Essays on Causal Inference and the International Investment Regime

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

Citable link: http://nrs.harvard.edu/urn-3:HUL.InstRepos:40050137

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
Abstract

In the last several decades, the growth of bilateral investment treaties (BITs) and the incorporation of investment provisions into regional trade agreements has led to the development of an international legal regime defining the rules states must follow with respect to investments made by foreign nationals. Among the most controversial elements of this regime are investor-state dispute settlement (ISDS) clauses, treaty provisions which allow foreign investors to initiate arbitration directly against states for alleged breaches of the treaty. ISDS proceedings are typically held outside of any national court and the arbitrators are selected from a small cadre of elite international legal professionals. Debates over the effectiveness of investment law and the legitimacy of ISDS raise a number of important empirical questions, such as whether investment treaties are actually effective in increasing foreign direct investment, whether developing countries are systematically disadvantaged by investor-state arbitration, and whether the arbitrators deciding these disputes are truly impartial or are influenced by extralegal factors.

The papers in this dissertation develop novel methodological tools in order to answer these three causal questions. The first paper evaluates the effect of bilateral investment treaties on foreign direct investment by United States firms using a differences-in-differences approach. It introduces a new differences-in-differences estimator for situations where outcomes are observed for more than two time periods. It shows how the conventional two-way fixed effects estimator yields misleading results when treatment effects persist over time and develops a new inverse propensity score weighting estimator that allows researchers to adjust for covariates without the need for restrictive assumptions on the time periods that can be affected by treatment. The second paper explores the problem of non-random attrition as applied to studies of legal disputes. It explains how analyses of dispute outcomes
will be misleading when the treatment of interest also affects the propensity of a case to settle early and drop out of the data. Building on the “principal stratification” framework, it develops a robust weighting method for adjusting for the bias due to attrition that is explained by observed covariates. It applies this approach to explain why richer countries appear to win more investor-state disputes and shows that most of the gap in observed win-rates can be explained by the fact that less wealthy countries are more likely to settle disputes early, which affects the distribution of cases that receive a final ruling. The third paper (co-authored with Sergio Puig) shows how experimental designs can supplement observational research for questions that cannot be answered from available observational data. Since investor-state arbitration mechanisms typically permit each party to appoint an arbitrator, questions have been raised about the impartiality of these party-appointed arbitrators. Unfortunately, because the vast majority of arbitral decisions are unanimous, evidence for bias is difficult to infer from case data alone. This paper presents results from a survey experiment of 257 arbitrators and arbitration experts in which respondents were asked to decide a hypothetical investor-state dispute. It finds that when arbitrators were told that they were appointed by one of the two parties to the dispute, they gave more favorable decisions to that party. These results suggest that professional arbitrators likely exhibit meaningful biases in their decision-making due to the party appointment mechanism.
## Contents

1 Introduction ................................................. 1

  2.1 Introduction ........................................... 11
  2.2 Setup and Theoretical Framework ....................... 16
  2.3 Bias from OLS with unit/time fixed effects ............ 25
  2.4 Inverse propensity weighting estimators for multi-period DID effects .......... 30
  2.5 Application: The effect of investment treaties on U.S. foreign direct investment ... 34
  2.6 Conclusion .............................................. 42

3 Why Rich Countries Win Investment Disputes ................. 45
  3.1 Introduction ........................................... 45
  3.2 Settlement and Arbitration ............................. 53
  3.3 Covariate adjustment for estimating treatment effects ........................................ 55
  3.4 Treatment effects under selective attrition .......... 61
  3.5 Data .................................................... 72
  3.6 Results .................................................. 79
  3.7 Conclusion .............................................. 90

4 Affiliation Bias in Arbitration: An Experimental Approach .... 93
  4.1 Introduction ........................................... 93
  4.2 Background ............................................ 96
  4.3 Evidence ............................................... 98
  4.4 Blind Appointments in Arbitration: Possibilities and Challenges ................. 118
  4.5 Conclusion .............................................. 122

Appendix A Supporting Materials for Chapter 2 .................. 123
Acknowledgments

Graduate school can be a very lonely and isolating experience. What has kept me going over the last six years has been the support of a community of scholars and good friends. I am deeply thankful for the wonderful cohort of fellow graduate students with whom I have shared this experience. They have been exceptional professional colleagues and brilliant co-authors and collaborators. My work has benefited immensely from discussions in workshops and from so many casual chats in the CGIS hallways. They have also formed the core of my social circle in Cambridge, without which this entire process likely would not have been possible. In particular, I would like to thank Pamela Ban, Peter Bucchianeri, Cosette Creamer, Dana Higgins, Connor Huff, Samuel Imlay, Daniel de Kadt, Mayya Komisarchik, Dominika Kruszewska, Christopher Lucas, Stephen Pettigrew, Melissa Sands, Jason Sclar, Kai Thaler, and Ariel White.

It is unlikely that I would have gotten to this point without the help of the members of my dissertation committee: Gary King, Beth Simmons, and Dustin Tingley. Working with each of them has been one of the highlights of the graduate school experience and I thank them for all of the time they spent reading and commenting on my work. They have been essential in helping me understand the broader picture of my research agenda and have kept me from getting stuck in the weeds for too long. I would especially like to thank them for their patience as I worked through the process of collecting the many disparate ideas contained in this dissertation and turning them into a written product. I have benefited greatly from interactions with so many of the faculty members in this department. I thank Matthew Blackwell, Jeff Frieden, Iain Johnston, Josh Kertzer, Adam Glynn, and Arthur Spirling for their invaluable feedback over the years.
I am thankful for the institutional and financial support provided by the Institute for Quantitative Social Sciences and the Weatherhead Center for International Affairs at Harvard. I am also grateful to be a part of an extended network of colleagues beyond Harvard. My undergraduate advisors Daniel Hopkins and Erik Voeten, who helped motivate my decision to go to graduate school, have remained an invaluable part of the entire graduate school process. Additional thanks go to my colleagues and collaborators Ryan Brutger and Teppei Yamamoto. I also owe a great deal to Sergio Puig who was instrumental in getting the study in Chapter 4 of this dissertation out and published.

Finally, none of this would have been possible without the unwavering support of my parents, Vladimir and Nataliya. Thank you.
Chapter 1

Introduction

The last several decades have seen radical changes in the structure of the international economy. Globalization has facilitated an expansion of cross-border economic flows. In addition to growth in trade, multinational enterprises (MNEs) are contributing to a global expansion of foreign direct investment (FDI). In contrast to trade flows, which are arms-length transactions between entities in two different economies, FDI involves a firm in one economy – referred to as the “home” economy or country – acquiring a direct interest in an affiliate in another economy – the “host” country. The main distinguishing element of FDI relative to other forms of passive investment is the intent of the home country entity to establish a “lasting interest” in the host country firm and to exercise some degree of management control (Organisation for Economic Co-operation and Development, 2008).

Since 1996, the world value of foreign investors’ net FDI positions abroad – the “outward stock” of FDI – has grown from roughly 15% of global gross domestic product (GDP) to around 35% in 2016 according to data from the UN Conference on Trade and Development (UNCTAD) (Figure 1.1). The vast majority of these investment flows originate from firms in wealthy, advanced industrial economies. However, both advanced and emerging markets have been recipients of FDI, particularly as developing countries have increasingly loosened restrictions on foreign ownership and investment over time (Pandya, 2014). Since FDI flows are seen as an important mechanism for technology trans-
fer, economic growth and development, many emerging market countries have made the promotion of foreign direct investment a policy priority (Ozawa, 1992).

![Figure 1.1: Inward and Outward FDI stocks as % of GDP](image)

**Notes:** Data from the UN Conference on Trade and Development (UNCTAD) Country groupings defined by World Bank income classifications.

Historically, investors have often been reticent to make direct investments abroad, particularly in emerging markets, due to issues of political risk. While governments have strong motivations to attract investors by promising secure property rights, once an investor has decided to undertake a particular project, the government may face new incentives to extract value from the investment. For many multinationals in the 1960s and 70s, this could mean nationalization and expropriation of the investment with minimal to no compensation (Kobrin, 1984). Even in the absence of explicit expropriation threats, governments may alter domestic policies in ways that make the investment no longer profitable. In the classic framework of Vernon (1971), any initial bargain to protect a foreign investor’s property rights struck ex ante is unenforceable without an external constraint and becomes “obsolete” ex post. Political risk, particularly in countries with weak constraints on government action, remains a significant obstacle to FDI to emerging markets especially when investors are unable to hedge risks.
In light of the challenges of cross-border investment governance, an international regime has emerged to define the rules that states should follow with respect to foreign investors. However, in contrast to the international legal regime governing trade, the investment regime has grown in a decentralized manner. There does not yet exist a single multilateral body for investment that is comparable to the World Trade Organization (WTO) in the realm of trade (Simmons, 2014). Instead, internationally binding rules defining investors’ rights have been implemented on a bilateral basis over time. Concomitant with the growth in FDI has been a similar expansion of bilateral investment treaties (BITs). These treaties, signed between pairs of states, obligate each party to grant investors who are nationals of the other party certain rights with respect to their investments. While BITs have evolved over time, the core provisions common to almost all investment agreements obligate states to refrain from expropriation without adequate compensation, to not discriminate against foreign investors relative to domestic investors (national treatment) and investors of other nationalities (most-favored nation treatment), to treat foreign investors fairly and equitably, to refrain from arbitrary measures and to grant foreign investments full protection and security (Salacuse, 2015).

BITs spread rapidly during the last several decades as developing countries, under pressure to attract highly mobile international capital, sought to adopt policies to increase their relative attractiveness to foreign firms (Elkins, Guzman and Simmons, 2006). Likewise, capital exporting countries saw BITs as a useful tool for protecting the investments of their nationals abroad and leveraged their bargaining power into binding investor-friendly commitments from their partners (Simmons, 2014; Allee and Peinhardt, 2014). As a result of demand from both capital exporters and importers, the number of BITs grew significantly during the 1990s, as shown in Figure 1.2. Moreover, in recent years, international investment law has also developed through preferential trade agreements (PTAs) as modern trade agreements incorporate investment protection provisions similar to those found in BITs (Büthe and Milner, 2014). With many developing countries having at least one BIT in force, this bilateral network of reciprocal investment protection obligations has evolved a de-facto legal regime defining the rules for treating foreign investors.
These legal commitments to protect investments have not been without controversy, with many international law scholars warning of a “backlash” against the regime (Waibel, 2010). Some countries have already sought to renegotiate or terminate past bilateral investment treaties. For example, the Ecuadorian government formally denounced its sixteen BITs in June of 2017 and the Indonesian government declared in March of 2014 that it would allow its current BITs to expire and renegotiate new agreements. Driving much of this backlash are the provisions for investor-state dispute settlement (ISDS) that are contained in many BITs. These ISDS clauses allow investors to bring claims directly against states for violations of treaty provisions. These arbitrations are conducted in an international legal forum, located outside of the host country’s legal system, and are decided by a panel of ad-hoc


2 See “Indonesia to terminate more than 60 bilateral investment treaties.” Financial Times. March 26, 2014. https://www.ft.com/content/3755c1b2-b4e2-11e3-af92-00144feabdc0.
arbitrators selected in part by the parties to the dispute.

Decisions rendered by these arbitral tribunals have meaningful consequences for both parties involved. Awards in favor of firms for violations of property rights can amount to sizeable portions of state budgets. One notable decision, *Occidental v. Republic of Ecuador*, awarded claimants $1.76 billion USD plus interest in damages – the largest award in ICSID history (Sabahi and Duggal, 2013).\(^3\)

And as Simmons (2014) points out, states that ratify BITs are very likely to face claims from investors. As the network of BITs has expanded, so too has the number of investment disputes initiated in a given year (Figure 1.3).\

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure13.png}
\caption{Number of treaty-based investor-state disputes initiated in each year}
\end{figure}

These disputes have provoked an adverse reaction among governments and publics. While early arbitrations dealt primarily with questions of expropriation and breach of contract, arbitrators are increasingly being asked to rule on questions regarding governments’ right to regulate as investors have be-

\(^3\)For comparison, according to the World Bank’s World Development Indicators, Ecuador’s nominal GDP in $USD in the year of the award – 2012 – was $87.6 billion.
gun using BITs to challenge health, environmental and other public interest regulations on the grounds that they affect the value of the firms’ investment (Pelc, 2017). These “indirect expropriation” claims have grown over time, and although they are more often than not unsuccessful, there are growing concerns regarding the possibility of a “regulatory chill” resulting from the mere threat of costly litigation (Janeba, 2016). Domestic audiences react negatively to the prospect of their governments being exposed to litigation from foreign actors (Brutger and Strezhnev, 2018), particularly as these claims are increasingly being brought against those wealthy advanced economies that advocated for BITs and ISDS in the first place.\footnote{One recent example is the Vattenfall v. Germany case in which a Swedish power company brought an arbitration claim against the German government after it announced it would phase out its nuclear power plants in the wake of the nuclear accident in Fukushima, Japan. This dispute sparked significant public protest over ISDS and in particular its inclusion in potential future EU trade agreements under consideration such as TTIP. EU trade commissioner Cecilia Malmström remarked that “ISDS is now the most toxic acronym in Europe.” See Ames, Paul. “ISDS: The most toxic acronym in Europe.” Politico E.U. September 17, 2015. https://www.politico.eu/article/isds-the-most-toxic-acronym-in-europe/.

For example, in a recent Op-Ed, U.S. Senator Elizabeth Warren argued “ISDS...wouldn’t employ independent judges. Instead, highly paid corporate lawyers would go back and forth between representing corporations one day and sitting in judgment the next...If you’re a lawyer looking to maintain or attract high-paying corporate clients, how likely are you to rule against those corporations when it’s your turn in the judge’s seat?” See: Elizabeth Warren, February 25, 2015 “The Trans-Pacific Partnership clause everyone should oppose.” The Washington Post.}

In addition to criticizing the sovereignty costs of ISDS, many political leaders have also attacked the impartiality of ISDS adjudicators themselves, arguing that it favors the interests of multinational corporations over those of states.\footnote{For example, in a recent Op-Ed, U.S. Senator Elizabeth Warren argued “ISDS...wouldn’t employ independent judges. Instead, highly paid corporate lawyers would go back and forth between representing corporations one day and sitting in judgment the next...If you’re a lawyer looking to maintain or attract high-paying corporate clients, how likely are you to rule against those corporations when it’s your turn in the judge’s seat?” See: Elizabeth Warren, February 25, 2015 “The Trans-Pacific Partnership clause everyone should oppose.” The Washington Post.} More generally, Trakman (2013) notes that “an underlying concern among some developing states...is that the ICSID was established by, and arguably in the interest of, wealthy countries and their investors abroad.” (606). Arbitral panels are constituted on an ad-hoc basis and the arbitration system has at its core a “market” of highly trained arbitration professionals who specialize in adjudicating such disputes (Rogers, 2004; Puig, 2014). Arbitrators do not retain permanent positions on tribunals – each arbitration panel is set up independently by the parties to a dispute and in accordance with the chosen set of arbitration rules. Supporters of the system argue that professional norms constrain overt expressions of favoritism and bias and encourage impartiality. Brower and Schill (2008) argue “arbitrators are impartial and independent dispute resolvers who ... are subject
to a number of mechanisms that can prevent private interests from taking precedence over public interests” (489) suggesting that arbitrators have incentives to cultivate reputations for impartiality in order to secure future appointments.

Nevertheless, there exist serious concerns over whether extralegal factors influence arbitral decisionmaking and what the implications are for the legitimacy of the ISDS regime. Anecdotal evidence from arbitrators suggests that arbitrators are uniquely conscious of the preferences of the party that appointed them. One particularly explicit example where a party’s promise of extralegal costs was acknowledged to have influenced an arbitrator comes from a story from Judge Abner Mikva who served as the United States’ apointee on the Loewen v. United States arbitration As described in Schneiderman (2010), the arbitration was one of the first to arise out of ISDS provisions contained in the North American Free Trade Agreement (NAFTA). Mikva recounted that after his appointment, he was explicitly warned by United States Department of Justice (DOJ) officials that a ruling against the United States could jeopardize NAFTA. While Mikva’s co-arbitrators were described as leaning towards the claimant, the case was ultimately decided in favor of the United States on jurisdictional grounds.

With so many potentially adverse consequences from investment claims, one might ask why states bother ratifying investment treaties in the first place. While many governments have expressed interest in renegotiating BITs in the wake of experience with ISDS (Haftel and Thompson, 2018), there has yet to be a wholesale abandonment of BITs, suggesting that states consider the general principles of the investment regime to be valuable for promoting foreign investment despite having concerns about investor-state dispute mechanisms. Unfortunately, finding evidence for the effect of BITs on actual FDI promotion remains a significant research challenge.

Many of the biggest questions surrounding the effectiveness and legitimacy of the international investment law are fundamentally empirical ones. This dissertation takes as its starting point three important causal questions in the study of the international investment regime: whether bilateral investment treaties increase foreign direct investment, whether ISDS tribunals are systematically biased against developing countries, and whether arbitrators themselves make different decisions based on the party that appoints them. In each of these cases, I argue that existing methodological tools are insuffi-
cient to provide good answers and can, in some cases, yield misleading inferences. Therefore, each of the chapters in this dissertation develops a new method or strategy for drawing valid inferences about each of these causal puzzles.

Chapter 2 considers the problem of inference in observational panel data, where researchers have repeated observations of units over time. Attempts to estimate the effect of BITs on FDI in this setting often rely on difference-in-differences (DID) identification strategies to adjust for possible unobserved factors that are associated with both BIT adoption and FDI. In settings with many time periods, these DID effects are estimated using two-way fixed effect (FE) regression models that include parameters for each unit and time period under analysis. I show that these models are biased for the true effect of treatment unless researchers assume that treatment effects do not persist beyond a single time period. Because of the speed at which large investment decisions are made, treaty effects are unlikely to manifest immediately. Therefore, two-way FE estimators will tend to underestimate the true effect of any treaty. I develop a new method for estimating DID effects, the “generalized DID estimator,” which avoids biases from the two-way fixed effects modeling approach while still allowing researchers to incorporate covariate adjustments through a straightforward weighting method. I apply this method to estimate the effect of adopting an investment treaty with the United States on the number of affiliate firms located in that country with U.S. parents, a measure of the extensive margin of bilateral FDI. I find that while the two-way FE estimator yields effect estimates statistically indistinguishable from 0, the generalized DID method suggests that BIT adoption does in fact boost U.S. multinational activity in a given country.

Chapter 3 shifts focus from the BIT regime in general to the investor-state dispute settlement system in particular. A stylized fact about the outcomes of ISDS disputes is that wealthier governments tend to win a larger share of cases that are brought against them compared to governments of countries with lower per-capita income, a statistic attributed by some critics to institutional bias. However, it is unclear how much of this is a result of ISDS tribunals themselves or the process by which disputes are selected into the sample. I argue that it is impossible to answer the question of whether arbitration tribunals are less favorable to poorer governments without first answering the question of whether poorer
countries are more or less likely to settle disputes early. Because the decision of the arbitrators is not observed for disputes that settle early, analyses of dispute outcome data implicitly control for a post-treatment variable: that a dispute failed to settle. This risks inducing post-treatment bias if treatment affects settlement and a common unobserved cause affects both outcome and settlement as well. I indeed find that wealthier countries are less likely to settle disputes early and are more likely to pursue claims to a final decision. I argue that this is consistent with a simple bargaining model of pre-trial settlement where a state’s resource capacity affects its willingness to settle early. Drawing on insights from the biostatistics literature, I then suggest that the most relevant causal quantity in a situation with post-treatment attrition is the treatment effect for those disputes that would never result in a settlement regardless of which type of country is involved. Unfortunately estimating effects for this sub-group is complicated by the fact that the set of cases that would never settle early is not directly observable. I illustrate a new method for estimating this causal quantity that adjusts for potential observed confounders of attrition and outcome using a robust weighting method. I apply this new approach to estimate the effect of a country’s wealth level in a given dispute on the probability that it wins the dispute and find that adjusting for attrition eliminates any statistically significant win-rate gap between rich and poor countries. I conclude that the pre-trial settlement process appears to be the main driver of the disparity in governments’ success rates in arbitration rather than the investor-state institution itself.

Chapter 4, excerpted from a paper co-authored with Sergio Puig and published in the Journal of Legal Studies, considers another puzzle of arbitration: whether arbitrators are in fact biased in favor of the party that appointed them. It is nearly impossible to answer this question using observational data since tribunals typically make unanimous decisions and each party appoints one member to the panel. There is no variation in outcome at the individual level except in rare cases of dissents. Building on the recommendation of Chilton and Tingley (2013) that “international law needs experiments,” we developed a new experiment that we embedded in a survey of international arbitration professionals. To our knowledge, this was the first survey experiment of its type conducted on this particular specialized population. The experimental vignette presented survey respondents with a hypothetical investor-state arbitration scenario in which we randomly varied the appointing party described in the vignette. We
found that arbitrators assigned to one of the parties were more likely to give favorable decisions to that party and replicated this result in a follow-up experiment. Our results provide additional evidence that appointing authorities exert significant influence over the arbitrators in a tribunal setting. We outline a potential “blinding” method that could be adopted by arbitral institutions to mitigate these biases without eliminating party appointments altogether and discuss some of the potential benefits and drawbacks.

For each chapter, I make publicly available the replication code and original datasets on the Harvard/MIT Dataverse (Strezhnev, 2018).
Chapter 2

Estimating the Effects of Investment Treaties with a New Generalized Difference-in-Differences Estimator

2.1 Introduction

Difference-in-differences (DID) estimators are an important tool for applied researchers studying causal effects with observational data. In settings where “as-if-random” treatment assignment assumptions are unreasonable due to unobserved common causes of treatment and outcome, DID is one of the most straightforward methods for adjusting for certain forms of selection bias. When researchers have access to additional observations of the outcome from periods where all units in the sample are untreated, any observed differences between these two groups can be attributed to underlying differences in the latent characteristics of the types of units receiving treatment and control. Assuming that this selection bias is the same in both periods, subtracting this auxiliary difference from the simple difference-in-means can correct the bias due to non-random assignment – hence the name difference-in-differences. Equivalently, DID designs can also be motivated by the idea of “de-biasing” within-
unit pre- and post-treatment comparisons when there are trends in the outcome over time that could account for changes between the two periods. By assuming that units that are treated would have the same underlying time trend in the absence of treatment as those units that never receive treatment, the difference-in-differences estimator subtracts the observed time-trend in untreated units from the naive pre-/post-treatment comparison.

The theory motivating the DID estimator has been primarily developed in the context of panels with only two time periods with a binary treatment assigned only in the second period. In this case, no additional modeling assumptions need to be made to estimate the effect as DID simply requires the estimation of four conditional expectations. However, in most actual applications of the difference-in-differences framework, researchers working with panel data will observe outcomes for more than two time periods. Units in the data typically also do not all initiate treatment at the same time and some may discontinue treatment after the initiation period. Building on the well-known result that the ordinary least squares estimator with unit and time fixed effects (FE) and an indicator variable for treatment is equivalent to the non-parametric difference-in-differences estimator in the special two-period case, popular research methods texts usually suggest researchers use this same two-way fixed effects estimator in the case of multiple time periods. Angrist and Pischke (2009) term this approach “Regression DD” and Bertrand, Duflo and Mullainathan (2004) note that this is the standard estimation strategy for most econometrics research that utilizes DID in panels with more than two time periods.

However, recent work has challenged the idea that the regression DD approach produces valid estimates of causal effects in the same way that the non-parametric DID estimator does in the two-period case. Imai, Kim and Wang (2018) note that the two-way fixed effects estimator itself does not correspond to any valid matching estimator and can often impute improper counterfactuals for treated units. Abraham and Sun (2018) and de Chaisemartin and D’Haultfoeuille (2018) find that the two-way fixed effects estimator places non-uniform weights on the treated units in the sample, resulting in effect estimates that may mischaracterize the sample when effects are heterogeneous across units even when researchers attempt to model effects or pre-treatment trends over time via leads or lags of the treatment variable. Borusyak and Jaravel (2017) argues that when effects are variable over time, two-way fixed
effects estimators will up-weight short-run effects and down-weight long-term effects.

In this chapter, I provide a simple, unifying, explanation for why the two-way FE estimator fails to estimate unbiased treatment effects by showing that the two-way FE estimator can be written as a simple uniform average of all possible two-period DID estimators in the sample. Some of these estimators are unbiased DID comparisons and require no additional assumptions beyond the standard “parallel trends” assumption. However, others are unbiased only if an additional assumption is made that treatment effects do not persist beyond a single period. This is because two-way FE considers time periods where two units are both exposed to treatment as a valid de-biasing second difference. However, even if two units are both under the same treatment status in a particular period, their overall exposures to treatment may differ. One unit will have been under treatment for a longer period of time than the other. Therefore, the difference in observed outcomes between those two periods cannot be attributed exclusively to “bias” as is the case for periods where both units have never been exposed to treatment. In addition to this bias resulting from the use of invalid second-differences, I show that the two-way FE estimator exhibits the same weighting problems as any multiple regression estimator (Aronow and Samii, 2016). This is a variation of the problem found by Imai and Kim (2017) for one-way unit fixed effects estimators – units for which treatment status is well predicted by the fixed effects receive less weight when averaging over the distribution of treatment effects. The result is that for a single treated unit, treatment effects for different time periods will receive non-uniform weights. Across units, treatment effects may be weighted in a way that is unrepresentative of a typical intervention, creating problems for interpretation and generalization beyond the sample.

I lay out a new framework for estimating treatment effects using DID beyond the two-period setting. In doing so, I build on a recent literature that attempts to generalize the DID approach to the setting with multiple time periods, persistent treatment effects and variable treatment uptake times. The “synthetic control” method of Abadie, Diamond and Hainmueller (2010) considers estimation of the entire effect trajectory for a single unit by re-weighting the pool of units never exposed to obtain a counterfactual time trend had that unit never been exposed to treatment. The weights are constructed such that the re-weighted pool of control units matches the treated unit with respect to observed covari-
ates and pre-treatment outcomes. Xu (2017) generalizes this approach to the case of multiple treated units by directly modelling the outcome among units not receiving treatment via an interactive fixed effects model and subsequently using this model to impute counterfactuals for treated units. While this method works well to address potential unobserved confounding, it requires strong functional form assumptions for the outcome model, which are difficult to validate. Moreover, increased model flexibility comes at the cost of higher variance and more taxing requirements on the data. I develop an alternative estimation approach that generalizes the DID framework without requiring any models for the outcome. It deviates from the “synthetic controls” approach by retaining the “parallel trends” assumption underlying the standard DID rather than conditioning on pre-treatment outcomes. I first define a new quantity of interest, the Average Cohort Treatment Effect on the Treated (ACTT), that corresponds to a well-defined intervention with respect to the distribution of units initiating treatment in a given sample. I then show how this quantity can be non-parametrically identified and estimated as a weighted average of two-period difference-in-differences estimates using a generalization of the parallel trends assumption. I then show how researchers can further relax the parallel trends assumptions by incorporating covariates through a generalization of the inverse propensity score weighting method for DID developed by Abadie (2005). This new weighting method is closest in spirit to recent work by Imai, Kim and Wang (2018) which likewise proposes matching on pre-treatment covariates to make difference-in-differences inferences more credible in a time-series context. Crucially, both of these approaches avoid a major pitfall associated with including covariates directly into a two-way fixed effects regression: bias due to conditioning on a post-treatment variable (Rosenbaum, 1984).

I apply this new method to answer an ongoing puzzle in the study of investment law: whether investment treaties signed by states are actually effective in promoting foreign direct investment. Sala-cuse and Sullivan (2005) describes the BIT regime as a “grand bargain” between capital importers and exporters. By tying the hands of capital importing countries and raising costs of expropriation or creating a hostile climate for FDI, BITs make credible a country’s commitment to protecting investors’ rights. This restriction of policy autonomy is the cost a capital importing country pays in exchange for the promised benefit of making it a more attractive investment target relative to its competitors.
Whether this second half of the bargain is truly upheld is a challenging empirical question. Existing work on this question with time-series cross-sectional data has shown mixed results. One reason for this may be due to the use of two-way fixed effects estimators to account for unit-fixed and temporal sources of confounding. While the DID framework is a powerful tool for addressing omitted variable bias in panel data analyses, its implementation via two-way FE may in fact be generating misleading results. Because the effects of treaty adoption are likely to manifest over a period of many years rather than instantaneously, estimates from two-way FE will be biased towards zero. Furthermore, because researchers will typically also include time-varying covariates into these models, inferences are also likely to suffer from post-treatment bias if effects persist over time.

I analyze a new dataset on the presence of United States multinational firms abroad based on statistics provided by the Bureau of Economic Analysis (BEA) from 1983 to 2013. Specifically, I look at the effect of investment treaties on the extensive margin of FDI: whether new firms are willing to enter a market after the entry-into-force of a bilateral investment treaty with the United States. I find that two-way fixed effects estimates substantially understate the average effect of BIT entry into force. While the fixed effect models suggest effects statistically indistinguishable from 0, the new generalized difference-in-differences approach shows that investment treaties boosted the number of U.S. firms in a market by about .2 log-points (a roughly 20% increase). Incorporating covariates into the two-way fixed effects model related to economic development, democracy, and other investment treaty commitments drives the point estimate to 0, suggesting post-treatment bias. Conversely, using the proposed weighting method to adjust for covariates does not change the point estimate substantially from the simple DID estimate. These results suggest that investment treaties do actually have an influence on firms’ investment decisions over the long-run and that two-way fixed effect estimators will tend to understate these effects due to their inherent biases.

The remainder of this chapter is structured as follows: Section 2.2 develops the theory behind causal effects in a panel data setting and the assumptions behind the difference-in-differences estimator. After summarizing the classical DID framework, I generalize these assumptions to the case with many time periods and define a new causal estimand, the Average Cohort Treatment effect on the
Treated (ACATT). I propose a straightforward non-parametric approach to estimation under a generalization of the parallel-trends assumption in many time periods. Section 2.3 then illustrates how the two-way fixed effects estimator is biased for the ACTT. It decomposes the source of the bias into two components: a bias due to the use of improper second-difference terms, and a bias due to the way in which OLS weights treatment effects, an analogue of the well-known “regression weighting” problem (Aronow and Samii, 2016). Section 2.4 explains how researchers can incorporate covariates into the proposed generalized DID estimator using an inverse propensity weighting approach that builds on the method described by Abadie (2005). Crucially, this method avoids problems with post-treatment bias that arise when including covariates that might be affected by treatment directly into a time-series regression model. Section 2.5 compares the difference between the two-way FE approach and the new estimator in an analysis of the effect of U.S. investment treaties on the activity of U.S. multinationals. Section 2.6 concludes with recommendations for applied researchers working in a panel data setting.

2.2 Setup and Theoretical Framework

To understand how a difference-in-differences design could work in a multi-period setting, it is important to clearly explain the quantity of interest being estimated. This section develops a theoretical framework for defining and understanding causal effects in a time-series context. Consider a sample of $N$ units each indexed by $i$. Each unit is observed over a total of $T$ time periods indexed by $t$. The research goal is to estimate the effect of some exposure or “treatment” on an outcome over a series of time periods. Each unit $i$ is assigned to some treatment history denoted $\vec{A}_i$, which is a $T$-length vector of the unit’s particular treatment status in each time period: $\vec{A}_i = \{A_{i1}, A_{i2}, \ldots, A_{iT}\}$. Denote the set of all possible treatment vectors as $A$. For the purposes of this chapter, I focus exclusively on the case where treatment in any given period is a binary indicator with $A_{it} = 1$ indicating that a unit is exposed to a particular treatment at time $t$ and $A_{it} = 0$ indicating that unit is not exposed. Furthermore, denote a unit’s partial treatment history as the sub-vector of treatments up to some time period $t$: $\vec{A}_{it} = \{A_{i1}, A_{i2}, \ldots, A_{it}\}$. A unit that has not yet initiated treatment at time $t$ will have a partial
treatment history of all zeroes $\vec{A}_{it} = \{0, 0, \ldots, 0\} = \vec{0}$. The observed outcome in time $t$ for unit $i$ is denoted $Y_{it}$. Likewise, the vector of all outcomes for unit $i$ is $\vec{Y}_i$ and the sub-vector up to time $t$ is $\vec{Y}_{it}$. Finally, let $\vec{X}_i$ be the $K$ by $T$ matrix of observed covariates for unit $i$ with $X_{it}$ the $K$-length vector of covariates observed at time $t$ and $\vec{X}_{it}$ the $K$ by $t$ matrix of covariates up to time $t$. Assume that $X_{it}$ is observed prior to the assignment of treatment in period $t$ but can be affected by treatment assigned in prior periods. For the purposes of illustration, I will focus here on identification without conditioning on covariates. However, Section 2.4 will discuss methods for covariate adjustment when the necessary identification assumptions only hold given some set of observed pre-treatment covariates.

I define causal effects using the conventional potential outcomes framework, also often referred to as the “Rubin Causal Model” (Neyman, 1923; Rubin, 1974) A causal effect is the change in the outcome that would be observed if a unit had been assigned to one treatment regime versus another. Since we only ever observe units under a single treatment regime, identifying a causal effect from data requires reasoning about counterfactuals and that researchers make assumptions about the treatment assignment process. This is often referred to as the “fundamental problem of causal inference” (Holland, 1986). Formally, let $Y_{it}(\vec{a})$ denote the potential outcome that we would observe for unit $i$ in time period $t$ if that unit were assigned to the particular treatment history $\vec{A}_i = \vec{a}$.\footnote{Note that by writing the potential outcome only in terms of the treatment vector for unit $i$, I am implicitly making what is often known as the Stable Unit Treatment Value assumption or SUTVA with respect to the units in the sample (Rubin, 1986). This assumption states that a unit’s potential outcomes only depend on the assignment of their particular treatment history and not on the treatment histories of other units. Often, this assumption is stated separately, depending on the particular theoretical treatment, but is also implied by the consistency assumption.}

Next, I make the assumption of “consistency” to connect the observed data to counterfactuals.

**Assumption 2.2.1 Consistency**

$$Y_{it} = Y_{it}(\vec{a}) \text{ if } \vec{A}_i = \vec{a} \quad (2.1)$$

Consistency states that the observed outcome for units with an observed treatment history equal to $\vec{a}$ is equal to that unit’s potential outcome had it been assigned to treatment history $\vec{a}$.$\footnote{In randomized experiments and studies where treatment is directly manipulated and assigned by a re-}
Because individual causal effects cannot be estimated due to the fundamental problem of causal inference, researchers typically focus on averages of effects. I define the “average treatment history effect” in some time period \( t \) as the difference in the expected potential outcome under assignment to two different treatment histories.

**Definition 2.2.2 Average Treatment History Effect**

\[
ATE_t(\vec{a}, \vec{a}^*) = E[Y_{it}(\vec{a}) - Y_{it}(\vec{a}^*)]
\]

Treatmenteffects can also be defined for sub-groups within the population. In the DID context, researchers focus on estimating the average treatment effect on the treated (ATT) since under the necessary identification assumptions, counterfactuals can only be imputed for treated units and not for the controls.

**Definition 2.2.3 Average Treatment History Effect on the Treated**

\[
ATT_t(\vec{a}, \vec{a}^*) = E[Y_{it}(\vec{a}) - Y_{it}(\vec{a}^*)|A_i = \vec{a}] 
\]

With a few sensible assumptions, it is possible to know, with certainty, that some treatment history effects are zero. Intuitively, if two treatment histories differ only in the treatment levels assigned in periods after some time period \( t \), then it is impossible for there to be a causal effect in period \( t \). For a manipulation in the future to affect an outcome in the past would violate known properties of time, which physicists understand as an asymmetric process which flows in a single direction. Cause temporally precedes effect, a property often described as “time’s arrow.”

Counterfactual reasoning also

---

3While the question of reconciling the perceived asymmetry of time with theories of the physical universe...
retains this time-asymmetric property (Lewis, 1979). An intervention made in the past has the potential to affect the present. But an intervention made in the present cannot change the past. I formalize this notion in a “no reverse causality” assumption, which I will show forms the basis of identification in the difference-in-differences setting.

**Assumption 2.2.4 No reverse causality**

\[ Y_{it}(\bar{a}) = Y_{it}(\bar{a}^*) \text{ if } \bar{a}_t = \bar{a}^*_t \]  

(2.4)

In other words, if two treatment histories are identical up to time \( t \), the potential outcomes in time \( t \) associated with those histories will also be the same, even if the histories differ in periods after \( t \). Under this assumption, it is possible to find comparisons between groups assigned to two different treatment histories where we know the causal effect must be 0. Therefore, any difference in observed outcomes can be attributed not to the effect of treatment, but rather to underlying differences between the types of units assigned to one history versus the other – the omitted variable bias.

### 2.2.1 Difference-in-differences with two time periods

With the no reverse causality assumption, it is possible to use observations from one period to “de-bias” the naive difference-in-means estimates of treatment effects in other periods where treatment varies. This facilitates identification in the DID setting with repeated observations of units. In this section I outline the assumptions behind the classic difference-in-differences estimator in the simplest setting with only two time periods \( (T = 2) \) and two possible treatment histories. Assume that at time 1, all units are under control. In time period 2, units can either initiate treatment or remain under control. Let \( a^1 \) denote the treatment history for units that are under control only in period 1 and initiate treatment in period 2, and \( a^2 \) denote always-control treatment history.

remains a serious puzzle in the field of theoretical physics, discussion of these complexities is far beyond the scope of this chapter. See Halliwell, Pérez-Mercader and Zurek (1996) for a review.
In this setting, there is only one non-zero treatment effect of interest, the ATT in period 2.

\[
\text{ATT}_2(a^1, a^2) = E[Y_{i2} (a^1) | \bar{A}_i = a^1] - E[Y_{i2} (a^2) | \bar{A}_i = a^1]
\]  \hspace{1cm} (2.5)

The first term can be identified directly from the data under the consistency assumption, as it is simply the expected outcome in period 2 for units assigned to treatment. However, the second is a counterfactual quantity that must be imputed from those observations assigned to the control history. If treatment assignment were completely randomized, there would be no differences in expectation between units under treatment and control except for the manipulated treatment condition. Therefore, the potential outcome under control for treated units would be equal to the average observed outcome for units receiving the control. In observational studies, researchers will typically invoke a conditional version of this assumption to obtain identification – that treatment is as good as randomized given some set of covariates. Under such a “selection on observables” assumption, the treatment effect can be estimated via typical covariate adjustment methods such as regression, sub-classification, matching or inverse propensity weighting (Imbens, 2004).

When there remain unobserved confounders of treatment and outcome, adjusting on observed covariates alone will still yield biased estimates of the treatment effect. Difference-in-differences designs relax the “selection on observables” assumption by allowing for the existence of unobserved, unit-fixed confounders of treatment assignment and outcome. Formally, the DID approach makes the assumption of “parallel trends”:

**Assumption 2.2.5 Parallel trends**

\[
E[Y_{i2}(a^2) - Y_{i1}(a^2) | \bar{A}_i = a^1] = E[Y_{i2}(a^2) - Y_{i1}(a^2) | \bar{A}_i = a^2]
\]  \hspace{1cm} (2.6)

This assumption states that in the absence of treatment, units assigned to the treated history would have the same linear trend in the outcome compared to units assigned to the control history. Under no reverse causality, we know that \(Y_{i1}(a^2) = Y_{i1}(a^1)\) since both treatment histories are identical up to
period 1. Therefore, under parallel trends, the ATT is identified non-parametrically by:

\[
\text{ATT}_2(a^1, a^2) = \left( E[Y_{i2}|A_i = a^1] - E[Y_{i2}|A_i = a^2] \right) - \left( E[Y_{i1}|A_i = a^1] - E[Y_{i1}|A_i = a^2] \right)
\] (2.7)

This estimator consists of a difference in two differences, hence the name difference-in-differences. For this chapter, I will refer to the “first difference” as the naive difference-in-means estimator in period 2 and the “second difference” as the bias correction estimated from the observed difference in outcomes between the treatment histories for period 1.\(^4\) If treatment is in fact randomly assigned, the second difference should be zero in expectation (as treated and control are in expectation the same on all pre-treatment covariates) and the expression will reduce to the typical difference-in-means estimator.

### 2.2.2 Difference-in-differences with multiple time periods

While the two-period difference-in-differences case is well-studied, extending the intuition from that case to multiple time periods is complicated by the absence of a comparable causal estimand or quantity of interest. Adding additional time periods expands the number of possible treatment histories for which an ATT can be defined. With \(T\) time periods and no restrictions on possible treatment histories, there are \(2^T\) possible unique treatment histories. As researchers add more and more time periods, purely non-parametric estimation with no additional restrictions becomes increasingly infeasible due to the “curse of dimensionality” as some treatment histories may only be observed for a handful of units while others may be never observed at all. Moreover, researchers are rarely interested in the effect of one particular history, but rather some sort of average of treatment history effects for the entire sample.

In this section, I define a general quantity of interest for difference-in-differences designs under cer-

\[^4\]An equivalent way of writing the DID estimator is to re-arrange expectations and first take the difference in expected outcomes for periods 2 and 1 for units assigned treatment, subtracting that from the difference in expected outcomes for the same time periods for those assigned to the control history. The version used here is preferrable for the purposes of this chapter as it makes clear the connection to the simple difference-in-means and clarifies the role of the second difference as a bias-correction.
tain limitations on the possible treatment histories. Specifically, I constrain units from reverting their treatment status once they initiate treatment. In many studies, it is reasonable to assume that treatment uptake can only go in one direction. Once a unit receives treatment, it is always under treatment until the final time period $T$. In studies of policy adoption, this is typically the case when the time period under consideration is relatively short. Imai, Kim and Wang (2018) refer to this as the assumption of “stable policy change” in a given intervention. This restricts the number of possible treatment histories to $T$ if it is also assumed that each unit is under control for at least one time period. However, units may not necessarily receive treatment all in the same time period. For example, governments may adopt similar policies at slightly different times with some units being leaders and others lagging behind. Following the terminology of (Abraham and Sun, 2018), I will refer to the groups of units initiating treatment at the same time as treatment “cohorts.” Each cohort corresponds to some value $C_i$, which denotes the last period under which that unit $i$ is under control. A unit’s treatment history is determined entirely by $C_i$. For a unit with $C_i = c$, $A_{it} = 0$ for all $t \leq c$ and $A_{it} = 1$ for all $t > c$. Units with $C_i = T$ never receive treatment and are always under the control condition.

To simplify the notation of treatment histories, let $a^c$ denote the treatment history associated with cohort $C_i = c$, $c \in \{1, \ldots, T\}$. $Y_{it}(a^c)$ is the potential outcome observed for unit $i$ in time $t$ if it initiated treatment at time $c + 1$. Define the Cohort Average Treatment effect on the Treated (CATT) in time $t$ as

$$CATT_i(c) = E[Y_{it}(a^c) - Y_{it}(a^T) | \vec{A}_i = a^c]$$ (2.8)

This corresponds to a natural effect of interest: the change in the expected outcome at time $t$ if a unit that initiated treatment at time $c + 1$ were instead never exposed. While researchers could plausibly be interested in other counterfactual comparisons with histories other than the “never treated” history, it is

---

5 Note that it is not possible to non-parametrically estimate the ATT for units that are always under treatment as there are no control periods that can serve as part of the de-biasing term.

6 In econometrics, these variable treatment uptake scenarios are often termed “event studies” as the units being studied experience some “event” at potentially different times and each unit may have a distinct set of pre-event and post-event observations (Abraham and Sun, 2018).
the most intuitive starting point for defining causal contrasts.

Under the no reverse causality assumption, $\text{CATT}_t(c) = 0$ for all $c \geq t$. Additionally,

$$ Y_{it}(a^c) - Y_{it}(a^T) = Y_{it}(a^c) - Y_{it}(a^j) \text{ for all } t \leq j \leq T $$  \hspace{1cm} (2.9)

In other words, the CATT for a given time period depends only on a unit’s treatment history up to time $t$. This allows treated units that initiate treatment at time periods after $t$ to act as control units for those that initiate treatment prior to $t$.

Identification in the multiple-period setting relies on a generalization of the parallel trends assumption.

**Assumption 2.2.6 Generalized parallel trends**

$$ E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i = c] = E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i \geq t] $$ \hspace{1cm} (2.10)

for all $t > c$, $t' \leq c$

This assumption states that, in the absence of treatment uptake at time $c$, the expected change in outcome between time $t$ and some past time $t' \leq c < t$ would have been the same for units that initiate treatment in period $c + 1$ and units that remain under control up until period $t$. With this assumption, no-reverse causality, and the assumption the CATT in time $t$ for cohort $C_i = c$ is non-parametrically identified by

$$ \text{CATT}_t(c) = \frac{1}{c} \sum_{t' = 1}^{c} [E[Y_{it}|C_i = c] - E[Y_{it'}|C_i = c]] - [E[Y_{it}|C_i \geq t] - E[Y_{it'}|C_i \geq t]] $$ \hspace{1cm} (2.11)

This generalized multi-period difference-in-difference estimator is essentially an average of $c$ “two-period” difference-in-differences estimators with the “treatment” period always equal to $t$ and the “control” period changing between all time periods prior to the cohort’s initiation of treatment. Note that this definition makes no additional restrictions on which time periods can be affected by treatment
so long as \( t \) is greater than \( c \). Consistent estimates of each conditional expectation can be obtained by directly substituting in the sample analogues.

Researchers studying the effects of a particular event or policy will typically not just be interested in the effect for a specific cohort and for a single time period. Treatment may have different effects in earlier periods relative to later ones and a natural way of aggregating this combination of short and long-term effects for a given cohort is to average the individual \( CATT_t(c) \) effects over all periods for which the cohort is exposed to treatment. Define the overall Cohort Average Treatment Effect on the Treated as

\[
CATT(c) = \frac{1}{T - c - 1} \sum_{t=c+1}^{T} CATT_t(c)
\]  

(2.12)

This aggregation also corresponds to a natural causal question of interest: what would have happened, on average, had a cohort never initiated treatment in those time periods where the treatment could have had an impact. Note, however, that the CATT for one cohort will cover a different set of time periods than a CATT for another time period. How then, should multiple CATTs be aggregated into a summary quantity. One approach is to only use a set number of post-treatment periods (denoted \( F \) for each cohort, as in Imai, Kim and Wang (2018). However, this approach requires that researchers make an additional up-front assumption about how many post-treatment periods are of interest. It also throws away data: units with more treated periods than \( F \) go partially unused, while units with fewer than \( F \) treated periods cannot be part of the analysis at all.

Here I define a causal estimand that does not require the user to specify a particular number of post-treatment periods, the Average Cohort Treatment Effect on the Treated (ACTT)

**Definition 2.2.7** Average Cohort Treatment Effect on the Treated (ACTT)

\[
ACTT = \sum_{c=1}^{T-1} CATT(c) Pr(C_i = c | C_i \neq T)
\]  

(2.13)

The ACTT corresponds to a weighted average of cohort treatment effects with the weights propor-
tional to the in-sample frequencies of each cohort (among cohorts receiving treatment). It has a natural interpretation as the CATT for a unit chosen randomly from the sample. To estimate the ACTT non-parametrically, it suffices to estimate each \( CATT(C_i) \) for each unit in the data and take the average across the sample. Moreover, inference is straightforward as it is possible to interpret this as a weighted average of all feasible two-period difference-in-differences estimators in the sample. Since this estimator is linear in \( Y \), valid standard errors can easily be calculated using bootstrapping methods, namely the block bootstrap which resamples units with replacement in order to preserve the outcome correlation structure within each unit (Bertrand, Duflo and Mullainathan, 2004).

2.3 Bias from OLS with unit/time fixed effects

Standard practice among researchers estimating DID effects in panels with many time periods is to fit an ordinary least squares regression with fixed effect parameters for both unit and time. In the two-period case with only two treatment histories, it is well known that the ordinary least squares regression estimator with time and unit fixed effects is identical to the non-parametric difference-in-differences estimator. Texts on causal inference typically will recommend using this same two-way fixed effects model in order to estimate difference-in-differences effects more generally, an approach Angrist and Pischke (2009) term “Regression DD” (pp. 223). The underlying regression model assumes the following data-generating process for \( Y_{it} \)

\[
Y_{it} = \alpha_i + \gamma_t + \beta A_{it} + \epsilon_{it}
\]  

(2.14)

where \( \alpha_i \) is a fixed effect parameter for each unit, \( \gamma_t \) is the fixed effect parameter for each time period, and \( \epsilon_{it} \) is a mean-zero error term. As before \( A_{it} \) is an indicator that takes on a value of 1 if unit \( i \) is under treatment at period \( t \) and 0 if it is not. Researchers will typically report the “average treatment effect” as the estimated coefficient \( \hat{\beta} \) on \( A_{it} \).

Unfortunately, interpreting this coefficient as a meaningful treatment effect requires very strong as-
sumptions on the way in which treatment can affect the outcome. And even when these assumptions are satisfied, the corresponding regression coefficient may not be representative of treatment effects in the sample or reflect an average over units in the sample that is substantively interesting to a researcher. I show here that the two-way fixed effects estimator can be written as a uniform average of all possible difference-in-differences comparisons in the sample.

**Proposition 2.3.1** The OLS estimate of $\beta$ in the two-way fixed effects model is equivalent to

$$
\hat{\beta} = \frac{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \left\{ [Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}] \right\}}{\sum_{t=1}^{T} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{1 - A_{it'} + A_{jt'}\}}
$$

(2.15)

where $\sum_{i:A_{it}=1}$ denotes a sum over all units (indexed i) where treatment status $A_{it} = 1$.

The complete proof is given in the appendix. In this section I will discuss the main intuitions behind it. Within every time period $t$, the two-way FE estimator starts by matching each treated unit to each unit under control. This corresponds to the “first-difference” in the difference-in-differences estimator. For each matched pair, it then iterates through all other time periods (denoted $t'$) and subtracts from the first-difference the “second-difference,” comprised of the outcomes in time $t'$ for the same pair of units.

It is in this “second difference” that the main source of bias is induced. There are three possible matches that can be found in the data for each treatment/control pair. First, a pair that is under treatment/control at time $t$ can be matched to a time period $t'$ where both units are under control. If treatment histories are restricted such that no unit can revert from treated to control, it must be the case that $t > t'$ and that units’ treatment histories are identical up to $t'$ – they are both always under control. Therefore, this is a valid second difference with respect to the DID estimator as there is no treatment effect of one history relative to the other at time $t'$. Second, a treatment/control pair can be matched to a time period where the unit under treatment remains under treatment and the unit under control remains under control. In this case, the difference-in-differences will cancel out as each treatment/control pair will appear once in a first difference and once in a second difference. Finally, treat-
ment/control pairs can be matched to a time period where both units are under treatment. Under the assumption that no unit reverts from treated to control, such periods are necessarily in the future \((t' > t)\). Additionally, the two units will not have the same treatment history up to \(t'\) since we know they differ at \(t\). Therefore, unlike the control/control case, these observations act as invalid second-differences because the treatment effect of one history versus the other is not guaranteed to be 0. Borusyak and Jaravel (2017) refer to this as a “forbidden extrapolation.” For these pairs of observations to act as valid second differences, it must be the case that the effect of treatment does not persist beyond a single period. In this case, the potential outcome for a unit depends only on its treatment assignment in period \(t\) and not on past assignments. With this additional assumption, there now exists no treatment effect for two units with the same treatment level at time \(t\) and periods where units are both under treatment can in fact serve as valid second differences for treatment/control comparisons in previous time periods. Unfortunately, such assumptions are highly restrictive and implausible in most settings.\(^7\)

Second, it is clear from the expression in Proposition 2.3.1 that, while the average over DID comparisons is uniform, treated units receive non-uniform weights. This is because the number of units matched to a treated unit in the first difference and as part of the second-difference varies from time period to time period. Cohorts for which there are many pre-treatment periods receive greater weight than cohorts with few pre-treatment periods. Within an individual cohort, future time periods receive less weight than past ones as the number of within-period controls decreases over time. This persists even if one were to eliminate the invalid second differences, suggesting that the problem of up-weighting short-term versus long-term effects is not just due to problem of invalid second differences. As a result, this weighting of units does not correspond to the ACTT as defined in the previous section since cohorts receive in-sample weights that are not necessarily proportional to their prevalence and time periods are not weighted uniformly within each cohort. This is a particular instance of a well-known property of multiple regression coefficients, the “regression weighting problem” (Aronow and Samii, 2016). In estimating \(\hat{\beta}\), units whose treatment status is well predicted by the covariates (in this

\(^7\)When there are no restrictions on treatment histories, a fourth type of match can occur: a treated-control pair is matched to a control-treated pair. In this case, the difference-in-differences counts twice for the average as the model assumptions imply that the difference between these two differences is equal to \(2\hat{\beta}\).
case, the fixed effect parameters), are down-weighted when calculating the average while those whose treatment status is poorly predicted receive greater weights. In the presence of effect heterogeneity both over time and across units, this can give misleading inferences.

To provide a concrete illustration of both sources of bias in the two-way FE estimator, I consider a simple numerical example with $N = 5, T = 3$. Figure 2.1 presents a hypothetical treatment history assignment. Two of the units (1 and 2) initiate treatment in period 2 which carries over to period 3. Units 3 and 4 only initiate treatment in period 3. Unit 5 receives no treatment. All units are untreated in period 1.\footnote{One could, of course, consider a much larger sample where the proportions of units assigned to each history remain the same. For ease of exposition, I limit the example to 5 but note that issues of bias are not simply confined to small samples. Rather, they are a function of the particular distribution of treatment histories in a given sample.}

| Unit | Time | | Unit | Time | | Unit | Time |
|------|------|------|------|------|------|------|
| 1    | 0    | 1    | 1    | 0    | 1    | 1    | 0    | 1    |
| 2    | 0    | 1    | 1    | 2    | 0    | 1    | 1    | 2    | 0    | 1    |
| 3    | 0    | 0    | 1    | 3    | 0    | 0    | 1    | 3    | 0    | 0    | 1    |
| 4    | 0    | 0    | 1    | 4    | 0    | 0    | 1    | 4    | 0    | 0    | 1    |
| 5    | 0    | 0    | 0    | 5    | 0    | 0    | 0    | 5    | 0    | 0    | 0    |

**Figure 2.1:** Illustration of valid and invalid second difference sets under a two-way fixed effects estimator

For a particular treated unit in period 2, that unit is matched to the three other observations under control in time 2 to construct the first difference term. However, the second-difference consists both of the valid differences from period 1, where no units are under treatment, and the invalid second dif-
ferences from period 3 for units 3 and 4 which are under treatment in time 3. Because the two-way FE estimator has no concept of “ordering” when it comes to time, it treats future and past periods as equivalent.

Figure 2.2 gives the implied weights on each treated unit for the two-way FE estimator. Note that for units 1 and 2 in period 3, the weights are negative, implying that those unit-periods are more often part of the second-difference term, acting as controls, than the first, acting as treated units. Even if we restrict the DID estimator to only those valid second differences, it does not solve the regression weighting problem. Effects in period 2 receive a greater weight when averaging than those in period 3, due to the reduced number of control observations in the third time period as illustrated in Figure 2.2.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Time</th>
<th>Time</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>−1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>−1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit</th>
<th>Time</th>
<th>Time</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Treatment assignment  Implied weights on cohorts  Valid second differences only

**Figure 2.2:** Implied weights on treated units under a two-way fixed effects estimator

Two-way fixed effects estimators will not yield a valid estimate of the treatment effect under parallel trends if treatment effects persist over time and are heterogeneous across cohort and time. Outside of the two-period setting, two-way FE relies heavily on the restrictions implied by the parametric model for $Y_{it}$. Two important restrictions are that $A_{it}$ only affects $Y_{it}$ and not future outcomes and that $\beta$ is a constant. In most applied settings, however, treatment effects are rarely instantaneous and the
impact of a particular intervention often takes many time periods to reveal itself. Additionally, units will typically respond differentially to treatment. Therefore, the underlying assumptions of the two-way fixed effects model are not plausible for most quantitative research. The alternative “generalized difference-in-differences” estimator outlined in this paper does not suffer from either of these drawbacks and is more appropriate in situations where treatment exhibits persistence after initiation.

2.4 Inverse propensity weighting estimators for multi-period DID effects

For many applications, even the parallel trends assumption is unlikely to hold unconditionally as factors associated with treatment history may also be associated with different pre-treatment paths. Therefore researchers will typically want to incorporate covariates in order to make the parallel trends assumption more credible. One of the reasons why two-way fixed effects models are so popular is that they facilitate easy inclusion of covariates as part of the outcome model. An alternative to regression modeling is inverse propensity weighting, which Abadie (2005) illustrates for the case of two-period difference-in-differences estimators. By up-weighting units with covariate profiles that are underrepresented among controls relative to treated units and down-weighting those that are overrepresented, weighting approaches allow for estimation when identification assumptions hold only conditionally.

In this section I extend this approach to the generalized difference-in-differences estimand with more than two outcome periods and treatment histories.

I again focus on identification of the $CATT_t(c)$, the cohort ATT for a particular time period $t$. Since this is identified by an average of two-period DID estimators, it is possible to apply the method in Abadie (2005) to each of these estimators in turn. Identification with covariates relies on a slightly weaker version of the parallel trends assumption.
Assumption 2.4.1 Conditional generalized parallel trends

\[ E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i = c, \vec{X}_{ic} = \vec{x}] = E[Y_{it}(a^T) - Y_{it'}(a^T)|C_i \geq t, \vec{X}_{ic} = \vec{x}] \] (2.16)

for all \( t > c, t' \leq c \)

In other words, the parallel trend from \( t' \) to \( t \) can vary depending on the value of \( \vec{X}_{ic} \), the covariate profile of unit \( i \) up to time \( c \).\(^9\) When all covariates are time-constant, the time index is unnecessary. However, this formulation allows for the presence of covariates that follow patterns over time that may account for the violation of the unconditional parallel trends assumption. Note also that only the covariate history up to \( c \) is part of the conditioning set even when \( t > c \). This is because values of \( X_{it} \) for periods after \( c \) are potentially affected by the treatment itself. Conditioning on variables potentially affected by the treatment risks post-treatment bias (Rosenbaum, 1984; Acharya, Blackwell and Sen, 2016; Montgomery, Nyhan and Torres, 2018).

Let \( Pr(C_i = c|\vec{X}_{ic}, C_i \in \{c, t, \ldots, T\}) \) denote the probability that a unit is in cohort \( c \) given both its covariate vector up to time \( c \) and knowing that \( C_i \) is either \( c \) or greater than or equal to \( t \). Under the weaker assumption of conditional parallel trends, the \( CATT_i(c) \) can be identified by

\[
CATT_i(c) = \frac{1}{c} \sum_{t'=1}^{c} \frac{E[I(C_i = c) - Pr(C_i = c|\vec{X}_{ic}, C_i \in \{c, t, \ldots, T\})]}{1 - Pr(C_i = c|\vec{X}_{ic}, C_i \in \{c, t, \ldots, T\})} \times E[Y_{it} - Y_{it'}|C_i = c, \vec{X}_{ic} = \vec{x}] \] (2.17)

where \( I(C_i = c) \) is an indicator that takes on a value of 1 if \( C_i = c \) and 0 otherwise. The reason for conditioning on \( C_i \in \{c, t, \ldots, T\} \) is that for each estimate of \( CATT_i(c) \), we are essentially subsetting the sample down to two sets of treatment histories, the one where treatment is initiated at \( c + 1 \) and the one where treatment is initiated after \( t \). The former corresponds to the “treated” group in a two-period DID while the no-reverse causality assumption implies that the latter are all equivalently “controls.”

\(^9\)It is also necessary here to make a “positivity” or covariate overlap assumption, that the probability of observing a given treatment history conditional on the covariates is not perfectly zero or one.
other words, we know that none of the matched control units will be ones that initiated treatment at \( t \) or earlier. With two periods and two treatment histories, we can apply weights in the vein of Abadie (2005). Intuitively, the treated units all receive a constant weight while the controls are re-weighted according to the ratio of the propensity that unit \( i \) would be treated and the propensity that it would be a control.

Under the assumption that no unit that initiates treatment reverts to being under control, the ratio of propensity scores \( \frac{Pr(C_i = c | X_{ic}^T, C_i \in \{c, t, \ldots, T\})}{1 - Pr(C_i = c | X_{ic}^T, C_i \in \{c, t, \ldots, T\})} \) can be written as a ratio of the probabilities that unit \( i \) initiates or does not initiate treatment in period \( c \) given that it has not initiated treatment at any previous point.

\[
\frac{Pr(C_i = c | X_{ic}^T, C_i \in \{c, t, \ldots, T\})}{1 - Pr(C_i = c | X_{ic}^T, C_i \in \{c, t, \ldots, T\})} = \frac{Pr(A_{ic} = 1 | A_{ic-1} = \vec{0}, X_{ic}^T)}{Pr(A_{ic} = 0 | A_{ic-1} = \vec{0}, X_{ic}^T)}
\]

(2.18)

with \( \vec{0} \) again denoting a vector of all zeroes. With a large enough sample, it is possible to estimate the weights separately for each individual \( CAT T_t(c) \) by fitting a logistic regression model on each subset of the data, predicting cohort membership using \( X_{ic}^T \). However, in practice this will be infeasible because only a few units will be a part of each cohort. Additionally, with many time periods, the dimensionality of \( X_{ic}^T \) will be very large if there are time-varying covariates present. An approach to simplifying the problem would be to fit a pooled logistic regression model for the probability that a unit initiates treatment in a particular time period given the covariates. Because pooling across time periods involves pooling over covariate profiles \( X_{it}^T \) with varying dimensionality, this method of estimating the weights requires making additional modeling assumptions that restrict the number of lagged previous periods that can enter into the model. One simple approach is to assume that only the current values of \( X_{it} \) affect the probability of treatment initiation at time \( t \). Then, one can estimate a pooled logistic regression to estimate \( Pr(A_{it} = 1 | A_{it-1} = \vec{0}, X_{it}) \) and obtain fitted values for the propensity scores. To account for varying propensities of treatment initiation over time, this model can incorporate a parametric time trend or time fixed-effect parameters. Inference can again be carried out using bootstrap methods, with the weighting model estimated repeatedly for each bootstrapped sample.
One advantage of the proposed weighting method as opposed to simply including confounders as regressors in the two-way FE linear model is that it avoids inducing post-treatment bias for long-term effects. Estimating a model of the form

$$Y_{it} = \alpha_i + \gamma_t + \beta A_{it} + \delta X_{it} + \varepsilon_{it}$$  \hspace{1cm} (2.19)

where $\delta$ is a vector of coefficients on each covariate in $X_{it}$ is common, but the coefficient on $\beta$ will only represent a meaningful causal effect if there are no effects of treatment that persist beyond the first period. This is because $X_{it}$ is potentially affected by and part of the causal effect of $A_{it-1}$. Controlling for $X_{it}$ blocks this causal pathway and may therefore attenuate treatment effects towards zero. Even if $X_{it}$ is not a causal mechanism for treatment, it is still unwise to control for consequences of treatment as bias will also be induced if there exist common causes of $X_{it}$ and $Y$ (Elwert and Winship, 2014).

An alternative to weighting is to simply match each treated cohort to a set of controls based on the pre-treatment covariates and possibly lagged outcomes, as is suggested by Imai, Kim and Wang (2018). Certainly, this is also a feasible approach here and the choice of covariate adjustment method will depend on how researchers choose to resolve the bias-variance trade-off. Relative to weighting, matching is typically inefficient (Abadie and Imbens, 2006) but may provide additional advantages in terms of making the covariate adjustment procedure more transparent.

$^{10}$Another issue with including covariates as part of the outcome model is that any time-invariant covariates will be perfectly colinear with the unit fixed effects and drop out of the model.
2.5 Application: The effect of investment treaties on U.S. foreign direct investment

Foreign direct investment (FDI) is an increasingly important vehicle for cross-border economic activity. Many firms look to situate elements of their production abroad and for governments looking to attract jobs and stimulate growth, competition for capital is fierce. In addition to direct incentives to foreign firms and domestic policy reforms, many governments have also looked to international legal arrangements to improve the attractiveness of their country to global capital. Bilateral Investment Treaties (BITs) have emerged as one of the most ubiquitous legal instruments used in the governance of cross-border investment. Competitive pressures from other states (Elkins, Guzman and Simmons, 2006) and economic downturns (Simmons, 2014) often push states to sign these agreements with major capital exporters. BITs typically commit states to refrain from expropriation or discriminatory treatment of foreign investors, among other investment protection obligations. To enforce these commitments, many BITs also contain provisions for Investor-State Dispute Settlement (ISDS) in which states pre-commit to allow foreign investors covered under the BIT to pursue binding arbitration in an international forum such as the International Centre for the Settlement of Investment Disputes (ICSID) in the event of a breach of the treaty. In addition to these separate investment treaties, many states are now party to regional trade agreements that include investor-state dispute settlement provisions, for example, Chapter 11 of the North American Free Trade Agreement (NAFTA).

By granting foreign investors secure property rights that can be enforced outside of a country’s domestic courts, BITs raise the cost of opportunistic expropriations. Kerner (2009) highlights two mechanisms through which BITs may, as a consequence, increase multinationals’ activities in a host country. First, they solve a time-inconsistent preference problem by committing states to maintaining property rights protections after an investment has been made, and thereby reducing investors’ anticipated risks ex-ante. Second, by imposing costs to violating property rights, BITs send a credible signal to uncertain foreign firms that ratifiers will respect investors’ property. This latter mechanism operates beyond
just investors that are able to access arbitration under the BIT and potentially affects overall FDI flows from non-covered source countries. Ultimately, BITs and their concommitant ISDS provisions function as a constraint on a government’s future policy flexibility, in theory providing predictability to foreign investors who are then expected to be more willing to undertake costly and illiquid investments.

Whether BITs are actually effective in promoting investment remains an elusive empirical question. Unfortunately for scholars, states do not enter into BITs at random. Moreover, even measuring FDI is itself a challenge, with varying definitions of MNE activity potentially generating variable empirical results. Different statistical modeling approaches, specifications, and strategies to control for omitted variable bias have yielded differing results. While early studies in the literature showed no effect on FDI flows (Hallward-Driemeier, 2003), subsequent work highlighted a positive association between BITs and FDI flows (Egger and Pfaffermayr, 2004), though other researchers have found that the effect is conditional on strong domestic institutions (Tobin and Rose-Ackerman, 2011). While more recent papers that attempt to better adjust for the temporal dynamics in panel analyses of FDI find positive support for the hypothesis that BITs increase FDI (Egger and Merlo, 2007; Busse, Königer and Nunnenkamp, 2010), other studies cast some doubt on the credible commitment story (Yackee, 2010).

One of the central challenges in estimating the causal effect of investment treaties is choosing an appropriate modelling strategy for observational data. The workhorse model for many of the existing analyses of BITs and FDI in the literature is the linear dynamic panel approach (Tobin and Rose-Ackerman, 2011) which models the outcome for a unit $i$ at time $t$, as a function of covariates ($X_{it}$), unit fixed effects $\gamma_i$ and lagged outcome values (typically by one period). The basic dynamic regression model is therefore of the form:

$$Y_{it} = \alpha Y_{it-1} + \beta X_{it} + \gamma_i + \varepsilon_{it}$$

where $\varepsilon_{it}$ is a mean zero random error.

This model assumes ignorability of treatment assignment in period $t$ conditional on covariates and the lagged outcome in $t-1$. Incorporating the lagged outcome term is treated as a means of accounting for a possible reverse-causal relationship between BIT initiation and FDI. Unfortunately, adjusting for
both outcome lags and unit fixed effects poses a number of statistical problems. It does not generally address the problem of confounding unless the model assumptions hold exactly. For example, if BIT ratification is predicted not by the 1-period lag, but rather by 3-, 4-, or 5-period lags, estimates will suffer from omitted variable bias (Bellemare, Masaki and Pepinsky, 2015). The validity of inferences depend heavily on the specific linear modeling assumptions chosen which can be difficult to validate.

In many cases, lags are omitted but additional fixed effect parameters for both time and unit are included.\textsuperscript{11} For example, the dyadic regressions in Busse, König and Nunnenkamp (2010), Berger et al. (2013) and Aisbett, Busse and Nunnenkamp (2016) employ additive regression models with dummy variables for both year and dyad (the unit of analysis). As discussed in this chapter, while such two-way fixed effect estimators are often motivated by an implicit difference-in-differences design, with the intent to adjust for unobserved time-constant confounders, they fail to do so in practice and require strong restrictions on how treatment can affect the outcome over time.

Many of the aforementioned studies also employ an instrumental variables strategy to address the problem of dynamic confounding. However, in all cases, the authors highlight the difficulty in justifying the exclusion restriction underpinning the instruments. Tobin and Rose-Ackerman (2011) admit that “good instruments are elusive, and weak at best” (15).\textsuperscript{12} Both Tobin and Rose-Ackerman (2011) and Busse, König and Nunnenkamp (2010) employ (among other IV approaches) a Generalized Method of Moments (GMM) strategy, instrumenting for BITs using lagged values of the independent variables. However, this strategy is only valid if lagged treatments only affect the outcome through their effect on current treatment – that is, if the outcome model is exactly true and there are no persistent effects over time – precisely the assumption that this chapter finds is highly problematic. In general, if the instrument affects the outcome via a mechanism outside of its effect on the treatment, instrumental variables estimates will be biased. Moreover, commonly used statistical tests for instrument

\textsuperscript{11}Typically, researchers choose either to include the lagged outcome as a regressor or to use fixed effects when using ordinary least squares. Incorporating both risks biased OLS estimates for the parameters of the dynamic model when the number of time periods is small (Nickell, 1981).

\textsuperscript{12}Tobin and Rose-Ackerman (2011) instrument BIT ratification with the number of BITs ratified by a country’s neighbors. However, this instrument would be invalid if there exist unobserved regional patterns in FDI and BIT ratification such that a country’s ratification patterns and its neighbors’ ratification patterns are affected by a common cause.
validity in time-series instrumental variables models often fail to detect exclusion restriction violations due to lack of power (Bazzi and Clemens, 2013). Determining whether an instrument is valid is a matter for theory, not for statistical testing.

In lieu of these strong parametric methods, I employ the generalized DID estimator outlined in this paper, adjusting for possible violations of the parallel trends assumption due to observed covariates using inverse propensity weighting. Because dyadic analyses are fraught with issues of cross-dyadic dependence, complex correlation structures and clustering, I focus on the effect of investment treaties with a single country: the United States. This is a particularly good case to consider given the United States’ long history of promoting BITs as a means of protecting its firms’ investments abroad and the emphasis U.S. BITs in particular placed on investor-state dispute settlement (Vandevelde, 1993). Additionally, all U.S. BITs in force contain provisions for investor-state dispute settlement and almost all of the United States’ recent preferential trade agreements have some sort of ISDS mechanism as well.

Focusing on U.S. foreign investment alone also helps address problems related to the measurement of FDI. Constructing a valid measure of multinational involvement in a host economy is a difficult task. The most common measures of MNC activity with the broadest coverage temporally and spatially are FDI flow and stock data. Unfortunately, these measures are also often the least theoretically applicable. As Kerner (2014) notes, FDI flows and stocks are derived primarily from balance of payments statistics gathered by governments and central banks. Ultimately flows reflect the cumulation of cross-border financial transactions. For many research questions, these measures are poor proxies for the extent to which MNCs are willing to make investments in production within a country. Flows can be particularly misleading as an observation of $0 flows can reflect the absence of foreign affiliates or it can indicate that firms’ repatriation of profits (net negative flow) matches the increase in firms’ foreign position (e.g. via reinvested earnings) (Kerner, 2014). FDI stock data perhaps provides a better measure of the value of foreign-owned capital in a particular country, but proper valuation of stocks (whether on market value or by historical value) depends on the research question of interest.

As an alternative, Kerner and Lawrence (2014) points to fine-grained measures of firm expenditures collected at a national level. The United States Bureau of Economic Analysis (BEA) releases data
from annual surveys of the worldwide activities of U.S. multinationals.\textsuperscript{13} These surveys are conducted annually on each foreign affiliate of a U.S. parent company and are required of all affiliates that exceed a certain size threshold. While the surveys themselves are confidential, aggregate information is made public on assets held by foreign affiliates of U.S. multinational firms on an annual level. Unfortunately, for many United States partner countries, even the aggregate data is suppressed for reasons of data privacy. This results in high levels of potentially non-random missingness in the data. However, the BEA survey data does not suppress information on the number of U.S. affiliates and this data is available for every country in which U.S. investment is reported. An advantage of focusing on raw counts of multinationals is that it allows me to directly assess the effect of bilateral investment treaties on the \textit{extensive} margin of foreign direct investment. That is, it specifically measures whether new firms are entering a market. It is difficult to distinguish from balance of payments or even aggregate asset data between existing firms increasing their investment position versus new entrants choosing to enter a market.

The dataset I assemble consists of observations of 157 countries over a period from 1983 to 2013. The outcome variable is the number of U.S. multinational affiliates with assets, sales or net income over $25 million in the host country in a given year as reported by the BEA’s US Direct Investment Abroad (USDIA) survey. Since the surveys only report data for countries in which any foreign direct investment was reported, not all countries are covered. Some country years are missing when no firms that are surveyed report any investment. When a country previously included in the survey is missing in a subsequent year, I impute the count of multinationals as 0 for that year. Of the 157 countries in the dataset, 46 entered into a bilateral investment treaty or a regional trade agreement (RTA) with investor-state dispute settlement provisions with the United States at some time during the period of observation.\textsuperscript{14} I omit all countries for which an investment treaty is in force for all time periods under observation.\textsuperscript{15} Data on BIT and FTA entry into force is obtained from the UN Conference on Trade

\textsuperscript{13}See https://www.bea.gov/surveys/diasurv.htm.
\textsuperscript{14}Only one additional U.S. BIT partner is omitted from the data due to missing covariate data: Grenada
\textsuperscript{15}These countries are Armenia, Bulgaria, the Czech Republic, Kyrgyzstan, Moldova, Mongolia and Slovakia.
and Development’s (UNCTAD) “International Investment Agreements Navigator.” For countries that enter into an investment agreement with the United States, I code the year of “treatment” initiation as the entry-into-force year if the entry-into-force month is prior to July. Otherwise, I code treatment initiation in the following year. Only one U.S. investment agreement has been formally terminated after initiation. Bolivia terminated its 2001 bilateral investment treaty in 2012. I therefore omit observations for Bolivia in 2012 and 2013 from the dataset. I obtain covariate information on the number of other BITs in force for a given country-year, using the same rules for determining starting year and again relying on the UNCTAD dataset. Data on Gross Domestic Product and real GDP per-capita is taken from the World Bank’s World Development Indicators (WDI) database. I also obtain a measure of distance between each country’s capital and Washington D.C. using latitude and longitude data from the WDI database API. Finally, I include two indicators of democratic governance from the Varieties of Democracy (V-Dem) project measuring electoral democracy and liberal democracy respectively.

Additionally, because the counts of firms differ across states by orders of magnitude, it is implausible that the parallel trends assumption will hold on an additive scale. Some countries have thousands of U.S. affiliates while others have zero. Because the parallel trends assumption is sensitive to the scale of the outcome (Athey and Imbens, 2006), to make the assumption more plausible, I transform the outcome to a logarithmic scale by taking the natural log of the raw counts plus 1 (to avoid issues with taking the log of 0).

Figure 2.3 displays the distribution of treatment uptake times for units in the dataset. Of note is the fact that there are few treatment cohorts with more than a single unit, making it infeasible to simply estimate the treatment effect for each unique cohort. Figure 2.4 plots the estimated treatment effects from both the two-way fixed effects model and the generalized difference-in-differences estimator with and without covariate adjustment. In the unadjusted results, the two-way fixed effects estimate is not statistically significant and is roughly half the size of the generalized DID estimate which is positive and statistically distinguishable from zero at $\alpha = 0.05$. This is consistent with the intuition that two-way FE estimators downweight longer term effects and upweight short-term effects.

16See http://investmentpolicyhub.unctad.org/IIA.
Figure 2.3: Distribution of investment treaties with the U.S. in force over time
investment treaties are unlikely to exhibit their full effects immediately and operate through longer-term channels, the two-way FE estimator may be biased towards zero. Using the new generalized DID approach, I find that, for the typical country that adopts an investment treaty with the U.S., the treaty increases the count of U.S. MNE affiliates operating in that country by about 0.2 log-points, or approximately a 20% increase relative to that country’s baseline.

To adjust for time-varying covariates in the outcome regression, I include in the two-way fixed effects OLS regression the linear combination of a country’s logged real GDP per capita measured at time $t$, logged real GDP also measured at $t$, whether that country is a member of the GATT/WTO at time $t$, the two V-Dem democracy indices (electoral and liberal democracy), and the log of the number of bilateral investment treaties that country has in force with non-U.S. partners at time $t$ plus 1. Including these covariates drives the estimated effect from the two-way fixed effects model even closer to 0.

I use a similar additive model to estimate the propensity of treatment initiation. I estimate a logistic regression and add to the set of covariates above the log of the distance between the country’s capital and Washington D.C. along with a linear time trend. Using this model, I predict the probability that
each unit would initiate treatment at time $t$ and use these weights to estimate the ACTT via the method described in section 2.4. Surprisingly, even with the weighting adjustment, the estimated treatment effect does not change substantially. The point estimate itself shifts by only about 0.02 log-points, with a slight increase in the standard error due to the weights. While this increase in variance does raise the corresponding p-value, the estimate is still statistically significant at $\alpha = .1$. Additionally, the sizeable gap between the generalized DID estimate and the two-way FE regression estimate illustrates the pitfalls of conditioning on time-varying covariates that are potentially post-treatment. When the time-varying covariates are included with a method that avoids this issue, the positive effect of BIT initiation remains.

Overall I find, contrary to recent survey evidence suggesting BITs are irrelevant to firms’ decisions to invest (Poulsen, 2010), that BITs increase the number of foreign affiliates from the BIT partner that operate in that country. While this estimate is limited to only U.S. BITs, it is not implausible that the effect generalizes to treaties with other capital exporting countries as investment treaties exhibit remarkable homogeneity across countries. It may well be that the effect of the BIT operates through subsequent policy changes in a host country that are ancillary to the treaty commitment itself. Therefore, even if firms do not directly respond to the presence or absence of an investment treaty when planning their investments, the treaty affects other variables that do factor into that decision. It is also possible that countries entering into an investment treaty with the United States are implementing a broader regime of capital-friendly policies that are difficult to disentangle from the BIT alone. Therefore, it is entirely within the realm of possibility that firms do not take BITs into account when investing, as Poulsen (2010) argues, but that effect of implementing a BIT does boost foreign direct investment overall.

2.6 Conclusion

This chapter addresses a major flaw in the way researchers typically implement difference-in-differences estimators in panel data settings. Ordinary least squares with two-way fixed effects, while valid when
there are two time periods and only two possible treatment assignment histories, is biased in the more general case of more than two time periods and treatment histories. Inference in this setting requires much stronger modeling assumptions in order to remain valid. I relax the most stringent of these assumptions, that treatment effects do not persist over time, to develop a non-parametric estimator under the constraint that units receiving treatment do not revert to control in subsequent time periods. I define a new quantity of interest, the Average Cohort Treatment effect on the Treated (ACTT), which corresponds to the average effect of initiating treatment for a unit randomly selected from the sample. I show that the ACTT is identified non-parametrically as a weighted average of two-period difference-in-differences estimates and provide a straightforward weighting method for relaxing the parallel trends assumption by conditioning on observed covariates.

One limitation of this analysis is that it considers only the case of the “static” two-way fixed effects model which does not include as regressors additional leads or lags of the treatment variable. It might then be argued that including such parameters would resolve any issues regarding treatment effects over multiple time periods as the model would include parameters for the “initial” effect of treatment and the treatment effect for subsequent periods as well. Unfortunately, the choice of the number of leads and lags remains up to the researcher, requiring an additional assumption about effect persistence. Moreover, including leads or lags of the treatment variable induces post-treatment bias in some of the coefficients in the model and not all parameters in the model will be causally interpretable in terms of counterfactual comparisons. This is particularly true if covariates are also included in the model since some time-varying covariates will be affected by model lags (Blackwell and Glynn, 2018). Additionally, as Abraham and Sun (2018) note, inclusion of leads and lags does not address issues of improper weighting of heterogeneous effects. Unfortunately, adding more parameters to the two-way FE model is not an adequate panacea for the problem identified in this chapter.

The method described in this chapter is also limited to the case where units do not revert to control after initiating treatment. While this reflects many situations encountered by social scientists, particularly in studies of policy implementation, some types of treatments exhibit reversion over time, especially when the number of time periods under consideration is large. For example, researchers study-
ing democratization have to consider the possibility of democratic backsliding. Whether a country that has always been an autocracy should be considered as having the same “treatment” condition as one that democratized and subsequently reverted to being a non-democracy is a substantive question for which the answer will depend on the particular research question being asked. Subsequent work should consider other approaches to reducing the dimensionality of this treatment history space without necessarily requiring persistence of treatment uptake, but also not restricting treatment effects to single periods as in the two-way FE estimator.
Chapter 3

Why Rich Countries Win Investment Disputes

3.1 Introduction

Over the last several decades, the proliferation of bilateral investment treaties (BITs) among states has granted many foreign investors access to a unique legal mechanism to enforce their property rights abroad. BITs, and increasingly many bilateral and multilateral trade agreements, commit states to protecting the investments of foreign nationals, obligating governments to maintain certain standards of treatment and to refrain from uncompensated expropriation. These treaties also often contain provisions allowing investors to directly seek damages for treaty violations through litigation before an ad-hoc international arbitration tribunal, a system known as investor-state dispute settlement (ISDS). While arbitration among investors and states is not new – contract based arbitrations date back to as early as 1864 (Yackee, 2016) – BITs have enabled the rapid expansion of ISDS claims by expanding the number of investor-state conflicts that could be brought to arbitration. By granting blanket consent to arbitration to broad classes of foreign investors, BITs have dramatically increased states’ exposure to litigation from foreign firms.
While capital-importing governments have sought out BITs as a means of attracting more foreign direct investment by creating a more certain legal environment for investors, the actual FDI-enhancing effect of investment treaties remains hotly debated (e.g. Neumayer and Spess, 2005; Haftel, 2010; Tobin and Rose-Ackerman, 2011). What is clear, however, is that by ratifying BITs, states have opened themselves up to a wave of litigation from investors (Simmons, 2014). Awards issued by arbitration tribunals can amount to a sizeable share of state budgets, with the average award for successful claimant firms amounting to about $76 million USD (Hodgson, 2014). Moreover, recent trends in investment litigation have expanded the scope of claims that are brought against states. While early investment arbitrations largely dealt with targeted expropriation or mistreatment of individual firms by governments, many recent arbitrations have instead challenged broad regulatory policies. Claimant investors allege that environmental, health, and other types of regulations violate the respondent state’s treaty obligations by diminishing the value of a firm’s investment (Pelc, 2017). The growth of regulation-centered arbitrations has generated significant concerns over the possibility that the threat of litigation can be used to limit the policymaking flexibility of governments.

Moreover, in contrast to other international institutions that grant private individuals standing to bring claims against states, like the European Court of Human Rights, ISDS is unique in that there is no formal court in which claims are adjudicated.¹ Rather, ISDS claims are litigated through a form of ad-hoc arbitration. The seat of arbitration is itself often de-nationalized and not governed by any individual state’s legal system. Arbitrators do not have permanent tenure and serve on a dispute-by-dispute basis. The parties themselves typically will each directly appoint one of the arbitrators to a tribunal. Arbitrators compete for re-appointment and the pool of investment arbitrators is comprised of a very small, elite, group of legal experts that rotate between sitting as arbitrators and working as counsel for the parties to a dispute (Puig, 2014). Investor-state arbitrations, borrowing from practices in commercial arbitration, are typically kept very private. Hearings are closed, and awards, when published, are

¹Efforts to develop more formal mechanisms of adjudication are being incorporated into some multilateral trade agreements, such as the recent Canada-E.U. CETA agreement. (See “Investment provisions in the EU-Canada free trade agreement (CETA)” http://trade.ec.europa.eu/doclib/docs/2013/november/tradoc_151918.pdf) However, the idea of an investment court remains very much in its infancy.
often partially redacted (Rogers, 2005). There exists no formal appellate process and there are few avenues for courts to review awards after they have been rendered (Laird and Askew, 2005).

It is no wonder that ISDS has become a target of substantial criticism from commentators, activists, and government officials. These critiques allege that the system grants too much power to private corporations at the expense of state sovereignty and legitimate democratic policymaking. Media accounts of ISDS have gone so far as to describe it as a “global super court that empowers corporations to bend countries to their will.” While the precise normative question of whether ISDS should exist in the first place is a matter for politics and far outside the scope of this paper, many critiques of ISDS rely on claims that can be empirically evaluated. If ISDS grants private corporations power to influence the behavior of states, the impact of this shift in power will not be uniform across states. Rather, we would expect that states less well-equipped to resist pressure from foreign firms will be the most impacted by this system. In other words, we would expect to see clear distributional effects apparent in the empirical data. One major proposition advanced by scholars is that ISDS is systematically biased against the interests of developing countries and in favor of Western capital-exporting states (Waibel, 2010; Trakman, 2013). As Schultz and Dupont (2014) declare: ISDS favors the “haves” over the “have-nots.” Indeed, it is the case that the majority of firms filing disputes are based in high-income states while low- and middle-income states are the typically the target or “respondent” in such claims. And as some governments, such as Bolivia, Venezuela, and Ecuador, that have been the target of many investment disputes withdraw from or revise their treaty obligations, allegations of institutional bias are frequently made as part of the justification (Brower and Blanchard, 2013, 709).

It is difficult to assess from dispute-level data whether ISDS as an institution is systematically more pro-corporate rather than pro-state, as this would require disentangling the bias in the legal provisions governing ISDS from how they are applied by the institution itself. Is the entire system of ISDS biased in its interpretation of the law, or does the problem lie with the legal obligations to which states have committed? Answering this question would necessitate some normative belief about what the “true” balance of corporate and state rights are under international law and the correct interpretation of the

law. But dispute resolution facilities exist precisely because there is no agreement about how the law should be interpreted.

A more tractable empirical question is whether the law is applied evenly across different types of respondent governments. In particular, are poorer respondent states with fewer available resources systematically disadvantaged in ISDS tribunals relative to wealthier governments? It is reasonable to state that in a fair system, identical actions by governments, giving rise to identical disputes, with identical facts, brought under identical legal standards, should receive the same outcome, regardless of whether the respondent state is a developed or developing country. However, research on other international courts suggests affinity biases may cause arbitrators to render different decisions based on attitudinal factors. For example, Posner and de Figueiredo (2005) find that judges in the International Court of Justice tend to favor governments with similar development level and political/cultural similarity.

![Graph](image)

**Notes:** $N = 383$. Lines denote 95% confidence intervals

**Figure 3.1:** Win-rates by respondent income level in 383 investor-state treaty arbitrations, 1987-2016

Existing work on the effect of respondent government development level on win-rates in ISDS reaches somewhat mixed conclusions. Franck (2009) finds that wealthy governments are no more likely to receive favorable awards. Franck (2014), considering a larger sample of disputes argues that differences in success rate between wealthy and poorer governments is attributable to differences in levels of democratic governance. In a further update of the investment arbitration data, Wellhausen (2016) finds that on average OECD governments tend to win more disputes. Behn, Langford and
Berge (2017) also argue that governance quality does not completely explain differences in win-rate among developing and developed governments. Finally Nunnenkamp (2017) argue that the disparity between rich and poor countries is mitigated by the types of arbitrator appointed to the panel. Figure 3.1 plots the average claimant win-rates by World Bank income category for all disputes considered analyzed in the chapter. While it is quite clear that claimants tend to win significantly fewer claims against high-income respondents compared to lower and middle-income respondents, the implications of this observed difference remain unclear. Does respondent nationality have a causal effect on dispute outcomes?

This chapter argues that answering whether nationality matters for outcomes first requires answering whether nationality matters for settlements. Empirical work on dispute outcomes has ignored a severe source of confounding and bias when considering the relationship between respondent government characteristics and win-rates in ISDS. Not all ISDS disputes reach the point where an arbitration tribunal renders an award. Many disputes settle or are withdrawn. Therefore, analyses of investment dispute outcomes implicitly condition on a post-treatment variable, that a dispute failed to settle. When the causal variable of interest affects the chances that a dispute settles, this can induce spurious correlations in the subset of the sample that is observed to have an award (Rosenbaum, 1984; Montgomery, Nyhan and Torres, 2018). Without a model for how some disputes settle, analyses of ISDS outcomes are likely to generate highly misleading results, even when researchers account for all potential confounders of respondent wealth and dispute outcome.

A long line of research in political science notes the importance of a state’s legal capacity for both bringing claims before international legal fora and successfully litigating them. Much of this work has considered legal capacity in the context of the World Trade Organization (WTO), but the results have similar implications for research on investment arbitration as well. Busch and Reinhardt (2003) argue that because developing countries may find litigation costly, they are often unable to compel concessions from weak defendants due to the lack of a credible threat to pursue a dispute. Busch, Reinhardt and Shaffer (2009) find in surveys of WTO delegations that legal capacity is a significant hurdle to pursuing further litigation for many governments. Likewise, Guzman and Simmons (2005) argue that
the absence of WTO claims from low-capacity governments against all but the highest-income governments suggests that there is a negative expected return for many governments to challenging smaller trade barriers due to the costliness of actually litigating a dispute. Davis and Bermeo (2009) note that litigation costs create an initial barrier for many developing countries in the WTO, but highlight that it may be overcome through experience with litigation.

The costliness of litigation in investment arbitration is well-known. Hodgson (2014) finds that the average costs of an investment arbitration treaty are around $10 million USD. This is a little under one-seventh of the size of an average award (among disputes where an award is issued). Moreover, cost recovery is uncertain in arbitration. While some tribunals will require that the loser of the dispute pays the costs of the winner, others will choose to have the parties pay their own way. There is currently no systematic guidance as to the method to follow, and arbitrators may choose one approach over the other due to idiosyncratic factors (Franck, 2010). A state that knows its case is strong cannot guarantee that litigation will be costless. In addition, the close-knit structure of investment arbitration places a premium on obtaining lawyers with the relevant legal expertise. Gottwald (2006) notes that while claimants almost always rely on one of the major law firms specializing in investment arbitration when bringing a claim, developing countries may lack similar in-house expertise. As a result, developing countries will also need to invest in hiring outside counsel or risk presenting a low quality defense. It is not simply that the cost of arbitration for a developing country may make up a larger fraction of its budget; developing countries typically need to invest more overall in order to properly defend themselves in arbitration proceedings.

Theoretical models of pre-trial bargaining suggest that disparities in legal capacity and variation in the costs of litigation can result in stark asymmetries between well-resourced and poorly-resourced states during the bargaining process with an equivalent claimant firm. Intuitively, governments for which litigation is comparatively less costly are more willing to gamble on fighting a dispute when uncertain about the claimant’s quality. I show that this is the case empirically as well. After adjusting for likely confounders, disputes with high-income respondents are about 22 percentage points more likely to receive a final award by a tribunal relative to low- and middle-income respondents.
Since the respondent state’s income level affects the propensity of settlement, among those cases that settle, researchers should expect to observe a spurious association between respondent income-level and outcome if there exists some third variable correlated with both settlement and outcome. Claimant quality is one such variable. High-quality claimants that expect to win are actually less likely to settle when respondents are uncertain over the quality of the claimant’s case. Because respondents need to balance the risk of over-compensating a weak claimant against the costliness of litigation they will tend to under-provide settlement offers, an example of the classic “market for lemons” problem in economics (Akerlof, 1970). Well-resourced respondents will tend to provide lower offers to claimants because they are more willing to tolerate going to trial. Therefore, the average case quality of the claimant should be lower among unsettled disputes against high-capacity respondents compared to low-capacity respondents. Essentially, those weak claimants that might be able to extract a settlement from a litigation-averse state have to fully litigate their dispute against a respondent with fewer barriers to going to arbitration.

Theory suggests that attrition will be both caused by the causal variable of interest, the respondent’s characteristics, and confounded with the outcome. This creates a problem for analyzing the empirical data. Indeed, this problem of post-treatment drop-out is common in other areas of research. For example, in medical studies, individuals under observation may die prior to the observation of some post-intervention outcome. This is sometimes termed attrition bias or “truncation-by-death” (Zhang and Rubin, 2003) – those individuals who die or drop out do not have well-defined values of the outcome of interest. Social scientists encounter similar situations. In general, when some outcomes are non-existent due to an intermediate event, the only well-defined average causal effect is the treatment effect on the sub-sample that would “survive” to follow-up regardless of treatment status, what Rubin (2006) terms the “Survivor Average Causal Effect” (SACE). Analyzing the data conditional on the observed survival outcome will result in biased estimates of the causal effect when treatment has some additional effect on the propensity that a unit reaches follow-up. Methods for estimating the SACE often require that researchers make additional assumptions. The approach used in this chapter, principal score weighting (Jo and Stuart, 2009; Ding and Lu, 2016; Feller, Mealli and Miratrix, 2017), uses
weights to adjust for imbalances in observed confounders of attrition and outcome that are induced by conditioning on survivorship. I illustrate a new method for estimating these weights that is more robust to mis-specification of the weighting model and improves balance between treated and control on relevant confounders. I find that after adjusting for both treatment confounding and post-treatment selection bias, high-income respondents do not win cases more often than low or middle-income at a rate that can be statistically distinguished from zero. In other words, the observed difference in outcomes shown in figure 3.1 can be reasonably explained by differential settlement rates between high-income and low/middle-income states. The overall findings of this chapter show that disparities in legal capacity among states have immense and counterintuitive consequences for the functioning of ISDS. Observed gaps in win-rates between developed and developing respondents appear to be attributable to a much more subtle difference in how legal capacity affects early settlement.

The remainder of this chapter is structured as follows. Section 3.2 outlines a theoretical model of bargaining over settlement between firms and states. It summarizes two key implications of the bargaining model: high respondent legal capacity reduces the rate of settlement, and the average case quality among disputes that fail to settle will be lower for high-capacity respondent governments. Section 3.3 discusses the empirical strategy for estimating the effect of legal capacity on settlement by adjusting for potential confounding variables. Section 3.4 explains how standard covariate adjustments alone are insufficient to estimate effects under attrition. It defines the “Survivor Average Causal Effect” as a quantity of interest and describes the assumptions necessary to estimate the effect using principal score weighting. Section 3.5 describes the data used in the analysis. Section 3.6 presents the results and analyzes the extent to which the covariate adjustment methods were sufficient to reduce imbalance. Section 3.7 concludes by outlining the implications of these results for policymakers and for the ISDS regime in general. It notes that efforts to improve the fairness and legitimacy of investor-state dispute settlement should pay attention to disparities in legal capacity among the parties, and, in particular, how these disparities may be exploited by claimants in order to compel early settlements.
3.2 Settlement and Arbitration

An extensive line of theoretical research in law and economics highlights the importance of legal costs and litigants’ resource endowments in explaining why some cases settle prior to trial. Because litigation is costly for both parties, each side has an incentive to reach an acceptable agreement rather than pay the costs of going to trial. However, when there is uncertainty between the parties, settlements are not guaranteed. Parties that settle too readily run the risk of paying off unworthy claims, while those that refuse to settle are forced to pay the additional costs of defending themselves in court. Anecdotal accounts from officials involved in ISDS litigation suggest that governments facing investment disputes are keenly aware of this risk-reward trade-off. For example, a recent report on ISDS by BuzzFeed News quoted Marie Talasova, a top lawyer for the Czech Republic’s Ministry of Finance, who stated “Every month I get a threat...We have to review the risks, how strong the claim is. We try to minimize the costs of the state.”

A strategic government facing an ISDS claimant wants to offer as small of a settlement as possible, but runs the risk that offers that are too small will be rejected by the claimant, resulting in costly arbitration proceedings.

This section discusses the implications of existing theoretical models of pre-trial bargaining for the relationship between legal capacity and propensity to settle. It argues that settlement rates should be higher for governments with high litigation costs as firms are less likely to accept settlement offers from wealthier governments. Intuitively, wealthy governments can force more claims to arbitration because the additional costs imposed on the government by litigating are comparatively smaller.

Consider the model proposed by Bebchuk (1984) where two risk-neutral litigants bargain over a settlement in the shadow of potential costly litigation. Assume the claimant has private information over the probability of their claim being successful. The respondent state does not observe this, only know-

---


4While Bebchuk (1984) models uncertainty primarily on the plaintiff/claimant side, the original paper notes that the model can be straightforwardly reversed with all comparative statics intact by allowing the defendant to be uncertain of the claimant’s likelihood of winning.
ing the distribution of the quality of potential claims.\textsuperscript{5} When a claim is brought, the state makes the claimant a settlement offer, which the claimant can choose to either accept or reject. In the context of this model, it is assumed that the claimant can credibly threaten litigation by rejecting the respondent’s offer. Therefore, if the claimant rejects, the dispute proceeds to litigation.\textsuperscript{6} When a dispute is litigated, each party pays some fixed cost.

Claimants will accept settlement offers only if they are equal to or greater than their expected value to litigation. Given a fixed award size and litigation costs, minimum settlement offers accepted by the claimant are increasing in the quality of the claimant’s case. The respondent must therefore balance two competing incentives in choosing a settlement proposal: conditional on acceptance, lower offers are better than higher offers, but lower offers are more likely to be rejected, resulting in the potentially worse outcome of litigation. Since the respondent does not directly observe the claimant’s case quality and reservation value, the respondent cannot tailor its offer to the specific case at hand. Instead, it may make an offer that is too low, thus forcing the claimant to resort to fighting the case in arbitration.

How do the parties’ litigation costs affect the likelihood that a settlement will succeed? On the respondent’s side, as litigation becomes more costly, the government becomes more willing to offer a larger settlement in order to increase the chance of acceptance since the costs to it of a rejected settlement are higher. Conversely, when litigation is relatively costless, a respondent state can force weak claimants that might otherwise secure a sufficiently high settlement into pursuing litigation, thus “revealing” the quality of their case. As the legal capacity of respondents grows, and their costs to litigation fall, their willingness to settle cases with plaintiffs should decrease since the amount that they would choose to offer in order to avoid a trial decreases. High capacity/low cost governments will give

\textsuperscript{5}This source of asymmetry is a more reasonable assumption for the investment arbitration context where firms are likely to know more about their own valuation and the details of their particular investment than governments do.

\textsuperscript{6}This assumption essentially implies that the expected value of litigation is always positive for the firm, which may not be the case for litigation threats aimed exclusively at obtaining a settlement. Since this chapter focuses primarily on those claims where proceedings have been initiated, and therefore the claimant has paid some initial cost, this assumption is plausible. However, if it were possible to see the entire universe of threatened disputes, it is likely that many would be ‘frivolous’ and aimed purely at extracting some payment from the state. See the model in Nalebuff (1987) which relaxes the assumption that the claimant is committed to litigation.
offers that satisfy only those claimants with a small reservation value – those with the weakest claims. All other claimants are forced to pursue litigation.

Differential settlement rates should therefore also result in ex-post differences in case quality among those respondents with high legal costs and those with lower barriers to litigation. The model predicts that high-quality claimants will fail to settle against both types of respondents. This is due to a type of “adverse selection” (Akerlof, 1970) in settlements as respondents give unacceptably low offers to high-quality claimants because they cannot directly observe the claimant’s type. Because respondents have to consider the possibility that the claimant could have a low-quality case, they make offers that hedge against the risk of over-compensating a weak litigant. Conversely, a low quality claimant is more likely to receive an acceptable offer from a respondent with high litigation costs. This is because the risk to the respondent of going to arbitration inflates the amount that they are willing to offer in bargaining. As the amount the respondent offers decreases with reductions in the respondent’s legal costs, some of these low quality claimants would receive unsatisfactory offers and instead proceed to arbitration. The model therefore predicts that among the cases that receive awards, the average claimant will be of higher quality in disputes with low capacity governments relative to governments with high legal capacity. Settlement induces a spurious negative correlation between legal capacity and the claimant’s win-rate that is attributable to high-capacity respondents being less willing to settle against weak claimants. This explains why the share of claims successfully won by states might differ significantly between high- and low-resource governments even under an entirely fair adjudication system.

3.3 Covariate adjustment for estimating treatment effects

Theory predicts that respondent income level has a negative causal effect on the probability that a dispute will reach an early settlement. Testing this prediction against the data requires developing a credible design for inferring causation. This section briefly reviews the assumptions necessary to estimate a causal effect in an observational study and discusses the strategy I will use to adjust for omitted vari-
able bias when estimating the effect of legal capacity on settlement. Formally, consider a sample of \( N \) disputes. For each unit, indexed by \( i \), researchers observe the realized value of a treatment of interest, denoted \( A_i \), and a vector of pre-treatment covariates \( X_i \). \( A_i \) is binary with 1 denoting the “treatment” condition: a high-income respondent and 0 denoting the “control” condition: a lower-income respondent. Researchers observe the intermediate outcome (in this study, settlement) \( S_i \) for all units. Let \( S_i = 1 \) denote a dispute that fails to settle (i.e. “survives” to the award stage) and \( S_i = 0 \) denote a dispute that is withdrawn or eaches a settlement. Researchers also observe a final outcome \( Y_i \) (whether the claimant firm wins the case) for all units with \( S_i = 1 \). I focus first on estimating the effect of \( A_i \) on \( S_i \) since \( S_i \) is observed for all units. The subsequent section will discuss the additional complications that arise when considering effects on partially observed outcomes.

Causal effects are defined using the Neyman-Rubin potential outcomes framework (Holland, 1986; Rubin, 1974). I define latent quantities for each unit with respect to both post-treatment variables, \( S_i \) and \( Y_i \). \( S_i(a) \) denotes the intermediate settlement outcome that would be observed for unit \( i \) if that unit, possibly contrary to fact, were assigned treatment level \( a \).\(^7\) Likewise, \( Y_i(a) \) denotes the final outcome that one would observe if unit \( i \) were assigned to treatment level \( a \). With a binary treatment, the causal effect of treatment on settlement for an individual \( i \) is \( S_i(1) - S_i(0) \). However, only one of these two potential outcomes can ever be observed (Holland, 1986). Therefore, researchers typically focus on estimating the average causal effect over units in the sample: \( E[S_i(1) - S_i(0)] \).

If respondents’ wealth level were randomly assigned, researchers could estimate the average causal effect without bias simply by comparing those disputes with high-income respondents and with disputes with low-income respondents. Unfortunately, in observational settings, the treatment is not randomly assigned. Rather, the characteristics of disputes can be expected to vary between disputes with high-income respondents and disputes with poorer respondents. When these characteristics are also predictive of the outcome of interest, unadjusted comparisons between treatment arms risk biased inferences – the classic omitted variable bias problem. The goal of any observational causal infer-

\(^7\)Writing the potential outcome this way implicitly makes the stable unit-treatment-value assumption (SUTVA) (Rubin, 1986). This assumption states that there are not multiple variations of the same treatment \( a \) and that a unit’s potential outcomes depend only on its treatment assignment (no interference).
ence design is to condition on as many potential confounders as possible such that it is plausible to assume that treatment is assigned as-if-random, conditional on the observed covariate vector $X_i$ (Imbens, 2004). In the context of this chapter, I assume that there are no unobserved variables omitted from the study that are correlated with host country wealth and with propensity to settle.

Selecting the right set of potential confounders requires theoretical knowledge about what a researcher expects will predict both treatment assignment and outcome. In this section I outline three main sources of likely confounding that must be adjusted for: claimant characteristics, dispute characteristics, and treaty characteristics. However, before defining which variables I choose as controls, it is important to note that some variables that might appear to be confounders are really plausibly part of the causal effect of interest. Existing analyses of respondent wealth and ISDS disputes have often included additional control variables related to other country-level characteristics, such as democracy (Franck, 2014) or institutional quality in general (Behn, Langford and Berge, 2017), with the goal of distinguishing the “effect” of development from the “effect” of democracy. However, depending on the causal question of interest, controlling for these variables may be inappropriate as they are plausibly post-treatment. In order to determine whether this is the case, it is necessary to specify what exactly the causal estimand of interest is in this study and on what population it is defined.

This study asks a very specific causal question: “On average, what would have happened in a dispute against a high-income respondent if it were instead brought against a lower-income respondent country country?”8 If the goal is to evaluate the bias of individual arbitration panels with respect to a country-level characteristic such as development, is it necessary to also control for variables like democracy or institutional quality? The answer is no.

This is because a variable like institutional quality is plausibly a part of the overall effect of interest and adjusting for it might induce post-treatment bias by blocking one mechanism by which treatment affects outcome. For example, if it is the case that wealth causes countries to have certain institutional

---

8For why the distinction between case-level and country-level manipulation matters, see Boyd, Epstein and Martin (2010) for a discussion of immutable characteristics and defining valid effects in the context of sex and judging. Greiner and Rubin (2011) discuss the broader question of what constitutes a valid causal quantity with respect to hard-to-manipulate variables.
arrangements that make arbitrators more favorable to that respondent’s arguments regarding, for example, the transparency of its regulatory decisionmaking, then that is still a form of country-level institutional bias. The arbitrator would make a different decision in a world where the case remained the same, but the country were different.

Suppose, however, that a researcher wanted to distinguish between developing country bias attributable to income level and bias attributable to other factors like a government’s institutional quality. This requires estimating a different type of causal effect, the “controlled direct effect” (Acharya, Blackwell and Sen, 2016) of manipulating the income level of a country under another intervention that somehow holds constant institutional quality. These types of causal mediation effects may more clearly define the specific hypothesis envisioned by the researcher. Unfortunately, estimation of such effects for this particular research question would likely result in high-variance estimates and/or heavy model-dependence due to the high co-linearity between development level and the institutional variable researchers want to hold constant (King and Zeng, 2005).

Respondent country characteristics are by definition not confounders of treatment assignment; they are a part of the treatment of interest as the manipulation is carried out by assigning a particular country to a dispute rather than altering the wealth level of a country (a much less coherent manipulation). However, there do exist a number of confounders that are correlated with a case being filed against a high rather than a low or middle income country that likely also influence the outcome. Failing to adjust for these confounding factors would lead to biased causal inferences as any observed correlation between respondent country and outcome could be attributable to a third “lurking” variable.

I consider three primary sources of confounding: claimant type, dispute type, and legal standards. For claimant type, first, I code the highest income level among all claimant nationalities as a general proxy for claimant resources. Additionally, two particular types of national claimants may be confounders of treatment, even among claimants from high-income countries. Claimants from the United States make up the majority of ISDS claimants in the dataset. The history of U.S. BIT negotiations suggests that these treaties were specifically targeted at developing countries with the aim of protecting both existing and future investment. Moreover, these BITs were particularly strict in the scope of
protections afforded to investors and the United States was largely unwilling to relax any such provisions Vandevelde (1988). Therefore, U.S. claimants might have access to a more advantageous network of BITs that also happens to be correlated with respondent country development level. Dutch investment treaties are also well known for being particularly generous to investors. Combined with the relative ease of incorporating in the Netherlands, some firms have been known to forum shop by locating in the Netherlands to take advantage of the availability of BIT litigation. This forum shopping among investors may be negatively correlated with the overall case quality as sufficiently blatant strategic incorporation by otherwise ineligible investors may lead a tribunal to reject a case out-of-hand (Kryvoi, 2010).

Dispute type concerns both the industry of the claimant and the type of incident out of which the dispute arises. Some types of disputes, for example, blatant expropriation, are easier to win than more complex challenges arising out of state regulatory policy. Dispute types are also not assigned evenly among governments as regulatory challenges have been particularly targeted against high-income governments in recent years Pelc (2017). I develop a comprehensive coding scheme to classify all available investment disputes with respect to the type of state action being challenged. The details of this coding are discussed further in the data section.

Finally, disputes differ between rich and poor countries because of the rules governing them. Allee and Peinhardt (2014) and Allee and Lugg (2016) note that wealthier and more powerful countries are more likely to obtain BITs with their preferred set of legal provisions. For example, governments are increasingly looking to include explicit reservations and “carve outs” for public health, environmental and other forms of regulations out of concerns that investment arbitrators may be construing vague obligations broadly in favor of claimant firms (Trakman, 2013). As a result, the international obligations of rich and poor governments with respect to a particular foreign investor may differ substantially. Differences in text are likely to contribute to differences in outcome. Trakman (2013) argues that investment arbitrators, trained predominantly in the field of commercial law, will tend to interpret treaty provisions literally, without regard to the specific circumstances of the state in question (609).

9For evidence on whether arbitrators tend to favor expansive interpretations of arbitration provisions, see (Van Harten, 2015).
Therefore, observed differences in rich country win rates may be explained by differences in the provisions of the treaties under which claims are brought.

One commonly used approach to adjusting for confounding in a non-randomized study is inverse propensity of treatment weighting (IPTW). The approach first estimates a model for units’ “propensity score” defined as the probability that a unit receives treatment given its observed covariates (Rosenbaum and Rubin, 1983). This is typically done by fitting a parametric regression model like a logistic regression that regresses treatment on the observed covariates. It then assigns weights to each observation based on the inverse of the estimated probability that the unit received the treatment that it did. Intuitively, the approach upweights observations that are underrepresented relative to what would be expected under randomized treatment assignment and downweights those that appear too frequently. When the propensity score model is correctly specified, the distribution of covariates should be balanced between treated and control groups in the re-weighted sample.

The advantage of the propensity score weighting method is that it does not require specifying a model for the outcome, as is the case for regression-based adjustments for confounding. Regression modeling can be a problem when the covariate space is particularly high-dimensional as slight changes to the specification of the regression model can lead to large changes in the estimated effects (Ho et al., 2007). Evaluating whether the “correct” model has been chosen can be difficult, and, with access to the outcome data, researchers may be susceptible to searching over the space of all possible regression models to find the one that supports their desired hypothesis (Rubin, 2001). In contrast, propensity score methods have built-in diagnostics to permit researchers to assess the quality of their model and the magnitude of any residual covariate imbalance without reference to the outcomes.\(^{10}\) Additionally, propensity score weighting avoids a common pitfall in multivariate regression analyses: the regression weighting problem. When researchers adjust for confounders by including them in a regression model, the coefficient on the treatment of interest no longer corresponds to an average treatment effect for the population of interest. Rather, the regression coefficient is a weighted average of individual treatment effects, with units with less predictable treatment assignment receiving greater weight. This results in

\(^{10}\)See Austin (2011) for additional discussion of the performance of propensity score weighting relative to regression methods.
regression estimates generating effects that are not necessarily representative of the average effect for the sample of interest (Aronow and Samii, 2016).

The particular method I use to estimate propensity scores for units is an extension of propensity score called the Covariate Balancing Propensity Score (CBPS) (Imai and Ratkovic, 2014). The intuition behind this refinement is that it explicitly incorporates the balancing property of propensity scores into the process for estimating them. In typical propensity score weighting methods, the search for a “correct” specification of the propensity score model requires repeatedly re-estimating models for treatment assignment and evaluating whether the covariate distributions in treated and control are roughly the same. This can be quite tedious as the space of potential models is often quite large. The CBPS approach eliminates the need for repeated model estimation and checking by including the covariate balance conditions as part of the objective function used in estimation. As a result, minor misspecifications of the logistic regression model are less likely to affect balance because the CBPS estimator directly penalizes weight estimates that result in high imbalance between treated and control groups. As shown in Section 3.6, the propensity score weighting model works remarkably well, reducing average imbalance on the covariates by about a factor of seven.

### 3.4 Treatment effects under selective attrition

The propensity score weighting approach outlined in the previous section is sufficient to allow estimation of the effect of respondent country wealth on the probability of settlement. The outcome of interest is well-defined for each dispute being considered. However, propensity score weights alone are not sufficient to adjust for bias in estimating the effect of wealth on the probability of winning the dispute. The theoretical argument in Section 3.2 should make researchers cautious about inferring too much from empirical patterns in observed awards without careful attention to the process by which some disputes are selected out. Simply comparing the win-rate for disputes with high-income governments with the win-rate for low-income governments in order to estimate the effect of respondent wealth falls prey to the classic problem of bias induced by conditioning on a “post-treatment” variable.
Settlement is an intermediate variable that causally follows the independent variable of interest, respondent income level. Theory suggests that the propensity for a dispute to be settled will be affected by the type of government it is brought against. If, counterfactually, the dispute were launched against a different state, then the chances of observing a settlement would change. Therefore, within the sample of disputes that failed to settle, there will be spurious differences between disputes against high-income and low-income respondents on covariates that are also predictive of settlement. Conditioning on a post-treatment variable breaks any balance between treated and control groups achieved through weighting ex-ante. This section outlines the problem of “attrition” in the context of early settlement of investment disputes. Building on work in the biostatistics literature, it illustrates a weighting strategy to correct for bias induced by drop-out that can be directly combined with existing methods of covariate adjustment in observational studies, such as IPT weighting.

Any analysis of outcomes of awards implicitly conditions on settlement failure. This is because the outcome – which party is declared the winner by the arbitrators – is properly defined only for those disputes in which the parties do not arrive at some settlement agreement. An arbitration tribunal by definition cannot issue a ruling for a dispute that has been withdrawn or has already been resolved by the agreement of the litigants. Conditioning on the observed value of whether a settlement occurs or not results in what is sometimes referred to as “collider” bias in the causal inference literature (Greenland, 2003).\footnote{See Cole et al. (2009) for a useful illustration of why this type of bias exists and how it differs from other forms of selection bias researchers can encounter.} Collider bias occurs when researchers control for an intermediate variable that is affected by treatment and there exists another variable that affects both the intermediate and the outcome of interest. Conditioning on a variable that is a common effect of treatment and the other variable creates a pathway for confounding that would not exist otherwise. In the bargaining model from the previous section, one such variable was the quality of the claimant’s case. Claimants with high quality cases were less likely to settle because, under uncertainty, respondents would not offer a large enough settlement to deter litigation. Claimants with high quality cases would also be (by definition) more likely to receive a favorable award from an arbitration tribunal. If respondent state wealth also nega-
tively affects the probability of settlement, then controlling for observed non-settlement introduces a spurious correlation between the respondent state’s income level and case quality (and thereby the outcome of interest). Since the respondent’s litigation costs increase the probability of settlement while the claimant’s case quality reduces the probability of settlement, respondents with high litigation costs will, ex-post, be associated with disputes with higher-quality claimants. These are the cases that failed to settle (i.e. had claimants that refused to accept a deal) despite the fact that settlement against high-cost respondents is easier to reach.

If the outcome were defined for all units, the solution would be straightforward – researchers could simply avoid controlling for whether the dispute settled or not. However, for this particular research question, this not an option as the outcome does not even exist unless the dispute fails to settle. The problem encountered in analyzing legal disputes under strategic settlements is analogous to the problem of “truncation-by-death” that appears in medical research (Zhang and Rubin, 2003). In studies of health outcomes such as quality of life measures that are conducted over an extended period of time, some respondents may drop out of the sample, possibly due to death or other factors. For these respondents, it is not simply that the outcome of interest is “missing,” it is “undefined” or “truncated.” Valid causal effects only exist for a subset of units – those that would survive until follow-up under either treatment condition (Rubin, 2006). When treatment affects the probability that a respondent is “truncated,” analyses of the outcome will be biased for this causal effect, even when treatment is randomly assigned. While methods to address this challenge have been primarily developed in the field of biostatistics, as Frumento et al. (2012) show, social scientists encounter many similar challenges. For example, in labor economics, analyzing the effect of a job training program on wages is complicated by the fact that wages are only defined among those respondents who are employed post-treatment.

Frangakis and Rubin (2002) outline a general approach for defining valid treatment effects with respect to post-treatment complications. They define a class of causal effects, termed “principal stratum effects” as treatment effects on the outcome within a particular grouping or stratum of units defined by the joint potential outcomes of the intermediate variable under all possible treatment conditions. The intuition is that, while the observed indicator of attrition or “survival” is a post-treatment quan-
tity, the set of all potential survival outcomes under all possible treatments is a latent, pre-treatment characteristic. If researchers were to somehow know the individual causal effects of the treatment on the intermediate, they could estimate effects on the outcome without post-treatment bias simply by looking at differences in outcome between treated and control among units known to have zero causal effect on the intermediate.

When the intermediate variable is a binary indicator for whether units “survived,”¹² and when treatment is binary, there exist four unique principal strata: those units that would survive to follow-up regardless of treatment (the “always-survivors” with \(S_i(1) = 1, S_i(0) = 1\)), those units that would survive only under treatment and not control (the “partial-survivors” with \(S_i(1) = 1, S_i(0) = 0\)), those that would survive only under control but not under treatment (the “partial-survivors” with \(S_i(1) = 0, S_i(0) = 1\)), and those that would never survive under either treatment condition (the “never-survivors” with \(S_i(1) = 0, S_i(0) = 0\)).

Zhang and Rubin (2003) and Rubin (2006) argue that the in the case of truncation, researchers cannot estimate average treatment effects for the full sample. \(E[\hat{Y}_i(1) - \hat{Y}_i(0)]\) is undefined for three of the four principal strata. The only stratum for which an ATE exists is the “always survivor” group.¹³ All other strata contain units for which one or both potential outcomes is missing or undefined.¹⁴ This conditional causal quantity is labeled the “Survivor Average Causal Effect” (SACE) and is defined

---

¹²For the purposes of this chapter, I use “survived” as shorthand for a unit that reaches the point at which the outcome is observed. In the context of early settlement, the “survivors” are disputes that did not reach a settlement before the arbitration tribunal issued an award.

¹³It is worth noting that principal stratification is applicable to many types of post-treatment complications, not just attrition. For example, in studies with imperfect compliance, instrumental variables methods are used to estimate effects for what is termed the “complier” stratum – the set of units that would take treatment when assigned treatment and take control when assigned control (Angrist, Imbens and Rubin, 1996).

¹⁴In theory, a type of causal effect can be defined for this sub-group. However, it is no longer just an average treatment effect since it requires considering the intermediate variable, \(S_i\), as another treatment variable that can be manipulated and “forced” to take on a value of 1 for all units. Such “controlled” treatment effect quantities are often considered in causal mediation analysis (Acharya, Blackwell and Sen, 2016). These effects may be conceptually difficult to justify because interventions on mediating variables are rarely well-defined. For a discussion of the conceptual challenges with such interventions on intermediate variables that arise in the literature on causal mediation, see VanderWeele and Vansteelandt (2009). Practically speaking, treating the intermediate as another manipulable treatment variable requires adjusting for all post-treatment confounders of the intermediate and outcome, which can be challenging when few post-treatment covariates are observed. In contrast, the approach described here only requires adjusting for pre-treatment covariates.
formally as

$$\text{SACE} = E[Y_i(1) - Y_i(0)|S_i(1) = S_i(0) = 1]$$  \hspace{1cm} (3.1)$$

or the average difference in potential outcomes under treatment and control for units that would always have an observable outcome under either treatment arm.

Unfortunately, principal strata are only partially observed. For any unit $i$, treatment assignment reveals $S_i(1)$ or $S_i(0)$, but not both. Table 3.1 illustrates the relationship between the observed quantities, treatment $A_i$, survivorship status $S_i$ and the possible principal strata to which that unit belongs. The observed data is only sufficient to rule out two of the four strata for each unit. Every unit with $S_i = 0$ is guaranteed to not be an always-survivor. However, the set of units with $S_i = 1, A_i = 1$ is a mixture of always-survivors and survivors-only-under-treatment. Likewise, the set of units with $S_i = 1, A_i = 0$ is also a mixture of always-survivors and survivors-only-under-control.

Table 3.1: Observed data and possible principal strata

<table>
<thead>
<tr>
<th>Observed Quantities</th>
<th>Principal Strata</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
<td>$S_i$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

A naive comparison of treatment and control arms conditional on survival will be biased for the SACE if there exists an effect of treatment on survival and potential outcomes are not independent of stratum membership. One assumption that is frequently made to simplify analyses is to assume that the treatment effect on the intermediate variable is monotonic.
Assumption 3.4.1 Monotonicity

\[
S_i(1) \geq S_i(0)
\]  

(3.2)

Monotonicity rules out the stratum of units that would survive under control but not under treatment. The result is that a subset of observed units are known to be always-survivors – those units under control \((A_i = 0)\) that nevertheless survive \((S_i = 1)\). Given the model of early settlement outlined in this paper, monotonicity is not an unreasonable assumption. For a fixed claimant, lowering the respondent’s costs to litigation lowers the size of the offers that they are willing to give claimants since the costs of rejection are lower. Because it would be irrational in the model for a claimant to reject a higher offer while accepting a lower offer, lowering the size of the offer given by the claimant can only increase the number of claimants that choose to follow through with litigation. While theoretical models do not necessarily capture all of the dynamics involved in the strategic process being modeled, they do provide some important intuition to allow researchers to assess the reasonability of empirical assumptions. If it is the case that governments are constrained from litigating due to legal capacity challenges, it is difficult to envision a situation where, all-else-equal, a less-constrained government would favor settlement while a more-constrained government with high litigation costs would prefer to fight the dispute in court.

With only three strata, the stratum proportions can be identified from the observed data (again, assuming ignorable treatment assignment).

Always-survivors: \(Pr(S_i(1) = S_i(0) = 1) = Pr(S_i = 1|A_i = 0)\)  

Survivors under treatment: \(Pr(S_i(1) = 1, S_i(0) = 0) = Pr(S_i = 1|A_i = 1) - Pr(S_i = 1|A_i = 0)\)  

Never-survivors: \(Pr(S_i(1) = 0, S_i(0) = 0) = 1 - Pr(S_i = 1|A_i = 1)\)

(3.4)

(3.5)

Under monotonicity, the average outcome for the control survivors is equal to the average potential outcome \(Y_i(0)\) for the always-survivors. The average outcome for treated survivors is a mixture of the
average potential outcome \( Y_i(1) \) for always-survivors and the partial-survivors.\(^{15}\)

\[
\widehat{SACE} = E[Y_i|A_i = 1, S_i = 1] - E[Y_i|A_i = 0, S_i = 1] \quad (3.6)
\]

\[
\widehat{SACE} = E[Y_i(1)|S_i(1) = 1] - E[Y_i(0)|S_i(1) = 1, S_i(0) = 1] \quad (3.7)
\]

Let \( \pi_{ab} = Pr(S_i(1) = a, S_i(0) = b) \). We can expand the first term by conditioning on the other potential outcome.

\[
\widehat{SACE} = E[Y_i(1)|S_i(1) = 1, S_i(0) = 1] \frac{\pi_{11}}{\pi_{11} + \pi_{10}} + E[Y_i(1)|S_i(1) = 1, S_i(0) = 0] \frac{\pi_{10}}{\pi_{11} + \pi_{10}}
\]

\[
- E[Y_i(0)|S_i(1) = 1, S_i(0) = 1] \quad (3.8)
\]

Subtracting off the true SACE and rearranging terms yields the bias:

\[
\text{Bias}(\widehat{SACE}) = [E[Y_i(1)|S_i(1) = 1, S_i(0) = 0] - E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]] \frac{\pi_{10}}{\pi_{11} + \pi_{10}} \quad (3.9)
\]

Intuitively, the bias is a function of a) the difference in potential outcomes between the always-survivor stratum and the “survive-only-under-treatment” stratum and b) the size of the “survive-only-under-treatment” stratum. When the treatment effect on the intermediate variable is small, the overall magnitude of any bias will be negligible. The direction of this bias will depend on whether the potential outcomes are higher or lower in the always-survivor stratum relative to the partial-survivor stratum. This is not directly knowable, but theory can help inform the direction for sensitivity analysis. From bargaining theory, there is strong reason to believe that the claimant’s probability of winning should be greater among those cases that would never settle compared to those that would settle under control.

\[
E[Y_i(1)|S_i(1) = S_i(0) = 1] > E[Y_i(1)|S_i(1) = 1, S_i(0) = 0] \quad (3.10)
\]

\(^{15}\)For simplicity and the purposes of illustration, I omit conditioning on \( X_i \) here as would be necessary in an observational study.
This is because the types of claimants that would never accept an offer from either respondent tend to have better quality cases since respondents will make offers that are too low in order to hedge against the possibility that the claimant is low quality (Akerlof, 1970). Those claimants that would accept some offers have weaker claims than those that would accept no offers since their expected pay-off to litigation is lower. Therefore, the naive estimator will tend to over-estimate the negative effect of higher respondent capacity on the probability that the claimant wins the dispute.

It is possible to then conduct a sensitivity analysis for the SACE by varying researchers’ beliefs about the size of this difference (Chiba and VanderWeele, 2011). Since there is no clear expectation for what a reasonable magnitude of confounding actually is, I omit this particular exercise. However, it is worth noting that in the worst case scenario, with a bounded, binary outcome, the largest possible difference in the two expectations in the bias term is 1. Therefore, the magnitude of the bias is equal to the share of partial survivors divided by the share of partial survivors or always-survivors. Researchers can obtain bounds for the true SACE (Imai, 2008). For this particular application, bounds are uninformative as the size of the partial-survivor stratum is large relative to the naive treatment effect estimate.

If principal stratum membership were independent of the potential outcomes, then analyses conditional on $S_i = 1$ would be unbiased for the SACE even in the presence of a treatment effect on the intermediate variable. Principal score methods, proposed by Jo and Stuart (2009) with recent developments in Aronow and Carnegie (2013), Feller, Mealli and Miratrix (2017) and Ding and Lu (2016), outline an approach for estimating stratum-specific effects using covariate adjustments. The key assumption motivating principal score adjustment is what is termed “principal ignorability” – that conditional on a set of covariates, principal stratum membership is independent of the potential outcomes. In other words, the observed control variables are sufficient to account for differences in the outcomes across principal strata. Adjustment is done by estimating the “principal score” for each unit, which is defined as the probability, conditional on $X_i$, of an observation appearing in a particular stratum. Sub-classification or weighting by the principal score can then be used to estimate the effect for the stratum of interest.

Principal score methods are intuitive for researchers to implement since they share many features
with commonly used propensity score methods for addressing treatment confounding. Just as propensity score methods adjust for potential confounders of treatment and outcome, principal score methods adjust for common causes of stratum membership and the outcome. Model diagnostics are similar as well – under a correctly specified principal score model, the covariate distributions should be the same across the re-weighted treated units and the control units (the latter having a known principal stratum membership).

While principal score methods were initially designed for estimating treatment effects under partial compliance for complier strata, the same principles can be extended to estimation of survivor effects, as both are simply different types of principal strata. Ding and Lu (2016) outline the particular weighting strategy for identifying the SACE. Define the always-survivor principal score for unit \( i \) as

\[
    w_i = Pr(S_i(1) = 1, S_i(0) = 1 | A_i, S_i, X_i)
\]

(3.11)

Under monotonicity, \( w_i \) is equal to 1 for control units with \( S_i = 1 \), and 0 for all units with \( S_i = 0 \). Therefore, researchers need only estimate model-based weights for treated survivors. For treated survivors, the weights are a ratio of probabilities

\[
    w_i = \frac{Pr(S_i(1) = 1, S_i(0) = 1 | X_i)}{Pr(S_i(1) = 1, S_i(0) = 1 | X_i) + Pr(S_i(1) = 1, S_i(0) = 0 | X_i)}
\]

(3.12)

\[
    w_i = \frac{Pr(S_i = 1 | A_i = 0, X_i)}{Pr(S_i = 1 | A_i = 1, X_i)} / \frac{Pr(S_i = 1 | A_i = 0)}{Pr(S_i = 1 | A_i = 1)}
\]

(3.13)

In other words, the treated units are re-weighted by their principal score conditional on their observed treatment/survivorship status. Higher weights are assigned to units that tend to be in the always-survivor group while lower weights are assigned to those that are unlikely to be always-survivors.

When there are few covariates that are all discrete, non-parametric estimates of the probabilities
can be easily obtained as the conditional principal score reduces to a ratio of observed propensities of survival given $X_i$. However, when there are many covariates, researchers will need to fit some form of parametric model for the principal score. Since under monotonicity, there are three possible principal strata, one approach is to jointly model membership in all three principal strata using a multinomial logistic regression. Because a unit’s principal stratum is a latent, unobserved, variable, the likelihood is often intractable to optimize directly. Instead, estimation routines typically rely on an iterative Expectation-Maximization algorithm (Aronow and Carnegie, 2013; Ding and Lu, 2016) that switches between predicting units’ principal stratum memberships given a set of parameter estimates and updating the parameter estimates given units’ predicted memberships until the parameter estimates converge to some optimum.

In general, these likelihood-based approaches are highly sensitive to model misspecification problems. The weights resulting from the models are only guaranteed to have the desired balancing properties when the parameterization is the “true” model. Indeed, with misspecified models, weighting can often exacerbate imbalance. Researchers, unfortunately, have no way of assessing whether they have indeed chosen the true model for the principal strata, and therefore often iterate between estimating models and checking balance until a model is found that reasonably reduces the imbalance. Therefore one way of improving the robustness of principal score estimates is to again use the intuition from the covariate balancing approach of Imai and Ratkovic (2014) and incorporate known moment constraints directly into estimation of principal score weights as well. Ding and Lu (2016) show that a property of correctly-specified principal score weights $w_i$ for the always-survivors is that for any function of a covariate $h(X_i)$

$$E [w_i h(X_i)|A_i = 1, S_i = 1] = E [h(X_i)|A_i = 0, S_i = 1]$$

(3.14)

In other words, the expectation of any function of $X_i$ in the re-weighted group of treated survivors should match the expectation of that function for control-group survivors. This suggests a straightforward set of balance constraints for each covariate. If one assumes a parsimonious additive logistic
form for the weights \( w_i(X_i) \)

\[
\begin{align*}
    w_i(X_i) &= \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}
\end{align*}
\]

then there are as many parameters to estimate in the model as there are covariates. The parameters can be estimated using method-of-moments using the sample analogues of the known population balance constraints for each covariate. While there is no optimal choice for what function of \( X_i \) should be optimized, an intuitive form is to simply choose \( h(X_i) = X_i \) (as in Imai and Ratkovic (2014)), therefore estimating weights that balance the sample means for each covariate.

To summarize, when estimating treatment effects under attrition, researchers need to consider whether treatment might have an effect on the propensity of a unit to drop out of the sample. If this is indeed the case, analyses of the final outcome will be biased even when treatment is randomized or the observed covariates are sufficient to adjust for confounding. Researchers should first consider determining the likely direction of the bias induced by selection and whether inferences made on the basis of an unadjusted analysis conditional on survival will be conservative or anti-conservative. Then, researchers can use additional covariate adjustment techniques, in the form of principal score weighting, to correct for imbalance on observed covariates induced by attrition. For the purposes of estimating the effect of respondent wealth on the probability that the claimant firm wins a dispute, I first adjust for the confounding of treatment using IPT weights estimated using the Covariate Balancing Propensity Score approach (Imai and Ratkovic, 2014). In the re-weighted sample, the association between treatment and covariates is broken, allowing estimation of the treatment effect on settlement using a simple difference-in-means. To then estimate the SACE on claimant win-rate, in the re-weighted sample, I estimate principal score weights for the treated survivors using method-of-moments to re-weight the treated survivors such that the means of each covariate are balanced between treated and control. The final weights for estimating the SACE are a product of the propensity score weights and the covariate-balanced principal score weights. The discussion of the results in section 3.6 will illustrate the changes in imbalance during each step of the process, highlighting how principal score weights substantially reduce covariate divergence between treated and control arms after conditioning.
on survival.

3.5 Data

In order to test the settlement and outcome hypotheses, I draw on a new dataset of investor-state treaty-based arbitrations compiled by the United Nations Conference on Trade and Development (UNCTAD). As of August 2017, this dataset was comprised of 767 known investment arbitration disputes brought under an investment treaty – including disputes formally registered with the International Centre for the Settlement of Investment Disputes as well as other arbitration fora and known ad-hoc arbitrations. This is one of the most comprehensive databases of investment dispute claims brought against states that is publicly accessible.

Currently, 524 of these disputes have known outcomes – either an award was rendered, the dispute was discontinued, or the parties reached a settlement. Notably, many of these awards are not available to the public. To the best of its ability, UNCTAD attempts to infer information about the final outcome of the dispute using third-party sources such as reporting from investment-centric news outlets like IAReporter. Therefore, while the exact content of the award is not available, general information about the winning party or the type of settlement can often be reliably inferred. For each of these disputes, I obtain data on the year of the dispute, the claimant bringing the dispute, the respondent state, the particular treaties under which the claim was brought, the institution with which the arbitration was registered, the industry classification of the claimant firm, the rules governing the arbitration, the final outcome, and a brief description of the substance of the dispute. I code the treatment of interest as the World Bank income category of the respondent state in the year the dispute is filed, coarsened to a binary indicator of whether the state is classified as a high-income or not based on World Bank thresholds for GNI per capita. The outcome is a binary indicator of whether the claimant received an award of damages, as provided by UNCTAD.

I combine this dataset on disputes with a secondary dataset, also published by UNCTAD, that maps

---

investment arbitration agreements with respect to the scope and depth of their investment protection provisions.\textsuperscript{17} As of August 2017, this dataset encompasses a large fraction of bilateral investment treaties (BITs) in force between states, and specifically, almost all BITs that have been cited as the basis for an investment arbitration. However, the dataset has yet to code many bilateral or multilateral trade agreements that incorporate investment-related provisions and arbitration.\textsuperscript{18} Therefore, I limit the analysis to BIT disputes exclusively. While a number of disputes have been brought under these types of agreements, notably the North American Free Trade Agreement (NAFTA) and the Energy Charter Treaty (ECT), the majority of arbitration claims are still raised under BITs. I am able to match 383 disputes to a coded Bilateral Investment Treaty. For disputes citing multiple treaties, I match the dispute to the oldest treaty in force.\textsuperscript{19}

The sample of disputes covers arbitrations initiated from 1987 to 2016. Of the 383 disputes with coded investment treaty provisions, 248 reached the stage where an award was rendered, while 135 were settled or discontinued. Among the 248 disputes with awards, Figure 3.1 plots the share of disputes where the claimant won (i.e. received an award of damages) by the World Bank income classification of the respondent state. Unadjusted, middle-income governments appear to lose about twice as many cases as high-income governments.

The quality of the claimant’s case is one of the most difficult confounders to adjust for. If it were easy to infer how successful a claimant is likely to be, no case would go to arbitration. Even if ex-post, a particular claim is obviously weak or strong, scholars have little way of knowing whether this was the case ex-ante. I attempt to account for the most likely differences in claimant quality by considering variation in industry type and in the type of violation being challenged. UNCTAD provides one or more industry classifications for nearly all disputes and I rely on this coding for the variables. However, the type of issue under dispute is more difficult to code. Existing work (Pelc, 2017) notes that one major qualitative difference among claims is the difference between investors challenging direct

\textsuperscript{17}Available at http://investmentpolicyhub.unctad.org/IIA.
\textsuperscript{19}In the few cases where claimants cited an out-of-force treaty and an in-force treaty, I match to the treaty currently in force.
takings by a government and investors challenging regulatory policy measures enacted by the government. One prominent recent example of a regulatory challenge is the pair of cases filed by tobacco multinational Phillip Morris against plain-packaging regulations enacted by Australia and Uruguay.\textsuperscript{20} While the tobacco regulations did not exclusively target any particular tobacco firm, Phillip Morris alleged that the legislation itself constituted a violation of states’ investment treaty obligations. Regulatory challenges typically claim that a particular government policy had the effect of expropriating their investment even if the government did not directly seize the property. Tribunals have often interpreted states’ obligations under non-expropriation provisions very broadly – in \textit{Metalclad v. Mexico}, the tribunal stated that actions involving “covert or indicental interference” that have the effect of “depriving the owner...of the use or reasonably-to-be-expected economic benefit of property even if not necessarily to the obvious benefit of the State” could constitute violations of a government’s treaty obligations.\textsuperscript{21} Indirect expropriation claims need not show intent nor benefit to the government, opening the door to a wider array of challenges from firms. Pelc (2017) suggests that the growth in disputes involving allegations of indirect expropriation is evidence of a significant shift in the ISDS regime, arguing that “the greatest portion of legal challenges in the investment regime today seeks monetary compensation for regulatory measures implemented by democracies” (560).

Unfortunately, relying on the violations alleged by the claimant as a proxy for dispute type suffers from two major challenges. First and foremost, allegations of “indirect expropriation” do not necessarily imply that the underlying dispute concerns a government’s regulatory policy. While almost all claimants challenging regulations do allege “indirect expropriation”\textsuperscript{22}, not all “indirect expropriation” disputes concern blanket regulations. Rather, allegations of indirect expropriation are often also a fea-


\textsuperscript{21}Metalclad Corporation v. The United Mexican States. (ICSID Case No. ARB(AF)/97/1). Award. August 30, 2000. para 103.

\textsuperscript{22}However, there do exist some disputes clearly over regulation that do not claim indirect expropriation. For an example, see Mesa Power v. Canada which challenged general changes in the regulatory structure of the Ontario government’s power purchasing program on fair treatment grounds rather the expropriation. \textit{Mesa Power Group v. The Government of Canada}. (UNCITRAL, PCA Case No. 2012-17). Award. March 24, 2016.
ture of the typical contractual disputes between firms and states. Revisions or breaches of concession contracts issued by a state to a firm for some form of service provision are a prominent example of this. In one of the largest ICSID awards in history, *Occidental Petroleum v. Ecuador*, the majority of arbitrators ruled that Ecuador’s cancellation of an oil exploration contract constituted a form of indirect expropriation, citing the definition in *Metalclad*.²³ Likewise, sovereign debt arbitrations, such as the *Abaclat, Alemanni, Ambiente Ufficio*, cases filed against Argentina in the wake of the 2001 Argentine debt crisis also allege that the government’s default on its bond contracts had an indirect expropriatory effect²⁴. Overall, whether the claimant alleges indirect expropriation is a poor classifier for whether a case actually concerns regulatory challenges.

The second more practical reason for not using data on claimant’s allegations is that for some disputes, there is no available data on precisely what the alleged breaches were. Because arbitrations are conducted in private and parties rarely discuss proceedings while they are ongoing, much of the data about the process of an arbitration is gathered after the fact as documents become publicized. For many investment disputes, the claimants’ alleged violations are coded retroactively based on the summaries contained in arbitral awards. Therefore, disputes that fail to settle are more likely to have data on alleged violations than those that settle before an award is issued. This creates an obvious post-treatment selection problem. Similar problems arise when gathering data on the amount of damages sought by the claimant.

What researchers do have for nearly all disputes is some knowledge about the general substance of the dispute. While litigants rarely release details about the precise treaty claims being alleged, it is typically possible to identify the state action that triggered the dispute, even when proceedings are kept highly confidential. The UNCTAD dataset, for all but a few disputes, contains a brief summary of the actions being challenged and the nature of the claimant’s investment, drawing on accounts provided by

specialized investment arbitration news services like IARreporter when documents from the arbitration are unavailable. Using these summaries, along with secondary news sources, I manually coded each completed dispute based on the type of government action using an eight-category typology. I also code two other elements of the dispute: whether the claimant is challenging the actions of the national government or a sub-national actor, and whether the dispute concerns the actions of a domestic court.

The primary division in this coding scheme distinguishes between disputes concerning specific firm-state disputes and those targeting more general government policies. I divide firm-state disputes into five categories: expropriation/takings, breach of contract, licensing disputes, conflicts arising out of criminal proceedings or a government’s law enforcement actions, and private/firm-to-firm disputes in which the government somehow became involved. Expropriation cases are limited to those where the actions of the government are explicitly aimed at taking the property of an investor, such as through an “expropriation law” enacted by a legislature. In such cases, whether an expropriation took place is rarely in dispute. Rather, the question before the tribunal involves determining whether adequate compensation was paid to the proper entities. Notably, some cases that clearly concern expropriatory actions still allege “indirect expropriation” due to the ownership relationship between the party bringing the claim and the entity being expropriated. For example in GAMI Investments v. Mexico, the claimant challenged the taking of a Mexican sugar firm on the grounds of an indirect shareholding interest. Breaches of contract concern disputes between state and firm that do not rise to the level of direct taking, but nevertheless involve some prior agreement between the parties, formal or informal. Licensing disputes involve the government’s revocation of or failure to grant necessary permits for business. Notably, these types of disputes are distinct from contractual breaches as they involve elements of regulatory policy – denial or revocation of a license is often done on the grounds of some public regulatory interest, such as concerns over environmental damage. However, these disputes are distinct from blanket regulatory policy as they are targeted at specific firms. Other investor-state claims focus not on breach of contract, but instead rather argue that the government’s conduct in criminal prosecutions of the claimants was unfair or politically motivated. Finally, a small subset of ISDS

---

claims arise primarily out of a prior dispute between two private entities in which, the claimant alleges, some government actions unfairly benefited the other firm. Often these disputes involve governments’ failure to enforce a prior private arbitration agreement.\(^{26}\)

Among regulatory disputes, I code cases into three types of categories: general regulation, taxation, and trade policy. General regulation includes a variety of measures taken by states to regulate markets. These can include, among other areas, tobacco regulation, chemical bans,\(^ {27}\) health insurance market reform,\(^ {28}\) price setting in energy markets,\(^ {29}\) and zoning policy.\(^ {30}\) Some disputes even allege damages due to a government’s failure to regulate as in *Anderson v. Costa Rica.*\(^ {31}\) Other policy challenges have been directed at broad taxation of particular industries\(^ {32}\) and even trade barriers like import bans, quotas, or tariffs that often have analogues in WTO disputes.\(^ {33}\)

Table 3.2 summarizes the distribution of coded disputes in the sample of 383 BIT arbitrations analyzed in this chapter. A plurality of disputes are indeed firm–state disputes over contractual arrangements, suggesting that arbitration is still primarily a tool for firms to challenge targeted government actions. While regulatory policy challenges are a notable component of investment arbitration, and certainly considerably more prone to spark public outrage and attention, it is inaccurate to say that BIT litigation has become entirely dominated by disputes filed for the purposes of chilling regulation. Rather, firms engaging in business with government agencies have found BITs a viable method for

\(^{26}\)For example in *Anglia Auto Accessories Ltd. v. Czech Republic,* the claimant alleged that it was unable to obtain payment for a previous arbitral award against a business partner due to delays in court proceedings, resulting in the loss of its investment. *Anglia Auto Accessories Ltd. v. Czech Republic.* (SCC Case No. V 2014/181) Final Award. 10 March 2017.


Table 3.2: Distribution of dispute type across 383 ISDS BIT claims

<table>
<thead>
<tr>
<th>Dispute type</th>
<th>Full-sample (N=383)</th>
<th>Disputes with awards (N=248)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contract</td>
<td>185 (48%)</td>
<td>120 (48%)</td>
</tr>
<tr>
<td>Criminal</td>
<td>19 (5%)</td>
<td>13 (5%)</td>
</tr>
<tr>
<td>Expropriation</td>
<td>70 (18%)</td>
<td>43 (17%)</td>
</tr>
<tr>
<td>Licensing</td>
<td>25 (7%)</td>
<td>16 (6%)</td>
</tr>
<tr>
<td>Private</td>
<td>7 (2%)</td>
<td>3 (1%)</td>
</tr>
<tr>
<td>Regulation</td>
<td>62 (16%)</td>
<td>42 (17%)</td>
</tr>
<tr>
<td>Taxation</td>
<td>11 (3%)</td>
<td>8 (3%)</td>
</tr>
<tr>
<td>Trade</td>
<td>4 (1%)</td>
<td>3 (1%)</td>
</tr>
<tr>
<td>Sub-national actor</td>
<td>27 (7%)</td>
<td>16 (6%)</td>
</tr>
<tr>
<td>Domestic court</td>
<td>26 (7%)</td>
<td>19 (8%)</td>
</tr>
</tbody>
</table>

Escalating private disputes to the status of international law violations by leveraging states’ treaty obligations. Interestingly, there is little difference in average settlement rates across dispute type, though this may mask heterogeneity by industry.

For treaty provisions, I extract features of treaties that expand or limit the types of policies that firms can challenge. Since the UNCTAD IIA Mapping project contains many possible features, many of which vary little across disputes in the sample, I focus on elements that expand or limit the sets of policies that firms can challenge. First, I consider variation in standards of treatment accorded to investors. BITs typically provide for general standards of treatment required by the investment treaty: absolute standards like “fair and equitable treatment,” and relative standards based on either treatment comparable with nationals or with foreign investors of third countries (most-favored nation). However, treaties vary in how they specify “fair and equitable treatment.” While some qualify the definition with reference to either a minimum standard of treatment under international law or domestic law, others leave the obligation broad and unqualified. Likewise, national-treatment and most-favored nation provisions specify whether the obligation extends to all phases of an investment (“pre-and-post-establishment”), or only after the investment has been established. Some treaties include explicit clauses outlawing performance requirements, such as local content or employment requirements, as a condition of an investment. Others reference obligations to accord full protection and security to all
investments, which may be qualified by references to domestic legal provisions. In addition to fair
treatment provisions, some treaties may also include general prohibitions on unreasonable, arbitrary or
discriminatory measures. Finally, BITs may incorporate “umbrella” clauses that treat breaches of con-
tract by states as a violation of international law, permitting investors to access international arbitration
even in the case of a purely contractual dispute.

Second, BITs are also increasingly including provisions that limit investors claims in areas involv-
ing regulatory policy and other matters typically seen to be within the purview of sovereign states.
These exceptions either restrict the scope of the types of investments covered by the treaty or allow the
government to derogate from its BIT obligations under certain exceptional circumstances. UNCTAD’s
IIA mapping project identifies three types of substantive limitations: taxation, subsidies, and govern-
ment procurement policy along with four types of exemptions: security exceptions, public health and
environmental exceptions, exceptions for prudential financial regulations, and other broad public pol-
icy exceptions.

Finally, I adjust for the rules used for arbitration. Most investor-state disputes are conducted under
one of two sets of rules: the International Centre for the Settlement of Investment Disputes (ICSID)
arbitration rules and the UN Commission on International Trade Law (UNCITRAL) rules. These two
sets of rules contain very different provisions relating to transparency in arbitral proceedings, the re-
viewability of awards, and, to some extent, the requirements for a tribunal to admit a claim (Jagusch
and Sullivan, 2010).

3.6 Results

Figure 3.2 plots the estimated average treatment effect of having a high-income respondent in a dis-
pute on the probability that the dispute will fail to settle. All weighting models estimate robust stan-

34Since there are comparatively fewer disputes with high-income respondents relative to low-income, I esti-
mate the average treatment effect on the treated (ATT) instead of the overall average treatment effect. This is
defined as the treatment effect averaged over the covariate distribution for the population of treated units (Stuart,
2010). Estimating the ATT is often more feasible than the ATE when there are few treated and many control
units as there are more potential counterfactuals among control units for the treated units than the converse. Ap-
standard errors using the standard sandwich estimator, following the approach in Austin (2013). The naive difference-in-means estimate is about 11 percentage points and not statistically significant at $\alpha = .05$. After re-weighting the sample to improve covariate balance, the estimated treatment effect nearly doubles and is statistically significant at $\alpha = .05$. On average, disputes with a high-income respondent state are about 22 percentage points less likely to settle relative to a comparable dispute with a low-income respondent state.

![Figure 3.2](image)

Notes: $N = 383$. Lines denote 95% robust confidence intervals.

**Figure 3.2:** Estimated ATT of assigning a high-income respondent versus a low or middle income respondent on the probability an investment dispute fails to settle.

Whether this estimate should be trusted depends on how effective the Covariate Balancing Propensity Score weighting was in reducing imbalance. Since all of the covariates are binary indicators, it is possible to visually inspect imbalance ex-ante and ex-post on a common scale using balance plots. Figures 3.3, 3.4, and 3.5 plot the difference in proportions between high- and low/middle- income respondent disputes prior to weighting and after. Imbalance on all covariates is dramatically reduced. As a summary measure, the average absolute deviation across the covariates before weighting was about .065. After weighting, the average divergence falls to 0.0088. In other words, weighting reduced the average imbalance by 87%. It is clear that balance on observed covariates is much improved as approaches to estimating the ATT fix the covariate distribution in the treated group and re-weight the controls to optimize covariate balance. This can improve efficiency since it is unnecessary to weight the smaller treated group, avoiding potentially extreme weights.
a result of the IPTW adjustment. Therefore, in the absence of any strong confounding variables, the observed weighted difference between high- and low/middle-income respondents in settlement rates can be interpreted causally. An equivalent dispute, under equivalent legal provisions, is about 22% more likely to result in a settlement when filed against a low or middle income government relative to a high-income government. Interestingly, regression estimates from an additive linear model with all covariates included, along with the combination of weighting and regression all yield similar estimates. However, this is not always the case, as will be shown in the analysis of the outcome data.

**Figure 3.3:** Covariate balance pre- and post- weighting – Claimant and dispute type
Figure 3.4: Covariate balance pre- and post- weighting – Industry of claimant
Figure 3.5: Covariate balance pre- and post- weighting – Applicable BIT provisions
While the weights generate a balanced sample among the set of all BIT disputes filed and completed, further conditioning on disputes with awards has the effect of breaking this balance. Figures
3.6, 3.7, and 3.8 plot the change in covariate balance between the full re-weighted sample and the re-weighted sample conditioning on whether the dispute reached the stage where an award was rendered. Average imbalance across the covariates rises to 0.04, an over 350% increase in imbalance relative to the full-sample. Clearly a selection effect driven by the treatment is at work here.

After estimating the principal score weights and re-weighting the sample, the average absolute imbalance between treated and control is reduced from 0.04 to 0.018, However, this improvement is not uniform across all covariates as figures 3.9, 3.10, and 3.11 illustrate. While most covariates see improvements in balance, there are a few on which imbalance is slightly worsened, namely claimant nationality. This is likely due to dimensionality problems when attempting to minimize imbalance on a
large number of covariates with only a small number of observations in the treated group. While these reductions in balance are comparatively small relative to the overall reduction in imbalance, this issue does point to some of the limitations of principal score adjustment in a high-dimensional covariate space.

**Figure 3.9:** Principal score weights improve balance – Claimant and dispute type
Figure 3.10: Principal score weights improve balance – Industry of claimant
Figure 3.11: Principal score weights improve balance – Applicable BIT provisions
$N = 383$. Lines denote 95% robust confidence intervals.

**Figure 3.12:** Estimated SACE of assigning a high-income respondent versus a low or middle income respondent on the probability the claimant firm wins the dispute

Figure 3.12 plots the estimated effects of having a high-income respondent on the probability the claimant wins the dispute using the different sets of weights. The estimate with no covariate adjustment suggests a strong negative effect of about 23 percentage points. Incorporating both the inverse-propensity score weights and adding the covariates directly to the regression model does little to change the estimated treatment effect. Since, the covariate balance in the IPT-weights-only sample is still better than in the raw sample, this estimate is arguably getting closer to the truth. However, it does not account for the selection bias induced by post-treatment settlement except insofar as one believes the outcome model to be correctly specified. Since it is rare that researchers get the regression function perfectly correct, there are strong reasons to suspect that this estimate is still biased. Incorporating the covariate-balanced principal score weights in addition to the propensity score weights further reduces the point estimate to about 10 percentage points. After adjustment, the effect of respondent income level on win probability is statistically indistinguishable from zero. While the small sample of existing disputes is insufficient for a very precise null estimate, there is good reason to believe that even the corrected estimate of the effect is exaggerated given theoretical expectations over what types of cases are selected out via settlement. If there is a lurking confounder of principal stratum membership and outcome, it is likely some component of the claimant’s case quality that is not fully captured by the dispute type and industry measures. I expect that this will induce artificially low estimates of the claimant’s win rate against high-income governments because the cases that fail to
settle only against high-income respondents should be weaker than those that would not settle against both high-income and low-income respondents. Overall, the results persuasively show evidence for a strong negative effect of respondent wealth on probability of settlement. The evidence for any advantage of capacity at the award stage, however, is weak to non-existent. It is unlikely that all sources of collider bias have been adjusted for in the analysis of dispute outcomes. Therefore, even the negligible, statistically insignificant difference between win-rates of high- and middle/low- income respondent governments can likely be attributed to attrition, rather than systematic bias by the tribunal.

3.7 Conclusion

For many policymakers and legal scholars, investor-state dispute settlement is at a crossroads. While ISDS provisions were ubiquitous in bilateral investment treaties signed during the last several decades, they have become a major point of contention in negotiations over new trade and investment agreements. Many critiques of ISDS allege that arbitration tribunals are systematically biased in favor of wealthy investors from capital-exporting countries and against the interests of developing countries and emerging markets. On face, the claim appears reasonable given stark disparities in win-rates between high-income and low-income countries.

This chapter critically evaluates claims of bias against developing countries in investment arbitration proceedings. It emphasizes that analyses of win-rates are fundamentally meaningless for assessing claims of bias without additional assumptions about the nature of the selection mechanism that leads some disputes to reach a settlement before a tribunal issues a decision. This is because considering only those cases that failed to settle results in a form of selection bias that can artificially induce a negative correlation between the win-rate and respondents’ characteristics when those characteristics also affect the propensity of settlement.

Using well-known theoretical models of pre-trial bargaining, this chapter shows that low rates of settlement should be expected among well-resourced countries relative to countries for which the cost of litigation is high. When respondents are uncertain over the claimant’s chances of successfully win-
ning a dispute, they will tend to issue smaller settlement offers to claimants with high quality cases – an example of adverse selection. While low-quality claimants will accept offers, claimants with better quality cases will tend to press disputes to a final award by a tribunal. A respondent government’s incentive to give as small of a settlement offer as possible is offset by the costs of litigation if the claimant chooses to reject the offer. Because developing country governments face higher litigation costs, they will tend to prefer reaching a pre-trial settlement against a larger set of claimants. As a result, there may exist a set of claimants that would reach a settlement with a litigation-averse low-income government that would not receive acceptable settlement offers from high-income governments. Under this selection process, high-income governments are able to force some of the claimants with weaker cases to proceed to litigation.

This appears to be the case in the data as well. Using a comprehensive dataset of 383 BIT investor-state disputes, the article shows that after accounting for many of the differences between cases with wealthy respondents and cases with lower-income respondents, respondent wealth has a causal effect on the propensity of a dispute to reach a settlement. Low and middle-income respondent states, all-else-equal, are about 22 percentage points more likely to settle a dispute prior to an award. The results show that resource disparities do influence the way in which arbitrations are conducted.

However, after accounting for some of the sources of the selection bias resulting from analyzing only those disputes that fail to settle, the empirical analysis finds no statistically discernable effect of respondent-level income on the probability of winning the dispute. There is no strong evidence that, on average, arbitral tribunals systematically favor high-income governments in their decisions after controlling for claimant type, dispute type and for the legal provisions under which the dispute was brought.

These results should help clarify debates over the legitimacy of international investment arbitration. While supporters of ISDS should feel somewhat vindicated by the absence of strong evidence for systematic bias in favor of rich countries at the tribunal-level, the results also highlight how significant inequities in a legal system can arise even under impartial adjudication simply because litigation is costly. While early settlement is a necessary feature of any legal system, when settlements can be com-
peled not because of an actual agreement between the parties, but rather due to massive disparities in legal capacity, the underlying fairness of the system is called into question. Advocates for the arbitration regime should pay much closer attention to the question of legal costs and consider reforms to arbitral institutions or that would address disparities in legal capacity and in the overall costs of litigating for developing country governments.
Chapter 4

Affiliation Bias in Arbitration: An Experimental Approach


4.1 Introduction

Arbitration has crept into nearly every corner of Americans’ lives. Scholars refer to this expansion as a “whole-scale privatization of the justice system” (Malin and Ladenson, 1992; Gilles, 2014). This trend goes beyond domestic legal disputes as arbitration has become a central element of economic interdependence. In particular, the growth of arbitration has put arbitration panels in a position to rule on transnational business transactions and key political questions like the ability of governments to tax their citizens or regulate health—long areas of sovereign prerogative (Born, 2012).

What makes this expansion potentially troubling is that systematic features of arbitration make it
difficult for arbitrators to be entirely unbiased in their decisions (Garth, 2001). Unfettered by precedent, deprived of strict uniform rules against conflicts of interest, and insulated from any judicial system, some argue that powerful corporations use arbitration to steer cases to friendly arbitrators incentivized by the prospects of sizeable earnings.

This debate is not completely new in the legal academy (Resnik, 2004; Fiss, 1984). What is new, however, is that recent critiques of arbitration come from insiders who otherwise defend this dispute settlement system as an efficient form of legal adjudication (Paulsson, 2010; Van Den Berg, 2010). These criticisms have focused on a particularly well-established feature of arbitration – the ‘right’ of disputants to appoint an arbitrator to a tribunal. While such unilateral party-appointments are well established in the field of arbitration, some now criticize the practice as a source of bias that negatively affects the impartiality and the legitimacy of arbitration proceedings.

Proposals for reform range from eliminating party-appointments altogether to introducing changes in the form of arbitrators’ selection, but most seem farfetched or unlikely as a practical matter (Giorgetti, 2013). In short, arbitration is designed to give parties greater control over the dispute resolution process relative to a more formal judicial setting, and parties would be hesitant to give up one of the primary tools for exercising that control—the ability to (partially) choose who hears the dispute. However, a recent proposal that may be more plausible would permit parties to appoint arbitrators, but would prevent nominees from knowing which party appointed them (Paulsson, 2013). This “blind appointment” approach attempts to relieve the arbitrator of the possible affiliation or allegiance effects resulting from the nomination process, but maintains unilaterally appointed arbitrators, considered a fundamental ‘right’ in arbitration (Brower and Rosenberg, 2013). The proposal is in part inspired by ideas ingrained in most legal systems (e.g., “justice is blind”), in some theories of justice (e.g., “veil of ignorance,”) and in the highest standards in scientific research (e.g., blind peer-review) (Robertson, 2010b), with the general principle being that impartiality can be enhanced by preventing a decision-

maker from accessing knowledge that might unduly affect the decision.

In this chapter, we use the blind appointments proposal as an entry point into the debate over implicit biases of legal actors, with three objectives in mind. First, we argue that an important issue in the debate over bias is to distinguish between two different mechanisms through which the nomination process affects arbitrators’ behavior: selection effects and affiliation effects. Prior observational research has shown that arbitrators tend to favor their appointing parties. However, because parties choose arbitrators in part based on how they expect the arbitrator will vote, appointment-driven bias cannot be inferred from these correlations alone. In order to infer evidence of an affiliation effect (the implicit bias of the arbitrator to favor the appointing party) we need to eliminate the selection effect (the strategic decision of the litigant to choose arbitrators likely to be favorable). Therefore, we develop a set of randomized survey experiments designed to isolate affiliation effects in which we directly controlled arbitrators’ assigned appointers.

Second, we present the results of these experiments. In short, we presented surveyed arbitrators with a hypothetical choice task related to an investment arbitration case. Participants were randomly told whether they were appointed by one of the parties, by joint agreement of the parties, or simply that they were appointed (without any information about the identity of the appointer). Based on a sample of 257 responses from arbitrators around the world, we found that arbitrators nominated by one of the two parties to the litigation tended to make decisions more favorable to that party compared to arbitrators appointed by the opposite party. We replicated this result in a follow-up experiment. Additionally, we found that arbitrators treated with the ‘blind appointment’ option exhibited similar responses to joint party appointees and tended to take positions in the middle of the two party-appointees.

Third, based on the results of the experiments, we argue that blinding could be a useful approach to reducing bias. The evidence we find strongly suggests that affiliation effects exist and while there are certainly difficulties in implementing such a proposal, we suggest that blinding could help ameliorate affiliation biases while still retaining the potential benefits of party-involvement in the appointment process.
The chapter proceeds as follows: the next section provides some background to arbitration and the debate over party appointments. Section 4.3 describes the design of the initial experiment and its subsequent replication and reports the findings of both. Section 4.4 discusses some of the limitations of our study and the challenges to adopting the blinding proposal. Section 4.5 concludes.

4.2 Background

4.2.1 Unilateral Party-Appointments: A Burgeoning Debate

A debate has emerged among members of the arbitration bar over unilateral party-appointments. Critics of this feature argue that the power and legitimacy of arbitration stems from an appearance of expertise, neutrality, and impartiality. Hence, the predisposition towards one party or the other that results from the system of appointments poses significant challenges to the main legitimating aspects of arbitration (Paulsson, 2010; Van Den Berg, 2010). It may also result in unnecessary antagonism, complicated compromises, and inconsistent decisions in the awards of tribunals that all undermine the goal of an independent, rule-based adjudicatory system.

Defenders of the use of party-appointed arbitrators, on the other hand, argue that the traditional party-appointment system is the “keystone” of arbitration because it gives the parties some “ownership” over the process (Perry, 2013). In addition, when tribunals render unanimous decisions, the presence of party-appointees may enhance the credibility of the award in the eyes of both parties since each litigant knows that a trusted appointee was willing to endorse the outcome. Being able to appoint an arbitrator, some argue, is one of the most attractive aspects of arbitration “as an alternative to domestic litigation” and its elimination would constitute a radical transformation, and potential devaluation, of arbitration (Brower and Rosenberg, 2013).

The debate over the convenience of party-appointments has been conducted almost exclusively with anecdotal evidence. However, a recent study using observational data has demonstrated some of the merits of each side of this debate. By comparing cases with and without party-appointments, Puig
(2016) shows that tribunals without party-appointed arbitrators tend to handle cases faster and settle more often. Their decisions tend to be unanimous and are less likely to be challenged by a subsequent annulment proceeding. Nonetheless, the rate of arbitrator challenges and resignations in tribunals without party-appointed arbitrators appears to be higher, and the higher rate of resignations in tribunals without party-appointed arbitrators tends to lower the expediency of the process.

What the debate has also shown is a lack of consensus among arbitration professionals for moving away from the system of party-appointments, limiting which reforms would be feasible.\textsuperscript{2} Blind appointments, however, may resolve some of the concerns that emanate from the party-appointment system, without eradicating the practice. Under a “blind appointment” system, parties to an arbitration can continue to appoint arbitrators, but a mechanism is introduced to ensure that nominees do not know who appointed them. This proposal maintains the fundamental arbitration feature of party control, but may help to mitigate the implicit bias and adversarial influence of explicitly known party appointments.

\subsection{4.2.2 Party Affiliation: Arbitrators and Implicit Biases}

Existing empirical evidence points to a number of different ways through which litigants can introduce biases in a legal process (Robertson, 2010\textsuperscript{b}). In the arbitration context, the nomination and appointment of arbitrators by the parties is a calculated decision by the litigants. When rules permit litigants to nominate a legal actor, they can ensure that such a person is not too independent minded by selecting someone who has shown reliability and the appropriate decision-making philosophy towards the relevant set of issues. Any influence in the process that result from such practice could be attributed mainly to a selection effect.

Yet, even in the absence of a selection effect, appointees may also find it difficult to maintain impartiality because of implicit preferences for their appointing party. We refer to this as an affiliation effect.\textsuperscript{2}

\textsuperscript{2}A survey of professionals shows “that there is general disapproval of the recent proposals calling for an end to unilateral party appointment.” See http://annualreview2012.whitecase.com/International_Arbitration_Survey_2012.pdf.
Modern psychology research has extensively documented the existence of such “implicit biases” in a variety of settings: from attitudes towards historically disadvantaged groups, to the effect of primes and context cues on expressed beliefs and actions (Greenwald and Banaji, 1995; Nosek and Riskind, 2012). These implicit effects on attitudes often operate outside of conscious cognitive modes—those expressing implicit biases may be entirely unaware of and unable to account for their influence. In this particular instance, an arbitrator may, despite her best intentions to remain unbiased, be nonetheless primed to favor their nominating party simply by knowing that they were selected by that party.

Blinding, or removing key information that may affect a legal actor’s decision, is commonly suggested as an effective debiasing strategy against affiliation effects in similar settings (Robertson, 2010a). In part, blinding is a preferred intervention because the “blind spot” tends to persist even after individuals think through and consider their biases. In fact, experimental research has shown a tendency for people to acknowledge bias more readily in others than in themselves, often increasing polarization in adversarial contexts. Hence, biases typically operate non-consciously, thereby leaving their influence hidden from direct introspection (Wilson and Brekke, 1994; Wilson, Centerbar and Brekke, 2002).

### 4.3 Evidence

#### 4.3.1 Context

Before turning to the design and results of our experimental approach, we describe some basics on the appointment of arbitrators. With some exceptions, arbitration tribunals are typically composed of three members—two party-appointed arbitrators, and a third arbitrator, usually the chair, appointed by one of the following three methods: a) by agreement of the parties in the proceedings, b) by agreement of the two party-appointed arbitrators; or c) by an independent designating authority (commonly, the institution administrating the proceedings). All arbitrators are supposed to be independent and impartial, including the two party-appointed arbitrators, and will often sign a declaration affirming their independence and impartiality.
As we discuss above, existing observational research on arbitral proceedings, such as analysis of dissent rates, is insufficient to differentiate between selection and affiliation effects. Moreover, little attention has been given to implicit biases in arbitral decision-making (Drahozal, 2004; Keer and Naimark, 2001). To our knowledge, no prior experiment of this nature exists using arbitrators as experimental subjects in an online survey context.

4.3.2 Design

Our pair of survey experiments were both based around a brief vignette describing a hypothetical investor-state arbitration. Elements of the vignette were randomized for each participant in order to study how arbitrators’ responses might vary across different scenarios. Our primary manipulation of interest is the arbitrator’s appointing party. Participants could be told that they were appointed by the Respondent, by the Claimant, by the parties (i.e., by agreement of the litigating parties), or simply that they were appointed to the tribunal (followed by a period), with no mention of any appointing method. This last condition is what we refer to as a “blind” appointment case. Figure 4.1 provides a sample of the vignette that we presented arbitrators in the first experiment, with the key manipulation of interest shown bold.

In the first experiment, the vignette included a few additional manipulations in order to examine other potential influences on arbitrator behavior. We randomly varied the implied resource endowments of the Claimant and the Respondent. The Claimant in the arbitration could be a firm headquartered in either a high- or middle-income economy while the Respondent state could be either a middle- or low-income economy. We also varied the type of ruling across four possible conditions reflecting different ways in which a case litigated under the applicable procedural rules in the vignette could be decided: a) the Respondent expropriated the Claimant’s property (Claimant wins), b) the Respon-

3Note that the surveyed arbitrator is not explicitly told that their appointing party was hidden from them. The exact text of the treatment simply reads: “You were appointed to the tribunal.” While arbitrators who are actually blinded in a dispute would likely know that they were blinded, we chose not explicitly mention blinding in this treatment as we did not want to present arbitrators with a potentially unfamiliar practice that is currently not part of the ICSID proceedings – the applicable rules in our vignette.
Imagine an investor-state dispute being conducted under the 2006 Arbitration Rules of the International Centre for Settlement of Investment Disputes (ICSID). The Claimant is a firm headquartered in a high-income economy. The Respondent is a country classified by the World Bank as a middle-income economy.

The Claimant alleged that the Respondent violated the provisions of a bilateral investment treaty to which the Respondent is a party. Among other arguments, the Claimant argued that the investor and its investments had been treated unfairly and that ultimately the Respondent expropriated the Claimant’s investment located within the Respondent’s territory. The underlying dispute concerns an infrastructure project undertaken by the Claimant under a concession contract with a governmental agency. The Respondent argued in response that the Claimant had violated provisions of the contract and that the investors received all compensation to which they were entitled.

You were appointed to the Tribunal [by the Respondent.]/[by the Claimant.]/[by the Parties.]/[.]. After careful consideration of the facts of the case, the tribunal unanimously decided that the Respondent unfairly treated and wrongfully expropriated the Claimant’s investment and that the Claimant is entitled to compensation.

In their submissions on costs, both parties have requested that the other party bear the costs of the proceedings in full, including legal fees and expenses. The counsels for both parties behaved professionally and ethically during the proceedings.

Figure 4.1: Sample experimental vignette showing key manipulation – Experiment 1.
dent did not expropriate the Claimant’s property (Respondent wins), c) the dispute is outside of the tribunal’s jurisdiction (Respondent wins on ‘technical’ grounds), or d) the Claimant’s claims are manifestly without legal merit (case dismissed summarily; i.e., Respondent decisively wins). Each of the conditions had equal probability of being assigned to any individual and the treatments were each randomized independent of one another. Table 4.1 summarizes the marginal distributions of the number of observations assigned to each treatment condition in the sample.

After being presented with the vignette, survey arbitrators were then asked how they thought the parties’ expenses in the dispute, including the cost of legal representation, should be apportioned in such a case. Arbitrators could choose to have one party reimburse the other for either all or some of their expenses or have each party bear their own expenses. Participants were then asked to briefly discuss the reasoning behind their decision in an open-ended question.

We chose in the first vignette to ask the surveyed arbitrators to rule on the allocation of costs rather than on the actual merits of a case for two important reasons. First, summarizing the arguments of the parties such that participants would have enough information to render an educated decision on the merits of a full case would require an impossibly long vignette. We were conscious of the fact that participants (busy lawyers and arbitration professionals) would likely not have the time or interest to spend hours on our survey. Prompting arbitrators to render a clear decision on costs given that the result is known allowed us to use a vignette that participants could easily read and analyze in a practical amount of time.

Second, we needed an outcome that would generate variation in responses. Had we chosen an outcome on which there is obvious and clear legal guidance, we would expect arbitrators to all reach more or less the same conclusion. In the context of the procedural rules applicable in our vignette, the ICSID 2006 Arbitration Rules, costs are an ideal outcome as there is little precedential guidance

---

4We present the marginal distributions of counts (rather than the full factorial joint distribution) as we are primarily interested in the marginal effects of each variable averaged over the distribution of the other treatments. While we do not have enough power to credibly estimate the effect of a particular unique combination of all four treatments, this is not the quantity we are interested in. We do have sufficient observations to estimate marginal effects for a single treatment. In this sense, our experiment is very similar to the “conjoint” multi-attribute choice experiment that is increasingly common in social science surveys (Hainmueller, Hopkins and Yamamoto, 2014).
Table 4.1: Summary of number observations assigned to each condition – Experiment 1 (N = 257)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appointer</td>
<td>Appointed by the Responder</td>
</tr>
<tr>
<td></td>
<td>77</td>
</tr>
<tr>
<td>Claimant wealth</td>
<td>135</td>
</tr>
<tr>
<td>Respondent wealth</td>
<td>121</td>
</tr>
<tr>
<td>Dispute outcome</td>
<td>Claimant wins on merits</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
</tbody>
</table>

Notes: Chi-squared tests for the marginal counts across all four conditions fail to reject the null that the counts are generated by discrete uniform distributions (p > .10).
for how they should be allocated. The arbitration rules grant significant discretion to tribunals in how they decide the parties should pay the costs. Some tribunals follow the principle that “costs follow the event,” and the losing party should compensate the winner’s legal fees. Others choose to have each party bear their own costs. While the former appears to be the most common approach, it is rare that costs are fully borne by one party or another. There is a meaningful amount of open-endedness in the decision such that it could plausibly be swayed by extra-legal considerations.

We conducted the survey in the Fall of 2015 and recruited participants by collecting publicly available e-mail addresses of arbitrators and lawyers specializing in arbitration (throughout the paper we refer to these participants as arbitrators). These addresses were largely obtained from lists published on the websites of arbitration institutions and the directories of other professional organizations. We chose to recruit from this specific set of potential arbitrators, rather than from a typical sample of individuals from the general population, in order to make our results as generalizable as possible to the population of potential arbitrators around the world. Moreover, competently completing the vignette itself requires some familiarity with international arbitration to understand in the first place, making typical convenience samples useless. While not all participants have served as arbitrators, based on survey responses given after we administered the vignette all respondents indicated some expertise in arbitration, either in investor-state or in commercial arbitration. To increase our response rate, we sent several reminders to participants, but promised only an advanced circulation of any publications summarizing the research as the reward. We obtained approval from the institutional review boards of both the University of Arizona and Harvard University.

---


6We are aware, however, that costs may appear to be an ancillary element of an arbitral award compared to the actual merits of a dispute. Therefore, in our replication experiment, we created a vignette to test for effects on how arbitrators would reason through how much should be awarded in damages. Please see below for further discussion on the topic.
4.3.3 Analysis and Results

A total of 538 individuals responded to the survey e-mail, but not all arbitrators completed the survey or answered all of the questions.\(^7\) We therefore only received 257 complete responses to the investor-state vignette. While this may appear to be a very low response rate, this does not invalidate our experimental design. Random assignment of treatment allows us to obtain an unbiased estimate of the treatment effect for arbitrators in the sample regardless of how individuals were selected into the sample.\(^8\) Moreover, if the factors driving non-response are not associated with individuals’ effect sizes we can treat our effect estimates in sample as representative of the average effect in the population. Nevertheless, we will thoroughly discuss the generalizability of the results given our sample composition.

In allocating costs, arbitrators are first and foremost driven by the outcome of the case. In practice, winning parties do not bear costs of losing parties and the vast majority of arbitrators followed this convention. In effect, there are only three reasonable outcome choices of interest that were available to arbitrators: losing party pays all, losing party pays some, and each party pays their own costs. However, likely due to measurement error on the part of the arbitrator, thirteen of our arbitrators indicated that they would have the winner reimburse the loser. While one could drop these individuals, doing so would risk inducing bias (Rosenbaum, 1984; Montgomery, Nyhan and Torres, 2018) as we would be controlling for a post-treatment variable, the survey response. To avoid post-treatment bias, we leave these arbitrators in the analysis (and treat them as choosing neither of the three outcome options), preferring to err on the side of slightly more measurement error in the outcome (and therefore variance) than systematic bias.

In our analysis, we collapse the outcome variable into these three relevant response categories. Folding over our outcome this way has the important benefit of improving estimation efficiency by reducing the total number of outcome categories considered and allowing us to pool results from vi-

---

\(^7\)In this experiment, we sent out 28,832 recruitment e-mails. However, some of these e-mails returned with a delivery failure message as the addresses were no longer active. Because arbitrators could choose to stop the survey at any time, we could not force responses for all questions in the survey.

\(^8\)Moreover, as we discuss in Table 4.1, the distributions of treatment assignment are not statistically distinguishable from uniform, suggesting that randomization held for the observed sample.
gnettes where the Claimant won and ones where the Respondent won. We therefore focus on estimating the average effect of being appointed by the winning versus the losing party, averaging over the separate Claimant and Respondent effects.

The affiliation bias hypothesis suggests that arbitrators who were told that they were appointed by the winning party will be more punitive towards the loser in terms of cost allocation than those who were told they were appointed by the losing party. To analyze the effect of being appointed by the dispute winner versus the loser, we simply take the difference in means for each of the three response indicators of interest between arbitrators assigned to each treatment condition. Because treatment is randomized, these differences-in-means are unbiased estimates for the causal effect of the appointer treatment on the probability that an arbitrator would select a particular response category.

Figure 4.2 plots the estimated treatment effect of winner versus loser appointment on the three outcome categories. On average, arbitrators were about eighteen percentage points more likely to award all costs to the winning party when they were appointed by the winner rather than the loser. However, assignment to the winning party did not appear to change arbitrators’ propensity to cost-shift in the first place. When an arbitrator chooses to follow the unwritten rule of “costs follow the event,” the winner’s appointee is more likely to punish the losing party by having them reimburse all of the winner’s costs while the loser’s appointee is more likely to protect their appointing side by having them pay only some of the winner’s costs. This is consistent with the affiliation bias story. While arbitrators do not completely advance their appointing party’s interests, when room for discretion arises, they


d| 9| While it would be possible to estimate separate effects for Claimant compared to Respondent appointees, we have insufficient power to meaningfully detect much of a difference between the two. To obtain greater precision, we find it more appropriate to estimate a single “pooled” winner vs. loser effect given the constraints of the sample size.

10| With a trichotomous outcome variable, another approach commonly used by researchers is to assume a model for an underlying latent variable that is then observed in a coarsened form. These ordered dependent variable models (such as ordinal logit or probit) would require us to make additional functional form assumptions relating the covariates to the outcome. Moreover, the proportional odds assumption for ordered logit may not be satisfied in this case as the effect of treatment is not constant across units. Our results suggest that arbitrators that would choose to split costs are unaffected by treatment, but those who would choose reimbursement are moved by the treatment to select a more punitive option. Because we have an experiment, any outcome modeling approaches (such as a logit model) are unnecessary at best and misleading at worst. See Volfovsky, Airoldi and Rubin (2015) for more on the challenges of estimating causal effects with multichotomous outcome variables.
appear to be more likely to choose outcomes that are more favorable to the side that appointed them.

![Diagram](image)

**Notes:** Thin lines denote 95% bootstrapped confidence intervals (5000 iterations). Thick lines denote 90% bootstrapped confidence intervals. Number of observations: 72 appointed by winning party, 55 appointed by losing party.

**Figure 4.2:** Estimated average effects of winning party appointment – Experiment 1

It is worth noting that these results are likely a very conservative test of affiliation effects. Arbitrators taking the survey are not actually participating in a months-long proceeding that may reinforce party allegiances, nor do they face any potential costs or benefits to how they rule in the vignette. The bias we are able to detect is purely implicit and inherent in the ‘role’ assignment. Moreover, if arbitrators are sensitive to the stigma of being perceived as not impartial, then they would likely try to attenuate any overt expressions of bias in their responses. If there is any form of social desirability bias influencing surveyed arbitrators, then it would likely bias the experiment against finding a difference between different treatment conditions. Nevertheless, we find strong evidence for a party affiliation effect.

Additionally, we wanted to consider whether blind appointees behave differently from those appointed jointly by the parties. Figure 4.3 plots the estimated differences-in-means for each cost allocation option – between the arbitrators who were told they were appointed by the parties and those who were not given any appointer information: the blinded condition. While in our sample, party appointees were slightly more likely to not cost shift, this difference is not statistically significant at any conventional rejection level. Overall the magnitude of any difference between these two groups
Figure 4.3: Estimated average effects of party appointment versus blind appointment – Experiment 1

appears to be quite small, suggesting that blinded arbitrators are likely to behave much like the joint party appointees when given discretion over some allocation between the parties. This suggests that blinding could have the benefits of having the parties agreeing to an arbitrator without facing the costs associated with such agreement.

One concern with the original experiment that was raised by a reviewer is that it may be difficult to generalize the observed behavior in response to a vignette on cost to real world arbitral decisions on actual questions that implicate the substance of the case. Because fees are a comparatively less significant component of the decision relative to the merits and damages and arbitrators are much less constrained in how they choose to award costs, it may be that observed biases dissipate when arbitrators consider questions of greater importance to the case. While our argument for choosing to look at costs is precisely that the absence of constraints (i.e., discretion) is most likely to generate affiliation bias and it is important to note that legal costs are not an insignificant component of the parties’ overall expenses in a dispute, we did consider it valuable to determine whether our result is robust to different choices of legal questions.

Therefore, we conducted a replication study in March of 2017 designed to evaluate how arbitrators would reason through a decision on damages, that is, how much should the Claimant be awarded given
that the tribunal has found the Respondent at fault. As in the first experiment, we chose to ask arbitrators to decide on a discrete question on damages to keep the vignette within reasonable length and the task as clearly defined as possible. Asking arbitrators to evaluate the entirety of an actual case would still be unreasonable in the time and space allotted for the experiment. However, a decision on damages provides a tougher test of the affiliation bias hypothesis. Arbitrators certainly face greater legal constraints and actual precedential guidance when evaluating how much a successful Claimant should be awarded in compensation from the Respondent. Moreover, the amount of damages received by a Claimant is certainly a central component of any arbitral dispute and highly salient to the litigating parties. Therefore, this vignette serves as an initial test of how well our affiliation bias finding carries over to the more substantive decisions that comprise an arbitral award.

Figure 4.4 outlines the vignette we used in this second experiment. In this vignette, a similar number of participants were presented with a similar setup as in the first experiment. Arbitrators were told that the tribunal to which they were appointed had decided unanimously that the Claimant investor was entitled to damages due to the Respondent state’s violation of the ‘fair-and-equitable-treatment’ standard of a hypothetical treaty. They were then tasked with deciding how much the Claimant should be owed in damages in terms of two possibilities: one proposed by the Claimant and another proposed by the Respondent in the case. The key legal question arbitrators had to evaluate was whether the Claimant should be awarded compensation for lost future profits (the Claimant’s argument) or whether the award should be based exclusively on the liquidation value of the firm (the Respondent’s argument). According to the vignette, both parties cited relevant ICSID precedents for their positions. For a frame of reference, we included a link to the precedents and an article describing the typical award amounts in ICSID disputes. As with the first vignette, our goal was to provide enough information that arbitrators would feel comfortable providing some answer but not so much information that there would be zero variation in arbitrators’ responses. We also re-tested our original question on costs.

Because our interest was in replication, in this vignette we only manipulated two elements: the appointing party and the amount of the submission on damages. As before, all treatments were randomly assigned independently of the surveyed arbitrators’ characteristics. However, to increase statistical
Imagine an investor-state dispute being conducted under the 2006 Arbitration Rules of the International Centre for Settlement of Investment Disputes (ICSID). The Claimant investor alleged that the Respondent state violated the provisions of a bilateral investment treaty to which the Respondent is a party. Among other arguments, the Claimant argued that the Respondent violated the treaty’s fair and equitable treatment provisions and mistreated the Claimant’s investment. You were appointed to the Tribunal [by the Respondent.]/[by the Claimant.]/[by the Parties.]/[].

After careful consideration of the facts of the case, the tribunal (you and your fellow arbitrators) unanimously decided that the Respondent unfairly treated the Claimant’s investment in violation of the treaty and that the Claimant is entitled to compensation.

You are now asked to decide on the amount of damages owed to the Claimant by the Respondent. The parties have agreed that the tribunal’s task is simply to pick one of the two positions of the parties’ experts and decide how the expenses should be apportioned in this dispute. In its relevant part, the bilateral investment treaty provides as follows:

1. A Tribunal may award monetary damages and any applicable interest, only.
2. A Tribunal may also award costs in accordance with the applicable arbitration rules.
3. A Tribunal may not order a Party to pay punitive damages.

[CONTINUES ON NEXT PAGE]

Figure 4.4: Sample vignette in the second experiment

power for detecting the affiliation effect, we chose to assign two-thirds of arbitrators to one of the party-appointed conditions rather than one-half as in the first experiment. We included a manipulation on damages to evaluate whether differences in the magnitude separating each of the parties’ proposals affected the choice of damages or moderated the affiliation bias effect. We wanted to ensure that our findings would not be driven exclusively by the particular values we chose for the Claimant’s and Respondent’s submissions. However, we found no evidence for effect modification or a statistically significant effect of the size of the proposed damages on the probability the arbitrator would choose either the Claimant’s or Respondent’s position. Table 4.2 summarizes the number of observations assigned to each treatment condition.

After reading the vignette, arbitrators were asked to choose the amount that the Respondent should
The Claimant has argued that they should be compensated for lost future profits that would have been realized had the measure not taken place plus interest. The Claimant justifies this claim on the grounds that the enterprise operated profitably for a period of almost three years prior to the violation. The Claimant cites Metalclad v. Mexico ICSID Case No. ARB(AF)/97/1 which considered a minimum presence of at least two or three years necessary for an award of future profits. The Claimant’s expert has calculated damages for US$50,002,100.00 based on the discounted cash flow value of the expected returns from the Claimant firm’s ten-year investment plan. The Respondent has argued that the enterprise had not operated for a sufficient period of time to establish itself as a “going concern” and that the ability of the enterprise to generate future earnings was uncertain and compromised. The Respondent cites Tecmed v. Mexico ICSID Case No. ARB(AF)/00/2, arguing that the tribunal in that dispute ruled that the Claimant’s operating history of two and a half years was insufficient to establish enough objective data on profitability to apply a discounted cash flow analysis. Therefore, any estimate of future profits would be highly speculative. The Respondent instead proposes that damages should be based on the liquidation value of the firm and their expert has calculated damages for [US$25,001,050.00, US$12,500,525.00, US$6,250,262.50].

In ICSID disputes, the average award for Claimants who are awarded damages is about US$45.6 million. The median award is US$10.9 million (See: Franck, S. D., & Wylie, L. E. (2015). Predicting outcomes in investment treaty arbitration. Duke Law Journal, 65(3), 494-527.). Throughout the proceedings, both disputing parties were cooperative and the counsels for both parties behaved efficiently, professionally and ethically. The parties have not agreed on how and by whom the expenses shall be paid.

Figure 4.4: Sample vignette in the second experiment (Continued)

pay to the Claimant in damages because of the treaty violation. We constrained our participants to choosing exclusively between the Claimant’s proposal and the Respondent’s proposal. In practice, investment arbitrators have freedom to choose the actual quantum awarded and rarely give the exact value proposed by either party. However, to increase statistical power in our experiment, reduce variance in the outcome, and increase the response rate by simplifying the task, we decided to limit arbitrators to two choices. This also forced arbitrators to consider the arguments of the parties instead of simply taking the shortcut of splitting the difference between the two proposals. Since we were
Table 4.2: Summary of number observations assigned to each condition – Experiment 2 (N=248)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Appointed by the Respondent</th>
<th>Appointed by the Claimant</th>
<th>Appointed by the Parties</th>
<th>Blind Appointment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appointer</td>
<td>88 (90)</td>
<td>79 (79)</td>
<td>45 (45)</td>
<td>36 (38)</td>
</tr>
<tr>
<td>Respondent’s proposed damages</td>
<td>US$25 Mil.</td>
<td>US$12.5 Mil.</td>
<td>US$6.25 Mil.</td>
<td></td>
</tr>
<tr>
<td>Respondent’s proposed damages</td>
<td>85 (86)</td>
<td>70 (73)</td>
<td>93 (93)</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Counts denote the number of observations among respondents who answered the question on damages. Counts in parentheses denote the number of observations among respondents who answered the question on costs. Chi-squared tests for the marginal counts across all four conditions fail to reject the null that the counts are generated by the distributions we specified for randomization (p > .10).

interested in seeing whether our original results on costs replicated, we also asked our arbitrators to decide on how the parties’ costs should be apportioned using the same outcome choices as in the first vignette.

We recruited our arbitrators from the same pool of e-mail addresses as in the first, removing any arbitrators who explicitly unsubscribed from the list or notified us that they were not interested in participating in any surveys. A total of 644 arbitrators consented to participate in the survey. However, as before, not all arbitrators finished the survey. Only 248 of the arbitrators who began completed the survey and responded to the question on damages.

Figure 4.5 plots the estimated effect of being appointed by the Claimant (the dispute winner) versus the Respondent (the dispute loser) on the probability of choosing the Claimant’s (high) proposed damages. On average, arbitrators appointed by the Claimant were about 15 percentage points more likely to choose the Claimant’s damages proposal compared to arbitrators appointed by the Respondent (p =

---

11 A total of 25,965 recruitment e-mails were sent out in the second wave. As in the first experiment, many addresses were inactive or unavailable. Because we did not store identifying information from the first wave, we could not exclude arbitrators from the first experiment. However, we do not think that there are likely any interference effects between experiments given the year and a half-long gap between them.

12 252 arbitrators finished the survey and answered the question on costs.
Figure 4.5: Estimated average effects of Claimant versus Respondent appointment on damages award – Experiment 2

Notes: Thin lines denote 95% bootstrapped confidence intervals (5000 iterations). Thick lines denote 90% bootstrapped confidence intervals. Number of observations: 88 appointed by the Respondent, 79 appointed by the Claimant.

While the result is just barely insignificant at the .05 level, we can reject the null at just about any slightly higher rejection threshold. The evidence from the follow-up is strongly suggestive that affiliation bias remains even for more substantive questions and beyond simply decisions on costs. However, it may be the case that the effect is attenuated somewhat by reduced discretion and greater precedential constraints associated with decisions on damages versus costs. A study with greater sample size and power would be needed to assess the magnitude of any such difference. As before, there does not appear to be a sizeable difference between blind and joint-party appointees in their choice of damage awards (Figure 4.6).

Figure 4.6: Estimated average effects of Claimant versus Respondent appointment on damages award – Experiment 2

Notes: Thin lines denote 95% bootstrapped confidence intervals (5000 iterations). Thick lines denote 90% bootstrapped confidence intervals. Number of observations: 45 appointed by the Parties, 36 blind appointees.

A promising result from our replication is that surveyed arbitrators’ answers to the costs question are nearly identical to those from our original study. In the replication, we only exposed arbitrators to
a “Claimant wins” condition (while in the original study most outcomes were conditions where the Claimant failed to obtain an award). Nevertheless, as shown in Figure 4.7, arbitrators appointed by the winner (Claimant) were about fifteen percentage points more likely to have the Respondent reimburse all of the Claimant’s costs. Again, the affiliation bias drives the magnitude of the cost award, but does not affect arbitrators’ initial decision on whether costs should follow the event.

![Graph](image)

*Notes:* Thin lines denote 95% bootstrapped confidence intervals (5000 iterations). Thick lines denote 90% bootstrapped confidence intervals. Number of observations: 90 appointed by Respondent, 79 appointed by Claimant.

**Figure 4.7:** Estimated average effects of winning party appointment – Experiment 2

Finally, we found in the replication experiment that blind appointees differed somewhat from joint appointees in how they apportioned costs (Figure 4.8). Joint appointees were significantly less likely to choose to have the Respondent reimburse all the Claimant’s costs relative to blind appointees. While we saw weak evidence of this in the first experiment, the difference between the two groups is much larger in the replication. However, because the sample size allocated to the blind and joint appointee conditions is smaller (since we wanted more power to detect the affiliation bias effect), it is important to not read too much into this finding. In general, with smaller samples, we can expect more variable and potentially more extreme estimates. Nevertheless, this surprising result does suggest one possible caveat to the observed similarity between blind and joint appointees that we observed in the first experiment – when one party wins outright, joint party appointees may have an aversion to extreme cost awards and favor a slightly more middle-of-the-road compromise relative to blind appointees.
Overall, the evidence across both experiments strongly suggests a meaningful affiliation effect when arbitrators are tasked with allocating some amount between the parties—a very fundamental aspect of arbitration. When given heavy discretion, as is the case for cost awards, party appointees tend to give the party that appointed them a more favorable outcome. Winning party appointees demand more from the loser while losing party appointees try to mitigate their appointer’s losses.

### 4.3.4 External Validity and Sample Composition

The experimental design allows us to isolate the treatment effect from selection biases common to observational analyses. Our estimates have high “internal validity.” We are confident, by design, that the observed association is not due to unobserved factors. Moreover, we obtained a large sample of a very selective group of legal actors—to our knowledge the largest set of experiments conducted in this population. However, it is also important to assess the “external validity” of our estimates—the extent to which our findings can be generalized to the population of interest.

In the case of our experiment, we want to generalize how arbitrators behaved in response to our vignette to the population of arbitrators serving on investment arbitration tribunals—the context of cur-
rent debate and the experiment. If non-response to our survey were random, then we could treat the observations as a random sample from the true population and our sample average treatment effect is an unbiased estimate of the population average treatment effect. Even in the case that non-response is correlated with arbitrators’ characteristics, our experiment still yields an unbiased estimate of the sample average treatment effect. This still captures the average effect of treatment within a theoretically interesting sub-population of the total set of arbitrators. While testing for non-random non-response is impossible, we assess how representative our sample is by comparing the background characteristics provided to us by survey arbitrators to those in the population of interest.

We consider three sets of covariates that we were able to measure for a large subset of our arbitrators: gender, country’s legal tradition, and employment background. Of our 257 arbitrators from the first vignette, 240 also provided information on these covariates. We then obtained a list of all 188 arbitrators that served on an ICSID tribunal constituted between 2010 and 2015 to act as our target population. For each of these variables, we calculated the proportion of arbitrators in each category and compared these target proportions to the proportions within our sample. Figure 4.9 compares the two covariate distributions.

Overall, the distribution of ICSID arbitrators in recent cases is predominantly male, consistent with prior accounts of the lack of women in arbitration appointments. Only 9.5 percent of arbitrators who served on at least one ICSID tribunal in the 2010-2016 period were women. This skew is also evident in our sample, which is very close to the population distribution. However, our sample does contain a slightly larger share of women arbitration experts—about 14 percent. Nevertheless, it is unlikely that this small discrepancy is sufficient to give misleading estimates, particularly as we find no statistically significant difference (p > .1) in treatment effect magnitude between men and women.

Arbitrator nationality is the second variable we considered. Consistent with empirical accounts of the distribution of arbitrator nationalities (e.g. Puig, 2014), the vast majority of arbitrators were nationals of European or North American countries. However, a fair number of arbitrators were also from of South American, Asian or African countries. Our sample is not exclusively comprised of nationals from a single state, which bodes well for the generalizability of our findings. One concern for our
Notes: Crosses denote proportion among ICSID arbitrators 2010-2015. Lines denote 95% confidence intervals.

Figure 4.9: Characteristics of arbitrators in experimental sample
sample is that because the survey was administered in English, we may be more likely to get arbitrators from English-speaking countries and, in particular, countries with English common law traditions. This could be an issue if arbitrators trained in different legal cultures approach the question of cost allocation differently. Indeed, the principle that “costs follow the event” or “loser pays” is extremely well-entrenched in some countries of the common-law tradition (Woodroffe, 1997).

Determining each arbitrator’s legal training is challenging given that nationality is not a perfect proxy—many arbitrators attend foreign law schools to obtain a masters or doctorate abroad after an original law degree. However, for the data that we were able to collect, it does appear to be a reasonable initial proxy. For each arbitrator who reported their nationality or nationalities we coded whether that country’s legal system has a common or civil law tradition using the dataset of La Porta, Lopez-de Silanes and Shleifer (2008). We did the same for the 188 arbitrators in our actual target population. Among ICSID arbitrators, slightly more than half come from exclusively civil law backgrounds while about 41 percent had common law backgrounds.

Within our sample, these proportions are essentially reversed, suggesting that arbitrators with English common law backgrounds were a bit more likely to respond to our survey. However, the magnitude of this difference is not particularly large. We also do not find evidence of effect heterogeneity for arbitrators with common versus civil law backgrounds. Testing for an interaction between our treatment does not yield a difference in effect that is statistically distinguishable from zero at commonly used thresholds (p > .1). While this does not disprove the possibility of effect heterogeneity, it does suggest that our observed evidence for bias is unlikely to hold for only one major legal tradition. Because any effect heterogeneity is likely small or non-existent, the difference between arbitrators’ legal origins in the sample and in the population does not meaningfully impact external validity.

Finally, we consider our arbitrators’ career backgrounds and expertise. Respondents to our survey had four options to indicate their current area of employment: private law, academia, government, or other. These categories were chosen to reflect the most common career backgrounds of international arbitrators as found by Costa (2011). Arbitrators who answered “other” were re-coded into one of the

\footnote{We treat the coding of “socialist” legal systems in La Porta, Lopez-de Silanes and Shleifer (2008) as another form of a civil law system.}
three remaining categories based on their open-ended response. Most arbitrators answering ‘other’ described themselves as independent arbitrators and were recoded as being in the private sector. A few also noted employment in an international organization and were recoded as public-sector employees (along with those selecting “government”). We also coded all 188 ICSID arbitrators in our target population into one of the three career categories based on their most recent area of employment using publicly available information (e.g. websites, CVs).

As shown in Figure 4.9, the majority of respondents in our sample work in the private sector. While this is also the case among ICSID arbitrators, a larger share of the target population originates in academia or the public sector. About 92 percent of our arbitrators indicated private sector employment compared to 69 percent in the ICSID group. While this does suggest a potential limitation of our sample, it is the case that arbitrators with backgrounds in the private sector comprise the largest sector of the overall international arbitration pool. Even if there is a difference in the response of private sector arbitrators relative to those employed in other areas, our experiment credibly identifies the treatment effect for a highly meaningful sub-group of arbitrators. Moreover, as with the other two factors considered, we do not find strong evidence that treatment effects differ between private sector and non-private arbitrators. Therefore, despite the over-representation of arbitrators with private sector employment backgrounds, our sample does not appear to be particularly idiosyncratic in a way that could give misleading results compared to the international arbitration population at large.

4.4 Blind Appointments in Arbitration: Possibilities and Challenges

Despite the current skepticism by commentators (Duarte, 2012; Giorgetti, 2013), our evidence suggests that blinding could be an effective debiasing alternative to correct the observed affiliation effects that result from the current practice of unilateral party appointments. Our methodological contribution shows that the pro-appointer attitudes among arbitrators cannot be explained solely by selection
effects. Therefore, blind appointments would very likely help to mitigate the overall level of party bias in arbitration.

There are other benefits with blind appointments. Chiefly, parties maintain the ability to make unilateral party appointments, one of the oft-cited advantages of arbitration compared to more conventional forms of litigation. Moreover, the practice of blinding need not be implemented solely for unilaterally party-appointed arbitrators but can be easily extended to all members of an arbitration tribunal. For the litigating parties, it may be possible to infer which side made each appointment (as one is known to each), but individual arbitrators will have significant uncertainty over the source of their appointment, having to consider the possibility not only of being a Claimant or a Respondent appointee, but also the potential of being a joint appointee or an appointee by the neutral arbitral institution.

To be sure, some limitations of our experimental approach are clear as there are significant implementation challenges to blinding in practice. Specifically, our experiment is unable to assess how exactly implicit biases of individual arbitrators affect outcomes of a collective body. Prior research on this question suggests that in deliberations, biases tend to be reinforced – an effect known as “bias accentuation” (Takada and Murata, 2014). Hence, the implicit biases of the two party-appointed members may not simply cancel one another out as often assumed, but rather become further entrenched throughout the arbitration process. As to the implementations challenges and limitations of blinding, in this final section we provide some final remarks.

4.4.1 Implementation Challenges

Best practices and professional norms can be developed to encourage blinding without introducing radical changes to the current practice of arbitration. However, among the most important limitations is the administration and enforcement of such a practice. Duarte (2012) notes that the “main problem with [blind appointments] is that it is not so reliable in practice, where it would be extremely easy to find out which party appointed who.” The most obvious issues stem from the fact that professional norms still consider it proper to interview the arbitrators in advance of the nomination (although some
arbitrators will refuse a pre-appointment interview) (Bishop and Reed, 1998; Carter, 2000). This is understandable given that substantial sums of money might rest in the hands of arbitrators. In fact, with limited tools to assess potential decision-makers (CVs, website information, or word-of-mouth), and without access to substantial decisions to assess judicial philosophy (especially in fields such as commercial arbitration) or, many times, guiding precedent, it may be advisable to allow some form of limited exchange between parties and potential arbitrators.

This particular challenge, however, should not be insuperable. For one, potential parties to contractual disputes can include specific guidelines in arbitration clauses on how to conduct such interviews without frustrating blind appointments. Parties could agree not to interview arbitrators at all or to agree to an interview phase that allows both parties to conduct preliminary interviews on all potential, previously identified candidates. Once each party has made a decision on the appointee, arbitrators can be notified with a joint communication from the parties to the litigation or from the arbitral institution, seeking at all times to ensure that nominees do not know which party nominated them.

Arbitral institutions can also issue appointment policies that address ‘blind’ interviewing. In fact, professional best practices on interviewing arbitrators are just emerging and could start including direction on how arbitral institutions can act as intermediates between the litigating parties and the potential nominee to prevent any indiscretion. \(^{14}\) Arbitral institutions can even facilitate the infrastructure for enabling the interviews that prevent the nominee from recognizing the interviewing party and counsel (e.g., use of an intermediary or written interrogatories).

A second obstacle is that some arbitration rules provide that the institution disclose the particular method of constitution of the tribunal.\(^{15}\) What exactly should be disclosed can be subject to interpretation. While there are reasons to opt for a broad interpretation (e.g., transparency of the process) we see very little value in arbitrators knowing the specific party making the appointment, because the arbitrator’s ethical and professional obligations are vis-à-vis both parties – not only the nominating party.


Moreover, such information could be concealed until the final decision is issued if transparency is a concern.

Finally, a more mechanical matter: arbitral institutions must transmit to each member of the tribunal any communication received from either party. In doing so, institutions could ensure that in transmitting information to the members of the tribunals the information on the source of appointment is redacted. Operational policies could even be established across arbitral institutions for requiring appointments made in a separate confidential document that is not shared with the nominees.

4.4.2 Limitations of Blinding in Arbitration

If implemented successfully, blinding could ameliorate affiliation effects. However, it remains to be seen whether that reform will substantially reduce the partisanship observable in many settings of arbitration like investor-state arbitration. There are many reasons for that, but the main one is that selection effects would still persist under blinding. Litigants are aware of the preferences of specific arbitrators, especially of repeatedly appointed arbitrators, and appoint this core group precisely because they are more predictable and effective in signaling a particular position, for example, by voting or dissenting (Puig, 2016). In fact, some arbitrators might prefer to signal their political preference clearly via their decisions and hence less willing to compromise and less reluctant to dissent when in disagreement with the majority.

As a more general point, blinding as a solution to bias should always take context into account. Some arbitration proceedings like investment arbitrations take place in a close-knit community of legal actors and repeat-players who interact routinely. In such contexts, arbitrators rely heavily on social capital and individual reputations to remain a part of the community. Once members of the profession develop a reputation, this information is passed on and translates into other nominations precisely because of such leanings. Hence, while the thick social structure of the arbitral community may be valuable for the implementation of a blind appointment proposal by facilitating the dissemination and development of new professional norms, it may also create additional hurdles by accentuating the selection bias and confidentiality challenges. Nevertheless, in the long run, blinding might help move
arbitration toward a more rule-based system.

4.5 Conclusion

Within the field of arbitration, the frequent use of party-appointed arbitrators is likely to result in litigant-induced biases. Disentangling selection effects (parties appointing friendly arbitrators) from affiliation effects (arbitrators changing behavior in response to their appointment) is particularly challenging using observational data. While the former is inherent in the system of party appointment, the latter could be ameliorated by “blinding” arbitrator – restricting party appointees from knowing the source of their appointment throughout the arbitration proceedings.

Through novel survey experiments, we provide a methodological solution to the problem of measuring affiliation effects in a world confounded by selection effects. As our survey results show, assignment to being appointed by one of the parties in a dispute directly changes the behavior of arbitrators. Hence, the appointment itself is the cause of some of the bias towards one’s appointing party. Apparent patterns of bias in real decisions are unlikely to be simply an effect of parties filtering their appointees on the basis of known, prior attitudes. Normatively, these results provide support for implementing “blinding” within arbitration proceedings. While the implementation challenges do not seem insuperable, the effectiveness of blinding may also be heavily dependent on the context of arbitration. Hence, this chapter should be a precursor for assessing affiliation bias and the impact of blind appointments in more realistic settings. We believe the proposed methodology is a significant contribution to exploring questions on judicial politics and the psychology of decision-maker biases.
Appendix A

Supporting Materials for Chapter 2

Proof for Proposition 2.3.1

The two-way fixed effects estimator assumes the following data-generating process

\[ E[Y_{it}|A_{it}] = \alpha_i + \gamma_t + \beta A_{it} \]

where \(\alpha_i\) denotes unit fixed effects, \(\gamma_t\) denotes time fixed effects and \(\beta\) is the quantity of interest.

The ordinary least squares estimates of the parameters: \(\hat{\beta}, \hat{\gamma}, \hat{\alpha}\) solve the least-squares optimization problem

\[ \hat{\beta}, \hat{\gamma}, \hat{\alpha} = \arg\min_{\beta, \gamma, \alpha} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \gamma_t - \beta A_{it})^2 \]
The first-order conditions for $\beta$

\[
0 = \sum_{i=1}^{N} \sum_{t=1}^{T} -2A_{it} \left( Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta}A_{it} \right)
\]

\[
0 = \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t \right) - \hat{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2
\]

\[
\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}
\]

For $\alpha_i$, the FOCs are:

\[
0 = \frac{1}{T} \sum_{t=1}^{T} -2 \left( Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta}A_{it} \right)
\]

\[
\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} - \frac{1}{T} \sum_{t=1}^{T} \hat{\gamma}_t - \hat{\beta} \frac{1}{T} \sum_{t=1}^{T} A_{it}
\]

And for $\gamma_t$

\[
0 = \sum_{n=1}^{N} -2 \left( Y_{it} - \hat{\alpha}_i - \hat{\gamma}_t - \hat{\beta}A_{it} \right)
\]

\[
\hat{\gamma}_t = \frac{1}{N} \sum_{n=1}^{N} Y_{it} - \frac{1}{N} \sum_{n=1}^{N} \hat{\alpha}_i - \hat{\beta} \frac{1}{N} \sum_{n=1}^{N} A_{it}
\]

Let $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}$, $\bar{Y}_t = \frac{1}{N} \sum_{i=1}^{N} Y_{it}$, $\bar{Y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}$, $\bar{A}_i = \frac{1}{T} \sum_{t=1}^{T} A_{it}$, $\bar{A}_t = \frac{1}{N} \sum_{i=1}^{N} A_{it}$, $\bar{A} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}$.

Then, re-write the first-order conditions

\[
\hat{\gamma}_t = \bar{Y}_t - \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_i - \hat{\beta} \bar{A}_t
\]
\[ \hat{\alpha}_i = \bar{Y}_i - \frac{1}{T} \sum_{t=1}^{T} \gamma_t - \hat{\beta} \bar{A}_i \]

Substituting one into the other

\[ \hat{\alpha}_i = \bar{Y}_i - \frac{1}{T} \sum_{t=1}^{T} \left[ \bar{Y}_t - \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_i - \hat{\beta} \bar{A}_t \right] - \hat{\beta} \bar{A}_i \]

\[ = \bar{Y}_i - \bar{\bar{Y}} + \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_i + \hat{\beta} \bar{A} - \hat{\beta} \bar{A}_i \]

Then substituting into the expression for \( \hat{\alpha}_i + \hat{\gamma}_t \)

\[ \hat{\alpha}_i + \hat{\gamma}_t = \bar{Y}_i - \bar{\bar{Y}} + \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_i + \hat{\beta} \bar{A} - \hat{\beta} \bar{A}_i + \bar{Y}_t - \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_i - \hat{\beta} \bar{A}_t \]

\[ = \bar{Y}_i + \bar{Y}_t - \bar{\bar{Y}} + \hat{\beta} \bar{A} - \hat{\beta} \bar{A}_i - \hat{\beta} \bar{A}_t \]

\[ = \bar{Y}_i + \bar{Y}_t - \bar{\bar{Y}} - \hat{\beta} (\bar{A}_i + \bar{A}_t - \bar{A}) \]
Substituting back into $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \left( \bar{Y}_{t} + \bar{Y}_{i} - \bar{Y} - \hat{\beta}(A_{i} + A_{t} - \bar{A}) \right) \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \bar{Y}_{t} - \bar{Y}_{i} + \bar{Y} + \hat{\beta}(A_{i} + A_{t} - \bar{A}) \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \bar{Y}_{t} - \bar{Y}_{i} + \bar{Y} \right) + \hat{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}(A_{i} + A_{t} - \bar{A})}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \bar{Y}_{t} - \bar{Y}_{i} + \bar{Y} \right) + \hat{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}(A_{i} + A_{t} - \bar{A})}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$

$$\hat{\beta} \left[ 1 - \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it}(A_{i} + A_{t} - \bar{A})}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2} \right] = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \bar{Y}_{t} - \bar{Y}_{i} + \bar{Y} \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left( Y_{it} - \bar{Y}_{t} - \bar{Y}_{i} + \bar{Y} \right)}{\sum_{i=1}^{N} \sum_{t=1}^{T} (A_{it})^2}$$
It is possible to expand the expression in the numerator to:

\[
Y_{it} - \bar{Y}_t - \bar{Y}_i + \bar{Y}
\]

\[
Y_{it} - \frac{1}{N} \sum_{i' = 1}^{N} Y_{i' t} - \frac{1}{T} \sum_{t' = 1}^{T} Y_{i t'} + \frac{1}{NT} \sum_{i' = 1}^{N} \sum_{t' = 1}^{T} Y_{i' t'}
\]

\[
\left[ 1 - \frac{1}{N} - \frac{1}{T} + \frac{1}{NT} \right] Y_{it} - \frac{1}{N} \sum_{i' \neq i} Y_{i' t} - \frac{1}{T} \sum_{t' \neq t} Y_{i t'} - \frac{1}{T} \sum_{t' \neq t} \left[ Y_{i t'} - \frac{1}{N} Y_{i t'} \right] + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i' t'}
\]

\[
\frac{NT - N - T + 1}{NT} Y_{it} - \frac{1}{N} \sum_{i' \neq i} \left[ Y_{i' t} - \frac{1}{T} Y_{i t} \right] - \frac{1}{T} \sum_{t' \neq t} \left[ Y_{i t'} - \frac{1}{N} Y_{i t'} \right] + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i' t'}
\]

\[
\frac{(N - 1)(T - 1)}{NT} Y_{it} - \frac{T - 1}{NT} \sum_{i' \neq i} Y_{i' t} - \frac{N - 1}{NT} \sum_{t' \neq t} Y_{i t'} + \frac{1}{NT} \sum_{i' \neq i} \sum_{t' \neq t} Y_{i' t'}
\]

Denote the normalizing constant \( C \) with

\[
C = \frac{(N - 1)(T - 1)}{NT} \frac{1}{\sum_{i = 1}^{N} \sum_{t = 1}^{T} A_{it} \left( A_{it} - \bar{A}_i - \bar{A}_t + \bar{A} \right)}
\]
Then write $\hat{\beta}$ as

$$\hat{\beta} = C \times \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \left\{ \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] - \frac{1}{N - 1} \sum_{t' \neq i} \left[ Y_{it'} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] \right\}$$

$$= C \times \sum_{t=1}^{T} \left\{ \sum_{i=1}^{N} A_{it} \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] - \frac{N}{A_{it}} \sum_{i=1}^{N} A_{it} \sum_{t' \neq i} \left[ Y_{it'} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] \right\}$$

Let $N_{t}^{(a)}$ denote the number of units that are under treatment $a$ in period $t$. Let $T_{t}^{(a)}$ denote the number of periods for which unit $i$ receives treatment $a$.

Note that in the second difference term, every unit under control in period $t$ appears here $N_{t}^{(1)}$ times. Likewise, every unit under treatment appears $N_{t}^{(1)} - 1$ times, excluding the period when it is in the first difference. Therefore, re-write the expression as

$$\hat{\beta} = C \times \sum_{t=1}^{T} \sum_{i=1}^{N} A_{it} \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} - \frac{N_{t}^{(1)} - 1}{N - 1} \left( Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right) \right]$$

$$+ (1 - A_{it}) \frac{N_{t}^{(1)}}{N - 1} \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right]$$

$$\hat{\beta} = C \times \sum_{t=1}^{T} \left\{ \sum_{i=1}^{N} \frac{N_{t}^{(0)}}{N - 1} A_{it} \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] - (1 - A_{it}) \frac{N_{t}^{(1)}}{N - 1} \left[ Y_{it} - \frac{1}{T - 1} \sum_{t' \neq t} Y_{it'} \right] \right\}$$

Re-arranging terms yields
\[ \hat{\beta} = C \times \sum_{t=1}^{T} \left\{ \frac{N_t^{(0)}}{N - 1} \sum_{i=1}^{N} Y_{it} A_{it} - \frac{N_t^{(1)}}{N - 1} \sum_{i=1}^{N} Y_{it}(1 - A_{it}) \right\} - \\
C \times \sum_{t=1}^{T} \left\{ \frac{N_t^{(0)}}{(N - 1)(T - 1)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} Y_{it'} - \frac{N_t^{(1)}}{(N - 1)(T - 1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} Y_{it'} \right\} \]

\[ \hat{\beta} = C \times \sum_{t=1}^{T} \left\{ \frac{N_t^{(0)}}{N - 1} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - \frac{N_t^{(1)}}{N - 1} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} [Y_{it} - Y_{it'}] \right\} - \\
C \times \sum_{t=1}^{T} \left\{ \frac{1}{(T - 1)(N - 1)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} Y_{it'} - \frac{1}{(T - 1)(N - 1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} Y_{it'} \right\} \]

\[ \hat{\beta} = C \times \frac{1}{(N - 1)(T - 1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} [Y_{it} - Y_{it'}] \right\} - \\
C \times \frac{1}{(T - 1)(N - 1)} \sum_{t=1}^{T} \left\{ N_t^{(0)} \sum_{i=1}^{N} A_{it} \sum_{t' \neq t} Y_{it'} - N_t^{(1)} \sum_{i=1}^{N} (1 - A_{it}) \sum_{t' \neq t} Y_{it'} \right\} \]

Incorporating the normalizing constant, we have a uniform average over all differences-in-differences.

\[ \hat{\beta} = C \times \frac{1}{(N - 1)(T - 1)} \sum_{t=1}^{T} \left\{ \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} [Y_{it} - Y_{it'}] - [Y_{jt} - Y_{jt'}] \right\} \\
\sum_{i=1}^{N} \sum_{i:A_{it}=1} \sum_{j:A_{jt}=0} \sum_{t' \neq t} \{1 - A_{it'} + A_{jt'}\} \]

129
Bibliography


