security Advisory Board's Technical Panel on Assumptions and Methods, has frequently recommended that OCACT make data and replication procedures available and improve their statistical methodology and uncertainty estimation. We demonstrate here that SSA regularly ignores this and many other recommendations of its scientific advisory panel, and this oversight provides important evidence on the evolution of biases inherent in SSA forecasting.

Section 5.1 outlines some of the methodological challenges involved in Social Security forecasts and then summarizes and extends our evidence from Chapter 4 on the biases in SSA's forecasts. Section 5.2 offers a hypothesis about how these systematic biases came about, supported by research from social psychology on situations like these where unnecessary and uncorrected biases are generated.

To gather information and evaluate our theories, we conducted a large number of semi-structured personal interviews with those in, and involved with, SSA at every level of government and the policy process. Since the politics of Social Security has become extremely polarized and highly conflictual, with SSA administrators deeply involved in many aspects of it, we keep the identities of our interviewees confidential. With this information, Section 5.3 summarizes the pressures on SSA that likely led to these forecasting biases. In order to reduce the possibility of bias going forward, Section 5.4 provides an overview of methods that formalize ad hoc, qualitative forecasts. Section 5.5 concludes. Appendix B gives details of the data used in this article, with a complete replication data set available at Kashin et al. (2015a).

5.1 Social Security Administration Forecasting

In this section, we briefly summarize and extend some of the empirical results from Chapter 4 about the performance of SSA's forecasting, convey the methodological challenges in the complex data used for Social Security forecasting, and outline some of the specific procedures SSA uses to forecast.

5.1.1 Social Security Forecasting Performance

We summarize our prior work in two points. First, despite many years of approximately unbiased forecasting, OCACT began issuing systematically biased forecasts, with overconfident assessments of uncertainty, after about the year 2000. Instead of following best practices and learning from these biases to improve subsequent forecasts, the biases have grown much larger over time. These biases led members of Congress, other policymakers, and the public to conclude each year that the Social Security Trust Funds would be on firmer financial footing than actually turned out to be the case, year after year.

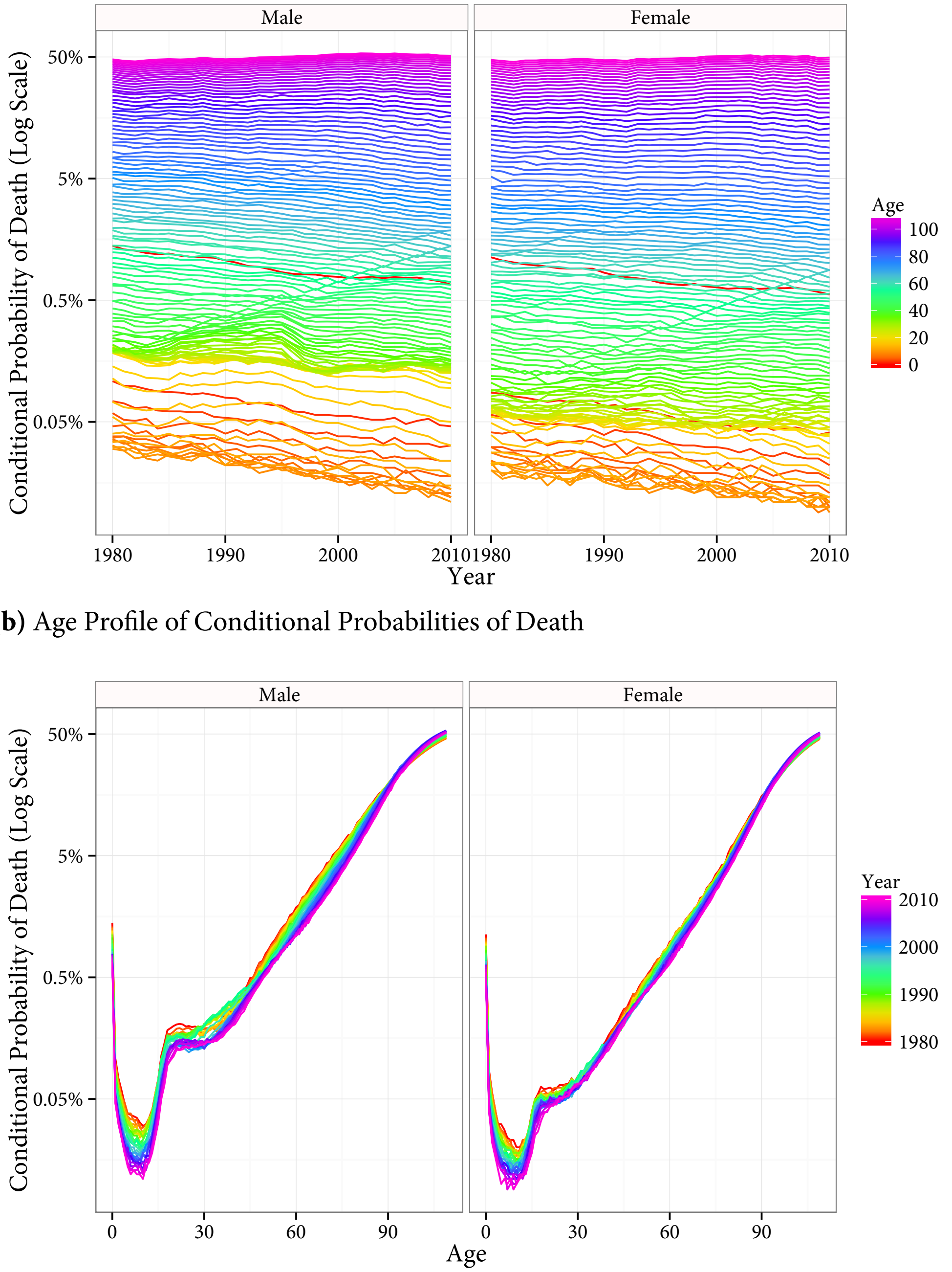
Second, OCACT does not share all its data, code, and replication information with the public, the scientific community, and other parts of SSA or the federal government. We also discuss here the contributing factor that OCACT's forecasting procedures are informal and qualitative, and fail to take advantage of the dramatic improvements in statistical modeling, and ways of formalizing informal procedures like these, that have been developed over the last quarter century.

5.1.2 Methodological Challenges

The statistical challenges in accurately forecasting Social Security are substantial. They involve modeling complex multivariate data structures and incorporating numerous details in modeling how this enormous and complicated government program is administered, run, and funded. We focus here on the crucial methodological challenge of forecasting mortality rates — a key input to SSA's financial forecasts. An accurate forecast of mortality is necessary to estimating the number of workers contributing payroll taxes and the number of retirees and disabled receiving benefits. To begin to provide a feel for the problem, consider Figure 5.1.

The key to this figure is simultaneously recognizing both the methodological challenges and the substantial information available to improve forecasting accuracy using the right model. First, consider the top row of Figure 5.1, which shows the probability of a male (top left panel) or female (top right panel) dying within one year, for each year of age (0 to 109, colored from red for the youngest ages to purple for the oldest ages) and for every calendar year from 1980 to 2010 along the horizontal axis. Small changes in any one line may contribute to large changes in the number of workers or the number of retirees. Thus, the highly regular aggregate patterns also contain important and apparently small variations that need to be modeled carefully. Each line in each graph is a time series plot representing different cohorts of people at the same ages at different times. One can also see important diagonal patterns (from bottom left to top right in each of the top panels) representing the continued experience of higher or lower mortality of the same birth cohort as it ages over time.

These two top panels in Figure 5.1 convey two other crucial facts that need to be taken into account in any serious model of the data generation process. First, although mortality is relatively smooth over time, forecasting any one of these lines would be less certain many years into the future. Second, there is considerable information in adjacent age groups that can be used to improve the forecasts if used appropriately. This second point can be seen even more clearly in the bottom two panels of the same figure which re-express the same data with age along the horizontal axis, and colors for each year. So instead of time series plots, these panels portray age profiles. The characteristic shape is common in demographic data across countries and time periods, indicating that the few years after birth are relatively risky, after which mortality drops, and then starting at about age 5-10 mortality inexorably increases, almost log-linearly, except for a bump that coincides with higher risk of accidental mortality among older adolescents and young adults.


(a) Time Profile of Conditional Probabilities of Death

(b) Age Profile of Conditional Probabilities of Death

Figure 5.1: Time and Age Profiles of Conditional Probabilities of Death (Human Mortality Database).

The opportunities in modeling mortality involve recognizing the powerful information available to build into forecasting methods. In fact, the patterns in Figure 5.1 turn out not to be unique to this time period or even the United States and so should be considered valuable information for model building, such as for constructing Bayesian priors. To convey how stable these patterns, are we constructed Figure 5.2, which gives the log-mortality age profile for 39 separate countries (lighter blue indicates later years). As is apparent, the general pattern from U.S. age profiles holds with remarkable generality across countries. Clearly, forecasting methods that ignore or do not formally encode this powerful information are, at best, highly inefficient.

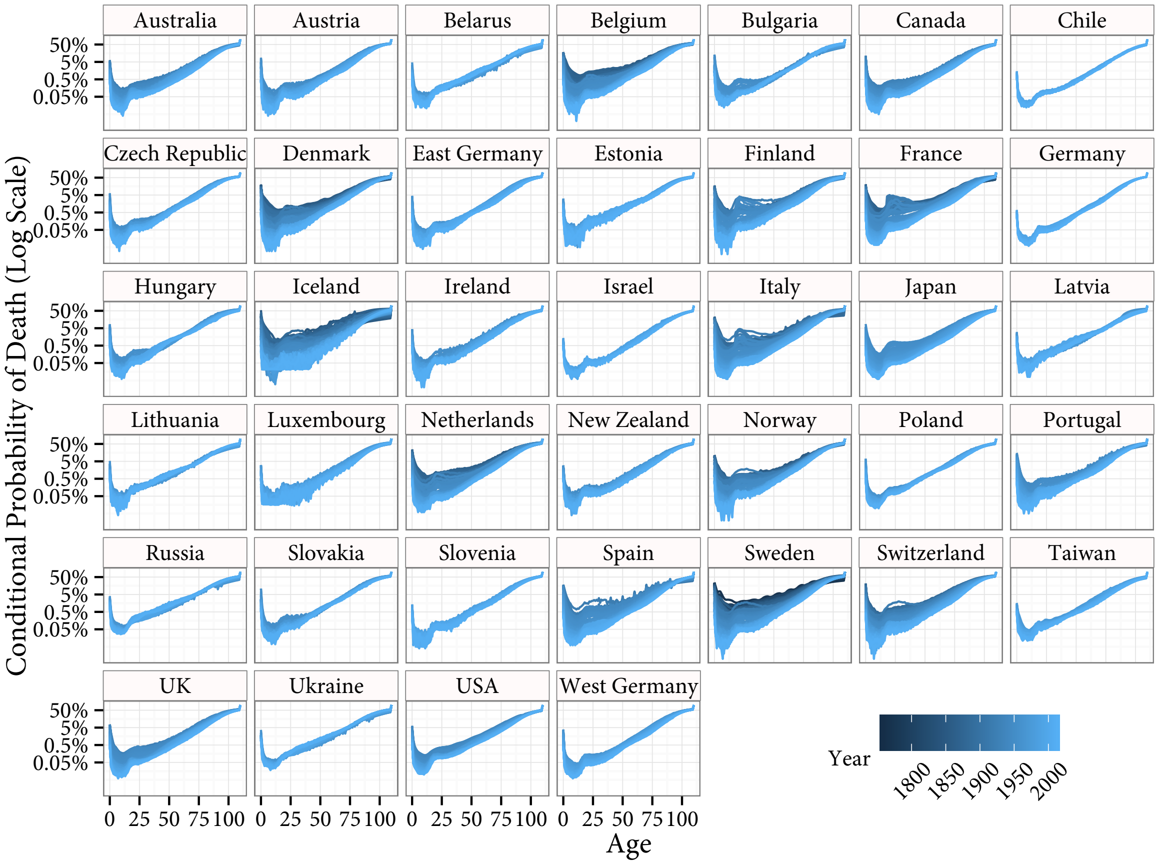


Figure 5.2: Conditional probabilities of death (Human Mortality Database).

The patterns in Figures 5.1 and 5.2 represent highly reliable demographic knowledge that forecasting methods should include. This information has been formally encoded in priors by smoothing across expected mortality in neighboring age groups and adjacent time points and their interactions (Girosi and King, 2008; King and Soneji, 2011), or via principal component modeling (Lee and Carter, 1992). However, instead of sophisticated statistical methods encoding this well-known demographic knowledge, SSA manually adjusts simple regression models of mortality on time to be consistent with the views of a committee of actuaries making qualitative judgments. Important covariates such as smoking and obesity are formally excluded from their statistical model, although they considered qualitatively.

Finally, we summarize patterns in log-mortality changes to illustrate a key feature of U.S. mortality data that will prove important in explaining the patterns of bias in SSA forecasts. We focus on age groups 65 and older, most of whom will be drawing Social Security benefits for as long as they are alive. To do this, we compute the change in log-mortality over successive 10-year time intervals for selected age groups. We thus regress log-mortality on time, with observations from $t - 9$ to t (repeated for each year t, $t = 1970, \ldots, 2010$) and plot in Figure 5.3 the coefficient on time as a measure of recent log-mortality changes (vertically) by year t (horizontally).

For example, 85-year old male mortality (left graph, blue line) declined an average of 0.3% (i.e., a -0.003 change in log-mortality) per year between 1991 and 2000. In stark contrast, the pace of mortality decline sharply increased after 2000 (highlighted by the shaded area of the graph). Mortality for this age group declined an average of 2.6% per year between 2001 and 2010. The shaded area in the graph, corresponding to the period since the year 2000, shows that pace of mortality decline quickened for every age group for both men and women since about 2000. We will return to this key result from Figure 5.3 in Section 5.2. These ever faster reductions in mortality may be due, in part, to greater use of statins that reduced cardiovascular disease and more widespread cancer screening. Whether or not one would regard this dramatic change as predictable ahead of time, the changes certainly became clear after a few years of unexpected declines.

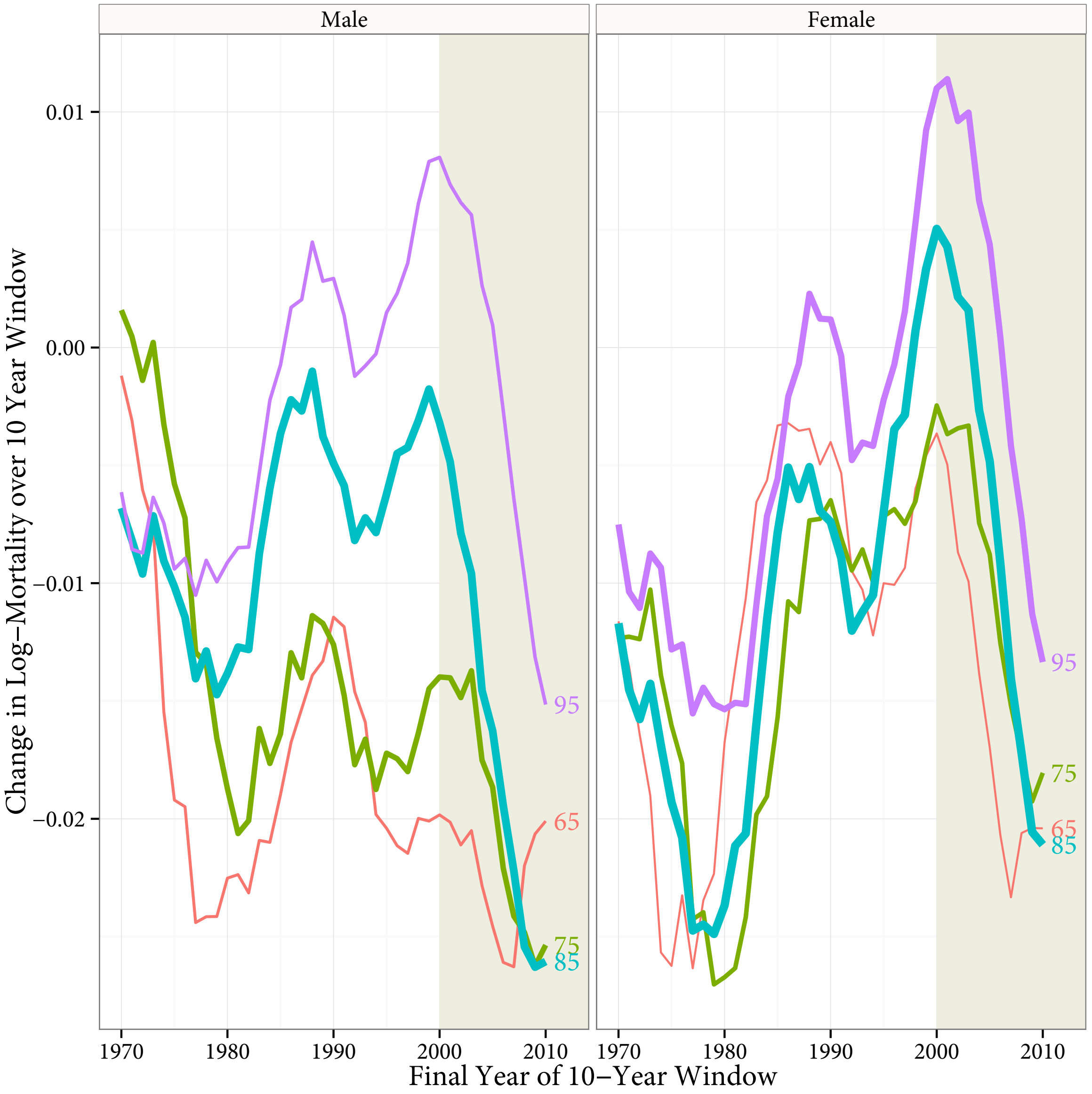


Figure 5.3: Changes in Log-Mortality over Successive 10-Year Windows. Each line is an age group, with the thickness of the line drawn proportional to the number of deaths in that age group. The confidence interval around each of the lines (which we do not add to the graph for clarity) is approximately ±0.004.

5.1.3 Forecasting Procedures

SSA employs forecasting procedures that require numerous ad hoc and interrelated qualitative judgments, in a manner very difficult for any human to do well Soneji and King (2012). For example, a critical aspect of the SSA forecast relies on the choice of 210 interrelated "ultimate rates of decline" in mortality rates across 5 broad age groups, 2 sexes, 7 causes of death, and 3 cost scenarios. Since 2012, the number of causes of death decreased to 5 and so the number of ultimate rates of decline that SSA must concurrently select became 150, which is still far too many to handle qualitatively. Rather than forecast mortality directly, SSA first assigns an annual rate of decline for each of the 75 years of its forecast. The rate of decline for the first two years of the forecast equals the historical rate of decline. The rate of decline for the next twenty-three years of the forecast linearly changes from the historical rate of decline to the subjectively chosen ultimate rate of decline. The ultimate rate of decline applies for the next fifty years of the forecast (the 26th through 75th years). SSA then imposes an additional step in its forecasts: if the historical rate of decline is negative, the rate of decline for the first two years of the forecast equals 75% of the historical rate of decline. Once SSA assigns rates of decline for each year in the forecast, it then iteratively multiplies the mortality rate in year t by its corresponding rate of decline. SSA then sums the forecasts across causes to produce a forecast for total mortality. Finally, SSA evaluates the quality of its overall forecast of total mortality well into the future. For example, if the age profile of the forecast in the year 2100 is not smooth or does not follow the ubiquitous shape of age profiles, SSA will readjust some of the ultimate rates of decline and reevaluate the quality of the updated total mortality forecast.

Formally, let $m_{a,t,c}$ represent the mortality rate for age group $a \in \mathcal{A}$, in year $t \in \mathcal{T}$, and for cause $c \in \mathcal{C}$. Let t_0 represent the year of the Trustees Report. Let $\hat{\beta}_{1,a,c}$ represent the estimated slope of a linear regression of the logarithm of historical mortality rates for age group a and cause c over the past twenty years as a function of time. Let $\gamma_{a,t_{\text{historical}},c}$ represent the historical rate of decline in mortality rate, which we estimate as $-\exp \hat{\beta}_{1,a,c}$. Let $\gamma_{a,t_{25},c}$ represent the ultimate rate of decline chosen by SSA, which applies for years t_{25} to t_{75} of the forecast. For years t_1 to t_{75}, the rate of decline is determined by the following conditional equation,

$$\gamma_{a,t_i,c} = \begin{cases} \gamma_{a,t_{\text{historical}},c}, & \text{if } i \leq 2 \text{ and } \gamma_{a,t_{\text{historical}},c} > 0 \\ 0.75 \times \gamma_{a,t_{\text{historical}},c}, & \text{if } i \leq 2 \text{ and } \gamma_{a,t_{\text{historical}},c} \leq 0 \\ \gamma_{a,t_{\text{historical}},c} + \frac{\gamma_{a,t_{25},c} - \gamma_{a,t_{\text{historical}},c}}{23}(i-2), & 3 \leq i \leq 24 \text{ and } \gamma_{a,t_{\text{historical}},c} < \gamma_{a,t_{25},c} \\ \gamma_{a,t_{\text{historical}},c} - \frac{\gamma_{a,t_{\text{historical}},c} - \gamma_{a,t_{25},c}}{23}(i-2), & 3 \leq i \leq 24 \text{ and } \gamma_{a,t_{\text{historical}},c} \geq \gamma_{a,t_{25},c} \\ \gamma_{a,t_{25},c}, & i \geq 25 \end{cases} \tag{5.1}$$

Finally, SSA iteratively multiplies mortality rates and rates of decline to forecast mortality rates, $m_{a,t_i,c} = \gamma_{a,t_i,c}\, m_{a,t_{i-1},c}$, for $1 \leq i \leq 75$.

Since SSA's task of forecasting involves the choice of a large number of highly interdependent parameters consistent with one another (210 until 2011 and 150 since 2012), it is inevitable that human beings—trying as hard as they possibly could to be fair and complete—will miss a great deal. Humans are not equipped with the memories large enough to keep all of the interrelated choices consistent. Although many estimates of ultimate rates of decline will at least be plausible, although probably not optimal, some will also not make sense. In Figure 5.4, we offer a few of some of the most problematic estimates we came across.

The top panel in this figure reports on deaths from diabetes in males and cancer in females. In both cases, the observed data (in dots on the left) indicate strong upwardly trending death rates, but the result of the parameter setting (the ultimate rate of decline) was an unnoticed and unjustified downward sloping curve. As implausible as this change in trend is, worse still may be that the (vertical) differences in death rates between the two adjacent age groups portrayed massive and implausible changes between the observed data (left dots) and forecasts (right lines).

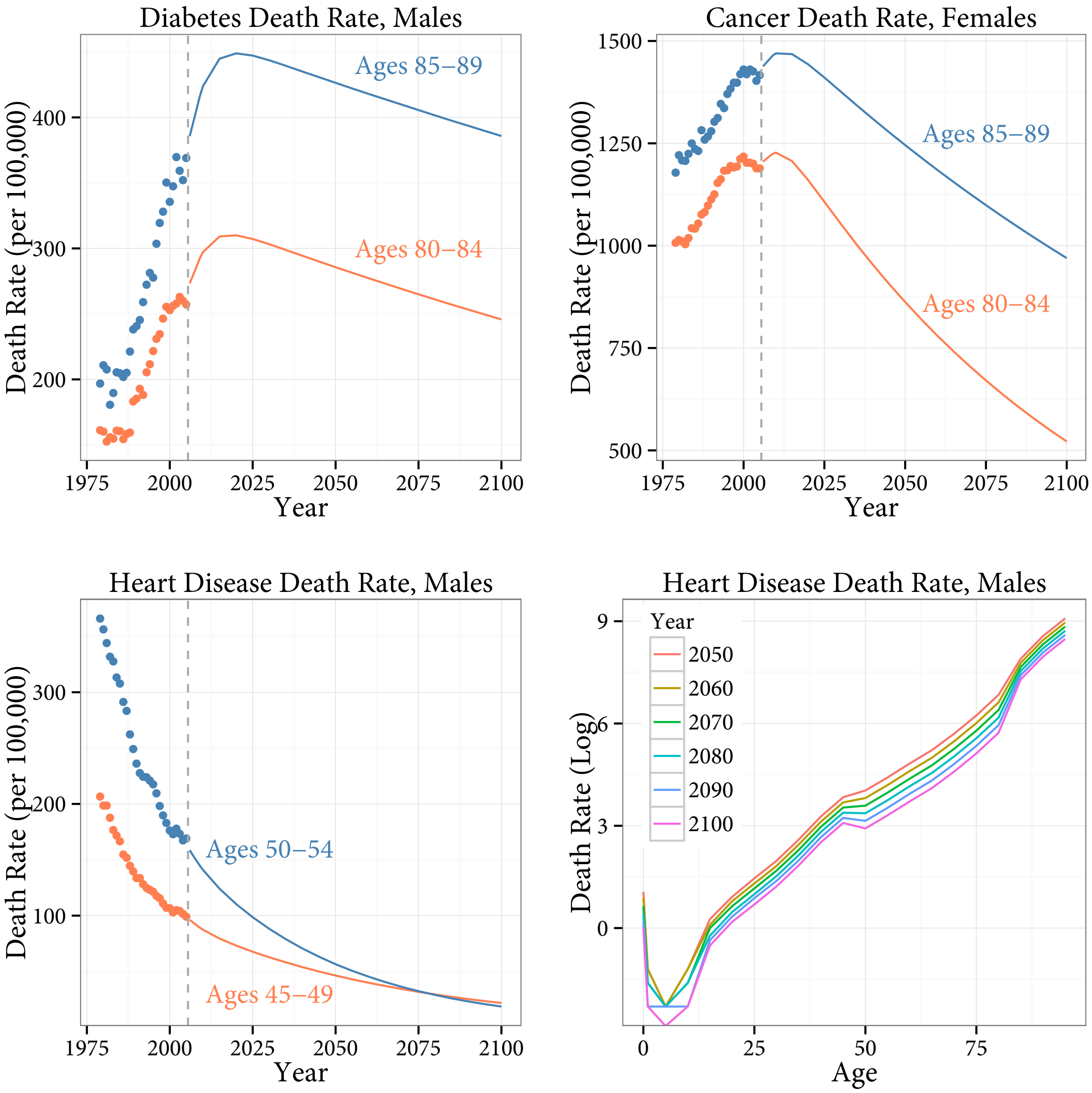


Figure 5.4: Selected Examples of Problems with SSA's Qualitative Parameter Selection in Point Estimates of Cause-Specific Mortality Forecasts.

For another example, the bottom row in Figure 5.4 reports on male heart disease rates. In the left graph, the out-of-sample data show the forecasts for two age groups crossing, so that after a time it will be supposedly safer to be age 50–54 than age 45–49; this, of course, makes no demographic or biological sense. Similarly, in the bottom right panel, we can see the age profile of the same data, along with a sharp drop at age 50–54 for all the forecast years, a pattern that would only make sense if there were some strange medical discovery that did not work until age 50 and ceased working at age 65. Clearly, if the experts at SSA could have focused on these numbers (as well as on all the others at the same time,) or if they had encoded their knowledge in formal, reliable, and informative statistical models, these highly implausible results that are contrary to well-known demographic knowledge would not have been part of official U.S. government forecasts.

Finally, Figure 5.5 gives a sample of two of the resulting SSA forecasts for male (left panel) and female (right panel) life expectancy. In this case, these are short term (1-5 year) forecasts, but our companion paper reveals similar results for longer forecasts. The figure is presented in terms of the residual (the forecasted SSA period life expectancy prediction minus the actual period life expectancy calculated from the Human Mortality Database; see `mortality.org`). In both panels, we can see approximately unbiased forecasts until about 2000 (i.e., residuals that vary around approximately zero) and then a sharp, trending, and continuing downward bias. The result after 2000 is that SSA forecasts people will die earlier than they actually do, making the Social Security Trust Funds appear healthier than they actually are.

In addition to the "best guess" intermediate cost scenario forecasts, SSA also forecasts life expectancy under low and high cost scenarios. While the interval formed by these alternative scenarios does not have a formal statistical basis, it is routinely used by the SSA to consider the range of plausible scenarios going forward. Recent Trustees Reports, for example, state that "alternatives I [low-cost] and III [high-cost] define a wide range of demographic and economic conditions". In addition to evaluating the residual for the "best guess" point estimate, Figure 5.5 shows the range of residuals correspond-

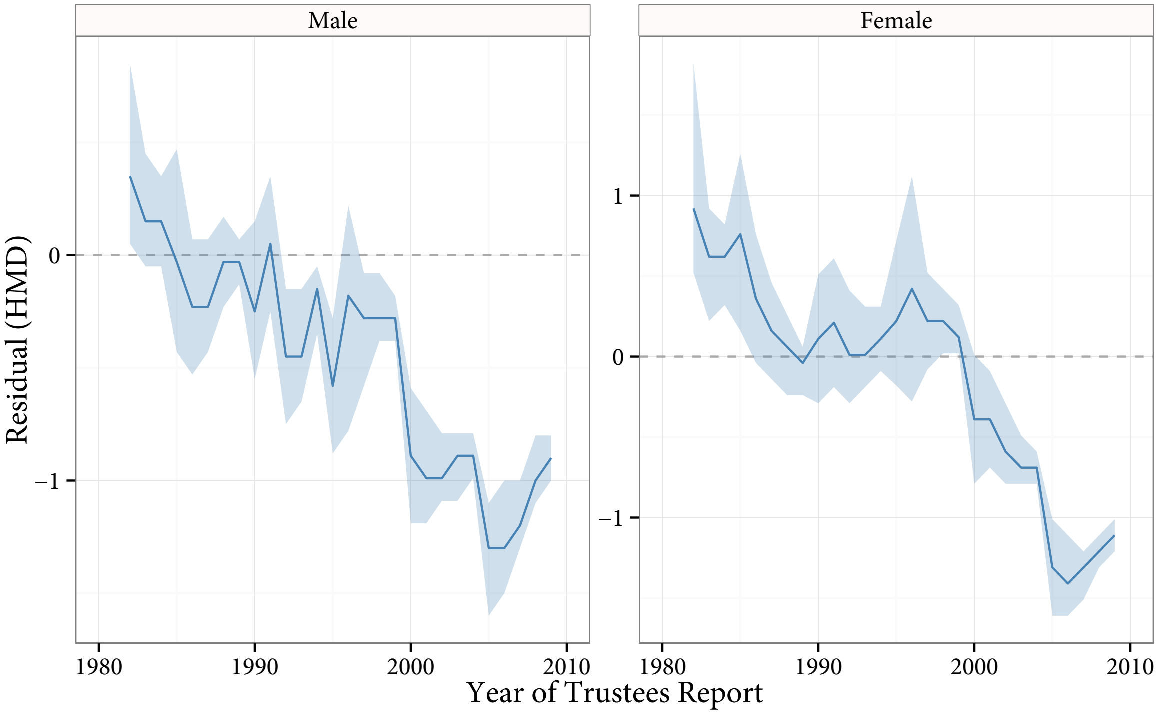


Figure 5.5: Residuals of Short-run Forecasts of Life Expectancy. Solid line represents residuals based on SSA's intermediate cost scenario forecasts. The shaded region represents the interval formed by SSA's low and high cost scenario forecasts.

ing to the high and low cost scenarios forecast by the SSA. If this interval covers zero, it means that the truth fell within the range of the bounds forecast by SSA. Post-2000, we see that the truth always fell outside this interval.

The overall pattern in Figure 5.5 is repeated in other mortality forecasts, as well as in the detailed financial forecast performance of SSA we give in Chapter 4. For example, the "cost rate" is one of the financial metrics the SSA forecasts and equals the ratio of the cost of the two SSA programs (Old-Age Survivors Insurance and Social Security Disability Insurance) to the taxable payroll for the year, expressed as a percentage. When comparing SSA's forecasts five years out to the eventual true observed value, we find that the average residual (defined as the difference between the forecast and the observed truth) from 1978 to 1999 was 0.10 percentage points. Post-2000, the average residual was -1.38 percentage points — a more than tenfold increase in the magnitude of the error. We find a comparable large increase in error when examining forecasts one year out to the eventual true observed value: 0.05 average residual for forecasts made in 1978-1999 and -0.52 average residual for forecasts made in 2000-2011. Similar results of approximate unbiased forecasts before 2000 and very substantially biased forecasts after 2000 exist for other SSA financial forecasts we studied, including the trust fund balance and the trust fund ratio. We present detailed figures with complete evaluations of SSA forecasting errors in our companion paper Chapter 4 and its associated replication data set (Kashin et al., 2015b).

5.1.4 The Uncertainty of Policy Scoring

In addition to producing the annual Trustees Reports, OCACT plays a singular role in American politics of scoring policy proposals put forth by members of Congress and public policy organizations. The scores are estimates of the impact of these proposals on the budget and the future of the Social Security Trust Funds. Because of the nontransparency of SSA forecasts, no other person or organization is able to produce fully independent estimates, and so all parties rely on these estimates in every major policy debate. Unfortunately, OCACT does not give uncertainty intervals for their point estimates. We fix this serious oversight here.

Since 1993, SSA has scored 105 policy proposals, including every major proposal to change the system by members of Congress and the White House. To provide rigorous uncertainty estimates, we note first that the uncertainty of counterfactual predictions, such as SSA's policy scores, always equals the sum of (a) the uncertainty of factual predictions — that is, forecasts under the current system — and (b) uncertainty due to what would happen if we change the system. Our results here quantify (a) directly and use these results as a lower bound on the total uncertainty of the policy scores.

OCACT gives estimates of the effect of policy proposals on the Trust Fund balance and on the cost rate. To estimate (a) for each, we take all forecast errors between 1 and 10 years out between 2000 and

2010, and compute the percentile of error at which each policy score appears. Policy scores that are statistically larger than zero should be larger than the 95th percentile of the forecast errors (i.e., corresponding to an $\alpha = 0.05$ significance level).

We plot all the forecast errors in Figure 5.6 for the Trust Fund balance (left) and cost rate (right). For each, we have policy scores for what will happen 10 years out (on the left of each box) and 75 years out (on the right). The uncertainty estimates we present here are lower bounds; however, because they are based on forecast errors 1–10 years out, they are much less tight lower bounds for the 75 year forecasts. Each dot represents one policy score; dots above the dashed line, appearing in red, are overwhelmed by forecasting error. The few policy scores that are significant at the 5% level appear in green at the bottom of the figure.

To be more specific, for the 75-year forecasts, the 95th percentile of forecast uncertainty for OCACT's policy scores is less than the estimated effect size of just 6 of the 68 proposal scores for Trust Fund balance and just 3 of the 23 proposal scores for the cost rate. Put differently, fewer than 10% of OCACT's 75-year-out policy scores made in the last two decades can be distinguished from random noise. None of the generally smaller 10-year-out policy scores can be distinguished from random noise. And this is under the most optimistic assumptions, with our figures being lower bounds as to the amount of uncertainty. If we were able to incorporate the likely larger uncertainty due to changing the system in ways that have never been observed (i.e., b above) or measure the uncertainty in real forecasts as far out as their policy scores are computed, the situation would be even more grim.

To be clear, the solution here is not for SSA to stop making policy scores, but rather for them to provide rigorous uncertainty estimates every time they make a prediction or counterfactual policy assessment.

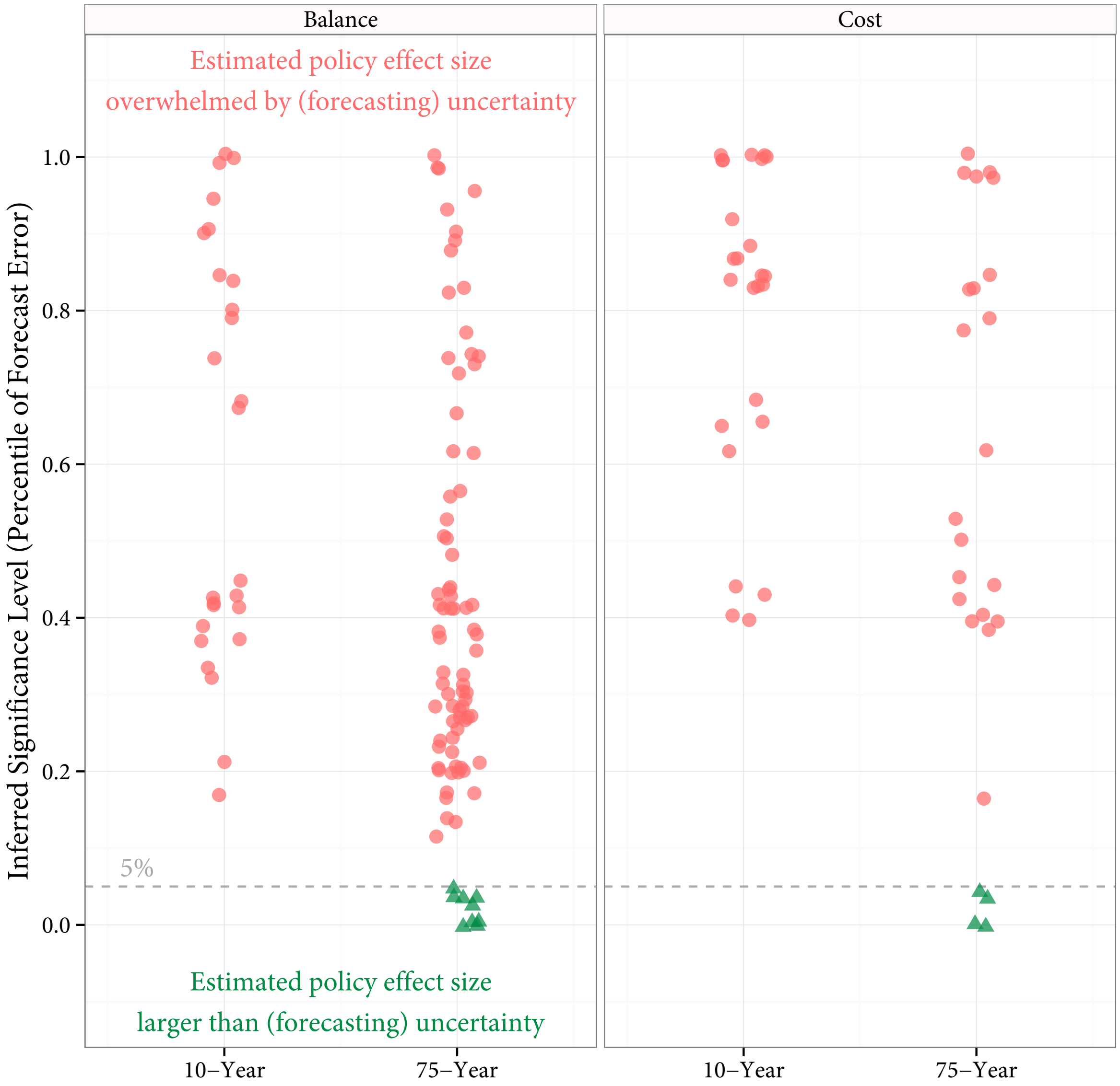


Figure 5.6: Empirically Inferred Significance Level of Estimated Policy Effect Size. Policy proposal scores for Trust Fund balance (on the left) and the cost rate (on the right), shown as red circles if forecast uncertainty overwhelms estimated policy effect size and as green triangles if forecast uncertainty (95th percentile, $\alpha = 0.05$) less than policy effect size. The left of each box are 10-year-out policy scores, and at the right are 75-year-out policy scores. All significance levels here are lower bounds on uncertainty, meaning that the reported α-level is likely higher than indicated.

5.2 The Origins of Social Security Forecasting Biases

We offer in this section a hypothesis about the origin of the systematic biases in SSA forecasting, which we summarized in Section 5.1. We begin by considering two possible but unlikely explanations, and then turn to our hypothesis. We present a variety of evidence for this hypothesis, including a powerful existing body of social psychological research and numerous interviews. Then in Section 5.3, we detail the internal and external pressures on OCACT, the weight of which is strongly consistent with the social psychological evidence for what generates biases in situations like these, and how to fix them.

5.2.1 A Possible but Unlikely Explanation

One logical possibility for increasing error rates in SSA forecasts is bad luck. No one predicted the onset of the Great Recession that began in December 2007, and so it would be unfair to hold SSA accountable for missing it. Additionally, between 2000 and the start of the recession, one might ask whether the small number of years of biased performance might well be due to random chance. This argument may be reasonable when evaluating the cost rate in isolation, but attributing the systematic patterns to SSA's errors in predicting life expectancy, and possibly other financial variables, is not consistent with that evidence. Moreover, SSA's errors in demographic forecasting go back significantly further than the Great Recession, and the observed values of demographic variables are highly smooth (with little change in the level of smoothness) over time. Finally, in our companion paper, we argue that the rise in unemployment that occurred during the Great Recession was not large enough to cause the corresponding decrease in mortality rates.

5.2.2 A Hypothesis

Our hypothesis for SSA's systematic forecasting errors begins with the observation portrayed in Figure 5.3: mortality and the change in mortality have been trending downward since 2000. Whether this pattern was predictable or not, it became increasingly clear after a few years that mortality rates were decreasing at a faster rate and, consequently, adults age 65 years and older were living longer lives and would be drawing Social Security retirement benefits longer than anticipated. At the same time, three separate phenomena caused SSA to not respond to these dramatic changes in the input data.

First, the *possibility of bias* exists because OCACT's forecasting procedures meet essentially all the major conditions for generating inadvertent biases established in the extensive and well-documented social psychological literature (Gilbert, 1998; Banaji and Greenwald, 2013; Kahneman, 2011; Wilson and Brekke, 1994). As we detail in Section 5.3.2, OCACT works hard to uphold its reputation and central position in policy debates, sometimes even to the extent of intentionally degrading the accuracy of its forecasts, by keeping them unchanged in the face of changing input data, to maintain its privileged central position. OCACT is represented at all major decision points and is open to talking to everyone in and out of government, but they see themselves—and attempt to maintain their position—as the sole judge and jury, and consequently routinely ignore recommendations of their scientific advisers. OCACT does not benefit from well-known internal procedures or external checks that could be imposed to avoid bias. OCACT relies heavily on informal decision-making procedures executed by committees composed of their staff with high levels of individual discretion and few formal procedures. In doing so, OCACT does not follow widespread objective and systematic administrative and statistical procedures that might prevent these problems. Their lack of transparency also leads to OCACT being the monopoly supplier of independent forecasts and policy proposal evaluations, with no one else in SSA, the government, or the public able to offer an alternative view or to help SSA by surfacing biases that may inadvertently arise. Although the Congressional Budget Office evaluates policy proposals, its solvency forecasting model assumes the veracity of SSA's demographic forecasts.

Second, a higher *probability of bias*, as well as the specific direction of bias, appears to have occurred because of new and unprecedented external political pressures on Social Security that began after about 2000. As we detail in Section 5.3.3, the Chief Actuary and his office found (or placed) themselves in the untenable position of simultaneously being the defender of Social Security, a supposedly unbiased arbiter between increasingly polarized political parties, and a defender of their own office's reputation. OCACT responded to this pressure by hunkering down and trying as hard as possible to resist change in response to political pressures. Although a laudable attempt, social psychological evidence indicates that ad hoc and qualitative procedures allow biases no matter how hard individuals try to avoid them. The particular direction of bias turned out to cause OCACT to be insulated not only from inappropriate political pressure but also from needed changes due to changing patterns in mortality and other inputs into the forecasting process.

The third and final component of our hypothesis is that we need not assume anything but *good intentions* of all employees of SSA. Our results are consistent with the oft-stated insistence of SSA officials that they try as hard as they can to be as unbiased and objective as possible with regard to external political or other pressures. Indeed, our interviewees indicate that the actuaries at OCACT, usually represented by Chief Actuary Stephen Goss, work hard to help those on both sides of every policy debate over Social Security. Some emphasized how hard the actuaries work by explaining that civil servants do not need to be at White House or Congressional policy meetings late into the night but, if that was when the discussions where happening and they could have influence, OCACT always occupied their seat at the table. OCACT jealously guards the independence it has been granted by Congress, despite pressures from members of Congress, the Administration, and other SSA officials. Moreover, OCACT does what it can to be helpful to those involved in crafting legislation while trying to avoid bias. OCACT is regarded as among the better forecasting groups in the federal government, and indeed most other countries with public retirement systems similar to SSA have no forecasting arm at all.

5.3 Pressures on the Social Security Administration

We now discuss social-psychological pressures, pressures internal to SSA, and external pressures on SSA.

5.3.1 Social-Psychological Pressures

Although perhaps counter-intuitive, Banaji and Greenwald (2013) and numerous others in the literature have shown that good intentions can coexist with a high probability of bias when human beings perform complex tasks with high levels of discretion over many individual decisions, little feedback on whether they made the right choice the last time, high levels of external pressure, and few external checks. Humans have limited access to their own mental processes, and their biases do not give rise to any self-evident subjective experience they can use to avoid the biases Wilson and Brekke (1994). Controlling one's own mental processes to avoid bias is often difficult or impossible, and most people vastly overestimate their ability to control their own mental processes and potential biases, even when explicitly told about documented biases in their own behavior. In fact, subject matter experts overestimate their ability to control their own personal biases even more than nonexperts.

Most importantly, attempting to reduce these biases by merely "trying harder," or replacing one person with another who is even more vigilant, will usually have little or no effect. In this regard, even "teaching psychology is mostly a waste of time" (Kahneman, 2011, p.170).

These psychological biases are exacerbated by problems with uncertainty estimates in the same situations. Experts who use informal qualitative approaches are typically overconfident. Indeed, the more prominent or central a role the forecaster has—and as the sole supplier of forecasts and policy evaluations, OCACT could hardly be more central—the more overconfident their statements (Tetlock, 2005). The conclusion of the psychological literature on estimating confidence levels qualitatively is clear: "do not trust anyone—including yourself—to tell you how much you should trust their judgment" (Kahneman, 2011, p.240).

Research shows that "Biases in human reasoning are of two general types: those that result from the failure to know or apply an explicit rule of inference — the *failure of rule knowledge or application* — and those that result from *mental contamination* (cases whereby a judgment, emotion, or behavior is biased by unconscious or uncontrollable mental processes)" (Wilson and Brekke, 1994, p.118). A key finding of this literature is that although mental contamination in individual judgment (i.e., "personal bias") is in many instances not correctable no matter how hard one tries to avoid it, errors due to the failure of rule knowledge or application can be fixed by learning or by instituting formal procedures.

Some of the problems can be avoided altogether by replacing qualitative judgments with formal statistical rules capable of being learned and applied objectively. Others can be corrected by imposing systematic procedures on qualitative decision making. Of course, deciding what formal statistical rules to apply are themselves also subject to mental contamination and other biases, especially when implemented by an individual or small group working together. However, these problems are vastly less likely to occur when different teams, with different backgrounds, perspectives, and preferences, check each other, as occurs in a well-functioning scientific community. This is a key advantage of following the replication standard, making data and forecasting procedures publicly available, and encouraging the scientific community and others to participate in helping to ensure that SSA's forecasts are as accurate as possible.

By this theory, avoiding future forecasting biases like those we document above will probably not be achieved by OCACT personnel working harder or trying to be less biased. The solution instead is to make organizational changes that:

1. Remove human judgment where possible by formalizing informal procedures, and in the process taking advantage of dramatic progress over the last quarter century in statistical modeling and data science (about which more in Section 5.4);

2. Institute structural procedures when qualitative judgments are still required. For example, as in double blind article reviews, it may be possible to elicit information from experts before they know the details of how the information they provide will affect the ultimate forecasts, and in ways that reduce the chances of "group think"; and

3. Share data and procedures with the rest of SSA, the scientific community, and the public so that any biases that inadvertently occur in steps (1) and (2) are detected and corrected, and so that groups in addition to OCACT can provide forecasts and policy evaluations.

5.3.2 Internal Pressures

The internal pressures on OCACT we detail in this section are related to the Office attempting to protect its central role in the Washington debate, to be useful to policymakers, and to be seen as important. We list here seven characteristics of OCACT which describe how it pursues these goals. These characteristics interact and overall portray hard working public servants trying to do their jobs without any substantive bias. Yet, they remain unprotected from biases because they have not adopted the well known procedures developed in social psychology, behavioral science, and statistics.

Island of Fairness

First, the self-conscious public stance of OCACT is as an island of fairness and objectivity amidst a storm of partisans, and so far as we can tell this is precisely what they attempt to do. The present Chief Actuary, Stephen Goss, regularly appears in public making earnest sounding but extreme claims about how unbiased he and OCACT are. For example, in a public address, Goss said "I'll take a bullet before I modify anything under any kind of political pressure, and that's just an absolute. My sense is that there are some jobs and you do whatever it takes and if people don't like it that's too bad.... We're giving it all we got, and objectivity, challenge everything, and no known bias, is always the mantra" (j.mp/GossCSPAN13). Similarly, one person who served as a Trustee told us emphatically "I have never seen a *single* instance of political pressure" in OCACT. And the feeling is mutual, as indicated by Goss' public statement that the Trustees "work in a really *truly* nonpartisan way" (j.mp/GossCSPAN6-13). Having public servants who try for this level of fairness is certainly ideal but, as indicated above, trying harder to be free from bias is an ineffective way to further reduce bias unless they begin to use the well-tested advice from the scholarly literature.

Monopoly Supplier of Evaluations and Forecasts

Second, as indicated in the introduction to this chapter, OCACT's nontransparency, lack of data sharing, and informal forecasting methods means it is the monopoly supplier of fully independent forecasts and evaluations of policy proposals. Goss says regularly that OCACT gives different projections under different "explicitly stated assumptions" (j.mp/GossCSPAN13); however, OCACT has full discretion to choose to evaluate or ignore any request to evaluate policy proposals under any assumptions other than those requested by Congress or the Administration. This stance may be consistent with the idea of OCACT being an island of fairness, and it may even be required, but it means that any bias inadvertently introduced, such as by choosing to evaluate only proposals with certain assumptions, is unlikely to be corrected. Moreover, SSA cannot take advantage of the central contribution of the idea of science, which is not merely "acting scientifically" but rather involves different groups checking on each other to do better than any group could on its own.

Although all final decisions are made by the Trustees, the scientific capacity to make and judge forecasts inside SSA resides almost exclusively with OCACT. Unfortunately, OCACT seems to act as a judge and jury, rather than a participant in the scientific process leading to the forecasts. Until the last two years, OCACT has offered little explanation as to why their forecasting practices differ so dramatically from the Technical Panels recommendations in many instances. In the 2012, 2013, and 2014 Trustees Reports, they addressed a small number of issues, mostly to declare the Technical Panels correct or incorrect on each issue, but with little serious engagement with, or respect for, the Panel's arguments or conclusions (j.mp/OCACT12, j.mp/OCACT13, and j.mp/OCACT14).

Consistency Bias

Third, most of our respondents emphasize that OCACT values consistency in forecasting over time above accuracy at any one time. That is, in the face of new predictive information or new methods of analysis, OCACT intentionally degrades forecast accuracy, biasing today's forecast towards yesterday's forecast. Congruent with this claim, OCACT forecasts tend to be much smoother over time than those from Medicare and others. There is of course a prudent aspect to this pattern, where a good forecaster tries not to overweight the last bit of new information.

However, many of our respondents prefer to explain this pattern via either personal or institutional explanations. The personal explanation attributes consistency bias to individuals, particularly the Chief Actuary, doing whatever he can to avoid having to admit his office was wrong. The institutional explanation attributes consistency bias to OCACT trying to emphasize its central role in the policy debate in Washington, since it recognizes that negotiation between the parties is easier when they agree on one set of consistent forecasts, even if they are wrong (or not known to be right). In support of either explanation, the bias in favor of consistency over time leads Chief Actuary Goss to be regularly described by our respondents as "intellectually biased, but not politically biased." Goss himself described this consistency bias: "the hard part is trying to balance the need to change on the basis of new ideas and understanding with the desire for consistency and stability over time" (Interview with Society of Actuaries, j.mp/GossSOA).

We can show exactly how this consistency bias occurs with an example that arose in multiple interviews we conducted. It conveys the high level of discretion allowed by OCACT's forecasting procedures (which may lead to this and other types of bias) and how consistency is implemented. When the Technical Panel strongly recommends changes in one of OCACT's myriad forecasting assumptions, they receive one of three responses: when the Chief Actuary had good evidence, he engaged the Panel and convinced them that no change was needed; when the Panel had better evidence, Goss ignored the Panel and did not change the assumptions (Autor and Duggan, 2006); finally, in the small subset of cases when the Panel pushed hard even though Goss was ignoring them, Goss changed the assumption in the direction the Panel wished (although often not as far as it wished) but then changed another, unrelated, assumption not at issue in the opposite direction to counterbalance the first and keep the ultimate solvency forecasts largely unchanged.

Several of our interviewees independently suggested that they thought Goss maintains a private list of assumptions that in his best judgment require change. However, instead of making these changes when they seem to him to be scientifically warranted—immediately—OCACT introduces them over a much longer time frame, at instances specifically chosen to counterbalance other changes in the world and pressures from the Technical Panel, all in order to keep the ultimate forecasts relatively consistent over time.

The issue here, again, is not the people but the procedures, since most of these interactions are informal, out of view of the public, and thus subject to potential unintended biases. These procedures could easily be changed by making them visible, and the forecasts, as a result would be easy to improve.

One defense of consistency bias that occasionally arises is that SSA's goal is forecasting 75 years out, and so it may not make sense to adjust forecasts in response to every new piece of information that comes in. From a Bayesian point of view, this argument is plainly false. And from the point of view of scientific evaluations, no evidence exists on the accuracy of 75-year Social Security forecasts, and so all that can be done is to evaluate shorter term forecasts where data exist. And finally, even if the shorter term evaluations are not relevant to 75-year forecasts, they are still vitally important to tens of millions of Americans who plan to receive benefits over shorter time horizons.

Ignoring Technical Panel Methodological Recommendations

Fourth, SSA's external scientific advisers (their Technical Panels) have long recommended that OCACT evaluate OCACT's forecasts, share their data, make their procedures and decisions transparent, and formalize their methods. The Congressional Budget Office, for example, routinely self-evaluates its own Social Security forecasts, although they rely on OCACT demographic forecasts as an input. (The technical panel consists of outside experts appointed by the Social Security Advisory Board. The panel assesses key demographic and economic assumptions, provides advice, but it does not independently make forecasts.) OCACT ignores or only partially follows most of these Technical Panel's recommendations. For example, on evaluating forecasts, the 2007 Technical Panel wrote:

> We believe that the accuracy of past projections should be the subject of routine reporting, either in the Trustees Report or in separate supplemental publications on methodological developments. There should be an analysis of the accuracy of past 10-, 20-, and 30-year projections similar to those periodically done by the Census, Bureau of Labor Statistics (BLS), and the Congressional Budget Office. The report should include a comparison of historical values with projected high-cost and low-cost scenario variants, noting how often each variable exceeded past projected outer bounds. (j.mp/SSATech07, p.4-5).

Other technical panels have also encouraged OCACT to share data and information with "different parts of SSA and within the larger research community" (j.mp/SSATech11, P.3). OCACT has made some information available, and the Technical Panel has been appropriately generous in complementing OCACT for some progress, but withholding even one link in the forecasting chain means that replication is impossible. The 2007 Technical Panel was unambiguous on this point:

> Throughout this report we call for more transparency in the models and data the actuaries use, as well as the assumptions that drive their results. This recommendation is perhaps the most important one we make. Only with more transparency can other social scientists…bring their intellect to bear on the many complex questions the Trustees and actuaries face.
>
> It is worth noting that all analytical agencies…must make assumptions about behavior or phenomena that are unobservable or immeasurable. In the process, the assumptions

> become deeply embedded and the models closely guarded; and these agencies are understandably reluctant to revisit their assumptions or reveal methods. Nonetheless, it is essential to do so. First, accountability requires it.... Second, and more important, ongoing comprehensive, external review can greatly assist the quality of the analytical exercise.
>
> Transparency of the OCACT models will require several developments: (1) providing more comprehensive documentation, (2) making data from SSA records more available, (3) creating explicit models where none exist, and (4) clearly explaining the processes employed. The ultimate test of transparency is whether the actuaries' results can be essentially replicated.

These issues have been raised in many ways in all the Technical Panel reports over the past 15 years. In 2011, SSA released some appendices with additional details, but they still do not meet the replication requirements of the recommendations.

Ignoring Technical Panel Substantive Recommendations

Fifth, OCACT and the Trustees ignore, or at best undervalues and underutilizes, important substantive recommendations by its Technical Panels, even those issued repeatedly. Others are given little more than token mentions or dismissals in the annual Trustees Report. On some points, the Trustees Reports directly contradict the conclusions of the Technical Panel and, in defense, the Trustees merely repeat identically worded assertions year after year in the annual Trustee Report without engaging the Technical Panel on the crucial issues raised. The Trustees and Technical Panel agree on many issues too, but the lack of engagement or mutual understanding is obvious.

Consider one important example highlighting both the lack of responsiveness of the Trustees to the Technical Panel and the high levels of discretion OCACT's actuaries have in making numerous informal and qualitative decisions—factors that create exactly the conditions for inadvertent bias more likely to occur. Part of OCACT's informal forecasting approach involves a committee choosing the large number of ultimate rates of decline in mortality discussed in Section 5.1. The committee then enters into complex discussions with many others in Congress, the Administration, and elsewhere, before making a final decision. (See Appendix B for more details.)

Since human beings are incapable of keeping so many moving parts in their heads at once, the results are suboptimal and are sometimes even logically inconsistent (Soneji and King, 2012). The Technical Panels have repeatedly recommended changing this approach, with little response from SSA (e.g., in 2012, the number of causes of death was reduced from 7 to 5):

> For the intermediate scenario, the Trustees make assumptions about 70 rates of decline (5 age groups × 2 sexes × 7 cause categories). Consideration of the low-and high-cost projections increases the total number of parameters to 210. The Trustees Report does not describe the process for arriving at this large number of assumptions. (2011 Technical Panel, p. 57; j.mp/SSATech11)
>
> The process of producing 70 assumptions about ultimate rates of decline by age, sex, and cause of death for each of three cost scenarios is not documented and appears to be informal. Simplifying the mortality assumption will considerably improve the transparency of the Trustees Report. [for emphasis, they also quote previous reports] (2011 Technical Panel, p.59)
>
> A model based on separate projections by cause of death over a long time horizon is both implausible and inconsistent with historical experience. There is little written explanation of how these assumptions were developed. (2011 Technical Panel, p.59, quoting the 2003 Technical Panel, p.38; j.mp/SSATech03)

In this situation, each of the 210 decisions for ultimate rates of decline is ambiguous; each decision depends on others and often in complicated ways. A clear definition of the objectively correct outcome does not exist, and during this process individuals in Congress, the administration, and OCACT have their own opinions, often pushing hard to get their favored outcome. The odds are high that an individual or committee in OCACT with an even very slightly favored outcome will inevitably, perhaps imperceptibly, bias the average decision in their favor. And the evidence indicates that hard working, professional and objective public servants will not be able to overcome this bias, no matter how hard they try, without instituting some type of known, open, and formal procedures.

The differences of scientific opinion between the scientists on the Technical Panels and the OCACT acting on behalf of the Trustees are often of considerable importance. For example, Figure 5.7 documents the four Technical Panel recommendations on the overall ultimate rate of decline of mortality (in red) and the choices made by OCACT and accepted by the Trustees (in turquoise).

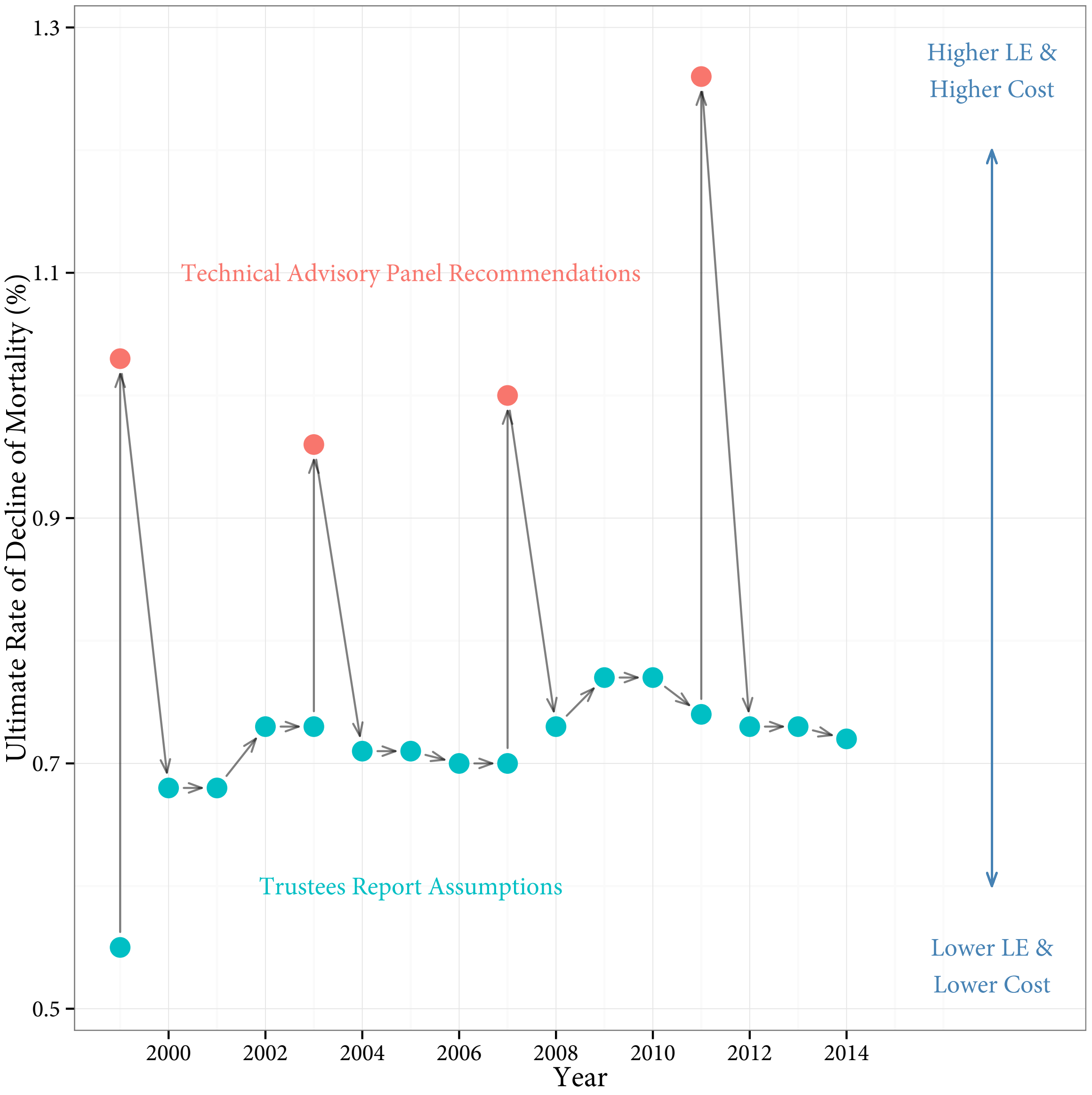


Figure 5.7: Technical Panel Recommendations of Ultimate Rate of Mortality Decline and SSA's Chosen Assumptions. Note: LE=life expectancy.

The data from the four Technical Panels in this figure cover the period during which we documented systematic bias in SSA's forecasts above. The results in this example demonstrate that the Trustees have consistently ignored the Technical Panel for the last 15 years. The differences in the figure are large, consistent, and all in the same direction. In which direction has SSA chosen to deviate from its own scientific advisers for so many years? In the same direction as the systematic bias shown above in shorter term forecasts: SSA chooses to make the Trust Funds look financially healthier than their scientific advisers think they are.

Informal Procedures that Increase Vulnerability to Bias

Sixth, many of OCACT's procedures have not changed in decades, despite scholarship that clearly demonstrates their suboptimality and likelihood of generating bias. Some of their procedures can be automated to eliminate bias; others cannot be automated and must remain qualitative, but could easily be changed to reduce bias.

For example, consider OCACT's informal procedures for choosing the 210 ultimate rates of decline (or 150 as of a few years ago). OCACT allows participants in the process to see the effect of their judgments about the value of input assumptions on the ultimate solvency forecasts while making the judgment. This means that the information they are trying to elicit is very likely to be contaminated by OCACT's known consistency bias in solvency projections (see the third point above) and possibly also any other pressure or preference.

Indeed, allowing this type of contamination is a qualitative example of a common quantitative forecasting mistake (Girosi and King, 2008, p.11–12), where a forecaster tweaks a statistical model until the results make the ultimate forecast consistent with their prior belief. The result, of course, is that the model itself contributes little or nothing to the process, and the forecaster is merely choosing the forecast he or she wanted initially.

Fortunately, eliminating this contamination would be straightforward and could occur in one of two ways. First, OCACT could retain their qualitative decision process and institute standard "debiasing" procedures such as blinding those providing information about the assumptions from their effects on solvency until after providing their information. Thus, OCACT would ensure that the process of eliciting information would be focused on only the information. Second, OCACT could instead formalize all available knowledge about their assumptions in a formal statistical methodology. The former would have the advantage of disrupting OCACT's practices as little as possible. The latter would change the process more, but would likely also substantially improve the quality of forecasts even after the effect of removing the bias.

Personnel Changes

Finally, we note the most recent change in personnel in OCACT leadership: Harry Ballantyne was the Chief Actuary from 1982 to 2000; Stephen C. Goss took over in 2001 and remains in the post. However, attributing the increasing biases that began coincident with this change in leadership is not as obvious as may seem, as Goss had worked as a civil servant in OCACT for 30 years prior to being promoted to Chief Actuary as have other actuaries in OCACT. Moreover, as we explain above, the social-psychological evidence indicates that a change in procedure, not different the specific individuals following the same procedures and or trying harder, is what would make a difference in reducing bias.

And in fact, the possibilities for bias were evident even before Goss took the helm. For example, Rosenblatt and DeWitt (2005, p.44) give an example of one type of influence from an earlier era:

> "The budget assumptions of any administration are often overly optimistic," said former chief actuary Harry Ballantyne. The political leaders, whether Democrat or Republican, believe that their economic and fiscal policies will produce positive results in the short term. They would like the Trustees' Report to reflect their optimism. "Many times there are some differences" between the actuaries and the political staff members, said Ballantyne. "The cabinet staff people would say, 'why can't we assume this?' We go back and forth, and sometimes changes are made."

When many such opportunities for influence, individual discretion, and lack of formality occur, the forecasting process becomes open to the possibility of inadvertent biases that no amount of individual effort is likely to solve. Instead, the solution here is for the personnel in charge to institute some of the proven procedures well known to avoid bias. Our hypothesis is that the internal pressures discussed in this section open up the possibility for bias, with probabilities for bias increasing as a result of the increase in external pressures we discuss in the next section.

5.3.3 External Pressures

The internal situation at OCACT within SSA described in Section 5.3.2 makes clear that inadvertent biases are at least a possibility. Such biases can occur because numerous small informal decisions are made by people without the benefit of formal statistical models, automated computer assistance, or formal procedures for how to impartially elicit informal or qualitative information. Since many of OCACT's general procedures have remained largely unchanged for so long (during which dramatic improvements in forecasting methodology, social psychology, and behavioral science have taken place), something else must have changed to increase the likelihood of bias beyond a theoretical possibility. The hypothesis we offer here is that of unusually intense political pressures from the outside, which began at about the same time, caused OCACT to resist all types of change, even when it was needed due to changing inputs.

To begin, we first lay out the partisan motives and political strategies underlying the ongoing debate over Social Security. In particular, the complicated politics seems to have played out in this domain in at least three ways. Through all of this OCACT was pushed and pulled from every direction, and during which the actuaries tried hard to resist change.

First, the standoff on Social Security has many conservatives preferring to change the system by creating personal savings accounts managed by the government (often called "privatization," especially by opponents) or cut back its benefits, and many liberals preferring to leave the program as is or to expand it. The conservatives' political strategy to achieve these goals seems to be to emphasize the forecast point estimates, most of which predict Trust Fund insolvency by the 2030s. They then characterize the situation as a crisis and call for early negotiations to "save the program," which would likely lead to their goal of cutting it back to some degree. In contrast, the liberals' political strategy seems to be to deny any crisis by emphasizing the size of the uncertainty intervals, which are wide enough so that insolvency might well be averted by waiting for the economy to improve. Waiting might also benefit the liberals since they would have a chance for the political tide to turn and possibly emerge with a more amenable or less powerful negotiating partner (as when the original Social Security Act was passed in 1935). Of course, this characterization does not account for all liberals and conservatives. In fact, both presidential candidates in the 2012 election took what we are describing as the "liberal" position — Obama presumably for the liberal reason above and Romney to focus voters on other issues.

Second, if the two sides in this debate took the point estimates seriously (i.e., the best guess about what will happen), they might well switch positions. That is, any reasonable forecast point estimate has the system going insolvent by about the 2030s. Instead of meaning the end of the program, insolvency would translate into benefits being cut by about 25%, which might be roughly what the conservatives are after. So, waiting might be the preferred conservative strategy, whereas insisting on early negotiation — when small and relatively painless changes in payroll taxes or retirement ages could ensure long term solvency — would be the liberal preference. Alternatively, since few involved in the political debate think it is reasonable to make changes to the system for those within a decade of retirement, waiting might well involve more taxes or fewer benefits, and so perhaps the conservatives should not want to wait. Overall, the intensity and complexity of politics here may explain some of the positions of Republicans and Democrats who, respectively, often do not fit the current conservative and liberal debating positions, and sometimes do not agree among themselves.

And finally, we could look at the issue from the perspective of time. Politicians who are trying to maximize their own self-interest need to do so within a time frame, which especially affects decision making about a program designed to work over many generations (Weaver, 1988, Ch.4). Republicans at any one point in time may want to delay shoring up Social Security in order to focus the immediate political debate on other issues, such as tax cuts. In contrast, Democrats, such as the Obama Administration taking office during an economic crisis, might also want to delay fixing long-term Social Security solvency problems while it deals with its own shorter term issues.

The result of these highly visible conflicting political forces has led to chaotic politics over the Social Security program. However, along with the general increasing partisan rancor in Washington around the turn of the millennium, the existing political divide over Social Security began to more cleanly separate and intensify. In the late 1990s, "the politics of Social Security in the United States entered a new phase as…privatization…came to dominate the debate" (Beland, 2005, p.165). As opposed to previous movements to reform Social Security, privatization requires a fundamental change in the structure of the program, eliminating or reducing the importance of the Trust Fund — and at least some of the need for actuaries and the OCACT to forecast its solvency — by instituting individual retirement accounts. As the conservatives increased their support for personal savings accounts, a debate about them emerged within the Clinton Administration. Although Social Security reform of some type was on their agenda, it was ultimately dropped because the Monica Lewinsky scandal used up too much political capital and attention. President Clinton, not wanting to divide his own supporters at this politically sensitive time, used his 1999 State of the Union address to oppose privatization (Beland, 2005; Beland and Waddan, 2012).

Around the same time, some began to argue that Social Security forecasts were too conservative, indicating that the Trust Funds were in less danger than claimed. This resonated with the Democrats, who even convinced the General Accounting Office to hire PricewaterhouseCoopers to investigate whether the "Social Security Board of Trustees' may overstate the Social Security deficit" (j.mp/pwcgao). PwC evaluated whether the actuarial methods used by OCACT to produce forecasts for the 1999 Trustees Report conformed to standards of actuarial practice, such as data sources and how measures were constructed; PwC did not evaluate or compare SSA forecasts with the eventual truth.

Then, in 2001, George W. Bush took office and, beginning with his inaugural address, became the first president since the inception of Social Security to call openly for a major structural reform in the program. Bush made numerous speeches supporting personal savings accounts (j.mp/GWBss). In 2001, Bush appointed the President's Commission to Strengthen Social Security — a body instructed to support his conclusions. The Commission developed and evaluated solvency proposals under a set of guiding principles provided by the Bush administration. While the September 11 terrorist attacks largely put Bush's social policy agenda on hold, its savings accounts were resurrected in 2005 in even stronger form (Altman, 2005; Beland, 2005; Beland and Waddan, 2012). As part of the initiative, Bush embarked on a series of town hall meetings that represented "perhaps the most extensive public relations campaign in the history of the presidency on behalf of reforming Social Security" (Edwards III, 2007, p.284). Sounding a favorite theme, President Bush said in a radio address that "the system is broken, and promises are being made that Social Security cannot keep" (1/15/2005, j.mp/BushRadio05). Overall, the Bush initiative "was the biggest and most concerted effort to overturn the program since its birth seventy years earlier and the only one directed from the White House itself" (Laursen, 2012, p.504).

These changes in the public debate put SSA and its actuaries under a new and unusually extreme form of political pressure. They were pulled in every direction — while always still trying to remain internally consistent and relatively constant over time (as described in Section 5.3.2). The Democratic minority on the Committee on Governmental Reform issued a partisan but compelling report with specific examples of changes in the language used by SSA as it supposedly bent to the will of the Bush White House (j.mp/DemsSS05). Contemporaneous news reports confirmed the Bush Administration's at least partially successful attempts to influence SSA communications to emphasize the financial unsustainability of the program (j.mp/NYTs1-05). In response to the claim that SSA forecasts were too conservative (and thus helping Bush's reform efforts), Charles Blahous (who was soon to become a Public Trustee of Social Security) distributed an extensive rebuttal for a public presentation (Blahous, III, 2007, 2010). Additionally, Chief Actuary Goss openly clashed with the Republican SSA Commissioner, and Goss was, in turn, strongly defended by Democratic lawmakers (see j.mp/GossVSreps).

The external political pressure on SSA summarized here became unprecedented at just about the time when SSA's systematic forecasting biases began to increase. It appears that OCACT reacted to the extraordinarily intense and chaotic partisan politics by redoubling the practices we discussed in Section 5.3.2 and trying to resist the political forces, but also inadvertently resisting genuine changes in input data. They engaged more in policy discussions to try to protect their position, resisted more than ever advice from their Technical Panel and others, maintained their consistency of forecasts over time even when evidence was building to the contrary, and generally remained in a pre-2000 world.

As the data indicated change and an increasing chorus of outside voices lobbied for it, the OCACT felt that the best way to retain their independence was to stay the course. Then, with the internal pressures and lack of open, transparent information and data sharing highlighted in Section 5.3.2, any biases that entered into the forecasting process, such as by ignoring the input data, were not likely to be detected or corrected.

The solution to these problems is not to find the perfectly unbiased and objective arbiter, since this person exists only in theory, but to open up hidden procedures for all to see and to encourage public and scientific scrutiny to help the system and its resulting forecasts improve. SSA forecasting methodology and procedures need to be modernized.

5.4 Formalizing Ad Hoc, Qualitative Forecasts

We provide an overview of how to fix the problems with SSA forecasting in Section 5.3.1. A key component of these suggestions is to replace ad hoc qualitative forecasting with formal statistical methods that include the key features of mortality data.

We conceptualize log mortality for each time period and age group as the dependent variable in a linear regression containing time trends and risk factors such as smoking and obesity as predictors. We can then express expert information on the smoothness of log mortality via a set of specialized Bayesian priors (Girosi and King, 2008). Unlike traditional Bayesian approaches, these priors are stated in terms of the expected value of the dependent variable, not in terms of coefficients. Such an approach reflects the knowledge demographers and public health experts actually have (since mortality is directly observable). In practice, placing priors on coefficients puts impossible constraints on the parameter space; priors on expected values implies priors on the coefficients that vary over the observations and, consequently, are much easier for experts to encode their prior information.

To begin, let a be the index age (for a total of A ages) and t be the index the year (for a total of T years). Suppose that the log conditional probability of death at a given age a and a given year t (denoted q_{at}) follows a normal distribution, where the mean of the distribution systematically varies as a function of covariates:

$$q_{at} \sim \mathcal{N}\left(\mu_{at}, \sigma_a^2\right) \tag{5.2}$$

and

$$\mu_{at} = \mathbf{Z}_{a,t}\beta_a, \tag{5.3}$$

where $a = 1, ..., A$, $t = 1, ..., T$. $\mathbf{Z}_{a,t}$ is a vector of exogenous covariates that typically contains a linear time trend, smoking prevalence lagged k years in both time and age, and obesity prevalence lagged k years in both time and age. Other covariates can be easily included as well. Time is included as a crude measure of technological change; improving this measure would be valuable, and measures exist, but more predictive measures have not been found. In more general matrix notation, $\mu = \mathbf{Z}\beta$, where $\mathbf{Z}$ is a block diagonal matrix (with $\mathbf{Z}_a$ forming the blocks) and β is a column vector formed by concatenating the A age-specific coefficient vectors.

As discussed, experts possess key information about the behavior that mortality forecasts should exhibit. This expert information is grounded in decades of carefully examining demographic data and the patterns within it. The typical Bayesian approach is to quantify such expert information in the specification of a prior on the regression, $\mathcal{P}(\beta|\theta)$, where θ is a hyperparameter or a smoothing parameter. Similarly, $\mathcal{P}(\sigma)$ is the prior distribution for the standard deviation.

We allow users to formulate their priors in terms of the expected value of log mortality (the dependent variable), subsequently backing out the corresponding prior on the regression coefficients: $\mathcal{P}(\mu|\theta) \Rightarrow \mathcal{P}(\beta|\theta)$, which then allows standard Bayesian computational techniques to be used. Since μ_i is a scalar and β is a vector, the many-to-one transformation seems impossible. However, if we restrict our attention to the subspace of $\mathbb{R}^{T\times A}$ where μ can be explained by the covariates $\mathbf{Z}$, $\mathbb{S}_{\mathbf{Z}} \subset \mathbb{R}^{T\times A}$ (which is the support of the prior), the transformation is directly invertible without additional assumptions beyond that $\mathbf{Z}$ is of full rank: $\boldsymbol{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mu$.

Then we only need to specify priors in a manner that accurately represents expert knowledge about the behavior of μ, the expectation of log mortality. We do this in a set of L statements. Each statement, written as $H_l[\mu]$ for $l \in [1, ..., L]$, is a *smoothness functional* (a map from a set of functions to the set of real numbers). The prior distribution on the expected value of log mortality may then be expressed in terms of these L smoothness functionals:

$$\mathcal{P}(\mu|\theta) \propto \exp(-\frac{1}{2}\sum_{l} H_l[\mu, \theta]) \tag{5.4}$$

where θ_l is a non-negative weight placed on the on lth smoothness functional. In practice, much of expert knowledge about the behavior of log mortality may be translated into at least three types of smoothness functionals. First, smoothness over age stipulates adjacent age groups should have similar expected values of log mortality. Second, smoothness over time stipulates nearby time periods should have similar expected values of log mortality. Third, smoothness over age and time stipulates adjacent age groups should have similar time trends of the expected values of log mortality.

For example, consider smoothness over age groups. Intuitively, we want adjacent age groups to have similar expected values of log mortality, μ. How can we quantify smoothness? Many simple encodings of smoothness turn out to be implausible; for example, using the straightforward squared differences in adjacent age groups as the core of a smoothness functional would imply a random walk over age. A better option is to calculate the squared second derivative of expected log mortality:

$$\left(\frac{\partial^2 \mu(a,t)}{\partial a^2}\right)^2. \tag{5.5}$$

The second derivative measures the curvature of $\mu(a,t)$ with respect to age. If $\mu(a,t)$ is a linear function with respect to age, then the curvature will be 0. Any deviation from a linear function has a larger squared second derivative, and thus is inherently less smooth. However, as we saw in Figure 5.2, certain non-smoothness is inherent in the observed data. For example, log mortality in age group 0–4 is usually very different from log mortality for ages 5–9. Similarly, the mortality bump in late adolescence and early adulthood is a ubiquitous non-smooth feature of the age profile. To better account for these empirical regularities in demographic data, it makes sense to measure smoothness in terms of deviations from a "typical" age profile:

$$\left(\frac{\partial^2 \mu(a,t) - \bar{\mu}(a)}{\partial a^2}\right)^2, \tag{5.6}$$

where $\bar{\mu}(a)$ is a typical age profile, such as calculated from the average over real data. (Using the average enables us to profess ignorance over the level of the data, formally using an improper prior, putting the mean in the null space, about which more below.) Now any linear deviation from the typical age profile is considered smooth. The greater the curvature of the deviation from the typical age profile, the less smooth the behavior of the dependent variable.

To obtain a single value of smoothness across all the ages and times in a forecast, we take the expected value of the squared second derivative with respect to age and time:

$$H_{\text{age}}[\mu, \theta] = \theta_{\text{age}} \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t) - \bar{\mu}(a)}{\partial a^2} \right)^2, \tag{5.7}$$

where $dw^{\text{time}}(t)$ and $dw^{\text{age}}(t)$ represent weights that can be used to assign greater importance to smoothness in certain ages and times than others. Then, we add a positive parameter θ_{age} that controls how influential this functional is vis-à-vis other functionals and the data. Finally, we encode the expert qualitative knowledge that the expected value of log mortality in adjacent age groups should be similar by merely specifying that $H_{age}[\mu, \theta]$ should be small.

Since the second derivative is indifferent between linear functions, our priors do not impose any specific linear relationship between age and the expected value of the dependent variable. To be specific, the null space of the prior is composed of all linear functions. Following this intuition, we can then write down the smoothness functionals for all three types of smoothness:

- Smoothness over age

$$H_{\text{age}}[\mu, \theta] = \theta_{\text{age}} \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t) - \bar{\mu}(a)}{\partial a^2} \right)^2, \tag{5.8}$$

- Smoothness over time

$$H_{\text{time}}[\mu, \theta] = \theta_{\text{time}} \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t)}{\partial t^2} \right)^2, \tag{5.9}$$

- Smoothness over age and time

$$H_{\text{age/time}}[\mu, \theta] = \theta_{\text{age/time}} \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{\partial^3 \mu(a,t)}{\partial a \partial t^2} \right)^2. \tag{5.10}$$

Our complete prior is then:

$$\mathcal{P}(\mu|\theta) \propto \exp(-\frac{1}{2}\{H_{\text{age}}[\mu,\theta] + H_{\text{time}}[\mu,\theta] + H_{\text{age/time}}[\mu,\theta]\}) = \exp(-\frac{1}{2}H[\mu,\theta_{\text{age}},\theta_{\text{time}},\theta_{\text{age/time}}]) \tag{5.11}$$

when μ lies in $\mathbb{S}_{\mathbf{Z}}$ and 0 otherwise. That is, we restrict the prior to only expected values of log mortality that may be explained using the covariates **Z**. If we restrict the prior on μ to just the subspace that spans the covariates, then we are able to express β in terms of μ:

$$\beta = (Z'Z)^{-1}Z'\mu. \tag{5.12}$$

Consequently, the prior on the expected value of log mortality can be translated into a prior on the regression coefficients (and where we use the superscript H^{μ}, following Girosi and King (2008, p.68), to emphasize that the density is derived using knowledge of μ):

$$\mathcal{P}(\beta|\theta) \propto \exp(-\frac{1}{2}H[Z\beta,\theta_{\text{age}},\theta_{\text{time}},\theta_{\text{age/time}}]) = \exp(-\frac{1}{2}H^{\mu}[\beta,\theta_{\text{age}},\theta_{\text{time}},\theta_{\text{age/time}}]) \tag{5.13}$$

The resulting prior is improper since, as $\theta \to \infty$, the prior collapses to a projection in the null space, not a single point. In this null space, the forecasts are entirely dependent on the likelihood and thus the data.

Inherent in this procedure is a tradeoff between smoothness and predictive accuracy. This tradeoff may be controlled by setting θ_{age}, θ_{time}, and $\theta_{\text{age/time}}$. For examples of forecasts from this type of model, see Girosi and King (2008), King and Soneji (2011), and Soneji and King (2012).

Model-based uncertainty estimates and credible confidence intervals for this forecasting procedure may be obtained using Gibbs sampling, whereby we sample from the posterior of β and calculate the forecast log mortality for each iteration. In order to account for model dependence — likely the greatest source of uncertainty — we can take a robust Bayesian approach by supplying a range of prior values where θ must lie (King and Zeng, 2004). And in order to more formally extract information about the hyperprior parameters from experts, we could use more formal methods for Bayesian elicitation.

If OCACT employed this statistical forecasting model, rather than its ad hoc qualitative forecasting model, it could encode its expert judgement in the selection of covariates, lag specification between covariates and mortality, and the choice of Bayesian priors. In doing so, OCACT would eliminate the informal procedures that increase vulnerability to bias and produce a transparent and scientific model that may be improved upon.

5.5 Concluding Remarks

In this study, we offer a possible explanation for the systematic biases we found in Social Security's demographic and financial forecasts (in Chapter 4). The possibility of bias arises because of the lack of professionally designed formal procedures in place to avoid them. Unnecessary informality and ad hoc qualitative forecasting approaches, lack of up-to-date statistical methods that could automate decisions considerably too difficult to manually make individually or collectively, and the absence of transparency (e.g., publicly available replication information) open up SSA to the possibility of bias. OCACT adds to the problems by insisting on the consistency of forecasts over time, in order to remain at the center of the policy debate, even when contradicted by strong trends in the data and SSA's Technical Panel experts.

Trying to resist the continuing intense political pressure is just what Americans would want of their government officials. Yet, the difficulty of the task and their suboptimal procedures caused the actuaries to hunker down and resist all dynamics that might lead them to modify their forecasts including genuine changes in patterns in demography, public health, risk factors, or medical technology. These dynamics combined with the perfect storm of political pressure, suboptimal forecasting procedures open to bias, and changes in the world (e.g., faster increases in life expectancy). Thus, as SSA continued to resist modifying their forecasts, doing so led to a much higher probability of bias.

Our study also identifies steps SSA can take to reduce the biases in its forecasts. SSA would benefit from instituting formal procedures within OCACT to learn from its forecasting errors and to reduce its biases. Additionally, SSA should seriously engage the issues raised by its own scientific advisers. As the Technical Panels emphasize, open evaluation of past performance is the best way to guarantee that forecasters learn over time, which is why open, repeated evaluation is common throughout other parts of government, commerce, industry, and academia. At least from now on, every SSA Trustees Report should routinely provide a comprehensive evaluation of prior forecasts.

Furthermore, SSA should also use what it learns from each evaluation to refine subsequent forecasts. They should openly share their data and methods with the public so that members of the scientific community can easily replicate SSA's forecasts and contribute to their improvement. And, importantly, they should institute structural barriers to prevent inadvertent bias in the form of more formalized and transparent statistical procedures that are also less subject to manipulation and mental contamination. As it happens, these procedures are not only more replicable and easier to share with others; they also enable SSA to take advantage of the spectacular advances in statistics, data analytics, demography, and machine learning over the past several decades. Adopting these procedures will improve the quality of SSA's forecasts, correct the obvious biases we unearthed above, and help ensure open and fully informed democratic debate.

Fair, transparent, and accurate forecasts afford members of Congress the ability to consider alternative assumptions as they debate policy proposals to preserve the solvency of Social Security. SSA's failure to follow well developed best scientific practices represents a significant squandering of public resources and hampers meaningful progress on policy changes to Social Security. If OCACT relinquishes its monopoly position as the sole provider of both demographic and financial forecasts and fully reveals its forecasting procedures, it will advance its own objective of producing the best and least biased forecasts possible. Moreover, this reform may also improve forecasts of the Medicare Trust Funds, which rely on SSA OCACT demographic forecasts as an input.

Humanity has not yet found a better way to learn than the collective efforts of the scientific community pursing the same goals, most of which would come as free effort to OCACT, SSA, and the U.S. government. Regardless of the causes, however, fixing these problems is not only crucial for SSA. For the future of Social Security, and even for American democracy, the alternative assumptions of those debating policy proposals must be based on fair, open, and accurate forecasts.

6

Capturing Business Power Across the States with Text Reuse

Political scientists have long puzzled over questions of business power in capitalist democracies like the United States. Businesses, some scholars have speculated, enjoy access to political resources that most ordinary citizens – and other interest groups – lack, and therefore carry disproportionate influence over the policymaking process (Lindblom, 1977; Block, 1977; Mills, 2000; Domhoff, 2006; Hacker and Pierson, 2010). Other political scientists have been more skeptical, arguing in the pluralist tradition that citizens, consumers, and other organized groups can check the power of business (Dahl, 1961; Smith, 1999; Trumbull, 2012). Inquires into the power that business commands have become even more timely in the contemporary era of rising economic inequality, as many scholars, politicians, and citizens speculate that concentration in economic resources is caused – and reinforced - by the political power of wealthy individuals and firms. Yet despite the importance of these issues, empirical research on the interaction between businesses and government has run up against considerable obstacles. Most centrally, scholars have struggled to develop systematic methods of observing and classifying what businesses want from the policymaking process, and then whether or not businesses actually exercised influence over those decisions.

In this paper, we develop a new dataset for studying the influence of business on public policy decisions across the American states. Compiling and digitizing nearly 1,000 leaked legislative proposals made by a leading business lobbying group in the states – the American Legislative Exchange Council (or ALEC) – along with digitized versions of all state legislation introduced or enacted between 1991 and 2013 (for a total of nearly 2.4 million bills and resolutions), we use text analysis methods adopted from computer science to examine the degree to which actual state legislation matches the language and concepts in the business-drafted proposals. Specifically, we calculate a measure of similarity between each state bill ever introduced or enacted over this period and each corporate model bill. These results offer a clear picture of where and when large American firms have had the most sway over state legislative activity. To our knowledge, we are the first scholars to amass this collection of all introduced and enacted state bills, and to examine their substantive content in a systematic manner. Moreover, as far as we are aware, we are the first scholars to use this scale of text analysis to assess where businesses - or any other interest groups for that matter - have been influential in American politics.

Using our new collection of state legislation and the text reuse methods, we test a variety of explanations for when and where state legislators introduce and enact business-authored bills. We find that different factors explain the success of firm lobbying at different stages in the policymaking process. In general, however, we find the most consistent evidence in support of policy capacity-based arguments about business power. Where legislators had fewer policy resources (for instance, where legislators had fewer staff members to rely upon for research assistance), they were much more likely to rely on corporate-drafted proposals from ALEC, introducing and enacting more bills that were based in whole or in part on ALEC model bills. This provides an important extension and confirmation of earlier research (Hertel-Fernandez, 2014b) that ALEC's power comes not from its campaign contributions or electoral activities, but rather by providing policy resources to otherwise under-resourced state lawmakers.

Apart from studying the determinants of bill reuse across the states, we also take a preliminary look at how business-drafted legislation might change substantively relevant outcomes across the states. Given that ALEC's proposals favor the policy preferences of firms and their executives, we hypothesize that ALEC bills ought to reduce the tax burden faced by wealthy individuals, and will ultimately increase the levels of inequality in the states. We find evidence that states that enact more ALEC-influenced bills do in fact have systematically lower top income tax rates and greater shares of their income flowing to the top income deciles – even after accounting for the strength of states' labor movements and partisan control of government. This analysis provides a validation that our measures of business-backed bill activity capture important features of state legislation.

Together, our analysis provides a number of important contributions to the study of business power in American politics, American political economy, federalism, and interest group influence more generally. First and foremost, our paper provides one of the first large scale analyses of business influence in American politics at the individual bill level. While other studies have offered either important case studies or historical narrative analysis of business power (Hacker and Pierson, 2010; Vogel, 1989; Culpepper, 2010; Trumbull, 2012), we are able to systematically test the theories emerging from that work across all fifty states for two decades with rich quantitative data. And more generally, the methodology and data we propose here provide an approach that other scholars interested in studying interest group power in the US states – and in other governments – could fruitfully use to capture the influence of policy-drafting groups. While there have been several recent efforts to use text analysis as a mechanism to study power, so far its applications have been limited. Thus, we hope that this paper spurs other scholars to consider how our approach could help their own work on interest group influence and power. Ultimately, however, our findings offer striking conclusions about the distribution of political power in the United States: business interests have broad influence over major legislative activities, and that influence has increased in recent years. Moreover, that influence has reshaped the landscape of state policy, with deep implications for the distribution of economic resources across the states.

6.1 Our Approach to Measuring Business Power Across the States

To study business influence across the states, we focus closely on one empirically and theoretically interesting business lobbying group: the American Legislative Exchange Council. Our interest in ALEC stems from the group's scope and strategy, as well as for more practical reasons that we outline below. Formed in 1973 by conservative political leaders, ALEC is a non-profit organization of nearly 2,000 state legislators (or just under a third of all state lawmakers) and around 200 private sector members, which mostly include firms, but also conservative activists and policy experts (this summary draws from Hertel-Fernandez (2014a,b). ALEC "works to advance limited government, free markets and federalism at the state level through a nonpartisan public-private partnership of America's state legislators, members of the private sector and the general public", according to the group's website.[1]

ALEC's main activities revolve around the production and dissemination of "model bills" and policy ideas. These model bills are drafted by ALEC's task forces, which each focus on a specific policy domain. Over the years, these task forces have ranged across many diverse issues, such as agriculture

[1] See http://www.alec.org/about-alec/.

regulation, housing policy, health insurance reform, taxes, election rules, and gun sales and ownership. ALEC's current task forces include groups dedicated to civil justice; commerce, insurance, and economic development; communications and technology; education; energy, the environment, and agriculture; health and human services; international relations; criminal justice; and tax and fiscal policy. The task forces are comprised of ALEC members, including state legislators, private sector firms, and representatives from conservative think-tanks and other advocacy organizations. For instance, the current heads of the Energy, Environment, and Agricultural task force include state representative Thomas Lockhart (from Wyoming) and Paul Loeffelman, director of public policy for American Electric Power, a major electric utility company operating in eleven states.[2]

Other corporate participants in ALEC's task forces have included many of the largest, most prominent firms in corporate America, including FedEx, UPS, AT&T, Visa, Kraft Foods, McDonalds, Amazon, Google, State Farm Insurance, Koch Industries, and Facebook. Together, these firms provide most of the financial support for ALEC's budget of approximately seven billion dollars per year; other financial support has come from conservative philanthropies, such as those associated with the Bradley, Koch, and Coors families. Why do companies invest such substantial amounts of time and money in ALEC? According to past investigative journalism and research, the reason is that companies gain substantial latitude in crafting the model bills that the group produces and disseminates to its legislative members (Nichols, 2011; PFA, 2011; Hertel-Fernandez, 2014a). While ALEC has claimed that the groups' legislative members ultimately have the final say over those proposals, past research has shown that corporate members have expanded privileges within the policy task forces, and records show that proposals supported by legislative members – but opposed by corporate members – are sometimes tabled (McIntire, 2012).

Given the power of individual firms within the organization, it should come as no surprise that the

[2] See www.alec.org/task-forces/energy-environment-and-agriculture/.

model bills produced by ALEC have historically been closely aligned with the preferences of the firms leading each task force, and also with the priorities of the conservative movement more generally. Energy companies have drafted model legislation aimed at scuttling efforts to address climate change and to promote renewable energy (Malewitz, 2013; Kasper, 2013); private prison contractors have participated in task forces that produced legislation privatizing prison services (Elk and Sloan, 2011; Sullivan, 2010); tobacco companies have promoted bills to prevent regulation of tobacco use and sales (ALEC, 1986a); gun manufacturers and resellers have promoted bills to liberalize gun ownership and self-defense laws Graves (2012a); Weinstein (2012); internet retailers like Amazon sought exemptions from state sales taxes Fis (2012); and health providers have promoted bills to weaken medical liability protections (ALEC, 2011, 1986b). Other longstanding legislative priorities of the group include efforts to weaken labor unions, perhaps most notably public sector unions, to cut social benefits and labor market regulations (like paid sick leave, minimum wages, and unemployment insurance), and to lowe taxes on wealthy individuals and businesses (Lafer, 2013).

On the other side of the ledger, state legislators also have much to gain from membership in the group. They obtain access to a vast archive of policy ideas in the form of model legislation that is already written for them. But perhaps equally importantly, they also gain access to the research assistance and expertise of ALEC's policy analysts and affiliated experts. ALEC helps its legislative members to introduce its model bills, organize legislative briefings and hearings (including offering expert witnesses to testify on behalf those bills), and also generate talking points and political intelligence to help obtain the necessary votes to pass bills. All of this assistance is made more appealing through multiple events each year where legislators (and their families) can learn about the specific policy proposals, and also attend junkets sponsored by ALEC's corporate members (such as wine and cheese tastings with liquor producers, gun shooting outings with the National Rifle Association, and so on; Graves (2012b)). Legislators' travel expenses and registration fees for these meetings are often sponsored by corporate members through a dedicated "scholarship fund".

ALEC has operated without much attention for most of its history; one is hard-pressed to find much coverage of the group in national outlets before the mid-2000s.[3] Indeed, it has deliberately attempted to foster that low salience, for instance, by explicitly exempting itself from lobbying or oversight regulations in some states (Abowd, 2012). However, in very recent years, the group has attracted considerable media scrutiny and backlash from progressive groups in light of its ties to several controversial policy proposals. ALEC had been promoting so-called voter ID laws that attracted the ire of groups representing racial minorities in 2010 (Hoffman, 2012). But a confluence of events raised the profile of the group even more in subsequent years. A number of newly elected Republican majorities began to introduce and enact ALEC bills aimed at weakening labor power, perhaps most notably in Wisconsin (CMD, 2011; Bottari, 2012; Taylor, 2013). Shortly thereafter, ALEC was tied to the controversy surrounding the shooting death of a young Florida teenager, Trayvon Martin. Martin's shooter initially claimed protection under Florida's "Stand Your Ground" law, which permits the use of lethal force in self-defense (Barry et al., 2012). Journalists discovered that ALEC had been promoting that Florida law as a model proposal (through its public safety task force led by the NRA and gun retailers). Progressive activists were quick to launch protests against the group and its members, pressuring firms and legislators to sever their ties with the organization – which many now have (Pilkington and Goldenberg, 2013; Shiner, 2013). Indeed, our own analysis has been facilitated in part by this increased attention and scrutiny, as we rely heavily on leaked documents that would have been otherwise unavailable to the public before the group began attracting so much controversy.

ALEC's model bills form the foundation of our approach to measuring business influence. We ar-

[3] Consider the contrast in media coverage between ALEC and the National Conference of State Legislators (NCSL), a non-partisan group of state legislators. According to Lexis Nexis archives, on average, only about four newspaper articles were written each year about ALEC between the 1970s and 1990s, while about ten times as many were written about the NCSL, and this was despite the fact that ALEC was growing rapidly during this period. Indeed, in no year between 1976 and 2011 did ALEC receive more than half of the newspaper coverage that the NCSL garnered. In 2012, however, ALEC received more coverage than the NCSL, and 22 times the average coverage it had received across all prior years.

gue that several features about ALEC's model bills offer us an important opportunity to gain leverage on questions of business influence. First, as previous research and investigatory journalism has shown quite persuasively, individual firms have played a central role in setting the legislative priorities of the organization and developing the specific policy proposals advanced through the model bills. We thus argue that ALEC's model bills provide a good proxy for business's policy preferences (see also Hertel-Fernandez (2014b) for a lengthier discussion on this point). Moreover, because these model bills were drafted behind closed doors and were not intended to be released to the public, we expect that these proposals will more closely approximate business's true, rather than strategic, preferences (on this issue, see Broockman, 2012; Hacker and Pierson, 2002).[4] Our general strategy is to systematically compare every bill that state legislators have introduced and enacted to each of these ALEC model bills to establish whether or not lawmakers relied on the business-drafted bills from ALEC when crafting policy.

6.2 Data

Our data in this study is comprised of two novel corpora: a dataset of model bill legislation from ALEC and similar left-wing groups, and a corpus of all introduced and enacted state legislation between 1995 and 2013.

6.2.1 Model Bills

To gather ALEC's model bills, we relied on a variety of sources. We first compiled and digitized the model bills that had been leaked by the Center for Media and Democracy, a left-leaning watchdog group that tracks ALEC.[5] Over the course of several years, we also accessed a number of state legislative libraries, as well as archives at the Library of Congress, the University of California—San Francisco,

[4] Note, however, that after the controversy from the Martin shooting, ALEC has made a limited number of its most recent model bills available online.

[5] These are available online: www.alecexposed.org.

and the University of California—Berkeley, to compile and scan older ALEC model bills that were not included in the Center for Media and Democracy's archives.[6] Subsequently, we labeled each bill with one of 16 policy areas. In all, we digitized and compiled close to 1,000 unique model bills. To the best of our knowledge, this is the most complete set of ALEC's proposals that exists outside of the organization's own records.

Furthermore, we collected model legislation from two groups that were set up as progressive counterweights to ALEC. We scraped and digitized model a total of 312 bills from the Center for Policy Alternatives (CPA) and the American Legislative and Issue Campaign Exchange (ALICE). The model legislation is on many of the same topics as ALEC and thus represents a good control group that we can make use of in classification.

6.2.2 State Legislation

Secondly, we constructed a novel dataset of all American state legislation introduced and enacted from 1995 to 2013. Although we have data for some states prior to 1995, the data is incomplete. The total corpus contains about 2.4 million bills and resolutions. For each bill, we have metadata that includes the legislative session, version, and primary sponsor. We are working to expand the metadata to include specific dates, co-sponsors, and committees.

6.2.3 Features

We first preprocessed all text by casefolding, stemming, removing stopwords, and removing all punctuation (other than hyphens). Next, we tokenized the text into n-grams. We ultimately used $n = 2$ and $n = 3$, although we experimented with both higher values of n as well as unordered skip n-grams. For each state bill, we calculated a set of vector-based similarity metrics across all ALEC proposals, includ-

[6] At USCF, we relied on the Legacy Tobacco Archives. At Berkeley, we relied on the People for the American Way Collection of Conservative Political Ephemera, 1980-2004.

ing overlap, the Jaccard coefficient, and cosine similarity. For simplicity, we only kept the ALEC bill with the highest bigram overlap as a candidate source of the state bill text when a state bill was matched to more than one ALEC proposal. Moreover, we dropped bills that shared fewer than 10 bigrams as possible candidates. Lastly, in addition to the similarity metrics, we also stored overlapping bigrams and trigrams for each state bill.

Next, we estimated the topics associated with each bill by fitting a latent Dirichlet allocation (LDA) model with $K = 50$ topics on unigrams, and extracted the distribution over topics for each bill (Blei et al., 2003). We experimented with both $K = 20$ and $K = 100$ topics, but found that $K = 50$ provides qualitatively more interpretable results. We also calculated the cosine similarity of the topic distributions between each state bill and its ALEC matches, and used this similarity as an additional feature for classification.

6.3 Methods

The goal of this paper is to identify ALEC-derived legislation in our state bill corpus. We thus take a supervised learning approach that builds upon past work classifying documents based on patterns of text reuse (Clough et al., 2002). To construct our training set, we would ideally take a representative sample of bills from our state legislation dataset and manually label each bill as ALEC-derived or not. Unfortunately, not only is this is a prohibitively time-consuming task, but human coding by non-experts is unlikely to be particularly trustworthy given the difficulty of reading and identifying the meaning of legislation. Instead, we opted to rely upon the identification of ALEC-derived bills in several states in recent years by liberal non-profit groups. For examples of state legislation that is not ALEC-derived, we use the liberal model bills from CPA and ALICE as we can be certain that any similarity between them is due to procedural language or topical similarity, but not due to a shared policy idea. Ultimately, we collected 131 texts that have been identified as ALEC-derived and 273 texts that are not ALEC-derived.

It is important to note that the resulting training set is rather small, and that the ”control” group is not a random sample from the entire corpus. Out of the labeled bills, we randomly selected a holdout set of 31 ALEC-derived bills and 55 non-ALEC-derived bills (roughly one-fifth of the labeled set). We use the remaining bills to tune our classifiers using 5-fold cross validation and use the holdout set to evaluate the performance of the optimal classifiers.

In addition to a limited training set, examination of the ALEC-derived bills identified by watchdog groups revealed quite variable patterns of text reuse. While some bills are near-duplicates of ALEC proposals, others do not share even a single trigram. For example, Virginia's H.B. 1331 introduced in 2010 draws nearly verbatim from ALEC's Council on Efficient Government Act (see Figure C.1). One does not need to be a policy expert to identify this bill as ALEC-derived given how significantly the Virginia bill has draw from the ALEC model. However, consider Virginia's H.B. 2314 or S.B. 10, also introduced in 2010 (see Figures C.2 and C.3). In these cases, there is minimal overlap of text, but substantial reuse of the underlying concepts from the original ALEC bills. In short, the variation in reuse patterns across state bills is very high. Given this variability, we decided to subset the classification process into two parts:

1. Identification of all state bills that share the same intent as ALEC bills.

2. Identification of all state bills that have blatant or near-verbatim reuse of ALEC language.

For task (1), we used the training set previously described above. Given that task (1) subsumes task (2) and there is no existing definition of blatant ALEC reuse, we turn to work by left-leaning non-profits that have been tracking ALEC's activities across the states. Conveniently, groups in Arizona, Michigan, and Virginia have carefully matched a number of introduced and enacted state bills to ALEC model legislation, and have also separated out ALEC-derived bills with nearly identical language (which they termed bills with ”ALEC DNA”) from those that merely share the same intent. Thus, for task (2), we could restrict our training set to these three states, using the additional information from these non-profit groups to classify bills as having ALEC DNA instead of ALEC intent.

For task (1), our general approach is to train a support vector machine (SVM) in two-stages. In the first stage, we used bigram and trigram Jaccard coefficients, cosine similarity metrics, and the cosine similarity between topic distributions from the LDA as features to train a radial kernel SVM. We set $C = 100$ and $\gamma = 0.20$ based on 5-fold cross validation. The intuition behind these features is that bill-ALEC pairs with higher lexical and topical similarity are more likely to exhibit the same policy intent. However, note that certain n-grams are more indicative of an ALEC bill than others. For example, the bigrams "castle doctrine" or "duty retreat" are much more suggestive of ALEC influence than bigrams such as "enact section" or "president congress". In order to capture words that are disproportionately associated with ALEC-derived documents, we first took all bills classified as ALEC-derived with a probability of 0.99 or higher in the first stage and added them to our labeled set. Then, we ran multinomial inverse regressions (MNIRs) (Taddy, 2013) separately by each ALEC policy domain on both the labeled and unlabeled bills, using an indicator for ALEC-derived bills in our labeled set as a covariate. We then calculated the sufficient reduction (SR) projections for each state bill and used those projections as additional features in the second-stage classification. The SR projections function as a low dimensional alternative to using the overlapping bigrams and trigrams as features directly. In our experience, such dimension reduction identified sensible bigrams and trigrams as being associated with ALEC, and also improved performance of our classifier. We then used this second stage SVM to classify bills as having the same intent as model ALEC legislation or not.

For task (2), we took all bills labeled as ALEC-derived in task (1). We then trained another SVM using this new training set with lexical and topical similarity metrics as features in order to separate ALEC intent bills into those with near-verbatim traces of ALEC language. Given the extremely small training set, we took an active learning approach to increase the number of labeled bills, sampling 10 of the bills classified with the most uncertainty, manually labeling them, and retraining the classifier on the expanded labeled set. We repeated this process 5 times.

6.3.1 Classification Performance

We evaluate performance of our SVM classifiers using the holdout set. For the task of identifying bills with ALEC intent, we correctly identify 25 of the 31 ALEC-derived bills (recall rate of 80.7%). We only incorrectly identify 2 out of the 55 non-ALEC-derived bills as sharing ALEC intent (precision rate of 96.4%). Given the non-representative nature of the labeled non-ALEC-derived bills, we cannot be sure that we will achieve a similarly low false positive rate on the entire corpus. To evaluate this, we randomly sampled 50 bills identified as sharing ALEC intent and coded them as sharing ALEC intent or not. Out of these bills, manual examination revealed that 28 were bills that shared ALEC intent (precision of only 56%). The low precision rate is likely due to difficulty this method has in identifying the same intent as ALEC. Manual examination of the bills classified as sharing ALEC intent repeatedly revealed bills that addressed charter schools or prisons - generally issues ALEC focuses on - but that did not fully share the exact policy intent as ALEC.

For the ALEC DNA classifier, we had an extremely small holdout set. We correctly classified all 5 ALEC DNA bills as having ALEC DNA. However, a 100% recall rate would be highly unrealistic had we recourse to a much larger test set. We then again sampled 50 random bills identified as sharing ALEC DNA and manually evaluated them. 42 out of the 50 bills had varying traces of ALEC language and were correctly coded (precision rate of 84%). Ultimately, given the much lower false positive rate for the ALEC DNA task versus the ALEC intent task, the results we present below are based on using bills classified as having ALEC DNA.

6.3.2 Related Work

The methods in this paper relate to several important pieces of scholarship on the detection of text reuse in computer science. Most approaches to the identification of text reuse have involved one or more of the following methods: vector space similarity metrics, string comparison, and sentence alignment Clough and Gaizauskas (2009). Clough et al. (2002), for example, examine text reuse from newswires in newspaper articles, testing the performance of three different metrics for text reuse detection: n-gram overlap, Greedy String Tiling, and sentence alignment. Training a Naive Bayes classifier using 10-fold cross validation on their corpus, they find that using all available features provides the most precise and reliable categorization of text into wholly derived, partially derived, and non-derived categories.

More recent approaches to plagiarism detection have attempted to identify text reuse in the context of "obfuscated" or rewritten text. Nawab et al. (2010), for example, use modified *n*-grams that accommodate deletions of single words or synonym substitutions. An overview of the approaches towards detecting paraphrased text in the Second International Competition on Plagiarism Detection and their respective efficacy is presented in Barrón-Cedeño et al. (2013). In another attempt to address slightly obfuscated text, Smith et al. (2013) use non-contiguous sequences of words, or skip n-grams, to deal with OCR errors and slight textual variation when detecting text reuse in 19th century newspapers. Our work builds upon these methods to examine text reuse in the legislative context - a challenging domain due to often-shared procedural language that is not indicative of more substantive text reuse. We thus depart from these other methods in using the content of shared language as additional features in our classifier.

In political science, text reuse has only recently been applied to the study of legislative activity. In order to trace the origins of final legislation to various proposed amendments, Huberty (2013) uses vector space distance metrics such as cosine similarity followed by nearest neighbor matching with replacement to identify coderivative pairs of amendment and bill sections. These matching texts can then be visualized or analyzed further via topic models to determine whether content on certain topics, or backed by certain legislators, is more likely to make its way into final legislation. Whereas Huberty (2013) seeks to trace how amendments make their way into a final bill, Wilkerson et al. (2015) are interested by the reuse of text across different bills in the 111th Congress. To find similar passages, they first restrict their analysis to only pairs of bill sections that share at least 5 overlapping word n-grams. Subsequently, they utilize the Smith-Waterman local alignment algorithm to identify the text two sections share in common - sections that have near-identical sequences of text receive higher scores. Finally, to deal with "false positives" due to shared procedural language, the authors train an SVM on a human labeled subset of bill alignments in order to predict substantive matches.

The goal of this paper differs from the preceding research in political science in several important ways. First, as opposed to Huberty (2013), we are interested in detecting similarity between bills and model legislation where there is no guaranteed match between a bill and ALEC model legislation, meaning that the use of nearest neighbor matching is not applicable. Wilkerson et al. (2015) seems closer to our proposed goal. However, the authors are interested in identifying long, nearly verbatim passages that are coderivative between bills. While this is a useful starting point for an analysis, detection of ALEC-derived legislation requires methods that are more robust to paraphrasing, reordering, and general obfuscation of reused text. Our paper attempts to identify coderived text beyond near-verbatim copying.

6.4 Mapping Business Influence Across the States

In this section, we present an initial summary of our measures of ALEC bill reuse across the states throughout the 1990s to 2013. One question is how ALEC bill activity has changed over time. To answer this question, we plotted the total number of introduced and enacted bills matching ALEC text by year, as a share of all introduced and enacted state bills; the results appear in Figure 6.1. There are several insights we can glean from this figure. First, state legislators are introducing and enacting more ALEC-derived legislation over time. In 1994, legislators turned to ALEC models for about 0.5 percent of all introduced bills; this reached a peak of 1.3 percent in 2010. Enacted ALEC-derived bills are generally more stable throughout the 1990s, but exhibit a sharp increase in the mid-2000s, reaching a peak of 1.1 percent in 2010. A second suggestive pattern from this figure is a decline in both introduction and enactments of ALEC-derived bills since 2011, consistent with the public backlash that ALEC has experienced in recent years.

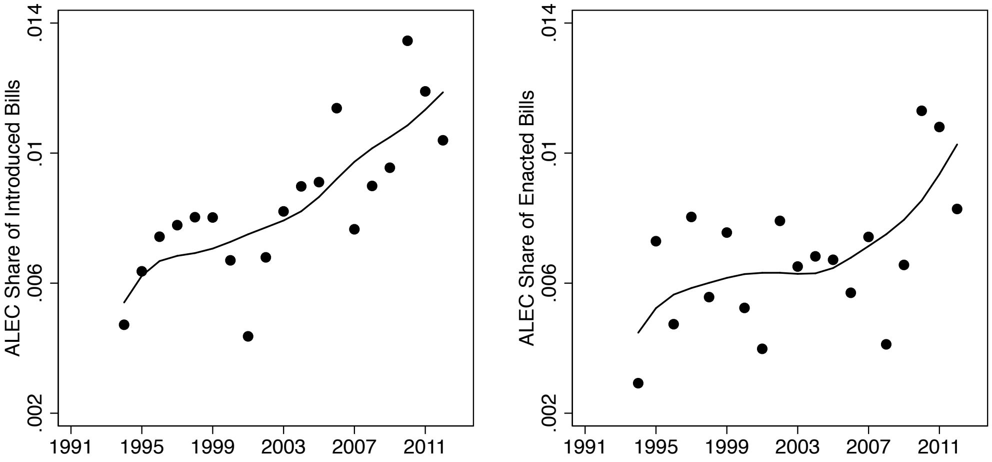


Figure 6.1: Enacted and introduced ALEC bill shares over time.

Another important question is how ALEC bill reuse varies across the states. Our data suggests that there are a number of states that have consistently relied on ALEC for legislation. Figures 6.2 and 6.3 plot these states, separating bill introductions from enactments, respectively. As the maps indicate, West Virginia, Arizona, Missouri, and Pennsylvania introduced the greatest number of bills that were derived from ALEC proposals, while Connecticut, Ohio, Minnesota, North Dakota, and Oklahoma introduced the fewest number of bills based on ALEC policy models. Turning to bill enactments, we see that Utah, Arizona, Idaho, and Oklahoma enacted the most number of bills drawing from ALEC models, while states like Maine, Massachusetts, Nebraska, and Connecticut enacted the fewest number of ALEC-derived bills. See also Figure 6.4 for a heatmap of bill introductions and enactments across the states and across years.

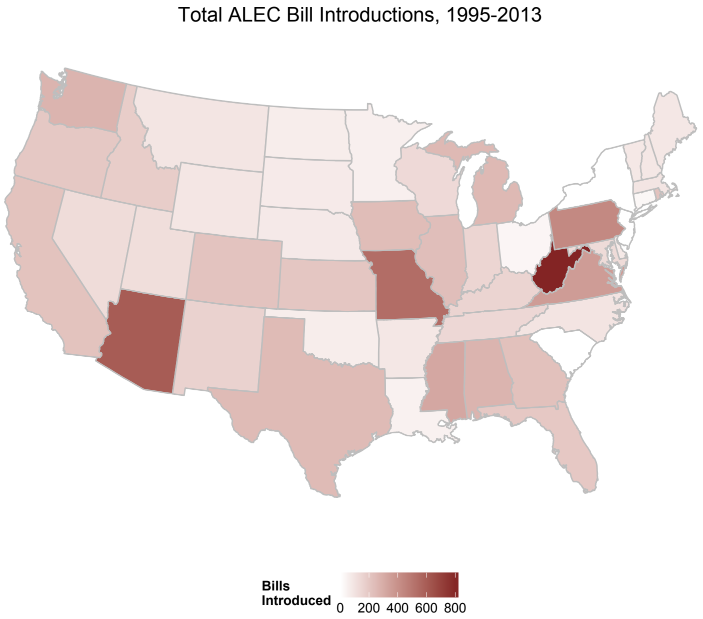


Figure 6.2: Total number of ALEC introduced bills from 1995–2013.

Of course, part of these differences across states might be due to the fact that some states simply consider and enact more bills each year as opposed to other states. To account for these underlying differences in legislative productivity, it is useful to consider ALEC-derived bill introductions and enactments as a share of all bills enacted and adopted by states each year. Examining this measure, we see that Kansas, Arizona, Missouri, Wyoming and West Virginia had the highest proportion of introduced bills that were derived from ALEC models, while Connecticut, Massachusetts, Louisiana, and Minnesota had the lowest introduction rates. Oklahoma, Kansas, Alaska, Wisconsin, and Arizona had the highest enactment rates, while Maine, Massachusetts, Connecticut, and Rhode Island had the lowest enactment rates. In general there was not much of a difference when we compared bill counts or

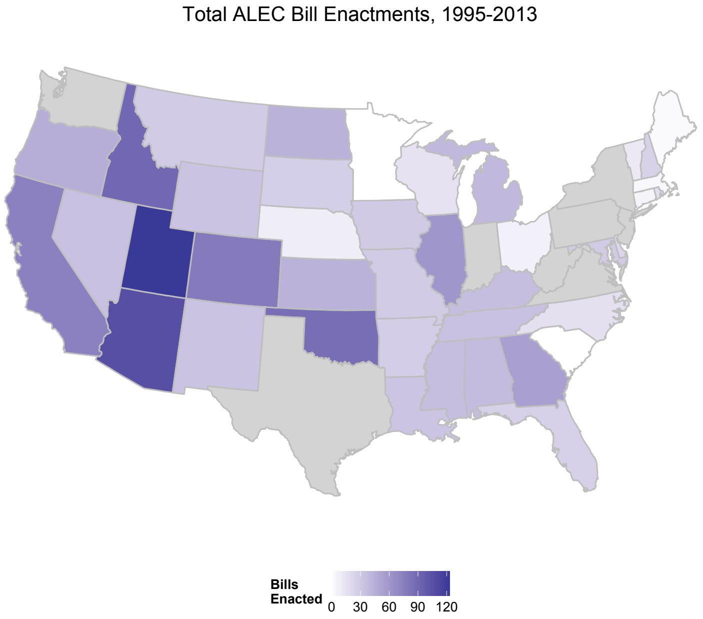


Figure 6.3: Total number of enacted ALEC bills from 1995–2013. We do not currently have enactment data on the grayed out states.

shares.[7]

A third interesting dimension of variation is by policy domain. These counts appear in Table 6.1. The vast majority of ALEC-derived bills (nearly half) were concentrated in just two policy domains: education and health care, followed by a distant third and fourth place for agriculture, energy, and the environment, and budget and tax policy. We see a similar story for enacted bills as well; the ranking of the top three policy domains is unchanged. Legislators were most likely to draw from ALEC's business-drafted education and health care bills. The heavy reliance on these ALEC models likely reflects the fact that education and health care policies dominate state budgets. For instance, in fiscal year 2012, states spent 25 percent of their budgets on K-12 education, 16 percent on Medicaid (state health insurance for

[7] Introduced and enacted shares and counts were highly correlated. The correlation coefficient for introduced counts and shares was 0.64 and the coefficient for enacted counts and shares was 0.69.

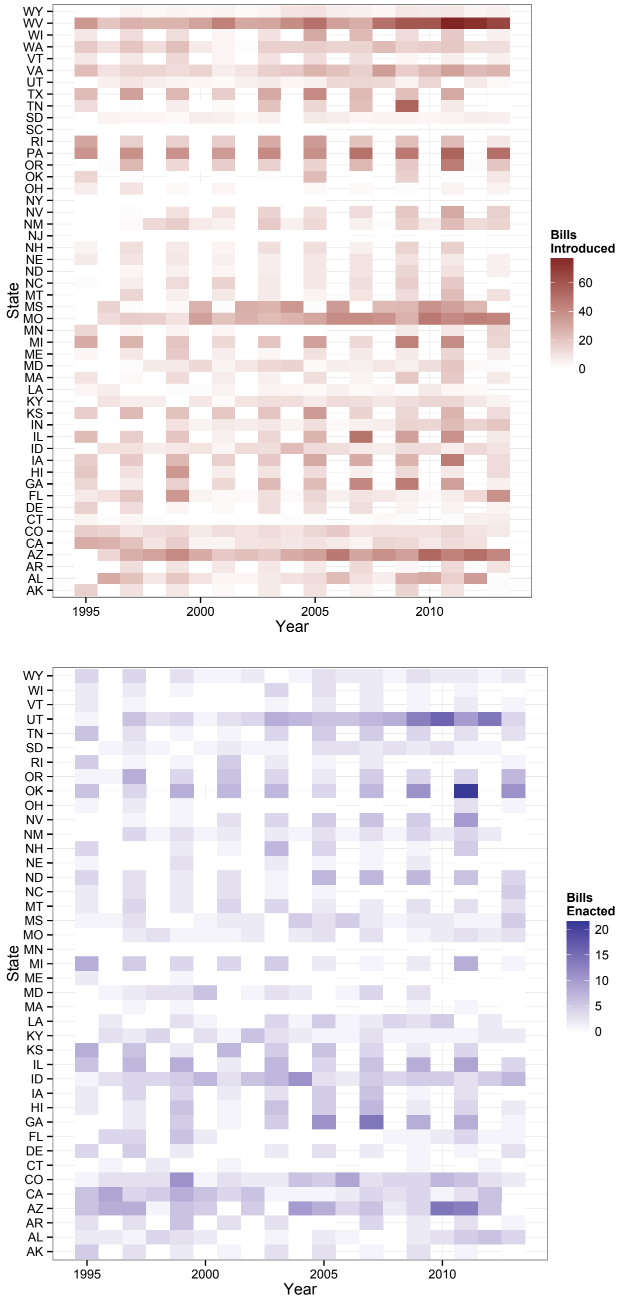


Figure 6.4: Heat maps of introduced and enacted bills. Note that we do not currently have enactment data for 10 states.

Table 6.1: Introduced and enacted bills by policy domain.

Policy Domain	*Introductions*	*Enactments*
Voting and Elections	46	5
Labor Unions	373	15
Housing	152	25
General Regulation	393	26
Guns	322	30
Social Welfare and Benefits	245	39
Transportation	252	42
Foreign Policy	88	52
Civil Justice	679	58
Criminal Justice	557	64
Budget and Taxes	710	66
Government Reform	686	71
Finance	229	72
Agriculture, Energy, and the Environment	734	167
Education	2065	414
Health Care	2839	427

the poor), and 13 percent on higher education (CBPP, 2014).

Which specific model bills from ALEC were most popular amongst state legislators? The top three proposals that legislators drew from for introduced bills were ALEC's comprehensive medical liability reform model, an education reform package, and a resolution on the importance of state sovereignty from the federal government. For enactments, the most popular model bills were the education reform package, medical liability reform proposals, and bills related to long-term care insurance. The comprehensive medical liability reform package, for instance, is intended "to address the rising cost of medical malpractice insurance that is imposing serious problems for [insert state name here]'s health care system". It includes a variety of provisions intended to reduce the ability of patients to sue their health care providers for health care-related injuries, such as limiting the monetary awards offered to plaintiffs.

6.5 Testing Theories of Business Influence in the Policy Process

What explains the patterns of business influence across the states and years that we have identified in the previous section? To answer that question, we test six sets of explanatory factors related to economic conditions; power resources and partisan ideology; policy capacity; campaign contributions; news media coverage; and venue shopping.

6.5.1 Economic conditions

Past research on the political influence of business interests has suggested that economic conditions matter greatly for the ability of business to intervene in the policymaking process. Early work on the structural "privileged" position of business suggested that since politicians are electorally motivated to maintain a healthy economy (e.g. Block, 1977; Lindblom, 1977; Mitchell, 1997), and since businesses are key to employment and economic growth, businesses should have the most sway over the policymaking process during periods of slow growth or downturns (see the review in Smith, 1999, pp. 843-6). Subsequent research, perhaps most prominently by David Vogel, has confirmed this intuition. Using a set of case studies from the postwar period, Vogel argued that "business has tended to lose political influence when the economy was performing relatively well and has become more influential when the performance of the economy deteriorated" (Vogel, 1989, pp. 8-9). According to this line of argument, ALEC should be most effective when states are in periods of high unemployment, which we operationalize with state unemployment rates from the Bureau of Labor Statistics.

6.5.2 Power resources and legislative ideology

The traditional power resources account views politics as a struggle between different economic classes. Consistent with this perspective, there are a number of robust correlations between union density, liberal control of government, inequality, redistribution, and social policy generosity across American states (Kelly and Witko, 2012) and the advanced economies (Bradley et al., 2003, p. 64). Writing in a similar vein, Jacob Hacker and Paul Pierson relate the rise in business power in the United States during the post-war era to the decline in countervailing union power (Hacker and Pierson, 2010). We measure several different aspects of power resource theory, including the power of organized labor (using data from UnionStats), the interest group most likely to check the power of business, partisan control of state governments (using data from Klarner, 2013), and the political ideology of state lawmakers (using data from Shor and McCarty, 2011).

6.5.3 Policy capacity

State governments vary dramatically in the level of resources offered to their elected officials to make policy – what is often described as "legislative professionalism" (e.g. Squire, 2007). In recent years, for instance, 16 states had legislators that only spent about half of their time working in the legislature, and were paid barely $20,000 per year. Moreover, 18 states only gave their legislators three or fewer full-time staffers, and, quite strikingly, nine states gave their members two or fewer full-time aides.[8] Previous work has argued that ALEC has exploited this weak policy capacity across the states strategically, focusing on providing model bills, research assistance, and political consulting precisely to those inexperienced legislators and lawmakers in states with weak legislative professionalism (Hertel-Fernandez, 2014b). In the absence of groups offering similar benefits as ALEC, weak state policy capacity, then, ought to increase legislators' demands for ALEC's services. To test this theory, we create an annual measure of legislative professionalism based on a standardized index of the time spent by legislators in session each year, legislator salaries (if offered by states), and spending on the legislature (we use

[8] See the National Conference of State Legislatures: http://www.ncsl.org/research/about-state-legislatures/full-and-part-time-legislatures.aspx and http://www.ncsl.org/research/about-state-legislatures/staff-change-chart-1979-1988-1996-2003-2009.aspx.

data from the Council of State Governments' Book of the States dataset).[9] We scale this index so that it ranges from zero to one.

6.5.4 Campaign contributions

Corporate-affiliated contributions dominate donations from unions by nearly three to one in PAC spending, and nearly 17 to one when looking at so-called "soft money" spending.[10] Yet despite the preponderance of corporate money in national and state campaigns, scholars have struggled to identify a clear effect of corporate giving on policy outcomes. Ansolabehere et al. (2003), for instance, have argued that contributions have only a weak relationship with legislator behavior, and argue that such electoral involvement should be a seen as a consumption good for firms. On the other hand, other scholars, citizens, and activists vehemently disagree with this conclusion, arguing that greater campaign contributions lead to changes in legislators' votes aligned with the preferences of the campaign donor (e.g. Lessig, 2011; Ferguson, 1995; Powell, 2012).

There are at least two explanations for these muddled findings on the importance of campaign contributions. One idea, proposed by Hall and Deardorff (2006), is that contributions are not intended to persuade legislators to change their preferences; rather, contributions should be seen as a signal of an interest groups' alignment with the legislator, and a promise of assistance with crafting and passing a bill. Another argument is that these analyses ignore the substantive content of legislation that is shaped well before final votes are taken, as well as neglecting other mechanisms for influence (see e.g. Powell, 2012). Consistent with this second critique, Lynda Powell finds that there is substantial variation across the states in the degree to which legislators report that contributions matter for the content and passage of bills.

[9] This measure is correlated with Squire's index of legislative professionalism at 0.90 for the years in which Squire's year overlaps with our measure.

[10] See https://www.opensecrets.org/overview/blio.php.

To the extent that contributions do in fact reflect a mechanism for influence, we should observe that states where business dominates political giving are more likely to introduce and pass ALEC-derived bills. We measure corporate giving by computing the ratio of total contributions from business associations to candidates for state legislatures to total contributions from labor unions to such candidates, using data from the Institute on Money in State Politics

6.5.5 News media coverage

In past work on ALEC, one of us has argued that a crucial strategy of the group is its relatively low profile and lack of public attention (Hertel-Fernandez, 2014a). Such low salience has prevented backlash from the public and progressive adversaries. These strategies fit well with prior work on business power, which has emphasized the advantages that businesses can command when they operate in domains of low salience with little media coverage (Culpepper, 2010; Layzer, 2012). To test salience-related theories of business power, we include a measure of the prevalence of the news media in each state and year: the share of workers employed as reporters or editors at news outlets using data from the Current Population Survey. Although crude, this measure has the advantage of being readily available and comparable for all of the years and states in our sample, which is not true of other measures, such as circulation statistics.

6.5.6 Venue shopping

Lastly, we examine the interplay between federal and state politics by considering theories of forum or venue "shopping". Given a high degree of decentralization and fragmentation in American politics, the policymaking authority for any given issue often can fall to different institutions. As a result, Frank Baumgartner and Bryan Jones have argued that interest groups seeking policy change will find the institutional setting that would be most likely to give the group a favorable hearing (Baumgartner and

Jones, 1993). The implication of this theory for our analysis is that businesses associated with ALEC will be more active at the state level during periods when the federal government is less amenable to conservative business interests (i.e., the federal government is controlled by Democrats). To account for this explanation, we include a dummy variable indicating if Democrats are in control of national government in a given year.

6.5.7 Methodological Approach

Crucially, we consider the importance of all of these factors at different stages in the policymaking process. While scholars of public policy have long distinguished between these stages, the study of business power has generally not explicitly tested to see whether the factors that shape firm influence are different when a legislative proposal is originally made, when it is deliberated by the legislative body, and when it is up for a final vote (but see Hall and Wayman, 1990). Specifically, we examine ALEC's influence on two outcomes: the number of introduced bills that correspond to ALEC proposals and the number of enacted bills that correspond to ALEC proposals. We adopt a relatively straightforward empirical setup, estimating negative binomial models.[11] We include state fixed effects in all specifications to account for state-specific, time-invariant factors that could confound our analysis, and include year fixed effects in some specifications to account for common time shocks.

6.6 Main Results

Model 1 in Table 6.2 presents results exploring the determinants of introduced bills across the states. Only power resource theories receive any support in this model. Union density is negatively related to the number of business-backed bills that a state enacted. Moving from the 5th to the 95th percentile of union density (an increase in union membership from five to 23 percent of employed wage and salary

[11] Poisson models produce similar results.

Table 6.2: Baseline models of ALEC bill introductions and enactments.

	Introduced Bills	*Enacted Bills*	*Introduced Bills*	*Enacted Bills*
	Model 1	Model 2	Model 3	Model 4
Union Density	-0.12**	-0.08	-0.02	-0.03
	(0.05)	(0.06)	(0.05)	(0.05)
Democratic Governor	0.19	0.01	0.12	-0.04
	(0.13)	(0.13)	(0.10)	(0.15)
Democratic Legislature	0.23	0.14	0.12	0.15
	(0.17)	(0.19)	(0.20)	(0.19)
Unemployment Rate	0.07	-0.02	0.13*	0.03
	(0.06)	(0.06)	(0.07)	(0.07)
Republican Conservatism	0.99	0.90	0.27	0.24
	(0.90)	(0.83)	(1.00)	(0.79)
Democratic Conservatism	0.01	-0.28	1.08*	0.23
	(0.46)	(0.45)	(0.56)	(0.52)
Legislative Professionalism	-0.83	-2.55**	-2.21*	-3.17***
	(1.22)	(1.02)	(1.19)	(1.08)
Media Density	-2.88	-0.68	1.83	1.89
	(2.11)	(2.96)	(1.76)	(2.94)
Democratic National Government	-0.01	0.17		
	(0.22)	(0.28)		
State Effects	YES	YES	YES	YES
Year Effects	NO	NO	YES	YES
N	384	328	384	328

Negative binomial models; robust errors clustered by state.
Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

workers) is predicted to reduce the number of business-drafted bills a state introduced from 24 bills to three bills.[12] To put these numbers in context, consider that the average state introduced ten business-authored bills (the standard deviation was 24 bills). No other explanatory factors are related to bill introductions at conventional levels of significance in this specification.

Model 2 turns to bill enactments. In this model, only legislative professionalism has any effect on business bill passage at conventional levels of statistical significance. Greater policy capacity, in the

[12] Difference significant at p=0.10. Holding all other variables constant at their observed values in this and all other predictions.

form of access to more staffers, more compensation, and longer sessions to deliberate over legislation leads legislators to enact to fewer business-drafted bills, consistent with previous analysis of ALEC and its lobbying efforts across the states. At the 95th percentile of legislative professionalism, a state is predicted to enact only one business bill, but is predicted to enact four at the 5th percentile of legislative professionalism.[13] Those are clearly sizeable effects given that the average number of enacted ALEC-derived bills was two (the standard deviation was also two).

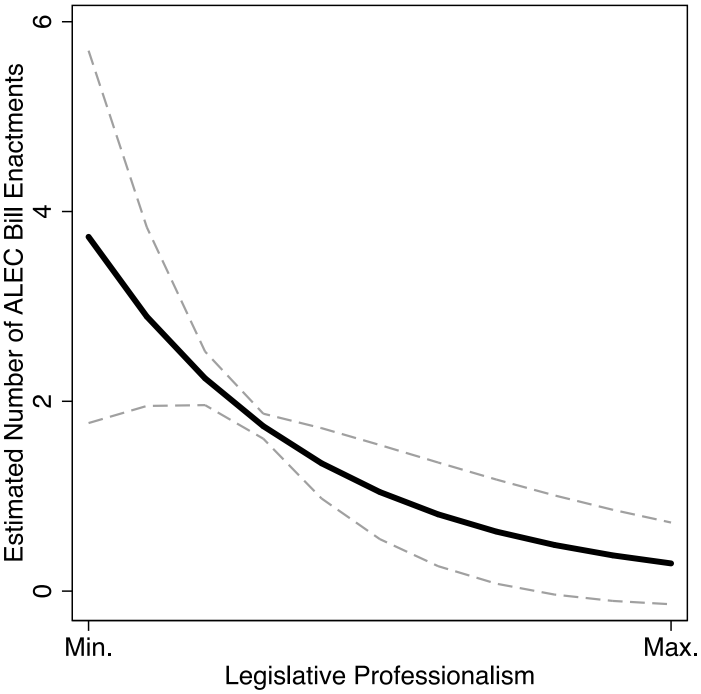


Figure 6.5: The estimated effect of legislative professionalism on ALEC bill enactments. Dashed lines indicate 95% confidence intervals. All other variables held at their observed values.

Thus far we have not accounted for confounding temporal effects that could affect all states equally aside from partisan control of national government. What happens when we include year fixed effects in our specifications to account for common shocks to all states? We show these results in Models 3

[13] Difference significant at p=0.04.

and 4 of Table 6.2.[14] In the year fixed effect models, we see a slightly different story for bill introductions. Union membership is no longer statistically significant at conventional levels of significance, likely because of the strong common trends in density across all states.

Three other factors, however, emerge as predictors of bill introductions: the unemployment rate, the ideology of the Democratic caucus, and legislative professionalism. Consistent with the hypothesis that economic downturns ought to empower business, we see that states with worse labor markets were more likely to introduce business-drafted bills. A move from the 5th to the 95th percentile of unemployment (or from three to seven percent unemployment) is predicted to increase the introduction of corporate-authored bills from eight to 14.[15] In a partial validation of power resource hypotheses, we see that more conservative Democratic caucuses were more likely to introduce ALEC model legislation, and the effect is quite large, though generally not significant at conventional levels of significance. Lastly, we see that in these year fixed effect models legislative professionalism emerges as a powerful predictor of bill introduction: a move from the 5th to the 95th percentile of policy capacity is predicted to increase bill introductions by about eleven bills.[16]

Unlike with the bill introduction models, the bill enactment models change little once we add year effects. Legislative professionalism continues to offer the most compelling explanation for how many ALEC-authored bills states enact each year. The effect here of a move in professionalism from the 5th to the 95th percentile is a decrease of about three bills, about the same effect size as in the models without fixed effects.[17]

We find no systematic evidence that states where businesses offer more electoral contributions in-

[14] We exclude the indicator for partisan control of national government in these specifications.

[15] Difference significant at *p*=0.08.

[16] Difference significant at p=0.10.

[17] Difference significant at p=0.05.

troduce or enact more ALEC bills.[18] Indeed, if anything states where businesses offered greater contributions relative to labor unions were less likely to enact ALEC bills (see results in Table 6.3). While this finding may appear strange on its face, it makes sense in light of earlier work examining the role of campaign contributions in state politics. Specifically, Powell (2012) has found that campaign contributions were more influential in states with stronger legislative professionalization; thus business may be giving more to electoral campaigns in states with greater policy capacity – precisely the state legislatures where ALEC is least active.

We are still assessing how ALEC's political logic varies across different policy issue domains, but there are some initially suggestive and interesting findings. For instance, while the density of news media coverage has no average effect on the enactment of business-backed legislation, we do find a strong and statistically significant effect of media density on the enactment of business-backed bills related to the reform of government processes and agencies.[19] These ALEC proposals tend to be the most technical – though they still have important consequences for state government. For instance, many of these bills involve privatization of state agencies or services, or cutting benefits and wages for public sector workers. A stronger news media presence in a state, then, may help to cover these otherwise obscure issues, raising their salience to citizens and mobilizing backlash to such proposals.

6.7 Does ALEC Influence Translate into Substantive Outcomes?

So far we have explored the determinants of ALEC legislative victories. But does the legislation that we track in our corpus of state bills translate into substantive state outcomes? While it is beyond the scope of this paper to fully examine the implications of ALEC legislation on state social, political, and

[18] Note that we do not include labor union density in these estimations since our contribution variable - the ratio of business to labor donations - already includes a measure of (relative) union strength.

[19] Moving from the 5th to the 95th percentile of news media density is predicted to decrease ALEC bill enactments by one bill, p=0.02.

Table 6.3: Models of ALEC bill introductions and enactments with campaign contributions.

	Introduced Bills	*Enacted Bills*
	Model 1	Model 2
Business/Labor Contributions	0.24	-0.71**
	(0.19)	(0.31)
Democratic Governor	0.13	-0.08
	(0.16)	(0.14)
Democratic Legislature	0.19	0.08
	(0.23)	(0.24)
Unemployment Rate	0.04	-0.01
	(0.07)	(0.06)
Republican Conservatism	1.33	0.83
	(1.45)	(0.84)
Democratic Conservatism	0.01	-0.81
	(0.54)	(0.50)
Legislative Professionalism	1.49	-2.94*
	(1.97)	(1.51)
Media Density	-2.72	0.58
	(2.31)	(2.94)
Democratic National Government	0.05	0.09
	(0.23)	(0.28)
State Effects	YES	YES
Year Effects	NO	NO
N	384	328

Negative binomial models; robust errors clustered by state.
Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

economic outcomes, it is worth assessing whether the business-drafted legislation we identify in our dataset is associated with changes in state outcomes of substantive interest as a validity check of our approach.

One of the central priorities of the organization, as we have explained above, is lowering state taxes, especially taxes on wealthy individuals and businesses.[20] Therefore we ought to expect that states that

[20] The Center on Budget and Policy Priorities has also hypothesized that ALEC legislation focused on budget and tax policy will lower taxes on wealthy individuals and businesses, thus lowering revenue and raising inequality, but did not empirically test this proposition (Williams and Johnson, 2013).

pass more ALEC legislation would have lower income tax rates on top earners. In addition, we also expect that the total effect of ALEC's model bills will be to redistribute resources towards wealthier individuals and firms, generating increased income inequality. This could happen through many different mechanisms, for instance, by weakening the power of labor unions to bargain collectively, by reducing labor market regulations and social benefits (especially the minimum wage), and by offering lucrative firm contracts and subsidies with state governments. Accordingly, we also measure the effect of ALEC legislation on income inequality in American states.

To test both hypotheses, we run several regressions with either the top tax rate paid by individuals on income or the share of income captured by the top 10 percent as outcomes (data on state top tax rates is from the National Bureau of Economic Research; data on state top income shares from Frank 2014). We include the number of ALEC bills passed by a state as the main explanatory variable, and also control for Democratic control of state government and union membership, since these are plausible confounders that could explain both ALEC's influence and tax rates on the wealthy, as well as levels of inequality. We include state and year fixed effects, as well as a lagged dependent variable to account for past values of the outcome variables, and cluster robust errors by state.[21]

Our results, which appear in Table 6.4, indicate strong support for our two hypotheses about the influence of ALEC on state outcomes. State governments that passed more ALEC legislation generally had lower tax rates on their wealthiest residents in the ensuing years (Model 1). Each ALEC bill enacted by a state lowered the top income tax rate paid by wealthy residents by about five percentage points over the long run, a considerable effect size given that the mean top rate was five percentage points, and the standard deviation was three percentage points.

Turning to Model 2, we see that states that passed more ALEC bills also had higher levels of inequality: each ALEC bill is predicted to increase the share of income captured by the top decile in a state

[21] We find similar results with and without state fixed effects, indicating that Nickell bias is not a concern.

Table 6.4: Effect of ALEC bill enactments on state top income tax rates and income inequality.

	Top Income Tax Rate	*Top 10% Share of Income*
	Model 1	Model 2
Lagged Outcome	0.70***	0.75***
	(0.09)	(0.05)
Enacted ALEC Bills	-0.01**	0.003**
	(0.01)	(0.00)
Union Density	-0.00	-0.00
	(0.01)	(0.00)
Democratic Governor	-0.06	-0.00
	(0.04)	(0.00)
Democratic Legislature	-0.07	0.00
	(0.05)	(0.00)
State Effects	YES	YES
Year Effects	YES	YES
N	677	677

Negative binomial models; robust errors clustered by state.
Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

by about a tenth of a percentage point, or about three percent of a standard deviation. To put those inequality estimates in context, consider that the median state enacted 29 bills over the period we examine – thus, enacting that many bills would increase the share of income flowing to the top decile by about three percentage points, or about one standard deviation of the inequality measure. Importantly, we find that the number of introduced bills in a state is not correlated with either the top income tax rates variable, nor the share of income flowing to the top income decile, suggesting that it is not simply increased ALEC activity in a state that is correlated with these measures, but actual bill enactments.

Together, these findings confirm that ALEC bills not only produce outcomes that are of interest to businesses and their executives, but that are also substantively relevant for important state policy outcomes as well. Indeed, they suggest that ALEC legislation may have contributed in part to rising inequality across the states over the 1990s and 2000s.

6.8 Limitations and Conclusions

The methodological approach of this paper was fairly successful in identifying bills that share traces of ALEC language (ALEC DNA), but much less successful in identifying bills with the same intent as ALEC. The current method relies on document-level overlap metrics and topic similarity for classification of bills that share ALEC DNA. Future work may examine sentence-level alignment approaches and build on recent work in the paraphrase identification literature to detect identical policy proposals in the legislative domain (see e.g. Das and Smith, 2009).

Though preliminary, these results suggest several important conclusions about business power across state governments. Most centrally, they suggest that policy capacity in the form of greater legislative professionalism provides the most consistent explanation for when and where business groups like ALEC are most successful enacting their proposals into law. This provides an important confirmation of earlier work studying ALEC's reach across the states (Hertel-Fernandez, 2014b). Legislators who are strapped for time, ideas, and research assistance are more likely to turn to ALEC, given that the group can offer precisely those resources that legislators lack in an accessible and appealing manner. Our quantitative analysis also points to the importance of labor power in offering a counterweight to the power of ALEC across the states, though these findings are less conclusive than those for legislative professionalism. Closely related, our analysis underscores the fact that ALEC's mode of influence across the states – exploiting weak policy capacity – is unrelated to more conventional forms of influence, especially campaign contributions. Electoral giving has captured the attention of many scholars, citizens, and political advocates, yet as our paper emphasizes, this is not how a substantial number of business-backed bills are being introduced and enacted across the states in recent decades. Thus, more work is needed to study these non-financial and non-electoral means of influence.

Our analysis and the data we have produced opens up a number of additional lines of inquiry. For instance, who are the individual legislators who are authoring these business-drafted bills? How do their own personal characteristics, such as their personal experience and occupation, influence their propensity to turn to business groups and to reuse corporate legislative proposals? To what degree are state legislators responding to the (perceived) opinions of their constituencies when they introduce business-drafted bills? Does ALEC bill activity explain part of the striking socioeconomic gaps in policy representation that have been identified by scholars such as Bartels (2008) and Gilens (2012)? For instance, ALEC has been responsible for promoting state measures to preempt cities from passing paid sick leave programs (at least ten states have passed such measures), despite the fact that large majorities of Americans support such policies (Smith and Kim, 2010; Bottari and Fischer, 2013).

Shifting from legislators to public policy, how does passage of business-backed bills change important state social and economic outcomes, like economic growth, inequality, poverty, and spending on specific programs? We have presented initial findings that ALEC model bills increase inequality, especially at the very top of the income distribution, but much more work is needed to understand this relationship. For instance, which kinds of ALEC bills are responsible for this relationship – is it only bills affecting taxes, or does legislation on other issues also shape the income distribution? These are all central questions whose answers would shed considerable light on the puzzle of rising income inequality in the United States, and how those disparities in economic resources are caused – and reinforced – by disparities in political access and influence between citizens and corporate interests.

Aside from our own interests in business power in the United States, we aim for our paper to serve as a guide for other scholars interested in interest group influence. For instance, we hope to eventually make our database of all state legislation introduced and enacted since the early 1990s available to the public. Scholars could use this dataset – the first publicly available version of its kind – to conduct their own analyses of interest group proposals across state government. And more generally, our methods of text analysis could be deployed to a variety of other contexts. Influence is at the heart of political analysis, and yet scholars have struggled to measure and quantify such relationships in a systematic manner. We think that the approach to text reuse analysis that we lay out in this paper can provide a valuable addition to scholars' toolboxes as they seek to uncover who governs across our societies.

A

Front-door Adjustment

This appendix provides proofs and additional information for the analyses in Chapters 1–3.

A.1 ATT Proofs

A.1.1 Large-Sample Bias of Front-door and Back-door Approaches

The asymptotic bias in the front-door approach for $E[Y(a_0)|a_1]$ is the following:

$$\begin{aligned}
B^{fd}_{a_1} &= \mu^{fd}_{0|a_1} - \mu_{0|a_1} \\
&= \sum_x \sum_m P(m|a_0, x) \cdot E[Y|a_1, m, x] \cdot P(x|a_1) \\
&- \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x, a_1) \cdot P(x|a_1) \\
&= \sum_x \sum_m P(m|a_0, x) \sum_u E[Y|a_1, m, x, u] \cdot P(u|a_1, m, x) \cdot P(x|a_1) \\
&- \sum_x \sum_u \sum_m E[Y|a_0, m, x, u] \cdot P(m|a_0, x, u) \cdot P(u|a_1, x) \cdot P(x|a_1) \\
&= \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x) \cdot E[Y|a_1, m, x, u] \cdot P(u|a_1, m, x) \\
&- \sum_x P(x|a_1) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|a_1, x)
\end{aligned} \tag{A.1}$$

Note that the bias will be zero when Y is mean independent of A conditional on U, M, and X (i.e., $E[Y|a_1, m, x, u] = E[Y|a_0, m, x, u]$) and U is independent of M conditional on X and a_0 or a_1 (i.e., $P(m|a_0, x) = P(m|a_0, x, u)$ and $P(m|a_1, x) = P(m|a_1, x, u)$). Hence, again it is possible for the front-door approach to provide an unbiased estimator when there is an unmeasured confounder.

The asymptotic bias of the back-door approach can be written as the following:

$$\begin{aligned}
B_{a_1}^{bd} &= \mu_{0|a_1}^{bd} - \mu_0 \\
&= \sum_x E[Y|a_0, x] \cdot P(x|a_1) - \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x, a_1) \cdot P(x|a_1) \\
&= \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|a_0, x) \cdot P(x|a_1) \\
&- \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|a_1, x) \cdot P(x|a_1) \\
&= \sum_x P(x|a_1) \sum_u E[Y|a_0, x, u] \cdot [P(u|a_0, x) - P(u|a_1, x)]
\end{aligned} \tag{A.2}$$

A.1.2 Front-door Adjustment with One-Sided Noncompliance

In the special case of one-sided noncompliance, the front-door estimator for ATT can be written as the following:

$$\begin{aligned}
\tau_{att}^{fd} &= E[Y|a_1] - \sum_x E[Y|a_1, m_0, x] \cdot P(x|a_1) \\
&= \sum_x E[Y|a_1, x] \cdot P(x|a_1) - \sum_x E[Y|a_1, m_0, x] \cdot P(x|a_1) \\
&= \sum_x P(x|a_1) \left\{ E[Y|a_1, x] - E[Y|a_1, m_0, x] \right\} \\
&= \sum_x P(x|a_1) \left\{ E[Y|a_1, m_1, x] \cdot P(m_1|x, a_1) + E[Y|a_1, m_0, x] \cdot P(m_0|x, a_1) - E[Y|a_1, m_0, x] \right\} \\
&= \sum_x P(x|a_1) \left\{ E[Y|a_1, m_1, x] \cdot P(m_1|x, a_1) + E[Y|a_1, m_0, x] \cdot [P(m_0|x, a_1) - 1] \right\} \\
&= \sum_x P(x|a_1) \left\{ E[Y|a_1, m_1, x] \cdot P(m_1|x, a_1) - E[Y|a_1, m_0, x] \cdot [1 - P(m_0|x, a_1)] \right\} \\
&= \sum_x P(x|a_1) \left\{ E[Y|a_1, m_1, x] \cdot P(m_1|x, a_1) - E[Y|a_1, m_0, x] \cdot P(m_1|x, a_1) \right\} \\
&= \sum_x P(x|a_1) P(m_1|x, a_1) \left\{ E[Y|a_1, m_1, x] - E[Y|a_1, m_0, x] \right\}
\end{aligned}$$

A.1.3 Front-door and Back-door Bias under One-sided Noncompliance

In the special case of nonrandomized program evaluations with one-sided noncompliance, the front-door and the back-door ATT bias can be written as the following, utilizing the fact that $P(M = 0|a_0, u) = 1$ and $P(M = 0|a_1, u) = 0$ for all u:

$$
\begin{aligned}
B^{fd}_{ATT} &= \mu_1 - \mu^{fd}_{0|a_1} - (\mu_1 - \mu_{0|a_1}) \\
&= \mu_{0|a_1} - \mu^{fd}_{0|a_1} \\
&= -B^{fd}_{a_1} \\
&= \sum_x P(x|a_1) \sum_u E[Y|a_0, m_0, x, u] P(u|a_1, x) \\
&- \sum_x P(x|a_1) \sum_u E[Y|a_1, m_0, x, u] P(u|a_1, m_0, x)
\end{aligned}
$$

Adding and subtracting $\sum_x P(x) \sum_u E[Y|a_0, m_0, u] \cdot P(u|a_1, m_0)$:

$$
\begin{aligned}
&= \sum_x P(x|a_1) \sum_u E[Y|a_0, m_0, x, u] \cdot [P(u|a_1, x) - P(u|a_1, m_0, x)] \\
&- \sum_x P(x|a_1) \sum_u \{E[Y|a_1, m_0, x, u] - E[Y|a_0, m_0, x, u]\} \cdot P(u|a_1, m_0, x)
\end{aligned}
\tag{A.3}
$$

Back-door bias under one-sided noncompliance is:

$$
\begin{aligned}
B^{bd}_{att} &= \mu_1 - \mu^{bd}_{0|a_1} - (\mu_1 - \mu_{0|a_1}) \\
&= \mu_{0|a_1} - \mu^{bd}_{0|a_1} \\
&= -B^{bd}_{a_1} \\
&= \sum_x P(x|a_1) \sum_u E[Y|u, a_0, m_0, x] \cdot [P(u|a_1, x) - P(u|a_0, x)].
\end{aligned}
$$

A.1.4 Alternative Expression for Front-door Bias under One-sided Noncompliance

The bias can be re-written further if we note that the imbalance in U can be written in terms of the mediator,

$$\begin{aligned} P(u|a_1,x) - P(u|a_1,m_0,x) &= P(u|a_1,m_1,x)\cdot P(m_1|a_1,x) + P(u|a_1,m_0,x)\cdot P(m_0|a_1,x) - P(u|a_1,m_0,x) \\ &= P(u|a_1,m_1,x)\cdot P(m_1|a_1,x) + P(u|a_1,m_0,x)\cdot [P(m_0|a_1,x) - 1] \\ &= P(u|a_1,m_1,x)\cdot P(m_1|a_1,x) - P(u|a_1,m_0,x)\cdot P(m_1|a_1,x) \\ &= P(m_1|a_1,x)\cdot [P(u|a_1,m_1,x) - P(u|a_1,m_0,x)], \end{aligned}$$

and the difference in expectations over Y can be written in terms of the potential outcomes under control,

$$\begin{aligned} &E[Y|a_0,m_0,x,u] - E[Y|a_1,m_0,x,u] \\ &\quad = E[Y|a_0,x,u] - E[Y|a_1,m_0,x,u] \\ &\quad = E[Y(a_0)|a_0,x,u] - E[Y|a_1,m_0,x,u] \\ &\quad = E[Y(a_0)|a_1,x,u] - E[Y|a_1,m_0,x,u] \\ &\quad = E[Y(a_0)|a_1,m_1,x,u]\cdot P(m_1|a_1,x,u) + E[Y(a_0)|a_1,m_0,x,u]\cdot P(m_0|a_1,x,u) - E[Y|a_1,m_0,x,u] \\ &\quad = E[Y(a_0)|a_1,m_1,x,u]\cdot P(m_1|a_1,x,u) - E[Y(a_0)|a_1,m_0,x,u]\cdot \left[\frac{E[Y|a_1,m_0,x,u]}{E[Y(a_0)|a_1,m_0,x,u]} - P(m_0|a_1,x,u)\right]. \end{aligned}$$

These equivalencies allow us to write the front-door bias under one-sided noncompliance as the following:

$$\begin{aligned}B^{fd}_{att} &= \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u E[Y|a_0,m_0,x,u]\cdot[P(u|a_1,m_1,x)-P(u|a_1,m_0,x)]\\ &+\sum_x P(x|a_1)\sum_u \Bigg\{E[Y(a_0)|a_1,m_1,x,u]\cdot P(m_1|a_1,x,u)\\ &-E[Y(a_0)|a_1,m_0,x,u]\cdot\left[\frac{E[Y|a_1,m_0,x,u]}{E[Y(a_0)|a_1,m_0,x,u]}-P(m_0|a_1,x,u)\right]\Bigg\}P(u|a_1,m_0,x).\end{aligned}$$

A.1.5 Front-door Bias Simplification

In order to improve interpretability of the bias formula in Section A.1.4 and establish comparability with the results for back-door bias in VanderWeele and Arah (2011), we offer a simplification of the front-door bias formula under one-sided noncompliance and an exclusion restriction. Under Assumptions 3 and 4 we write front-door bias as:

$$\begin{aligned}B^{fd}_{ATT} &= \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u E[Y|a_0,m_0,x,u]\cdot\underbrace{[P(u|a_1,m_1,x)-P(u|a_1,m_0,x)]}_{\varepsilon}\\ &+\sum_x P(x|a_1)\sum_u P(m_1|a_1,x,u)\underbrace{\left[E[Y(a_0)|a_1,m_1,x,u]-E[Y(a_0)|a_1,m_0,x,u]\right]}_{\eta}P(u|a_1,m_0,x)\end{aligned}$$

Furthermore, under Assumptions 6, 7, and 8, we can simplify the above expression as:

$$\begin{aligned}B^{fd}_{ATT} &= P(m_1|a_1)\cdot\varepsilon\cdot\sum_x P(x|a_1)\sum_u E[Y|a_0,m_0,x,u]\\ &+P(m_1|a_1)\cdot\eta\cdot\sum_x P(x|a_1)\sum_u P(u|a_1,m_0,x)\end{aligned}$$

Assuming U is binary as in VanderWeele and Arah (2011):

$$B^{fd}_{ATT} = P(m_1|a_1)\left[\varepsilon \cdot \sum_x P(x|a_1)\underbrace{\left(E[Y|a_0, m_0, x, U = 1] - E[Y|a_0, m_0, x, U = 1]\right)}_{\gamma} + \eta\right]$$

Assuming that γ does not depend on x as in VanderWeele and Arah (2011):

$$B^{fd}_{ATT} = P(m_1|a_1)\left[\gamma \cdot \varepsilon + \eta\right]$$

A.1.6 Front-door Bias within Levels of x

Alternatively, let's return to the expression for front-door bias under one-sided noncompliance from Section A.1.3:

$$B^{fd}_{att} = \sum_x P(x|a_1)\sum_u E[Y|a_0, m_0, x, u]\underbrace{[P(u|a_1, x) - P(u|a_1, m_0, x)]}_{\xi} \tag{A.4}$$

$$+ \sum_x P(x|a_1)\sum_u \{\underbrace{E[Y|a_0, m_0, x, u] - E[Y|a_1, m_0, x, u]}_{\omega}\}P(u|a_1, m_0, x). \tag{A.5}$$

ξ can be rewritten as:

$$\begin{aligned}
\xi &= P(u|a_1, x) - P(u|a_1, m_0, x) \\
&= P(u|a_1, m_1, x)P(m_1|a_1, x) + P(u|a_1, m_0, x)P(m_0|a_1, x) - P(u|a_1, m_0, x) \\
&= P(u|a_1, m_1, x)P(m_1|a_1, x) + P(u|a_1, m_0, x)[P(m_0|a_1, x) - 1] \\
&= P(u|a_1, m_1, x)P(m_1|a_1, x) - P(u|a_1, m_0, x)P(m_1|a_1, x) \\
&= P(m_1|a_1, x)[P(u|a_1, m_1, x) - P(u|a_1, m_0, x)].
\end{aligned}$$

We can also expand ω as:

$$\begin{aligned}
\omega &= E[Y|a_0, m_0, x, u] - E[Y|a_1, m_0, x, u] \\
&= E[Y|a_0, x, u] - E[Y|a_1, m_0, x, u] \\
&= E[Y(a_0)|a_0, x, u] - E[Y|a_1, m_0, x, u] \\
&= E[Y(a_0)|a_1, x, u] - E[Y|a_1, m_0, x, u] \\
&= E[Y(a_0)|a_1, m_1, x, u]P(m_1|a_1, x, u) + E[Y(a_0)|a_1, m_0, x, u]P(m_0|a_1, x, u) - E[Y|a_1, m_0, x, u] \\
&= E[Y(a_0)|a_1, m_1, x, u]P(m_1|a_1, x, u) - E[Y(a_0)|a_1, m_0, x, u] \cdot \{\frac{E[Y|a_1, m_0, x, u]}{E[Y(a_0)|a_1, m_0, x, u]} - P(m_0|a_1, x, u)\} \\
&= P(m_1|a_1, x)\Big\{E[Y(a_0)|a_1, m_1, x, u]\frac{P(m_1|a_1, x, u)}{P(m_1|a_1, x)} \\
&\quad - E[Y(a_0)|a_1, m_0, x, u] \cdot \frac{\frac{E[Y|a_1, m_0, x, u]}{E[Y(a_0)|a_1, m_0, x, u]} - P(m_0|a_1, x, u)}{P(m_1|a_1, x)}\Big\}.
\end{aligned}$$

We note that the bias can be written as scaled by the compliance proportion within levels of x ($P(m_1|a_1, x)$).

We can thus rewrite front-door bias under one-sided noncompliance as:

$$\begin{aligned}B^{fd}_{att} = \sum_x P(x|a_1)P(m_1|a_1,x)\sum_u \Bigg[& E[Y|a_0,m_0,x,u]\cdot[P(u|a_1,m_1,x) - P(u|a_1,m_0,x)] \\ & + \Bigg\{E[Y(a_0)|a_1,m_1,x,u]\frac{P(m_1|a_1,x,u)}{P(m_1|a_1,x)} \\ & - E[Y(a_0)|a_1,m_0,x,u]\cdot\frac{\frac{E[Y|a_1,m_0,x,u]}{E[Y(a_0)|a_1,m_0,x,u]} - P(m_0|a_1,x,u)}{P(m_1|a_1,x)}\Bigg\}P(u|a_1,m_0,x)\Bigg].\end{aligned}$$

This expression looks similar to the expression for front-door bias under one-sided noncompliance in Section A.1.4, but we have rewritten the formula here so that one can express the bias within levels of x.

A.1.7 Front-door Bias Under Constant Compliance Rates Across Values of u

We prove that Assumption 9 and binary M implies that $\xi = 0$:

$$\begin{aligned}P(m_1|a_1,x,u) &= \frac{P(u|a_1,m_1,x)\cdot P(m_1|a_1,x)}{P(u|a_1,x)} \\ 1 &= \frac{P(u|a_1,m_1,x)}{P(u|a_1,x)} \\ P(u|a_1,x) &= P(u|a_1,m_1,x)\end{aligned}$$

Since M is binary, by similar logic as above we know that $P(u|a_1,x) = P(u|a_1,m_0,x)$.

Therefore:

$$\begin{aligned}
\xi &= P(m_1|a_1, x)[P(u|a_1, m_1, x) - P(u|a_1, m_0, x)] \\
&= P(m_1|a_1, x)[P(u|a_1, x) - P(u|a_1, x)] \\
&= 0
\end{aligned}$$

Under Assumption 9, we can simplify front-door bias to:

$$\begin{aligned}
B^{fd}_{att} = \sum_x P(x|a_1)P(m_1|a_1, x) \sum_u \Bigg[& E[Y(a_0)|a_1, m_1, x, u] \frac{P(m_1|a_1, x, u)}{P(m_1|a_1, x)} \\
& - E[Y(a_0)|a_1, m_0, x, u] \cdot \frac{\frac{E[Y|a_1, m_0, x, u]}{E[Y(a_0)|a_1, m_0, x, u]} - P(m_0|a_1, x, u)}{P(m_1|a_1, x)} \Bigg] \cdot P(u|a_1, x).
\end{aligned}$$

A.2 Front-Door Difference-in-Differences Proofs

A.2.1 No Large-Sample Bias in the Front-door Difference-in-Differences Estimator

First define $\tau_{att,x} = E[Y(a_1)|a_1, x] - E[Y(a_0)|a_1, x]$. It is well known that $\tau_{att} = \sum_x \tau_{att,x} P(x|a_1)$. Therefore in order to show that τ_{att}^{fd-did} has no bias, we need only show a lack of bias for $\tau_{att,x}$ within levels of x. If Assumptions 11 and 12 hold, then the front-door difference-in-differences estimator has no large-sample bias:

$$\begin{aligned}
\tau_{att,x}^{fd-did} &= \tau_{att,x,g_1}^{fd} - \frac{P(m_1|a_1, x, g_1)}{P(m_1|a_1, x, g_2)} \tau_{att,x,g_2}^{fd} \\
&= \tau_{att,x} + B_{att,x,g1}^{fd} - \frac{P(m_1|a_1, x, g_1)}{P(m_1|a_1, x, g_2)} \tau_{att,x,g_2}^{fd} \\
&= \tau_{att,x} + B_{att,x,g1}^{fd} - \frac{P(m_1|a_1, x, g_1)}{P(m_1|a_1, x, g_2)} B_{att,x,g_2}^{fd} \\
&= \tau_{att,x} + B_{att,x,g1}^{fd} - B_{att,x,g1}^{fd} \\
&= \tau_{att,x}
\end{aligned}$$

A.3 National JTPA Study

Our analysis makes use of the following samples in the National JTPA Study: experimental active treatment group, experimental control group, and the nonexperimental / eligible nonparticipant (ENP) group. We restrict our attention to the 4 *service delivery areas* at which the ENP sample was collected: Fort Wayne, IN; Corpus Christi, TX; Jackson, MS, and Providence, RI. We also only examine 2 target groups: adult males and adult females. Note that the active treatment group for our purposes means receving any JTPA service, even though the services actually received from the JTPA varied across individuals.[1]

The raw data and edited analysis files are available as part of the National JTPA Study Public Use Data from the Upjohn Institute. The covariates for the experimental sample are available through the background information form (BIF) and the covariates for ENPs are available through the long baseline survey (LBS). The experimental samples completed the BIF, which contains demographic information, social program participation, and training and education histories, at the time of random assignment. The ENPs completed the LBS anywhere from 0 to 24 months following eligiblity screening. Unlike the BIF which mostly covers the previous year in terms of labor market experiences, the LBS covers the past 5 years prior to the survey date and thus provides a much richer portrait of labor market participation. Moreover, experimental control units at the 4 ENP sites also received the long baseline survey, completed 1-2 months after random assignment. Heckman et al. (1998), Heckman and Smith (1999), and related works rely on the detailed labor force participation data and earnings histories in LBS to identify selection bias by comparing the experimental control units to the nonexperimental control units. Unfortunately, treated units were never administered the LBS and we have no detailed labor force participation data for multiple years prior to random assignment. Moreover, no one survey in-

[1] The National JTPA Study classified services received into the following 6 categories: classroom training in occupational skills, on-the-job training, job search assistance, basic education, work experience, and miscellaneous.

strument was administered to all three of the samples we are using in this analysis, yielding issues of noncomparability. The limited set of covariates we use in the conditioning sets in our analysis have all been established to be comparable by verifying their values across the BIF and LBS for the experimental control group, which completed both surveys at the 4 ENP sites.

The dataset we end up using in our analysis was obtained in communication with Jeffrey Smith and Petra Todd. It is the dataset used in the estimates presented in Section 11 of Heckman et al. (1998) and contains all three samples we use in our analysis. It also contains compliance information for the experimental treated group sample. The covariates we utilize in our analysis have been cross-checked against the raw data from the Upjohn Institute. There are also additional covariates in the Heckman et al. (1998) data that have been imputed using a linear regression as described in Appendix B3 of their paper.

The outcome variable we use in the analysis is total 18-month earnings in the period following random assignment (for experimental units) or eligiblity screening (for ENPs). The monthly total earnings variable available from the public use data files is the `totearn` variable. The data covers months 1-30 after random assignment (denoted as $t+1$ to $t+30$, where t is the time of random assignment). The data also includes data for t, the month of random assignment. Note that this variable is not raw earnings data, but was constructed by Abt Associates from the First and Second Follow-up Surveys, as well as based on data from state unemployment agencies, for the initial JTPA report.[2] Please consult Appendix A of Orr et al. (1994) for description of the First Follow-up Survey, Second Follow-up Survey, and earnings data from state unemployment insurance agencies and Appendix B of the same report for construction and imputation of the 30-month earnings variables. The Narrative Description of the National JTPA Study Public Use Files also contains description of the imputation process (see http://www.upjohninst.org/erdc/njtpa.html).

[2] One of the major imputations was a decision to divide raw earnings by a `shares` variable which adjust earnings reported for incomplete months (due to the timing of the interviews) to full monthly earnings.

In our analysis, we rely upon the monthly total earnings variable in the dataset we obtained from Jeffrey Smith and Petra Todd. We have verified the earnings data used in the calculation of the program impact from this dataset against the earnings variables in the public use data and they match exactly except for a few individuals where Heckman et al. (1998) have imputed missing monthly data. This applies to around 1% of observations and thus is unlikely to substantively change any results. A unit-by-unit comparison of earnings across the raw data and the data we are using can be obtained from us upon request.

The full dataset we obtained contains 1478 treated units, 649 experimental control units, and 667 ENPs for adult males. For adult females, there are 1734 treated units, 830 experimental control units, and 1340 ENPs. These numbers already exclude individuals without any earnings records. We follow the sample restrictions in Heckman et al. (1998) to reduce the full dataset to the final sample (see Appendix B1). We impose an age restriction of 22 to 54 years old on the experimental samples to match the ages of the ENP sample. We then omit individuals who are missing data on race and date of eligibility. Finally, we impose a *rectangular restriction* based on quarterly earnings. For experimental control and the ENP samples, we require (i) at least one month of valid earnings prior to random assignment (for experimental controls) or prior to eligibility screening (for ENPs), denoted as $t = 0$, (ii) valid earnings data at $t = 0$, and (iii) at least one month of valid earnings data in months $t + 13$ to $t + 18$. For the treatment group, we impose only restriction iii. The final sample sizes are presented in Table A.1.

Table A.1: Sample Sizes Before and After Imposing Sample Restrictions. The treated units are broken up into compliers (C) and noncompliers (NC). Control denotes experimental control and ENP denotes the eligible nonparticipants.

	Adult Males				Adult Females			
	Treated		Control	ENP	Treated		Control	ENP
	C	*NC*			*C*	*NC*		
Pre-restriction	843	635	649	667	953	781	830	1340
Post-restriction	834	622	523	384	934	765	706	852

Even after imposing the rectangular restriction on earnings, some individuals had missing earnings data for some months. In the construction of the 18-month total earnings variable, we mean impute the missing months using the average of the individual's available monthly earnings. Details on the extent of missingness are available from authors upon request.

A.3.1 Analysis by Marital Status

In the analysis in Chapters 2 and 3, we subset the observations by marital status. The final sample sizes (by marital status) are presented in Table A.2. We cross-checked the covariates we utilize in our analysis against the raw data, available as part of the National JTPA Study Public Use Data from the Upjohn Institute. We established that all covariates in our conditioning sets are identical. However, the marital status variable that denotes whether individuals are currently, or were once, married was imputed as described in Appendix B3 of Heckman et al. (1998) and thus does not exactly match the raw data. We treat all individuals with values of the marital status variable that fall between 0 and 1 (non-inclusive) as married. We note that any given coding scheme is highly unlikely to alter results since only 3% of observations fall in this range.

Table A.2: Sample sizes for adult males by marital status. The treated units are broken up into compliers (C) and noncompliers (NC). Control denotes experimental control and ENP denotes the eligible nonparticipants.

	Treated		Control	ENP
	C	*NC*		
Non-single	484	304	274	292
Single	350	318	266	92

A.4 Florida Voting Data

To construct our population of eligible voters, we examine individuals that have appeared in one of four voter registration snapshots: book closing records from 10/10/2006, 10/20/2008, and 10/18/2010, as well as a 2012 election recap record from 1/4/2013. This yields a total population of 16,371,725 individuals that we are able to subset by race (Asian / Pacific Islander, Black (not Hispanic), Hispanic, White (Not Hispanic), and Other). Note that the Other category contains individuals who self-identify as American Indian / Alaskan Native, Multiracial, or Other, as well as individuals for whom race is unknown. In cases where race changes across voter registration snapshots for the same voter, we use the latest available self-reported race. Such changes affect only 1.1% of observations. The breakdown of the the population by race is presented in the rightmost column of Table A.3.

We use voter history files from 08/03/2013 to subset the population by voting mode in each election. The voter history files required pre-processing before we could use them for estimation. As mentioned in Gronke and Stewart (2013) and Stewart (2012), there is an issue of duplication of voter identification numbers within the same election. In some cases, this duplication is rather innocuous because the voting mode is identical across records. In these cases, we simply remove duplicate records and include the voter in our analysis. In other cases, voters are recorded as both having voted in a given election and not having voted (code "N"). In these cases, we assume that the voter did indeed cast a ballot and use that code. Finally, there are a few instances in which a voter is recorded to have voted in multiple ways. For example, a voter history file may indicate that a voter voted both absentee and early at a given election. While Gronke and Stewart (2013) indicates that voters may legitimately appear multiple times in the voter history file, this makes the task of stratifying by voting mode difficult. As a result, we choose to exclude individuals who are recorded to have voted using more than one mode. When analyzing the 2008 election subsetting by 2006 voting modes, we exclude 385 individuals. The corresponding numbers for analysis of the 2012 election subsetting using 2008 and 2010 voting groups are

1951 and 2998, respectively. These figures are dwarfed by the sample sizes and thereby highly unlikely to exert any serious effect upon our estimates.

We also made several choices regarding the definition of voting modes. Specifically, we classified anyone who voted absentee (code “A”) and whose absentee ballot was not counted (code “B”) as having voted absentee. We classified anyone who voted early (code “E”) and anyone cast a provisional ballot early (code “F”) as having voted early. Finally, we classify anyone who voted at the polls (code “Y”) and cast a provisional ballot at the polls (code “Z”) as having voted on election day. We do not use the code “P”, which indicates that an individual cast a provisional ballot that was not counted since we cannot ascertain whether it was cast on election day, early, or as an absentee voter.

Another difficulty with the data is defining the eligible electorate and thus individuals who did not vote. While the voter history files have a code “N” for did not vote, most individuals who do not vote are not present in the voter history files at all. For example, for the 2008 election there were no “N” codes at all in the voter history files. Therefore, we count an individual as not having voted in a given election if they appeared in the voter registration files at one point but are either not present in the voter history file for that election or are coded as “N”.

Table A.3: Voting modes as percent of population in 2006, 2008, and 2010 elections. Population is defined as anyone who has appeared in voter registration records from 2006-2012. Note that percentages of individuals who did not vote, whose provisional ballots were not counted, or who are dropped due to conflicting voting modes are not shown.

	2006			2008			2010			
Race	Early	Absentee	Election Day	Early	Absentee	Election Day	Early	Absentee	Election Day	Total
Asian	2.87	2.61	12.21	15.15	10.45	20.43	4.75	5.69	13.76	233664
Black	2.81	1.85	16.86	27.67	7.14	16.97	6.41	4.24	18.57	2159473
Hispanic	2.46	2.83	12.46	15.05	8.60	22.72	4.11	6.01	13.37	2049683
White	5.89	5.60	23.16	14.48	13.57	25.32	7.39	9.05	20.63	11179293
Other	2.60	2.43	11.97	13.67	8.12	19.31	3.87	4.19	12.26	749612

A.5 Early Voting Results

Table A.4: Rate of compliance (percent) with early voting program by prior voting mode and race in 2006-2008, 2008-2012, and 2010-2012 transitions (e.g., for the 2006-2008 transition, prior voting mode is based on voting behavior in 2006, while compliance rate is proportion voting early in 2008). Note that percentages of individuals who did not vote, whose provisional ballots were not counted, or who are dropped due to conflicting voting modes are not shown.

	2006-2008			2008-2012			2010-2012		
Race	Early	Absentee	Election Day	Early	Absentee	Election Day	Early	Absentee	Election Day
Asian	55.43	12.14	27.59	37.48	8.50	12.68	57.16	9.12	24.65
Black	71.53	18.66	53.70	49.12	12.35	19.48	67.91	15.05	43.65
Hispanic	56.70	8.98	29.50	33.18	6.81	11.02	53.03	7.31	21.63
White	53.87	9.19	22.98	40.54	7.69	12.33	56.69	7.11	20.41
Other	55.42	10.43	28.10	37.15	7.33	11.06	57.43	8.28	24.50

Table A.5: Front-door estimates for EIP effect by race for 2006-2008, 2008-2012, and 2010-2012 transitions (e.g., for the 2006-2008 transition, prior voting mode is based on voting behavior in 2006 and EIP estimate is for 2008). All estimates use county fixed effects. 99% block boostrapped confidence intervals are reported in brackets.

Race	2006-2008	2008-2012	2010-2012
Asian	0.1548	0.1587	0.1476
	[0.1421, 0.1857]	[0.1608, 0.1778]	[0.14, 0.1823]
Black	0.2239	0.2101	0.1793
	[0.1923, 0.2455]	[0.1899, 0.2325]	[0.1582, 0.2191]
Hispanic	0.1442	0.1305	0.128
	[0.132, 0.2037]	[0.1131, 0.1664]	[0.1096, 0.1821]
White	0.1292	0.1539	0.1325
	[0.1064, 0.1564	[0.136, 0.173]]	[0.1131, 0.1563]
Other	0.1699	0.1756	0.1623
	[0.1421, 0.2021]	[0.1608, 0.1897]	[0.14, 0.1869]

Table A.6: Front-door difference-in-differences estimates for EIP effect by race for 2006-2008, 2008-2012, and 2010-2012 transitions (e.g., for the 2006-2008 transition, prior voting mode is based on voting behavior in 2006 and EIP estimate is for 2008). Estimates are reported across different group of interest and differencing groups. All estimates use county fixed effects. 99% block boostrapped confidence intervals are reported in brackets.

Race	Group of interest - Differencing group	2006-2008	2008-2012	2010-2012
Asian	EIP-Absentee	0.0535 [0.028, 0.0785]	0.0169 [0.01, 0.0233]	0.0528 [0.0352, 0.0757]
	EIP-Election Day	0.0569 [0.0398, 0.0801]	0.0225 [0.0177, 0.0303]	0.0501 [0.0344, 0.0739]
	Election Day-Absentee	-0.0027 [-0.0093, 0.0023]	-0.0023 [-0.005, 1e-04]	0.001 [-0.0036, 0.0048]
Black	EIP-Absentee	0.0922 [0.0658, 0.1134]	0.028 [0.0181, 0.0375]	0.0692 [0.0538, 0.0934]
	EIP-Election Day	0.0694 [0.0544, 0.0811]	0.0185 [0.0104, 0.0261]	0.0502 [0.0346, 0.078]
	Election Day-Absentee	0.0152 [-0.0022, 0.0286]	0.0032 [-0.0014, 0.0077]	0.0101 [0.0021, 0.0177]
Hispanic	EIP-Absentee	0.0532 [0.0445, 0.0782]	0.0162 [0.0126, 0.0245]	0.0377 [0.027, 0.0703]
	EIP-Election Day	0.0454 [0.0385, 0.0758]	0.0082 [0.003, 0.0193]	0.0346 [0.0257, 0.0624]
	Election Day-Absentee	0.0037 [-0.003, 0.0059]	0.0025 [-9e-04, 0.0046]	0.0012 [-0.0014, 0.0054]
White	EIP-Absentee	0.033 [0.0201, 0.049]	0.0098 [0.0024, 0.0171]	0.0349 [0.0228, 0.0495]
	EIP-Election Day	0.0452 [0.0334, 0.0615]	0.0218 [0.0168, 0.0278]	0.0426 [0.0316, 0.0565]
	Election Day-Absentee	-0.0054 [-0.0071, -0.0038]	-0.0038 [-0.0056, -0.002]	-0.0029 [-0.0046, -0.0013]
Other	EIP-Absentee	0.0477 [0.0305, 0.0707]	0.0169 [0.0096, 0.0228]	0.0482 [0.0315, 0.0652]
	EIP-Election Day	0.0558 [0.042, 0.0726]	0.0191 [0.0149, 0.024]	0.0493 [0.037, 0.0646]
	Election Day-Absentee	-0.0042 [-0.0078, 0]	-9e-04 [-0.0034, 0.0015]	-4e-04 [-0.0048, 0.0037]

A.6 Get Out the Vote Experiments

To define the scope of the experiments we replicate, we start with the list of GOTV experiments in Green et al. (2013) and subset to those studies that use phone GOTV treatments.[3] Furthermore, we only include studies for which we can readily locate replication data. Specifically, this means that the replication data was available in one of three places:

1. Author's site
2. Journal site
3. Yale's Institution for Social and Policy Studies (ISPS) data repository

We are thus left with a total of 8 replicable journal articles, spanning 19 GOTV experiments.

[3] The only exception is Arceneaux et al. (2006), which was absent from Green et al. (2013) because they counted the 2002 Iowa and Michigan experiments under the Gerber and Green (2005a) article.

Table A.7: Overview of GOTV Phone Experiments

Source	Experiment	Population restriction?	HH size	Election year	Election type	Volunteer?
Arceneaux et al. (2006)	Iowa Michigan		<=2	2002	Midterm	
Arceneaux et al. (2010)	Illinois		<= 5	2004	Presidential	
Gerber and Green (2005b)	New Haven	No students	<= 2	1998	Midterm	
Gerber et al. (2010)	Michigan	Voted in Nov. 2004 & 2006, not August 2006		2010	Primary	
Nickerson (2006)	Albany Boulder Student Eugene Student Stonybrook	Student voters registered on campus (ages 18-30)		2000	Presidential	✓
	Boulder Vendor Eugene Vendor	Ages 18-30		2000	Presidential	✓
	Boston			2001	Local	✓
Nickerson (2007)	Local Professional National Professional Volunteer	Ages 18-26		2000	Midterm	 ✓
Nickerson et al. (2006)	Michigan	Ages 18-35, not Republican		2002	Gubernatorial	✓
Panagopoulos (2011)	2 Weeks Before 3 Days Before 4 Weeks Before			2005	Municipal	

Table A.8: Sample Size and Compliance of GOTV Phone Experiments

Source	Experiment	N_{treat}	Compliance rate
Arceneaux et al. (2006)	Iowa	78000	0.3511
	Michigan	66475	0.3415
Arceneaux et al. (2010)	Illinois	6569	0.5185
Gerber and Green (2005b)	New Haven	6626	0.3470
Gerber et al. (2010)	Michigan	8448	0.4718
Nickerson (2006)	Albany	804	0.6157
	Boulder Student	653	0.7213
	Eugene Student	705	0.7433
	Stonybrook	680	0.8867
	Boulder Vendor	1143	0.3447
	Eugene Vendor	953	0.4900
	Boston	1209	0.5542
Nickerson (2007)	Local Professional	4849	0.3081
	National Professional	54948	0.3920
	Volunteer	54017	0.4427
Nickerson et al. (2006)	Michigan	10547	0.6426
Panagopoulos (2011)	2 Weeks Before	2000	0.7395
	3 Days Before	2000	0.5185
	4 Weeks Before	2000	0.7420

A.7 Oregon Health Insurance Experiment

We replicate results from Tables 5-6 and 8-10 of Finkelstein et al. (2012). Data is subset to those individuals who responded to 12-month post-randomization survey. There are a total of 11808 treated units in the study, 8365 of which are compliers and 3429 are non-compliers. There are 11933 control units that we use to establish the experimental benchmark. We condition on birth year, gender, whether or not English is the primary language, whether or not the individual signed themselves up, whether the individual applied on the first day, whether or not the individual has a phone, whether or not an individual has a P.O. box, and a dummy for being in a metropolitan statistical area (MSA) when fitting conditional expectation functions to treated units and treated non-compliers using OLS.

A.8 Front-door Adjustment for ATE

For an outcome Y and a treatment/action A, we define the potential outcome under a generic treatment as $Y(a_1)$ and the potential outcome under control as $Y(a_0)$. The ATE is defined as $E[Y(a_1)] - E[Y(a_0)]$. In what follows we discuss the asymptotic bias in estimating $E[Y(a_0)]$ and the ATE.

A.8.1 Asymptotic Bias in Estimating $E[Y(a_0)]$

As in VanderWeele and Arah (2011), we assume that $E[Y(a_0)]$ can be equated to the expectations over observed outcomes by conditioning on observed covariates X and unobserved covariates U. For simplicity in presentation we assume that X and U are discrete, such that

$$\mu_0 = E[Y(a_0)] = \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x) \cdot P(x), \tag{A.6}$$

but continuous variables can be easily accommodated.[4] We also assume that probabilistic assignment holds such that there is a positive probability of both a_1 and a_0 for all values of U and X among all units (Rubin, 2010).

If we have measured a set of post-treatment variables M, the front-door adjustment can be written as the following:

$$\mu_0^{fd} = \sum_x \sum_m P(m|a_0, x) \sum_a E[Y|a, m, x] \cdot P(a|x) \cdot P(x), \tag{A.7}$$

where these sums are taken over values of x, m, and a with positive probability. The asymptotic bias for

[4] We also note that for formulas of this type throughout the paper, when any of the densities take the value zero (e.g., $P(u|x, a_1) = 0$ or $P(x|a_1) = 0$) we mean the entire term to be zero.

$E[Y(a_0)]$ can be written as the following (see Appendix A.8.4 for a proof):

$$
\begin{aligned}
B_0^{fd} &= \mu_0^{fd} - \mu_0 \\
&= \sum_x P(x) \sum_m \sum_u P(m|a_0, x) \sum_a E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \\
&- \sum_x P(x) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|x)
\end{aligned}
\tag{A.8}
$$

In this supplement as in Chapter 1, we will use the term bias to refer to asymptotic bias. The bias is zero when the following two conditions hold:

Assumption 13 (Y is mean independent of A conditional on M, X, and U)

$E[Y|a, m, x, u] = E[Y|a_0, m, x, u]$ for all a, m, x, u.

Assumption 14 (U is independent of M conditional on A and X)

$P(m|a_0, x) = P(m|a_0, x, u)$ for all m, x, u and $\sum_a P(u|a, m, x) \cdot P(a|x) = P(u|x)$ for all m, x, u.

The result for μ_1 is analogous. Therefore, as demonstrated in Pearl (1995), it is possible for the front-door approach to provide an unbiased estimator of ATE, even when there is an unmeasured common cause of A and Y. However, note that unlike the presentation in Pearl (1995, 2000, 2009), the presentation here does not require the definition of potential outcomes beyond those originally used to define the ATE. In other words, this presentation is agnostic as to whether SUTVA holds with M as a treatment variable.

A.8.2 Asymptotic Bias for ATE

The front-door formula ATE can be written as:

$$
\mu_1^{fd} - \mu_0^{fd} = \sum_x P(x) \sum_m [P(m|a_1, x) - P(m|a_0, x)] \sum_a E[Y|a, m, x] \cdot P(a|x), \tag{A.9}
$$

with the bias written as the following (see proof in Appendix A.8.4):

$$
\begin{aligned}
B^{fd}_{ATE} &= \mu^{fd}_1 - \mu^{fd}_0 - (\mu_1 - \mu_0) \\
&= \sum_x P(x) \sum_m [P(m|a_1, x) - P(m|a_0, x)] \sum_a \sum_u E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \\
&- \sum_x P(x) \sum_u \sum_m \{[P(m|a_1, x, u) - P(m|a_0, x, u)]E[Y|a_1, m, x, u]\} \, P(u|x) \\
&- \sum_x P(x) \sum_u \sum_m \{[E[Y|a_1, m, x, u] - E[Y|a_0, m, x, u]]P(m|a_0, x, u)\} \, P(u|x)
\end{aligned}
\tag{A.10}
$$

Note that the last line is zero when Assumption 13 holds. When Assumption 14 holds as well, B^{fd}_{ATE} can be shown to be zero similarly to B^{fd}_0.

In order to compare the bias in the front-door approach to the standard back-door approach, we write the back-door formula for ATE based on observed covariates as the following:

$$
\mu^{bd}_1 - \mu^{bd}_0 = \sum_x P(x)[E[Y|a_1, x] - E[Y|a_0, x]], \tag{A.11}
$$

and the bias of the back-door formula as the following (see Appendix A.8.4 for a proof), which is very similar to the formula presented in VanderWeele and Arah (2011):

$$
\begin{aligned}
B^{bd}_{ATE} &= \mu^{bd}_1 - \mu^{bd}_0 - (\mu_1 - \mu_0) \\
&= \sum_x P(x) \sum_u \{[P(u|a_1, x) - P(u|x)] \cdot [E[Y|a_1, x, u] - E[Y|a_0, x, u]]\} \\
&- \sum_x P(x) \sum_u \{[P(u|a_1, x) - P(u|a_0, x)] \cdot [E[Y|a_0, x, u]\}
\end{aligned}
\tag{A.12}
$$

There are two important general facts to note about the comparison between B^{fd}_{ATE} and B^{bd}_{ATE}. First, it is quite possible that the front-door ATE bias will be smaller than the back-door ATE bias even when the aforementioned front-door independence conditions do not hold exactly. Second, because both estimators are defined within levels of the observed covariates X, it is possible to form hybrid estimators that utilize the front-door estimator for some values of X and the back-door estimator for other values of X. In order to develop some intuition about when the front-door approach would be preferred to the back-door approach (perhaps within a level of X), we next consider the special case of linear Structural Equation Models with constant effects (SEMs) and a scalar M.

A.8.3 Special Case: Linear Structural Equation Models

If we assume additive linear models with constant effects for Y and M, then:

$$E[Y|a, m, x] - E[Y|a, m', x] = \kappa(m - m'), \tag{A.13}$$

which is constant in a and x, and:

$$E[M|a_1, x] - E[M|a_0, x] = \lambda(a_1 - a_0), \tag{A.14}$$

which is constant in x. This allows us to write the front-door ATE as the following (proof in Appendix A.8.5):

$$\mu_1^{fd} - \mu_0^{fd} = \kappa\lambda(a_1 - a_0) \tag{A.15}$$

Therefore, when we assume additive linear models, the front-door formula for ATE simplifies to a product of multiple regression coefficients. If we also assume that Y is an additive linear model in a, x, and u, then $E[Y|a_1, x, u] - E[Y|a_0, x, u] = \tau(a_1 - a_0)$ and the ATE simplifies as well:

$$\begin{aligned}\mu_1 - \mu_0 &= \sum_x \sum_u \tau(a_1 - a_0) \cdot P(u|x) \cdot P(x) \\ &= \tau(a_1 - a_0)\end{aligned} \tag{A.16}$$

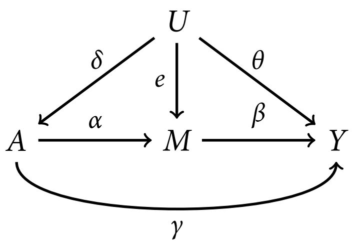


Figure A.1: Structural Equation Model

In order to present the ATE bias in the front-door approach, it will be helpful to present a simplified linear structural equation model with constant effects for these variables. This is defined by the path diagram in Figure A.1. For simplicity in presentation, independent error terms have been removed from the graph, we have assumed that there are no measured conditioning variables, and we have assumed that A, M, U, and Y are scalars. Note that when $a_1 - a_0 = 1$, the ATE τ can be written as the following for this model:

$$\tau = \alpha\beta + \gamma \tag{A.17}$$

When $a_1 - a_0 = 1$, the front-door formula is the following (see Appendix A.8.5):

$$\mu_1^{fd} - \mu_0^{fd} = \alpha\beta + \alpha\theta e \frac{V(U|A)}{V(M|A)} + \beta e\delta \frac{V(U)}{V(A)} + e^2\delta\theta \frac{V(U)}{V(A)} \frac{V(U|A)}{V(M|A)} \tag{A.18}$$

and the difference between the front-door formula and the ATE is the following:

$$B_{ATE}^{fd} = \alpha\theta e \frac{V(U|A)}{V(M|A)} + \beta e\delta \frac{V(U)}{V(A)} + e^2\delta\theta \frac{V(U)}{V(A)} \frac{V(U|A)}{V(M|A)} - \gamma \tag{A.19}$$

Therefore, the front-door formula will equal the ATE when the first three terms equal γ. In other words, when the bias for the indirect effect equals the direct effect. A special case of this is the situation when $e = 0$ and $\gamma = 0$, and this can itself be seen as an example of the front-door criterion within the context of SEMs.

For comparison, the back-door formula and bias can be written as the following (see Appendix A.8.5):

$$\mu_1^{bd} - \mu_0^{bd} = \alpha\beta + \gamma + (\beta e\delta + \theta\delta)\frac{V(U)}{V(A)} \tag{A.20}$$

$$B_{ATE}^{bd} = (\beta e\delta + \theta\delta)\frac{V(U)}{V(A)} \tag{A.21}$$

When comparing the back-door and front-door bias within SEMs, we first notice that both share the $\beta e\delta\frac{V(U)}{V(A)}$ terms. This represents the $A \leftarrow U \rightarrow M \rightarrow Y$ path. The key comparison is between the bias terms unique to the front-door formula ($\alpha\theta e\frac{V(U|A)}{V(M|A)} + e^2\delta\theta\frac{V(U)}{V(A)}\frac{V(U|A)}{V(M|A)} - \gamma$) and the bias term unique to the back-door formula ($\theta\delta\frac{V(U)}{V(A)}$). Roughly speaking, the front-door bias can be smaller than the back-door bias when e and γ are small or when the front-door bias terms cancel. Notice as well that the front-door and back-door bias will be equal when $\theta = 0$ and $\gamma = 0$, which is equivalent to saying that there is no direct effect from A to Y or from U to Y. Therefore, another general case where the front-door will be preferred to the back-door is when U is largely mediated by M, and the bias from the common term is ameliorated by the direct effect ($\beta e\delta\frac{V(U)}{V(A)} - \gamma$).

A.8.4 Asymptotic Bias Proofs

Front-door Bias

The large-sample bias for the front-door formula of $E[Y(a_0)]$ can be derived as the following:

$$\begin{aligned}
B_0^{fd} &= \mu_0^{fd} - \mu_0 \\
&= \sum_x \sum_m P(m|a_0, x) \sum_a E[Y|a, m, x] \cdot P(a|x) \cdot P(x) - \sum_x \sum_u E[Y|a_0, x, u] \cdot P(u|x) \cdot P(x) \\
&= \sum_x \sum_m P(m|a_0, x) \sum_a \sum_u E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \cdot P(x) \\
&- \sum_x \sum_u \sum_m E[Y|a_0, m, x, u] \cdot P(m|a_0, x, u) \cdot P(u|x) \cdot P(x) \\
&= \sum_x P(x) \sum_m \sum_u P(m|a_0, x) \sum_a E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \\
&- \sum_x P(x) \sum_m \sum_u P(m|a_0, x, u) \cdot E[Y|a_0, m, x, u] \cdot P(u|x)
\end{aligned} \tag{A.22}$$

The large-sample bias for the front-door formula for ATE can be derived as the following:

$$
\begin{aligned}
B_{ATE}^{fd} &= \mu_1^{fd} - \mu_0^{fd} - (\mu_1 - \mu_0) \\
&= \sum_x P(x) \sum_m [P(m|a_1, x) - P(m|a_0, x)] \sum_a E[Y|a, m, x] \cdot P(a|x) \\
&- \sum_x P(x) \sum_u \{E[Y|a_1, x, u] - E[Y|a_0, x, u]\} \cdot P(u|x) \\
&= \sum_x P(x) \sum_m [P(m|a_1, x) - P(m|a_0, x)] \sum_a \sum_u E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \\
&- \sum_x P(x) \sum_u \sum_m \{E[Y|a_1, m, x, u] \cdot P(m|a_1, x, u) - E[Y|a_0, m, x, u] \cdot P(m|a_0, x, u)\} \cdot P(u|x) \\
&+ \sum_x P(x) \sum_u \sum_m P(m|a_0, x, u) \cdot E[Y|a_1, m, x, u] - \sum_x P(x) \sum_u \sum_m P(m|a_0, x, u) \cdot E[Y|a_1, m, x, u] \\
&= \sum_x P(x) \sum_m [P(m|a_1, x) - P(m|a_0, x)] \sum_a \sum_u E[Y|a, m, x, u] \cdot P(u|a, m, x) \cdot P(a|x) \\
&- \sum_x P(x) \sum_u \sum_m [P(m|a_1, x, u) - P(m|a_0, x, u)] \cdot E[Y|a_1, m, x, u] \cdot P(u|x) \\
&- \sum_x P(x) \sum_u \sum_m \{E[Y|a_1, m, x, u] - E[Y|a_0, m, x, u]\} \cdot P(m|a_0, x, u) \cdot P(u|x)
\end{aligned}
\tag{A.23}
$$

Back-door Bias

The back-door formula for ATE based on the observed covariates is the following:

$$\mu_1^{bd} - \mu_0^{bd} = \sum_x P(x) \cdot \{E[Y|a_1, x] - E[Y|a_0, x]\},$$

and the large sample bias of the back-door formula is the following:

$$
\begin{aligned}
B_{ATE}^{bd} &= \mu_1^{bd} - \mu_0^{bd} - (\mu_1 - \mu_0) \\
&= \sum_x P(x) \cdot \{E[Y|a_1, x] - E[Y|a_0, x]\} \\
&- \sum_x P(x) \sum_u \{E[Y|a_1, x, u] - E[Y|a_0, x, u]\} \cdot P(u|x) \\
&= \sum_x P(x) \sum_u \{E[Y|a_1, x, u] \cdot P(u|a_1, x) - E[Y|a_0, x, u] \cdot P(u|a_0, x)\} \\
&- \sum_x P(x) \sum_u [E[Y|a_1, x, u] - E[Y|a_0, x, u]] \cdot P(u|x)
\end{aligned}
$$

Adding and subtracting $\sum_x P(x) \sum_u P(u|a_1, x) \cdot E[Y|a_o, x, u]$:

$$
\begin{aligned}
&= \sum_x P(x) \sum_u \{E[Y|a_1, x, u] - E[Y|a_0, x, u]\} \cdot P(u|a_1, x) \\
&- \sum_x P(x) \sum_u [E[Y|a_0, x, u] \cdot [P(u|a_1, x) - P(u|a_0, x)] \\
&- \sum_x P(x) \sum_u \{E[Y|a_1, x, u] - E[Y|a_0, x, u]\} \cdot P(u|x) \\
&= \sum_x P(x) \sum_u \{E[Y|a_1, x, u] - E[Y|a_0, x, u]\} \cdot [P(u|a_1, x) - P(u|x)] \\
&- \sum_x P(x) \sum_u E[Y|a_0, x, u] \cdot [P(u|a_1, x) - P(u|a_0, x)]
\end{aligned}
\tag{A.24}
$$

A.8.5 Linear SEM Proofs

Front-door Formula

When writing the front-door formula for ATE within linear SEMs, note that $\sum_m [P(m|a_1,x) - P(m|a_0,x)] = 0$, so if we choose a reference value of m', then we can include the quantity $-\sum_a E[Y|a,m',x] \cdot P(a|x) \cdot P(x)$ which is constant in m. If we further assume additive linear models for Y and M, then $E[Y|a,m,x] - E[Y|a,m',x] = \kappa(m-m')$, which is constant in a and x, and $E[M|a_1,x] - E[M|a_0,x] = \lambda(a_1 - a_0)$ which is constant in x.

$$\begin{aligned}
\mu_1^{fd} - \mu_0^{fd} &= \sum_x P(x) \sum_m [P(m|a_1,x) - P(m|a_0,x)] \sum_a E[Y|a,m,x] \cdot P(a|x) \\
&= \sum_x P(x) \sum_m [P(m|a_1,x) - P(m|a_0,x)] \sum_a \{E[Y|a,m,x] - E[Y|a,m',x]\} \cdot P(a|x) \\
&= \sum_x P(x) \sum_m [P(m|a_1,x) - P(m|a_0,x)] \sum_a \kappa(m-m') \cdot P(a|x) \\
&= \sum_x P(x) \sum_m [P(m|a_1,x) - P(m|a_0,x)] \kappa(m-m') \\
&= \sum_x P(x) \sum_m [P(m|a_1,x) - P(m|a_0,x)] \kappa m \qquad \text{(A.25)} \\
&= \sum_x P(x) \kappa \sum_m [mP(m|a_1,x) - mP(m|a_0,x)] \\
&= \sum_x P(x) \kappa \{E[M|a_1,x] - E[M|a_0,x]\} \\
&= \sum_x P(x) \kappa\lambda(a_1 - a_0) \\
&= \kappa\lambda(a_1 - a_0)
\end{aligned}$$

Therefore, when we assume additive linear models, the front-door formula for ATE simplifies to a product of multiple regression coefficients. If we also assume that Y is an additive linear model in a,

x, and u, then $E[Y|a_1, x, u] - E[Y|a_0, x, u] = \tau(a_1 - a_0)$ and the ATE simplifies as well.

We can express κ and λ in terms of covariances:

$$\begin{aligned}\kappa &= \frac{Cov(Y, M|A)}{V(M|A)} \\ &= \beta + \theta e \frac{V(U|A)}{V(M|A)}\end{aligned} \tag{A.26}$$

$$\begin{aligned}\lambda &= \frac{Cov(M, A)}{V(A)} \\ &= \alpha + e\delta \frac{V(U)}{V(A)}\end{aligned} \tag{A.27}$$

Within the linear SEM the following covariance relationships hold (we omit uncorrelated errors in these expressions as is typically done with SEM graphs since they do not affect the derivations):

$$\begin{aligned}Cov(Y, M|A) &= Cov(\beta M + \gamma A + \theta U, M|A) \\ &= \beta Cov(M, M|A) + \theta Cov(U, M|A) \\ &= \beta V(M|A) + \theta Cov(U, \alpha A + eU|A) \\ &= \beta V(M|A) + \theta e V(U|A)\end{aligned} \tag{A.28}$$

$$\begin{aligned}Cov(M, A) &= Cov(\alpha A + eU, A) \\ &= \alpha V(A) + eCov(U, A) \\ &= \alpha V(A) + eCov(U, \delta U) \\ &= \alpha V(A) + e\delta V(U)\end{aligned} \tag{A.29}$$

Therefore, when $a_1 - a_0 = 1$, the front-door formula is

$$\begin{aligned}\mu_1^{fd} - \mu_0^{fd} = \lambda\kappa &= (\alpha + e\delta\frac{V(U)}{V(A)})(\beta + \theta e\frac{V(U|A)}{V(M|A)}) \\ &= \alpha\beta + \alpha\theta e\frac{V(U|A)}{V(M|A)} + \beta e\delta\frac{V(U)}{V(A)} + e^2\delta\theta\frac{V(U)}{V(A)}\frac{V(U|A)}{V(M|A)}\end{aligned} \tag{A.30}$$

and the difference between the front-door formula and the ATE is the following:

$$\begin{aligned}B_{ATE}^{fd} = \lambda\kappa - \tau &= \alpha\beta + \alpha\theta e\frac{V(U|A)}{V(M|A)} + \beta e\delta\frac{V(U)}{V(A)} + e^2\delta\theta\frac{V(U)}{V(A)}\frac{V(U|A)}{V(M|A)} \\ &- \alpha\beta + \gamma \\ &= \alpha\theta e\frac{V(U|A)}{V(M|A)} + \beta e\delta\frac{V(U)}{V(A)} + e^2\delta\theta\frac{V(U)}{V(A)}\frac{V(U|A)}{V(M|A)} - \gamma\end{aligned} \tag{A.31}$$

Back-door Formula

The back-door formula and bias can be described in terms of the following covariance relationships:

$$\begin{aligned}Cov(Y, A) &= Cov(\beta M + \gamma A + \theta U, A) \\ &= \beta Cov(M, A) + \gamma Cov(A, A) + \theta Cov(U, A) \\ &= \beta(\alpha V(A) + e\delta V(U)) + \gamma V(A) + \theta Cov(U, eU) \\ &= \beta(\alpha V(A) + e\delta V(U)) + \gamma V(A) + \theta\delta V(U)\end{aligned} \tag{A.32}$$

$$\begin{aligned}
\mu_1^{bd} - \mu_0^{bd} &= \frac{Cov(Y,A)}{V(A)} \\
&= \alpha\beta + \gamma + (\beta e \delta + \theta\delta)\frac{V(U)}{V(A)} \qquad (A.33) \\
B_{ATE}^{bd} &= (\beta e \delta + \theta\delta)\frac{V(U)}{V(A)}
\end{aligned}$$

B

Evaluating Social Security Forecasts

This appendix provides additional information for the analyses in Chapters 4–5.

B.1 Data Sources

Demographics We obtain observed period life expectancy data for 1982–2010 from the Human Mortality Database (HMD, mortality.org). We collected all life expectancy forecasts published in the annual Trustees Reports 1982–2010. In reports prior to 2001, SSA published life expectancy at birth and at age 65 forecasts for males and females projected in 5-year intervals for a total of 75 years into the future. Post-2001, supplementary single-year tables are included online. Our sources are Tables 11 of Trustees Reports 1982-1991, Table II.D.2 of Trustees Reports 1992-2000, and Table V.A3 of the supplemental single-year tables of Trustees Reports 2001-2010. We calculate residuals as the difference between SSA's "best guess" projection (intermediate-cost scenario / alternative II) and the observed life expectancy from HMD. To calculate the uncertainty interval, we use the minimum and maximum values of projected life expectancy across the three scenarios.

We obtain observed total fertility rate (TFR) data for 1982–2010 from the Human Fertility Database (HFD, humanfertility.org). We collected all TFR forecasts published in the annual Trustees Reports 1982–2010. In reports prior to 2001, SSA published TFR forecasts in 5-year intervals for a total of 75 years into the future in the same tables as life expectancy (see preceding paragraph). For 2001 and onwards, supplementary single-year Table V.A1 of each Trustees Report contains TFR forecasts. We calculate residuals and uncertainty intervals in an analogous manner to life expectancy.

Finally, we collected all forecasts of net legal immigration for 2005 and 2010 published in the 2000–2010 Trustees Reports, available on page 62 of the 2000 Trustees Report and in Table V.A1 of the 2001–2010 Trustees Reports. The observed levels of net legal immigration are available in Table V.A1 of the 2014 Trustees Report.

Observed Life Tables We obtain observed conditional probability of death and period life expectancy data used in Figures 5.1, 5.3, and 5.2 from the Human Mortality Database (HMD, 2015). Figure 4.1 plots conditional probabilities of death in the United States from 1980–2010 for males and females separately. Figure 5.2 plots conditional probabilities of death across 39 countries for both genders combined. The availability of the data ranges from as early as 1751 for Sweden until 2013 at the latest. Complete details of data availability by country are available at j.mp/HMDavailability. Figure 5.3 plots the changes in conditional probabilities over successive 10-year intervals from 1960–2010. We present the results by age (65, 75, 85, and 95) and sex. This change is calculated as the slope of a regression of the log of the conditional probability of death on time for each ten year period. For example, the value for 1970 is the slope in the aforementioned regression for the 10-year window from 1961-1970. In the plot, we weight each age in proportion to the number of deaths for that age in 2010.

SSA's Forecasts of Cause-Specific Mortality We obtained observed (1979–2005) and forecast (2006–2100) cause-specific mortality rates under SSA's intermediate cost scenario from personal communication with Felicitie Bell on May 15, 2009.

Ultimate rates of decline of mortality Ultimate rates of decline (UROD) are single number summaries of the myriad long-term mortality assumptions made by the Trustees by sex, age group, and cause of death. They represent the average annual percent reduction in age-adjusted central death rates for the last 50 years of the 75-year projection window. We gather the intermediate-cost scenario ("best guess") URODs assumed by 1991-2013 Trustees Reports and the URODs recommended by the four Technical Panels commissioned by the Social Security Advisory Board (quadrennially starting with 1999). For the 1999 TR, we average the male and female UROD as reported in the 2003 Technical Panel Report (Social Security Advisory Board Technical Panel, 2003). Page 71 of the 2001 TR gives the UROD assumed by 2000 and 2001 TR. Table II.C1 of TR 2002-2009 give the assumed UROD. UROD for 2010 TR is found on page 80 of the report. For 2011, we average the male and female UROD - 0.75% and 0.73%, respectively - as reported in Table 2.3 of Office of the Chief Actuary (2013). For 2012, 2013, and 2014 Trustees Reports, the UROD are found in Table 2.2 of the respective *Long-Range Demographic Assumptions* (Office of the Chief Actuary, 2012, 2013, 2014).

UROD recommendations made by the 1999 and 2003 Technical Panel are available in Table 3 of the 2003 Technical Panel Report (Social Security Advisory Board Technical Panel, 2003). The recommended UROD by the 2007 Technical Panel is available in Table 1 of their report (Social Security Advisory Board Technical Panel, 2007). Although the 2011 Technical Panel doesn't make its recommendations in terms of UROD, *The Long-Range Demographic Assumptions for the 2013 Trustees Report* converts the 2011 Technical Panel's life expectancy recommendation into an UROD.

Financials For 1978–2012, we take the observed cost rate and balance from Table IV.B1 and the observed trust fund ratio from Table IV.B3 of the 2013 Trustees Report (j.mp/2013tables). We calculate residuals as the difference between SSA's "best guess" projection (intermediate-cost scenario / alternative II) and the historic statistics from SSA. Note that for 1981–1990, Trustees Reports have two intermediate-cost scenarios: alternative II-A and II-B. For these years, we follow subsequent Trustees Reports and use II-B as the "best guess" projection. To calculate the uncertainty interval, we use the minimum and maximum values of projected life expectancy across the three scenarios (four for 1981-1990).

Sources for SSA projections published in the annual Trustees Reports (TR) from 1978 to 2012:

Cost rate: Table 26 of TR 1978, Table 27 of TR 1980-1980, Table 28 of TR 1982, Table 29 of TR 1982-1983, Table 30 of TR 1984-1985, Table 28 of TR 1986, Table 26 of TR 1987-1991, Table II.F.13 of TR 1992-2000, Table IV.B1 of the supplemental single-year tables of TR 2001-2012.

Balance: For Trustees Reports from 1986, balance projections are available from the same tables as the cost rate projections. In 1978-1982 TR, the Trustees project the same scheduled tax rate across the cost scenarios (available in Table 25 for TR 1978, Table 26 for TR 1979-1980, Table 28 for TR 1981, and Table 29 for TR 1982). We subtract the cost rates across the scenarios from the scheduled income rate to obtain the range of balance projections. For 1983-1984 TR, the income tax rate varies slightly across the cost scenarios and is only published for the two intermediate projections (Table 27 for TR 1983 Table 28 for TR 1984). We are thus unable to evaluate coverage of SSA's uncertainty intervals for projections made in these two reports.

Trust fund ratio: Table 28 of TR 1978, Table 29 of TR 1979-1980, Table 31 of TR 1981, Table 32 of TR

1982-1983, Table 33 of TR 1984-1985, Table 31 of TR 1986, Table 29 of TR 1987, Table 31 of TR 1988-1990, Table 32 of TR 1991, Table II.F.19 of TR 1992-1994, Table II.F20 of TR 1995-2000, Table IV.B3 of the supplemental single-year tables of TR 2001-2012.

Unanticipated costs We calculate the unanticipated costs due to demography by estimating the number of unanticipated OASI beneficiaries for a given Trustees Report and projection year. In order to estimate the number of unanticipated beneficiaries, we determine the proportionate change in age-specific mortality rates that would correspond to the forecast error in life expectancy for each Trustees Report and projection year. For each age 65 years and older, we multiply the age-specific population count by the difference between the counterfactual age-specific mortality rate necessary to achieve the projected life expectancy and the observed age-specific mortality rate. This product represents the number of unanticipated beneficiaries for a given age and sex. We then sum across all ages for males and females to estimate the total number of unanticipated beneficiaries. For example, the 2005 Trustees Report projected year 2010 life expectancy at age 65 for males to be 16.6 years while observed life expectancy in 2010 equaled 17.9 years. This forecast error of 1.3 years corresponds to a 18.75% increase in observed age-specific mortality rates, or 150,844 unanticipated male beneficiaries.

To obtain unanticipated cost due to under-estimating life expectancy, we then multiply the number of unexpected OASI beneficiaries by the average annual benefits per OASI beneficiaries in constant 2010 dollars. We calculate average annual benefits per OASI beneficiary in current dollars by dividing total OASI expenditures (j.mp/OASIexpend) by total OASI beneficiaries (j.mp/OASIbenif). We convert average annual benefits from current dollars to constant 2010 dollars using the CPI-U series available from the Bureau of Labor Statistics (bls.gov/data).

We calculate total unanticipated costs as the difference between observed total OASDI expenditures in current dollars (j.mp/OASDIexpend) and projected OASDI expenditures. We convert unanticipated costs from current dollars to constant 2010 dollars using the CPI-U series available from the Bureau of Labor Statistics. Forecasts of total OASDI expenditures in billions of current dollars for 2000-2010 Trustees Reports are found in Table III.B3 of 2000 TR, Table VI.E9 of 2001-2002 TR, Table VI.F9 of 2003-2004 TR, and Table VI.F8 of 2005-2010 TR.

B.2 Demographic Results Using SSA Mortality Data

In Figures B.1–B.3 in this Appendix, we replicate the analyses in Figures 1–3, respectively, from Chapter 4 by replacing mortality data from the Human Mortality Database with that computed by SSA. Only small differences between the sources of data are revealed by this analysis.

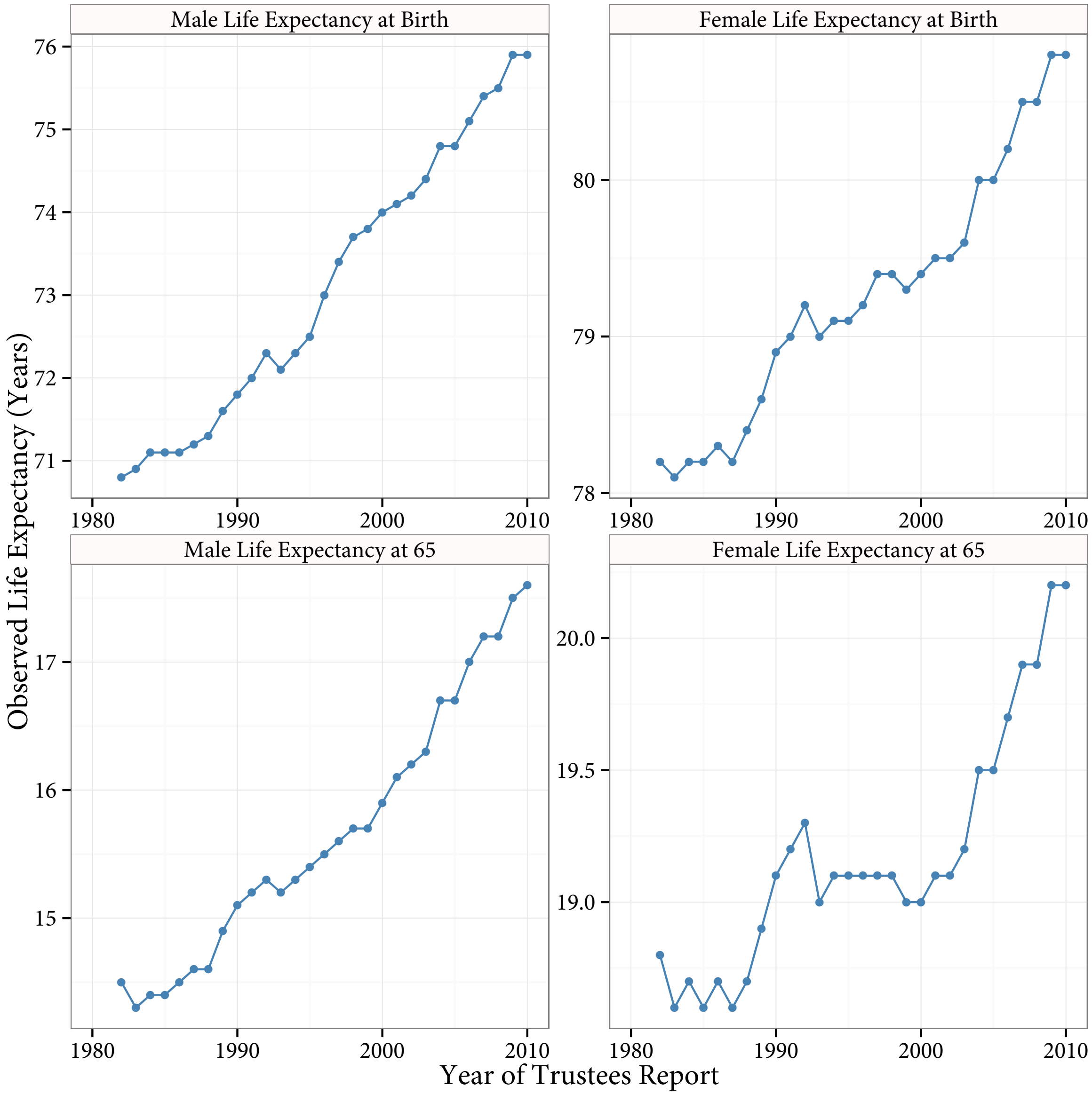
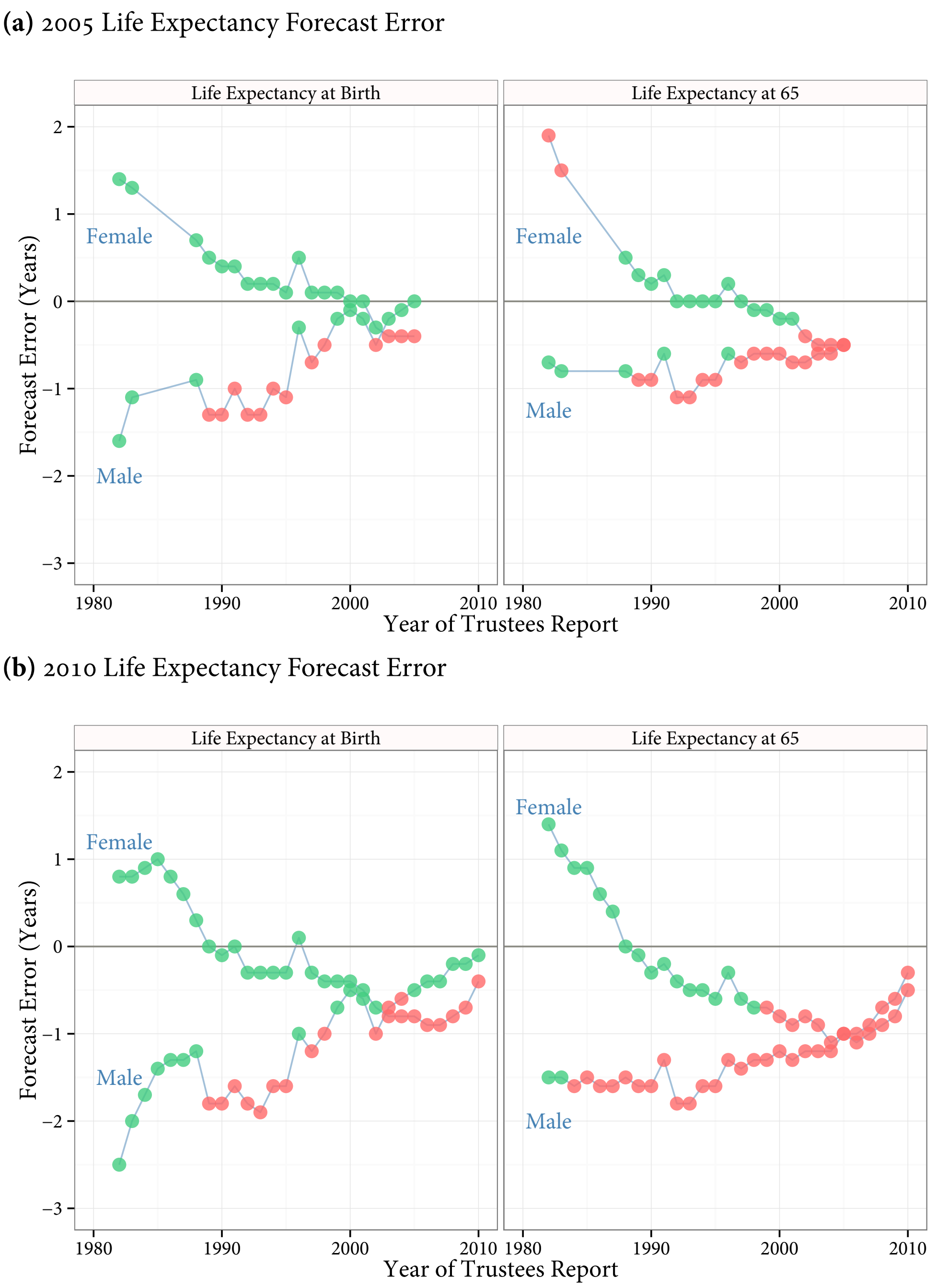


Figure B.1: Observed Period Life Expectancy (Source: SSA). As described in the text, "period life expectancy" for a year is a single-number summary of all the age-specific mortality rates for that same year and is interpreted as the average number of years a person could expect to live if he or she experienced the mortality rates of a given year over the course of their life.



(a) 2005 Life Expectancy Forecast Error

(b) 2010 Life Expectancy Forecast Error

Figure B.2: Forecast Error of Life Expectancy in 2005 (panel a) and 2010 (panel b) by Trustees Report, with data from SSA. Dots are colored green when truth falls within SSA uncertainty intervals. Dots are colored red when the truth falls outside SSA uncertainty intervals.

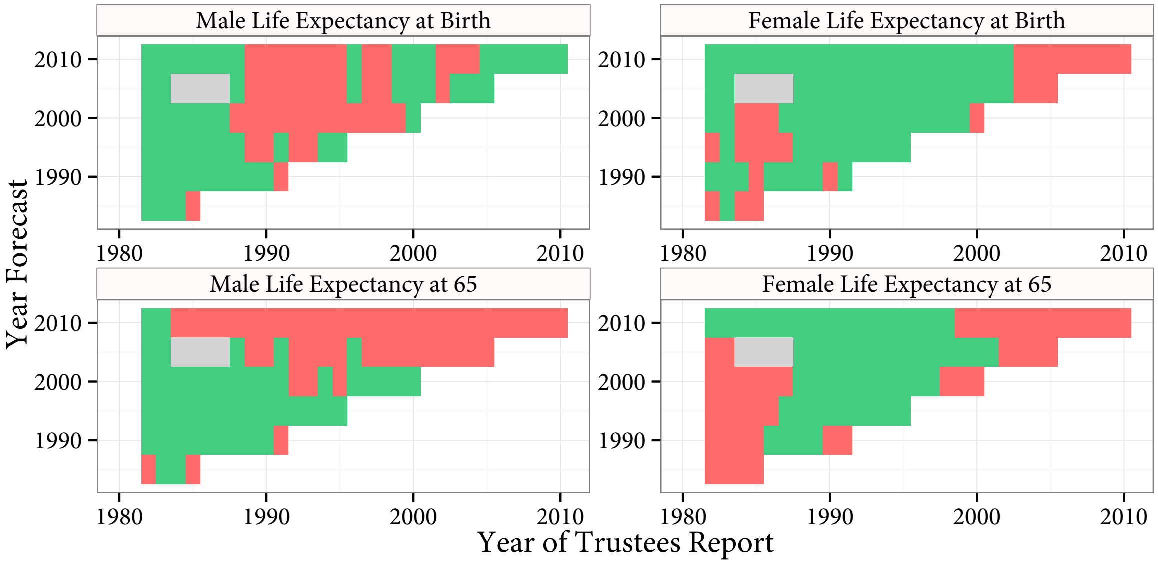


Figure B.3: Uncertainty Interval Coverage, with data from SSA. Green indicates uncertainty interval covered the truth, red indicates that it did not, and gray indicates that SSA did not provide an interval.

B.3 Life Expectancy Forecast Errors for 1985–2010

We summarize in Figures B.4 and B.5 the forecast error for all available years of life expectancy (forecasts for years 1985–2010). The four graphs in this figure correspond to the four demographic variables from Figure 4.1. Again, each graph plots forecast error (vertically) by the year of the Trustees Report (horizontally). Each graph also includes a smoothed line (locally weighted scatterplot smoothing [LOESS]) weighted by proximity of the forecast (allowing for the fact that forecasting years farther into the future should be more difficult), along with a 95% confidence interval. The general pattern from the 2005 and 2010 demographic forecasts in Figure 4.2 can be seen here as well.

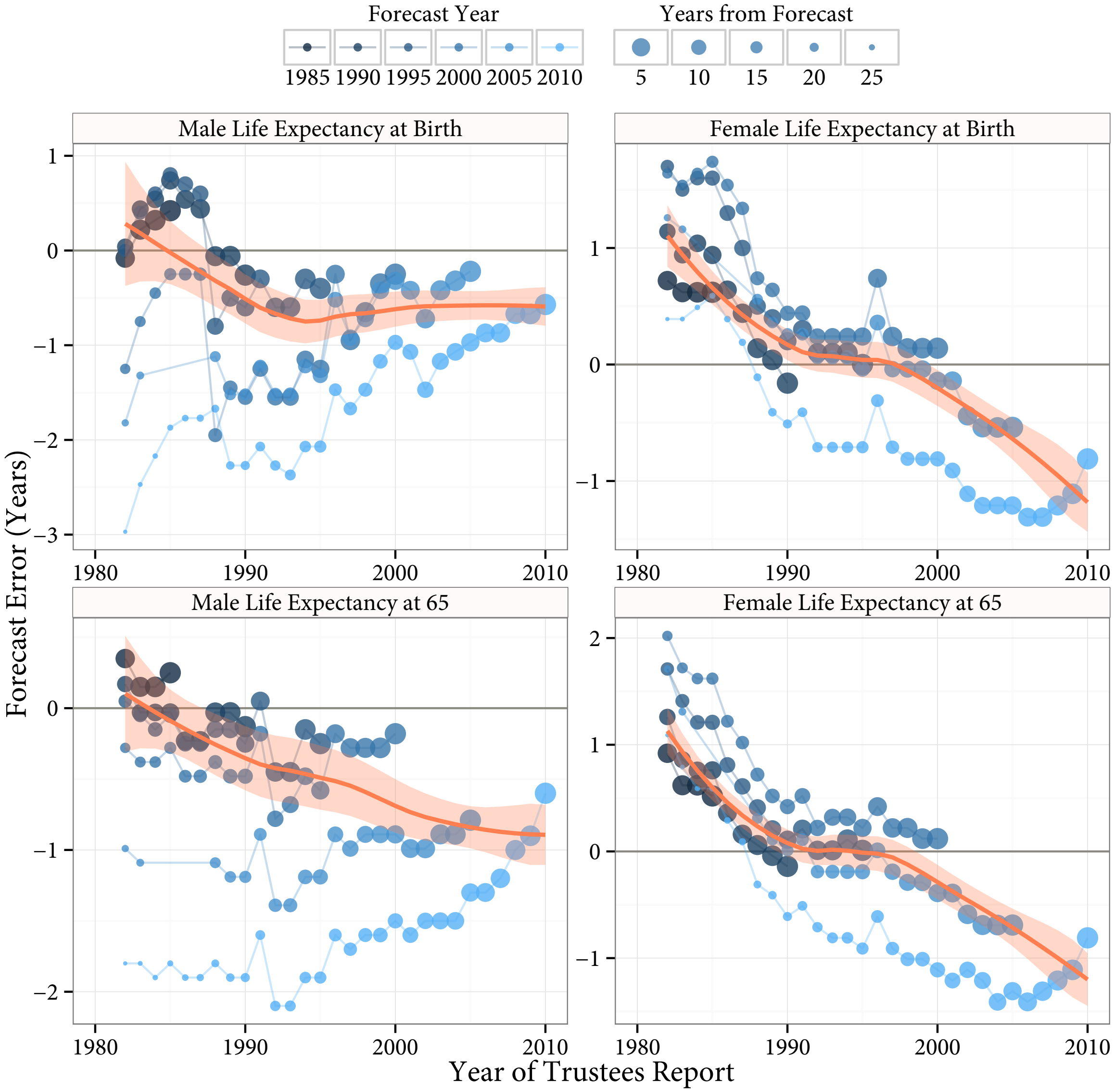


Figure B.4: Life Expectancy Forecasts: Errors by Trustees Report Year by Forecast Year, with data from HMD. Larger dots indicate forecasts nearer to the date the forecast was made. Smoothed line weighted by proximity to year forecast.

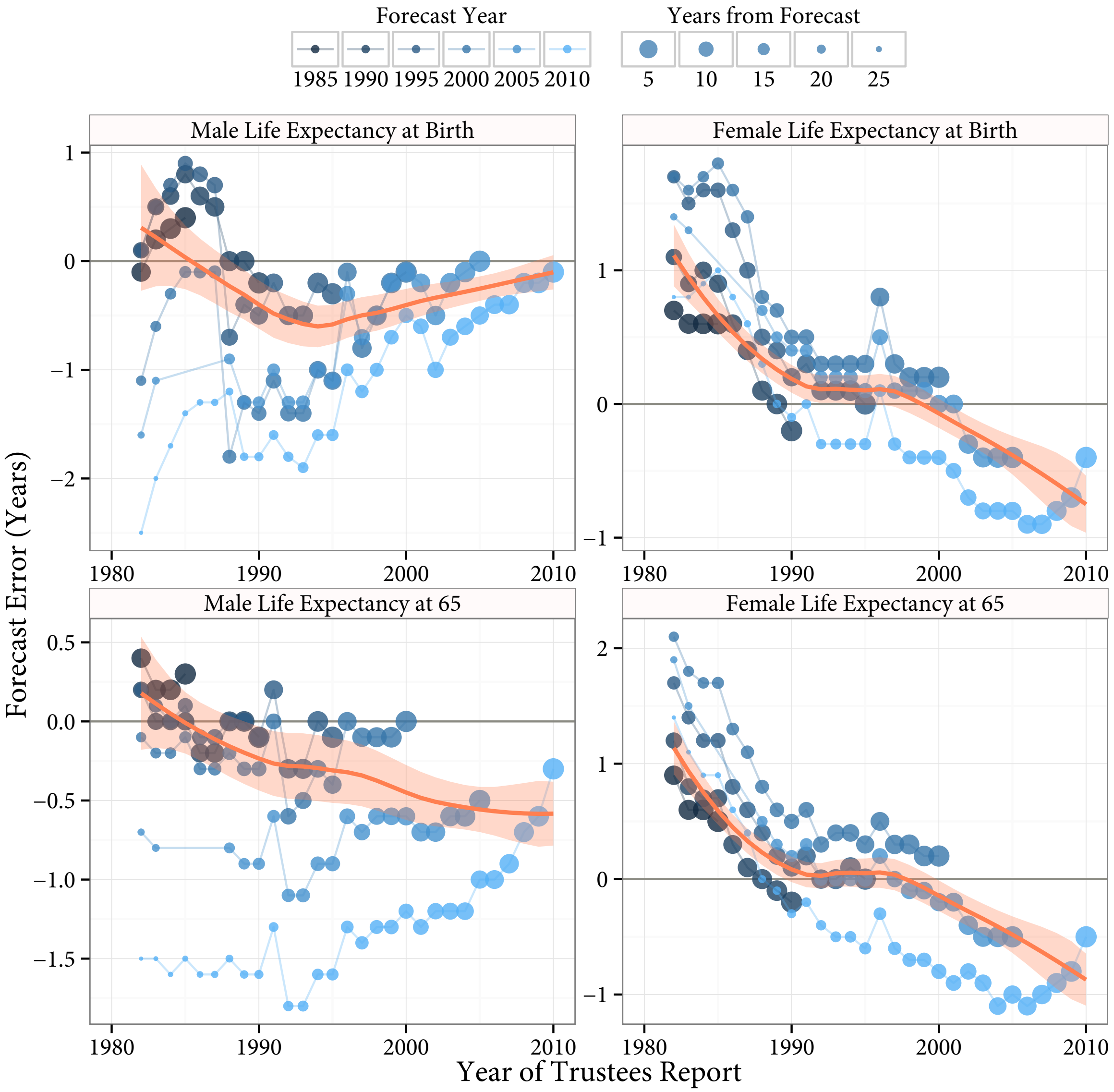


Figure B.5: Life Expectancy Forecasts: Errors by Trustees Report Year by Forecast Year, with data from SSA. Larger dots indicate forecasts nearer to the date the forecast was made. Smoothed line weighted by proximity to year forecast.

B.4 Total Fertility Rate Forecast Errors

Our results for the analysis of total fertility rate forecasts appear in Figure B.6 for data from the Human Fertility Database and Figure B.7 for SSA data. The two figures are quite similar. Results in both are relatively unremarkable, except for consistent overconfidence nearer to the date forecast. See for example the bottom right graph in both figures which has disproportionate numbers of red squares near the diagonal on the graph (where the year of Trustees Report is close to the year forecast).

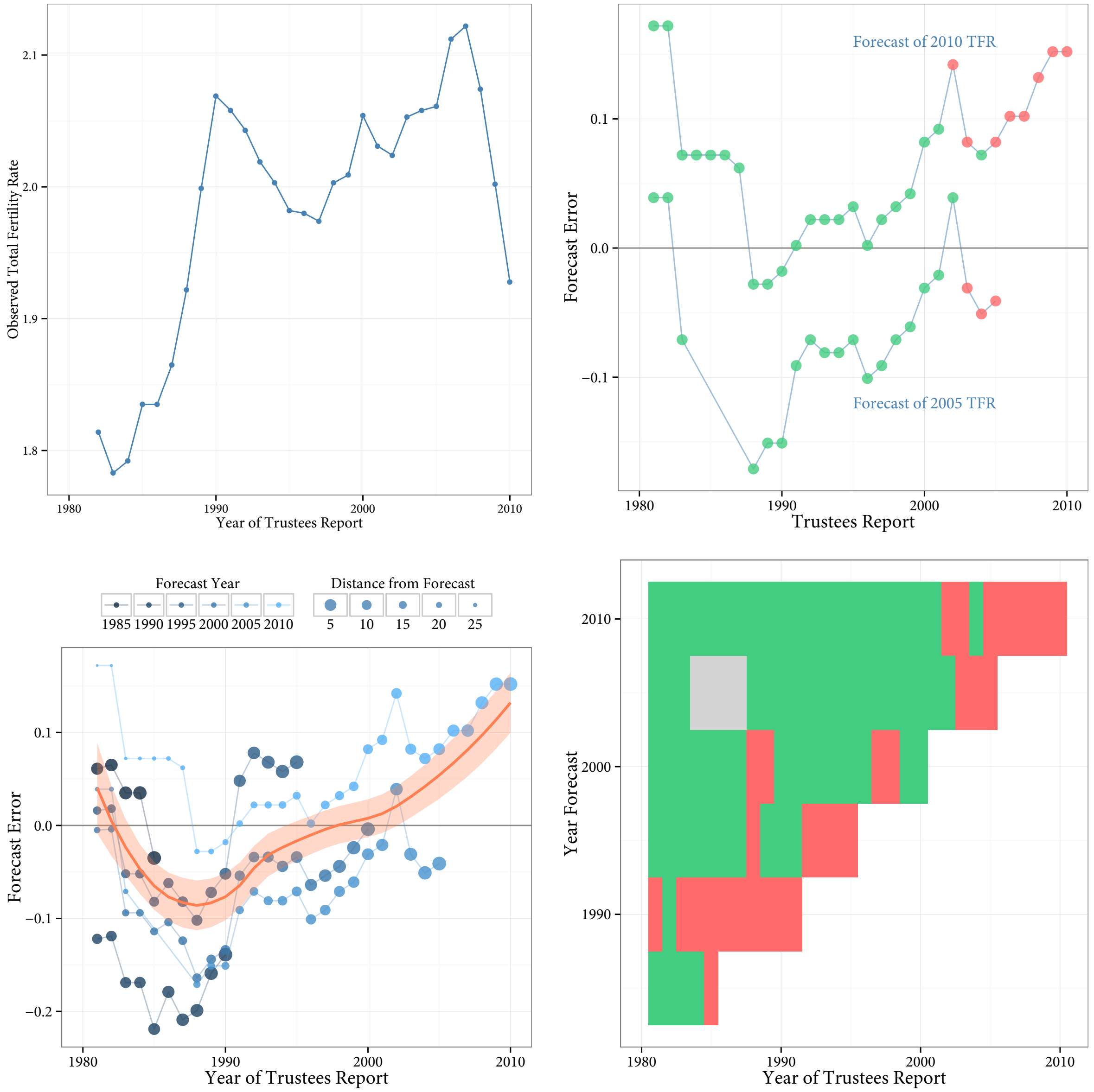


Figure B.6: Clockwise from top left: a) Observed total fertility rate from Human Fertility Database; b) Forecast error of total fertility rate for 2005 and 2010 by Trustees Report. Dots are colored green when truth falls within SSA uncertainty intervals. Dots are colored red when the truth falls outside SSA uncertainty intervals; c) Errors by trustees report year by forecast year. Larger dots indicate forecasts nearer to the date the forecast was made. Smoothed line weighted by proximity to year forecast; d) Uncertainty interval coverage. Green indicates uncertainty interval covered the truth and red indicates that it did not.

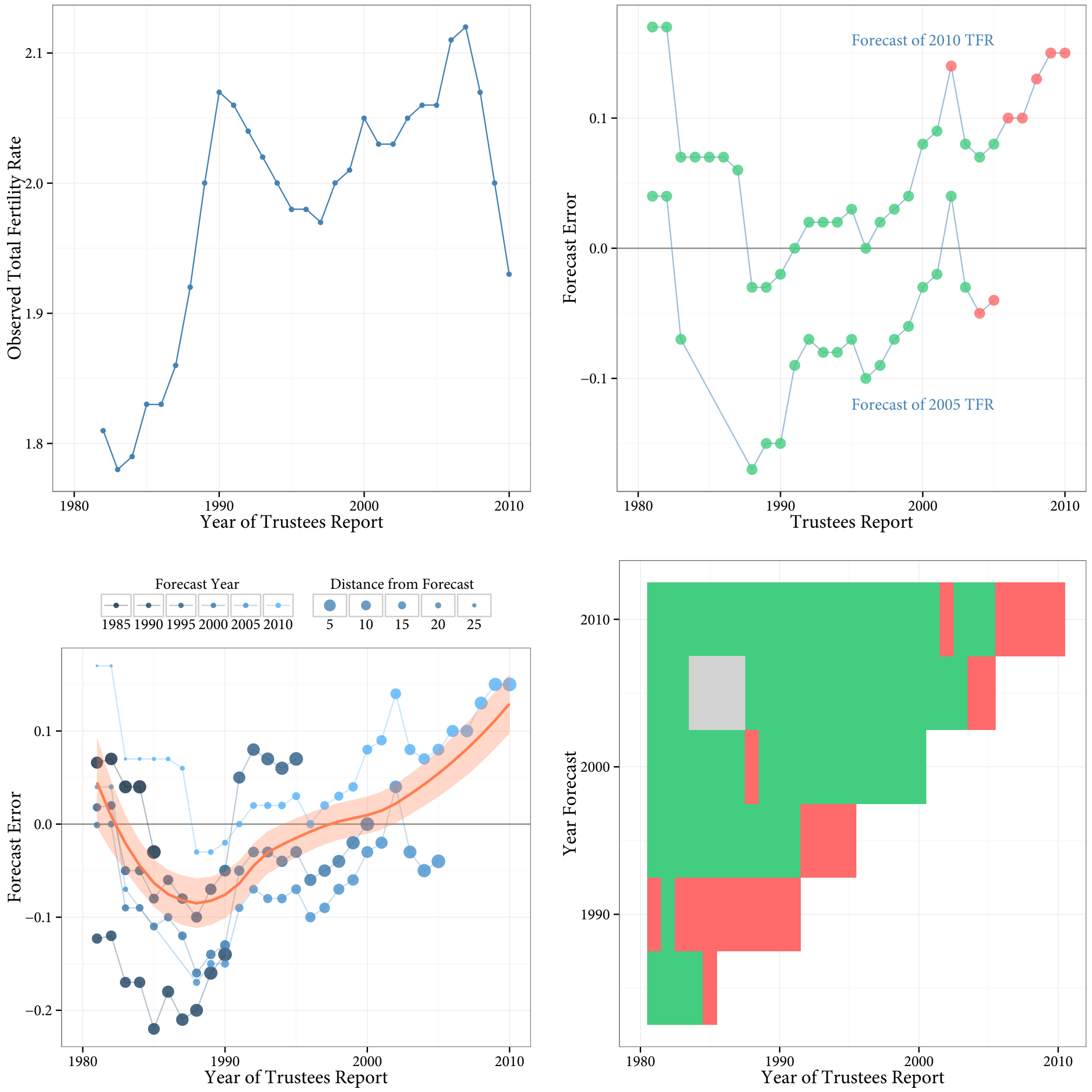


Figure B.7: This figure replicates the analyses in Figure B.6 by replacing fertility data from the Human Mortality Database with that from SSA.

B.5 Migration Forecast Errors

Our results for the analysis of net legal immigration forecasts appear in Figure B.8. For a brief discussion of this figure, see “A Note About Fertility & Immigration” in Section 1 of Chapter 4.

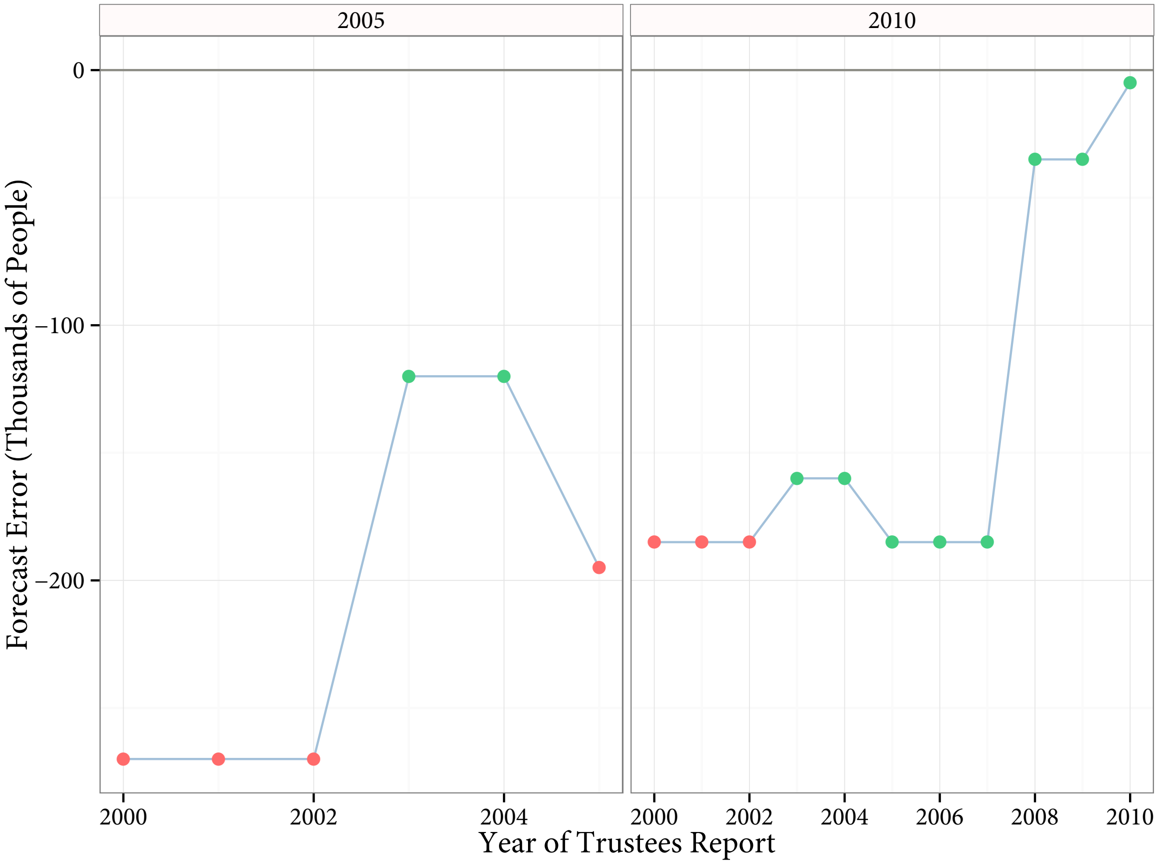


Figure B.8: Forecast Error of Legal Migration in 2005 (panel a) and 2010 (panel b) by Year of Trustees Report. Dots colored green when truth falls within SSA uncertainty intervals. Dots colored red when the truth falls outside SSA uncertainty intervals.

C
Text Reuse

This appendix provides additional information for the analyses in Chapter 6.

C.1 Bill Excerpts

VA H.B. 1331 (2010)

a proposal to outsource having a projected cost of more than 10 million in any fiscal year shall require **1** an initial business case analysis conducted by the state agency and submitted to the council the governor the president of the senate and the speaker of the house of delegates at least 60 days before a solicitation is issued the council shall evaluate the business case analysis and submit an advisory report to the state agency the governor the president of the senate and the speaker of the house of delegates when the advisory report is completed but at least 30 days before the agency issues the solicitation **2** a final business case analysis conducted by the state agency and submitted after the conclusion of any negotiations at least 30 days before execution of a contract to the council the governor the president of the senate and the speaker of the house of delegates

ALEC's Council on Efficient Government

a proposal to outsource having a projected cost of more than ten million dollars in any fiscal year shall require **1** an initial business case analysis conducted by the state agency and submitted to the council the governor the president of the senate and the speaker of the house of representatives at least sixty days before a solicitation is issued the council shall evaluate the business case analysis and submit an advisory report to the state agency the governor the president of the senate and the speaker of the house of representatives when the advisory report is completed but at least thirty days before the agency issues the solicitation **2** a final business case analysis conducted by the state agency and submitted after the conclusion of any negotiations at least thirty days before execution of a contract to the council the governor the president of the senate and the speaker of the house of representatives

Figure C.1: Comparison between VA H.B. 1331 (2010) and ALEC's Council on Efficient Government Act is presented below. Differences are highlighted in red.

VA H.B. 2314 (2010)

guidelines for scholarship foundations a a scholarship foundation as defined in section 58 1-439 25 and included on the list published annually by the department in accordance with the provisions of section 58 1-439 27 shall disburse annually at least 90 percent of its tax-credit-derived funds for qualified educational expenses through scholarships to eligible students

ALEC's School Tax Credit

Administrative Accountability Standards. All scholarship granting organizations shall: (5) ensure that at least 90 percent of their revenue from donations is spent on educational scholarships, and that all revenue from interest or investments is spent on educational scholarships;

Figure C.2: Comparison between VA H.B. 2314 (2010) and ALEC's School Tax Credit Act is presented below. Only overlap of greater than 2 words is highlighted in blue.

VA H.B. 10 (2010)	Freedom of Choice in Health Care
No resident of this Commonwealth, regardless of whether he has or is eligible for health insurance coverage under any policy or program provided by or through his employer, or a plan sponsored by the Commonwealth or the federal government, shall be required to obtain or maintain a policy of individual insurance coverage. No provision of this title shall render a resident of this Commonwealth liable for any penalty, assessment, fee, or fine as a result of his failure to procure or obtain health insurance coverage.	The people have the right to enter into private contracts with health care providers for health care services and to purchase private health care coverage. The legislature may not require any person to participate in any health care system or plan, nor may it impose a penalty or fine, of any type, for choosing to obtain or decline health care coverage or for participation in any particular health care system or plan.

Figure C.3: Comparison between VA S.B. 10 (2010) and ALEC's Freedom of Choice in Health Care Act is presented below. This is an example of highly rewritten ALEC-derived legislation that shares the same intent. Only overlapping bigram is highlighted in blue.

References

(2011). *ALEC: The Voice of Corporate Special Interests in State Legislatures*. People for the American Way Foundation.

(2012). *Amazon's Membership Demonstrates Corporate Control of ALEC Agenda*. The Center for Media and Democracy, Madison, WI.

(2015). Human Mortality Database. University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at `http://www.mortality.org` or `http://www.humanmortality.de` (data downloaded on January 31, 2015).

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Abowd, P. (2012). Alec gets a break from state lobbying laws. *Mother Jones*.

ALEC (1986a). Clearing the air: The environmental tobacco smoke debate. In *The State Factor*, volume 12. University of California, San Francisco: Legacy Tobacco Archives.

ALEC (1986b). Risk and the civil justice system: The crisis in tort law. Berkeley, CA: University of California, Berkeley Bancroft Library: People for the American Way Collection, Carton 6, Folder 16.

ALEC (2011). *The State Legislator's Guide: Tort Reform Boot Camp*. American Legislative Exchange Council, Washington, D.C.

Altman, N. J. (2005). *The Battle for Social Security: From FDR's Vision to Bush's Gamble*. John Wiley & Sons.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Pischke, J. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ansolabehere, S., de Figueiredo, J. M., and Snyder, J. M. (2003). Why is there so little money in u.s. politics? *Journal of Economic Perspectives*, 17(1):105–30.

Arceneaux, K., Gerber, A. S., and Green, D. P. (2006). Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14:37–62.

Arceneaux, K., Gerber, A. S., and Green, D. P. (2010). A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological Methods & Research*, 39:256–282.

Autor, D. H. and Duggan, M. G. (2006). The growth in the social security disability rolls: A fiscal crisis unfolding. *Journal of Economic Perspectives*, 20(3):71–96.

Balke, A. (1995). *Probabilistic counterfactuals: semantics, computation, and applications*. PhD thesis, University of California, Los Angeles.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92:1171–1176.

Banaji, M. R. and Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Random House LLC.

Barry, D., Kovaleski, S. F., Robertson, C., and Alvarez, L. (2012). Race, tragedy and outrage collide after a shot in florida. *The New York Times*.

Barrón-Cedeño, A., Vila, M., Martí, M. A., and Rosso, P. (2013). Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *Computational Linguistics*, 39(4):917–947.

Bartels, L. (2008). *Unequal Democracy: The Political Economy of the Gilded Age*. Princeton University Press, Princeton, NJ.

Baumgartner, F. and Jones, B. (1993). *Agendas and Instability in American Politics*. University of Chicago Press, Chicago, IL.

Beland, D. (2005). *Social Security: History and Politics from the New Deal to the Privatization Debate*. University Press of Kansas.

Beland, D. and Waddan, A. (2012). *The Politics of Policy Change: Welfare, Medicare, and Social Security Reform in the United States*. Georgetown University Press.

Blahous, III, C. P. (2007). Have the social security trustees been too conservative? A Presentation at the American Enterprise Institute.

Blahous, III, C. P. (2010). *Social Security: The Unfinished Work*. Hoover Institution Press.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Block, F. (1977). Block, fred. *Socialist Revolution*, 33:6–28.

Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F., Lin, W., and Bos, J. (1997). The Benefits and Costs of JTPA Title II-A Programs. *Journal of Human Resources*, 32:549–597.

Bloom, H. S., Orr, L. L., Cave, G., Bell, S., and Doolittle, F. (1993). The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months. Bethesda, MD.

Bottari, M. (2012). Alec in wisconsin: The hijacking of a state. *The Huffington Post*.

Bottari, M. and Fischer, B. (2013). Efforts to deliver 'kill shot' to paid sick leave tied to alec. *The Huffington Post*.

Bradley, D., Huber, E., Moller, S., Nielsen, F., and Stephens, J. D. (2003). Distribution and redistribution in postindustrial democracies. *World Politics*, 55:193–228.

Brady, H. E., Collier, D., and Seawright, J. (2006). Toward a pluralistic vision of methodology. *Political Analysis*, 14:353–368.

Broockman, D. (2012). The 'problem of preferences': Medicare and business support for the welfare state. *Studies in American Political Development*, 26:83–106.

Burden, B. C., Canon, D. T., Mayer, K. R., and Moynihan, D. P. (2014). Election laws, mobilization, and turnout: The unanticipated consequences of election reform. *American Journal of Political Science*, 58(1):95–109.

CBPP (2014). Policy basics: Where do our state tax dollars go?

Chalak, K. and Halbert, W. (2011). Viewpoint: An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics*, pages 1–51.

Clough, P. and Gaizauskas, R. (2009). Corpora and Text Re-use. In Lüdeling, A., Kytö, M., and McEnery, T., editors, *Handbook of Corpus Linguistics*, Handbooks of Linguistics and Communication Science, pages 1249–1271. Mouton de Gruyter.

Clough, P., Gaizauskas, R., Piao, S. S., and Wilks, Y. (2002). Measuring Text Reuse. *Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics*, pages 152–159.

CMD (2011). *ALEC Bills in Wisconsin*. The Center for Media and Democracy, Madison, WI.

Collier, D. and Brady, H. E. (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield, Lanham, MD.

Cox, D. R. (1960). Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society, Ser. B*, 22:172–176.

Cox, D. R. and Wermuth, N. (1995). Discussion of 'Causal diagrams for empirical research'. *Biometrika*, 82:688–689.

Culpepper, P. D. (2010). *Quiet Politics and Business Power: Corporate Control in Europe and Japan*. Cambridge University Press, New York, NY.

Dahl, R. A. (1961). *Who Governs? Democracy and Power in an American City*. Yale University Press, New Haven, CT.

Dale, A. and Strauss, A. (2009). Don't forget to vote: Text message reminders as a mobilization tool. *American Journal of Political Science*, 53:787–804.

Das, D. and Smith, N. A. (2009). Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proc. of ACL-IJCNLP*.

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Diamond, P. and Hausman, J. (1984). Individual retirement and savings behavior. *Journal of public economics*, 23(1-2):81–114.

Diamond, P. A. and Orszag, P. R. (2005). *Saving social security: A balanced approach*. Brookings Institution Press.

Domhoff, G. W. (2006). *Who Rules America?: Power and Politics, and Social Change*. McGraw-Hill, New York, NY.

Doolittle, F. and Traeger, L. (1990). *Implementing the National JTPA Study*. Manpower Demonstration Research Corporation.

Edwards III, G. C. (2007). *Governing by Campaigning: The Politics of the Bush Presidency*. Longman Publishing.

Elk, M. and Sloan, B. (2011). The hidden history of alec and prison labor. *The Nation*.

Feldstein, M. (1974). Social security, induced retirement, and aggregate capital accumulation. *Journal of political economy*, 82(5):905–926.

Ferguson, T. (1995). *Golden Rule: The Investment Theory of Party Competition and the Logic of Money-Driven Political Systems*. University of Chicago Press, Chicago, IL.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.

Fitzgerald, M. (2005). Greater convenience but not greater turnout: The impact of alternative voting methods on electoral participation in the united states. *American Politics Research*, 33:842–867.

Fraker, T. and Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, 22:194–227.

George, A. and Bennett, A. (2005). *Case studies and theory development in the social sciences*. Mit Press.

Gerber, A. S. and Green, D. P. (2005a). *The Annals of the American Academy of Political and Social Science*, 99(2):142–154.

Gerber, A. S. and Green, D. P. (2005b). Correction to gerber and green (2000), replication of disputed findings, and reply to imai (2005). *The American Political Science Review*, 601:301–313.

Gerber, A. S., Green, D. P., Kaplan, E. H., and Kern, H. L. (2010). Baseline, placebo, and treatment: Efficient estimation for three-group experiments. *Political Analysis*, 18:297–315.

Gilbert, D. T. (1998). Ordinary psychology. In Gilbert, D. T., Fiske, S. T., and Lindzey, G., editors, *The Handbook of Social Psychology*, volume 2, pages 89–150. McGraw Hill, New York.

Gilens, M. (2012). *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton University Press, Princeton, NJ.

Girosi, F. and King, G. (2008). *Demographic Forecasting*. Princeton University Press, Princeton. http://gking.harvard.edu/files/smooth/.

Glynn, A. and Quinn, K. (2011). Why Process Matters for Causal Inference. *Political Analysis*, 19(3):273–286.

Gramlich, E. (1996). Different approaches for dealing with social security. *Journal of economic perspectives*, 10(3):55–66.

Granados, J. A. T. (2005). Increasing mortality during the expansions of the us economy, 1900–1996. *International Journal of Epidemiology*, 34(6):1194–1202.

Graves, L. (2012a). *Backgrounder: the History of the NRA/ALEC Gun Agenda*. The Center for Media and Democracy, Madison, WI.

Graves, L. (2012b). *Buying Influence: How the American Legislative Exchange Council Uses Corporate-Funded 'Scholarships' to Send Lawmakers on Trips with Corporate Lobbyists*. Common Cause, The Center for Media and Democracy, and DBA Press.

Green, D. P., McGrath, M. C., and Aronow, P. M. (2013). Field experiments and the study of voter turnout. *Journal of Elections, Public Opinion and Parties*, 23(1):37–62.

Gronke, P., Galanes-Rosenbaum, E., and Miller, P. (2007). Early Voting and Turnout. *PS: Political Science and Politics*, XL.

Gronke, P., Galanes-Rosenbaum, E., Miller, P. A., and Toffey, D. (2008). Convenience voting. *Annual Review of Political Science*, 11:437–455.

Gronke, P. and Stewart, C. (2013). Early Voting in Florida. Paper presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Gronke, P. and Toffey, D. K. (2008). The psychological and institutional determinants of early voting. *Journal of Social Issues*, 64(3):503–524.

Hacker, J. S. and Pierson, P. (2002). Business power and social policy: Employers and the formation of the american welfare state. *Politics and Society*, 30(2):277–325.

Hacker, J. S. and Pierson, P. (2010). *Winner-Take-All Politics: How Washington Made the Rich Richer–and Turned Its Back on the Middle Class*. Simon & Schuster, New York, NY.

Hainmueller, J. and Hazlett, C. (2014). Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. *Political Analysis*, 22:143–168.

Hall, R. L. and Deardorff, A. V. (2006). A new state-level panel of annual inequality measures over the period 1916 - 2005. *The American Political Science Review*, 1001(1):69–84.

Hall, R. L. and Wayman, F. W. (1990). Buying time: Moneyed interests and the mobilization of bias in congressional committees. *The American Political Science Review*, 84(3):797–820.

Hanmer, M. J. (2009). *Discount Voting: Voter Registration Reforms and Their Effects*. Cambridge University Press.

Harris, S. (1941). *Economics of Social Security*. McGraw-Hill, New York.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66:1017–1098.

Heckman, J., Ichimura, H., and Todd, P. (1997). Matching as an econometric evaluation estimator evidence from evaluating a job training program. *Review of Economic Studies*, 64:605–654.

Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The Economics and Econometrics of Active Labor Market Programs. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics, Volume III*. Elsevier Science North-Holland.

Heckman, J. J. and Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, 9(2):85–110.

Heckman, J. J. and Smith, J. A. (1999). The pre-programme earnings dip and the determinants of participation in a social programme: implications for simple programme evaluation strategies. *Economic Journal*.

Hellerstein, J. K., Neumark, D., and Troske, K. R. (1999). Wages, productivity, and worker characteristics: Evidence from plant-level production functions and wage equations. *Journal of Labor Economics*, 37(3):409–446.

Herron, M. C. and Smith, D. A. (2014). Race, party, and the consequences of restricting early voting in florida in the 2012 general election. *Political Research Quarterly*.

Hertel-Fernandez, A. (2014a). Corporate Interests and Conservative Mobilization Across the U.S. States.

Hertel-Fernandez, A. (2014b). Who passes business's 'model bills'? policy capacity and corporate influence in the u.s. states. *Perspectives on Politics*, 12(3):582–602.

Hoffman, J. (2012). Colorofchange.org and advocacy: The alec campaign. *Nonprofit Quarterly*.

Hotz, J. V. (1992). Designing an evaluation of the job training partnership act. In Manski, C. F. and Garfinkel, I., editors, *Evaluating Welfare and Training Programs*. Harvard University Press.

Huberty, M. (2013). Applying Natural Language Processing for Computer Assisted Analysis of Legislative History: The LegHist Package for R. Working Paper.

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19.

Imbens, G. and Rubin, D. (1995). Discussion of 'Causal diagrams for empirical research'. *Biometrika*, 82:694–695.

Joffe, M. M. (2001). Using information on realized effects to determine prospective causal effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 759–774.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kashin, K., King, G., and Soneji, S. (2015a). Replication data for: Explaining systematic bias and nontransparency in u.s. social security administration forecasts. UNF:6:967llFHgiywsHWWp1cVg9A== http://dx.doi.org/10.7910/DVN/28323 Harvard Dataverse [Distributor] V1 [Version].

Kashin, K., King, G., and Soneji, S. (2015b). Replication data for: Systematic bias and nontransparency in u.s. social security administration forecasts. UNF:5:10erGFXQ0Bu9bcMFU5/t2A== http://dx.doi.org/10.7910/DVN/28122 Harvard Dataverse [Distributor] V1 [Version].

Kasper, M. (2013). How a powerful group of corporations quietly tried to roll back clean energy standards, and failed miserably. *Think Progress: Climate Progress*.

Kaufman, S., Kaufman, J. S., and MacLehose, R. F. (2009). Analytic bounds on causal risk differences in directed acyclic graphs with three observed binary variables. *Journal of Statistical Planning and Inference*, 139:3473–87.

Keele, L. and Minozzi, W. (2013). How much is minnesota like wisconsin? assumptions and counterfactuals in causal inference with observational data. *Political Analysis*, 21(2):193–216.

Kelly, N. J. and Witko, C. (2012). Federalism and american inequality. *Journal of Politics*, 74(2):414–426.

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, 28(3):443–499. http://j.mp/jCyfF1.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press, 1 edition.

King, G. and Soneji, S. (2011). The future of death in america. *Demographic Research*, 25(1):1–38. http://j.mp/iXUpBv.

King, G. and Zeng, L. (2004). Inference in case-control studies. In Chow, S.-C., editor, *Encyclopedia of Biopharmaceutical Statistics*. Marcel Dekker, New York, 2nd edition. http://gking.harvard.edu/files/abs/1s-enc-abs.shtml.

Klarner, C. (2013). State Partisan Balance Data. `http://www.indstate.edu/polisci/klarnerpolitics.htm`. Accessed 03/10/2013.

Korenman, S. and Neumark, D. (1991). Does marriage really make men more productive? *The Journal of Human Resources*, 26(2):282–307.

Kuroki, M. and Miyakawa, M. (1999). Identifiability Criteria for Causal Effects of Joint Interventions. *J. Japan Statist. Soc.*, 29(2):105–117.

Lafer, G. (2013). The legislative attack on american wages and labor standards, 2011–2012.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620.

Laursen, E. (2012). *The People's Pension: The Struggle to Defend Social Security Since Reagan*. AK Press.

Layzer, J. (2012). *Open for Business: Conservatives' Opposition to Environmental Regulation*. MIT Press, Cambridge, MA.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lee, R. D. and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419):659–675.

Lessig, L. (2011). *Republic, Lost: How Money Corrupts Congress - and a Plan to Stop It*. Twelve, New York, NY.

Lindblom, C. (1977). *Politics and markets: the world's political economic systems*. Basic Books, New York, NY.

Lu, J. and Wong, W. (2011). Mortality improvement in the usa: Analysis, projections and extreme scenarios. Technical report, Society of Actuaries, Schaumburg, IL.

Malewitz, J. (2013). Renewable energy incentives survive lobby attack. *Stateline*.

Manski, C. F. (2013). *Public policy in an uncertain world: analysis and decisions*. Harvard University Press.

McDonald, M. P. and Popkin, S. L. (2001). The Myth of the Vanishing Voter. *American Political Science Review*, 95:963–974.

McIntire, M. (2012). Conservative non-profit acts as a stealth business lobbyist. *The New York Times*.

Meyerson, N. (2014). Social security: What would happen if the trust funds ran out? Technical report, Congressional Research Service, Washington, DC.

Mills, C. W. (2000). *The Power Elite*. Oxford University Press, New York, NY.

Mitchell, N. (1997). *The Conspicuous Corporation: Business, Public Policy, and Representative Democracy*. University of Michigan Press, Ann Arbor, MI.

Nawab, R. M. A., Stevenson, M., and Clough, P. (2010). Detecting Text Reuse with Modified and Weighted N-grams. University of Sheffield lab report for PAN at CLEF 2010.

Nichols, J. (2011). Alec exposed. *The Nation*.

Nickerson, D. W. (2006). Volunteer phone calls can increase turnout: Evidence from eight field experiments. *American Politics Research*, 34:271–292.

Nickerson, D. W. (2007). Quality is job one: Volunteer and professional phone calls. *American Journal of Political Science*, 51:269–282.

Nickerson, D. W., Friedrichs, R. K., and King, D. C. (2006). Partisan mobilization campaigns in the field: Results from a statewide mobilization campaign in michigan. *Political Research Quarterly*, 59:85–97.

Office of the Chief Actuary (2012). The long-range demographic assumptions for the 2012 trustees report. Technical report, Social Security Administration. http://j.mp/OCACT12.

Office of the Chief Actuary (2013). The long-range demographic assumptions for the 2013 trustees report. Technical report, Social Security Administration. http://j.mp/OCACT13.

Office of the Chief Actuary (2014). The long-range demographic assumptions for the 2014 trustees report. Technical report, Social Security Administration. http://j.mp/OCACT14.

Orr, L. L., Bloom, H. S., Bell, S. H., Lin, W., Cave, G., and Doolittle, F. (1994). The National JTPA Study: Impacts, Benefits, And Costs of Title IIA. Bethesda, MD.

Panagopoulos, C. (2011). Timing is everything? primacy and recency effects in voter mobilization campaigns. *Political Behavior*, 33:79–93.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition.

Pilkington, E. and Goldenberg, S. (2013). Alec facing funding crisis from donor exodus in wake of trayvon martin row. *The Guardian*.

Powell, L. W. (2012). *The Influence of Campaign Contributions in State Legislatures: The Effects of Institutions and Politic*. University of Michigan Press, Ann Arbor, MI.

Primo, D. M., Jacobmeier, M. L., and Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, 7:446–459.

Ramsahai, R. (2012). Supplementary variables for causal estimation. In Berzuini, C., Dawid, A., and Bernardinelli, L., editors, *Causal Inference: Statistical Perspectives and Applications*. Wiley and Sons.

Rosenblatt, R. and DeWitt, L. (2005). The role of social security's chief actuary. *Contingencies*, pages 40–45. http://j.mp/ChiefA.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data: The fisher randomization test," by d. basu. *Journal of the American Statistical Association*, 75:591–593.

Rubin, D. B. (2010). Reflections stimulated by the comments of shadish (2010) and west and thoemmes (2010). *Psychological Methods*, 15(1):38–46.

Ruhm, C. J. (2000). Are recessions good for your health? *The Quarterly Journal of Economics*, 115(2):617–650.

Samwick, A. (1999). Social security reform in the united states. *National Tax Journal*, 52(4):819–842.

Shaw, C. (2007). Fifty years of united kingdom national population projections: how accurate have they been? *Population Trends*, 128:8–23.

Shiner, M. (2013). After trayvon martin verdict, durbin pushes 300 companies on 'stand your ground' laws. *Roll Call*.

Shor, B. and McCarty, N. (2011). The ideological mapping of american legislatures. *The American Political Science Review*, 105(3):530–51.

Shpitser, I. and Pearl, J. (2006). Identification of conditional interventional distributions. Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI).

Smith, D. A., Cordell, R., and Dillon, E. M. (2013). Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers. *IEEE Workshop on Big Data and the Humanities*.

Smith, J. A. (1994). Sampling Frame for the Eligible Non-Participant Sample. *Mimeo*.

Smith, M. (1999). Public opinion, elections, and representation within a market economy: Does the structural power of business undermine popular sovereignty? *American Journal of Political Science*, 43(3):842–863.

Smith, T. W. and Kim, J. (2010). *Paid Sick Days: Attitudes and Experiences*. NORC/University of Chicago.

Social Security Advisory Board Technical Panel (2003). 2003 technical panel on assumptions and methods. Technical report, Social Security Advisory Board. http://j.mp/SSATech03.

Social Security Advisory Board Technical Panel (2007). 2007 technical panel on assumptions and methods. Technical report, Social Security Advisory Board. http://j.mp/SSATech07.

Soneji, S. and King, G. (2012). Statistical security for social security. *Demography*, 49(3):1037–1060. http://j.mp/Qvla7N.

Squire, P. (2007). Measuring state legislative professionalism: The squire index revisited. *State Politics & Policy Quarterly*, 7(2):211–227.

Stewart, C. (2012). Declaration of dr. charles stewart iii. State of Florida vs. United States of America.

Stuckler, D., Meissner, C., Fishback, P., Basu, S., and McKee, M. (2011). Banking crises and mortality during the great depression: evidence from us urban populations, 1929–1937. *Journal of Epidemiology and Community Health*.

Sullivan, L. (2010). Prison economics help drive ariz. immigration law. *NPR Morning Edition*.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108.

Taylor, C. (2013). The rise and fall of alec nation. *The Journal Sentinel*.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Princeton.

Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence*, pages 567–573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Tian, J. and Pearl, J. (2002b). On the identification of causal effects. In *Proceedings of the American Association of Artificial Intelligence*.

Trumbull, J. G. (2012). *Strength in Numbers: The Political Power of Weak Interests*. Harvard University Press, Cambridge, MA.

VanderWeele, T. and Robins, J. (2009). Signed directed acyclic graphs for causal inference. *JR Stat Soc B*.

VanderWeele, T. J. (2008). The sign of the bias of unmeasured confounding. *Biometrics*, 64(3):702–706.

VanderWeele, T. J. (2009). On the relative nature of overadjustment and unnecessary adjustment. *Epidemiology*, 20(4):496–499.

VanderWeele, T. J. and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, 22(1):42–52.

Vogel, D. (1989). *Fluctuating Fortunes: The Political Power of Business in American Politic*. Beard Books, Washington, DC.

Weaver, R. K. (1988). *Automatic government: The politics of indexation*. Brookings Institution Press.

Weinstein, A. (2012). How the nra and its allies helped spread a radical gun law nationwide. *Mother Jones*.

Wilkerson, J., Smith, D., and Stramp, N. (2015). Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science*. doi: 10.1111/ajps.12175.

Williams, E. and Johnson, N. (2013). *ALEC Tax and Budget Proposals Would Slash Public Services and Jeopardize Economic Growth*. Center on Budget and Policy Priorities.

Wilson, T. D. and Brekke, N. (1994). Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin*, 116(1):117.

Winship, C. and Harding, D. (2008). A Mechanism-Based Approach to the Identification of Age-Period-Cohort Models. *Sociological Methods & Research*, 36(3):362.

Wolfinger, R. E., Highton, B., and Mullin, M. (2005). How postregistration laws affect the turnout of citizens registered to vote. *State Politics & Policy Quarterly*, 5:1–23.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176.