Why Rich Countries Win Investment Disputes: Taking Selection Seriously

Anton Strezhnev*
Draft†

September 22, 2017

Abstract

Investor-state dispute settlement (ISDS) is a rapidly growing field of international law, due mainly to a proliferation of investment treaties among states that grant foreign investors the right to pursue binding arbitration against states for alleged violations of property rights. However, many have argued that poorer countries are systematically disadvantaged in arbitration since the costs of litigation and possible arbitrator bias in favor of wealthy capital exporters make it harder for these states to defeat investors' claims. This article argues that while high-income governments tend to win about 20% more investment disputes than low- or middle-income governments, this disparity in win-rates can be explained by differences in the rates of early settlement between rich and poor governments. I find that developing country governments with high costs to litigation are about 22 percentage points more likely to settle a given dispute. This generates a selection effect among those disputes that are actually litigated. Investors with low-quality cases that could otherwise extract an acceptable settlement from a poorer respondent state are forced to litigate against wealthier governments, inducing a spurious positive correlation between respondent state income level and success rate. After adjusting for this selection process using a novel weighting approach, I find that wealthier respondent countries are no more likely to obtain a favorable ruling from an arbitration tribunal than poorer governments. Overall, while resource disparities among litigants do matter for outcomes in ISDS, observed differences in win-rates are not primarily due to the biases of arbitrators. Rather, wealthier governments have greater bargaining power when negotiating over settlements with investors, which affects the types of disputes for which arbitrators ultimately render awards.

*Harvard University, Department of Government astrezhnev@fas.harvard.edu.

†The author thanks Matthew Blackwell, Marc Ratkovic, Dustin Tingley, Gary King, and participants at the 2017 New Faces in Political Methodology Conference at Penn State for helpful comments on previous versions of this paper.

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 protect the investments of foreign nationals, obligating them 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 by states 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, indeed contract based arbitrations date back to as early as 1864 (Yackee, 2016), BITs have enabled the rapid expansion of ISDS claims by expanding the scope of claims that could be brought to arbitration. By granting blanket consent to arbitration to classes of foreign investors over treaty violations, 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 ISDS 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 exproproiation or mistreatment of individual firms by governments, a wave of more recent arbitrations have instead challenged broad regulatory policies. Claimant investors have alleged that environmental, health, and other 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 organizations like the European Court of Human Rights, which also grant private individuals standing to bring claims against states, ISDS is unique

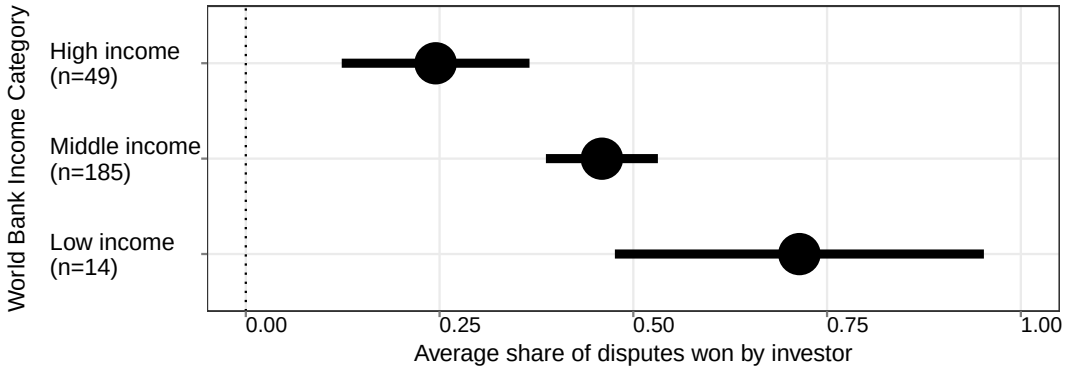
in that there is no formal court in which claims are adjudicated.[1] Rather, ISDS claims are litigated through a form of ad-hoc arbitration. The seat of arbitration is itself often de-nationalized and not governed by any indivdual state's legal system. Arbitrators do not have permanent tenure and serve on a dispute-by-dispute basis. The parties themselves typically 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 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."[2] 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. One major proposition 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 "respondent" states. 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).

[1]Efforts 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.

[2]Hamby, Chris. "The Court that Rules the World" *Buzzfeed News*. August 28, 2016. `https://www.buzzfeed.com/chrishamby/super-court`.

It is difficult to assess whether ISDS is systematically more pro-corporate rather than pro-state bias, as this would require some normative belief about what the "correct" balance of corporate and state rights are under international law. A more tractable empirical question is whether poorer respondent states are 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.



$N = 383$. Lines denote 95% confidence intervals

Figure 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 1 plots the average claimant win-rates by World Bank income

category for all disputes considered analyzed in the article. 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 article 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 the 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. 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* successfuly 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. This is the case empirically as well. After adjusting for likely confounders, this article shows that 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 with 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). Because well-resourced respondents will tend to

provide even lower offers, the average case quality of the claimant should be lower among unsettled disputes against high-capacity respondents compared to low-capacity respondents. Essentially, 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.

This problem of post-treatment drop-out is common in other areas of research. For example, in some areas of medical research, individuals under observation may die prior to the observation of some auxiliary 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 auxiliary outcome. 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, the "Survivor Average Causal Effect" (SACE) (Rubin, 2006). Analyzing the data conditional on the observed survival outcome will result in biased estimates of the causal effect when treatment has some effect on the propensity that a unit reaches follow-up. Methods for estimating the SACE require additional assumptions. The approach used in this article, principal score weighting (Jo and Stuart, 2009; Ding and Lu, 2016; Feller, Mealli and Miratrix, 2017), uses weights to adjust for imbalances in observed covariates induced by attrition. I find that after adjusting for confounding and post-treatment selection, the effect of high-income versus low/middle-income respondent on the probability that the claimant wins the dispute is statistically indistinguishable from zero. In other words, the observed difference in outcomes shown in figure 1 can be reasonably explained by differential settlement rates between high-income and low/middle-income states. The overall findings of this article 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 article is structured as follows. Section 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 discusses the empirical strategy for estimating the effect of legal capacity

on settlement by adjusting for potential confounding variables. Section 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 4 describes the data used in the analysis. Section 4 presents the results and analyzes the extent to which the covariate adjustment methods were sufficient to reduce imbalance. Section 4 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.

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."[3] 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

[3]Hamby, Chris. "The Billion Dollar Ultimatum." *Buzzfeed News*. August 30, 2016. `https://www.buzzfeed.com/chrishamby/the-billion-dollar-ultimatum`.

for 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.[4] Assume the claimant has private information over the probability of their claim being successful. The respondent state does not observe this, only knowing the distribution of the quality of potential claims.[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.[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

[4]While 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.

[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 details of their particular investment than governments do.

[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 article 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.

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 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. As the amount the respondent offers decreases with reductions in the respondent's legal costs decrease, 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's will be of higher capacity 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 a fair adjudication system.

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 from the data. This section briefly reviews the assumptions necessary to estimate a causal effect in an observational study and discusses the strategy I use to adjust for omitted variable bias when estimating the effect of legal capacity on settlement. Formally, consider a sample of N disputes. For each unit, indexed by i, we observe the realized value of a treatment of interest, denoted A_i. and a vector of pre-treatment covariates $\mathbf{X_i}$. A_i is binary with 1 denoting the "treatment" condition: a high-income respondent and 0 denoting the "control" condition: a lower-income respondent. We 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. We 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). We 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 we 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 can 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 be-

[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).

tween treatment arms risk biased inferences – the classic omitted variable bias problem. The goal of any observational causal inference 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 $\mathbf{X_i}$ (Imbens, 2004). In the context of this article, 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 the 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 arrangements that make arbitrators more favorable to that respondent's

[8]For 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 differences between 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.

arguments for, for example, the transparency of its regulatory decisionmaking, then that is still a form of country-level 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 institutions and bias attributable to other factors. This requires estimating a different type of causal effect – the "controlled direct effect" (Acharya, Blackwell and Sen, 2015) of wealth under another intervention holding constant institutional quality. These types of causal mediation effects more clearly define the specific manipulation 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 high co-linearity between development and the institutional variable resarchers want to hold constant (King and Zeng, 2005).

Respondent country level characteristics are by definition not confounders of treatment assignment; they are *a part* of the treatment of interest as what is being manipulated is the respondent country involved in a given dispute, not the wealth level of a respondent country. However, there do exist a number of confounders that are correlated with a case being filed against a wealthy rather than a poor 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 claimants 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. 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 jurisdiction (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. Disputes 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).[9] 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). 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

[9]For evidence on whether arbitrators tend to favor expansive interpretations of arbitration provisions, see (Van Harten, 2015).

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. This 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 pit-fall 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 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 incorporate 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

[10]See Austin (2011) for additional discussion of the performance of propensity score weighting relative to regression methods.

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 mis-specifications 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 4, the propensity score weighting model works remarkably well, reducing average imbalance on the covariates by about a factor of seven.

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 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 (Rosenbaum, 1984; Montgomery, Nyhan and Torres, 2016). 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 the. Conditioning on a post-treatment variable breaks any balance between treated and control groups achieved through weihgting 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 already decided 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).[11] 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. 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 negatively 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, ex-post, respondents with high litigation costs will be associated with disputes with higher-quality claimants. These are the cases that failed to settle (i.e. had claimants that refused to take the 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 long-term outcomes such as quality of life, some respondents may

[11] 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.

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 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 quantity, 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 with the same individual causal effect on the intermediate variable.

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

[12]For the purposes of this article, 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.

cannot estimate average treatment effects for the full sample. $E[Y_i(1) - Y_i(0)]$ 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 formally as

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

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 1 illustrates the relationship between the observed quantities, treatment A_i and 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-treatment.

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

[13] It 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).

[14] In 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 maipulated and "forced" to take on a value of 1 for all units. Such "controlled" treatment effect quantities are often considered in causal mediation analysis. 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, 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.

Observed Quantities		Principal Strata	
A_i	S_i		
1	1	$S_i(1) = 1, S_i(0) = 1$ (Always survivor)	$S_i(1) = 1, S_i(0) = 0$ (Survivor under treatment)
0	1	$S_i(1) = 1, S_i(0) = 1$ (Always survivor)	$S_i(1) = 0, S_i(0) = 1$ (Survivor under control)
1	0	$S_i(1) = 0, S_i(0) = 0$ (Never survivor)	$S_i(1) = 0, S_i(0) = 1$ (Survivor under control)
0	0	$S_i(1) = 0, S_i(0) = 0$ (Never survivor)	$S_i(1) = 1, S_i(0) = 0$ (Survivor under treatment)

Table 1: Observed data and possible principal strata

that the treatment effect on the intermediate variable is monotonic.

Assumption 1. *Monotonicity*

$$S_i(1) \geq S_i(0)$$

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 repsondent'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).

$$\begin{aligned}
\text{Always-survivors: } & Pr(S_i(1) = S_i(0) = 1) = Pr(S_i = 1|A_i = 0) \\
\text{Survivors under treatment: } & Pr(S_i(1) = 1, S_i(0) = 0) = Pr(S_i = 1|A_i = 1) - Pr(S_i = 1|A_i = 0) \\
\text{Never-survivors: } & Pr(S_i(1) = 0, S_i(0) = 0) = 1 - Pr(S_i = 1|A_i = 1)
\end{aligned}$$

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]

$$\begin{aligned}
\widehat{SACE} &= E[Y_i|A_i = 1, S_i = 1] - E[Y_i|A_i = 0, S_i = 1] \\
\widehat{SACE} &= E[Y_i(1)|S_i(1) = 1] - E[Y_i(0)|S_i(1) = 1, S_i(0) = 1] \\
\widehat{SACE} &= E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]\frac{\pi_{11}}{\pi_{11} + \pi_{10}} + E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]\frac{\pi_{10}}{\pi_{11} + \pi_{10}} - E[Y_i(0)|S_i(1) = 1, S_i(0) = 1]
\end{aligned}$$

Therefore, the bias of the naive difference-in-means estimator under ignorability is

$$\begin{aligned}
\text{Bias}(\widehat{SACE}) &= E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]\frac{\pi_{11}}{\pi_{11} + \pi_{10}} + E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]\frac{\pi_{10}}{\pi_{11} + \pi_{10}} - E[Y_i(1)|S_i(1) = 1, S_i(0) = 1] \\
&= E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]\frac{-\pi_{10}}{\pi_{11} + \pi_{10}} + E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]\frac{\pi_{10}}{\pi_{11} + \pi_{10}} \\
&= [E[Y_i(1)|S_i(1) = 1, S_i(0) = 0] - E[Y_i(1)|S_i(1) = 1, S_i(0) = 1]]\frac{\pi_{10}}{\pi_{11} + \pi_{10}}
\end{aligned}$$

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.

$$E[Y_i(1)|S_i(1) = S_i(0) = 1] > E[Y_i(1)|S_i(1) = 1, S_i(0) = 0]$$

[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 possiblity 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 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 magnitude of the bias is equal to the share of partial survivors divided by the share of partial survivors or always-survivors. Therefore, 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 $\mathbf{X_i}$, 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 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 identfiying the SACE. Define the always-survivor principal score for unit i as

$$w_i = Pr(S_i(1) = 1, S_i(0) = 1 | A_i, S_i, \mathbf{X_i})$$

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

$$w_i = \frac{Pr(S_i(1) = 1, S_i(0) = 1 | \mathbf{X_i})}{Pr(S_i(1) = 1, S_i(0) = 1 | \mathbf{X_i}) + Pr(S_i(1) = 1, S_i(0) = 0 | \mathbf{X_i})} \Big/ \frac{Pr(S_i(1) = 1, S_i(0) = 1)}{Pr(S_i(1) = 1, S_i(0) = 1) + Pr(S_i(1) = 1, S_i(0) = 0)}$$

$$w_i = \frac{Pr(S_i = 1 | A_i = 0, \mathbf{X_i})}{Pr(S_i = 1 | A_i = 1, \mathbf{X_i})} \Big/ \frac{Pr(S_i = 1 | A_i = 0)}{Pr(S_i = 1 | A_i = 1)}$$

Like the propensity score, the principal score is a balancing score. When the model is correctly specified, the distribution of covariates across strata should be equal.

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

This permits a very straightforward assessment of model quality. As with propensity scores, comparing covariate balance after weighting should allow researchers to assess whether weights have reduced bias. This model checking is a necessary element of the estimation process since often strong modeling assumptions typically need to be made in order to reliably estimate the weights. When there are few covariates that are all discrete, non-parametric estimates of the probabilities can be easily obtained. However, when there are many covariates, researchers will need to fit some form of regression model of survival on the covariates and treatment.

One method is to simply split the sample and fit one model for survivorship among untreated

units for the numerator and a model for survivorship among the treated. However, when there are few treated units, the weighting models can result in unstable weights, particularly when there are few observations in either the treated or control group. An alternative approach is to fit a pooled regression model with a limited number of covariate-treatment interactions such that the weights are allowed to vary across units.[16] By placing constraints on the types of permitted interactions between treatment and covariates, this approach yields less variable weights. However, it can also induce potential bias by restricting the way in which covariates $\mathbf{X_i}$ can relate to the stratum probabilities. When searching over the space of reasonable models, researchers should use the covariate balance conditions to determine whether adding more covariates or interactions into the model will reduce or increase treatment/control imbalance.

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 fit a series of candidate logistic regression models for the principal scores. I select the model that most reduces the mean absolute divergence among covariates. The final weights for estimating the SACE are a product of the CBPS weights and the principal score weights. The discussion of the results in section 4 will illustrate the changes in imbalance during each step of the process, highlighting how principal score weights, while insufficient to bring imbalance to zero, nevertheless substantially reduces covariate

[16]Note that estimating a model with no treatment-covariate interactions yields weights that are constant across all observations.

divergence between treated and control arms.

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).[17] 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 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.[18] As of August 2017, this dataset encompasses a large fration of bilateral

[17] Available at `http://investmentpolicyhub.unctad.org/ISDS`.
[18] Available at `http://investmentpolicyhub.unctad.org/IIA`.

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.[19] 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.[20]

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 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. However, the type of issue under dispute is more difficult to code. Existing work (Pelc, 2017) notes that one major qualitatve difference among claims is the difference between investors challenging direct 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 challenging plain-packaging regulations .[21] While the tobacco regulations did not exclusively target any particular tobacco firm, Phillip Morris alleged that

[19]For a discussion of the coding methodology see `http://investmentpolicyhub.unctad.org/Upload/Documents/Mapping%20Project%20Description%20and%20Methodology.pdf`.

[20]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.

[21]*Philip Morris Asia Limited v. The Commonwealth of Australia.* (UNCITRAL, PCA Case No. 2012-12). Award on Jurisdiction and Admissibility. December 17, 2015. ; *Philip Morris Brands Sàrl, Philip Morris Products S.A. and Abal Hermanos S.A. v. Oriental Republic of Uruguay*, (ICSID Case No. ARB/10/7). Award. July 8, 2016.

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 *Metalclad v. Mexico*, the tribunal stated that actions involving "covert or indicidental interference" that have the effect of "depriving the owner...of the use or reasonably-to-be-expected economic benefit of property even if not necssarily to the obvious benefit of the State" could constitute violations of a government's treaty obligations.[22] 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"[23], not all "indirect expropriation" disputes concern blanket regulations. Rather, allegations of indirect expropriation are often also a feature 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*.[24] 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

[22] *Metalclad Corporation v. The United Mexican States*. (ICSID Case No. ARB(AF)/97/1). Award. August 30, 2000. para 103.

[23] 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. *Mesa Power Group v. The Government of Canada*. (UNCITRAL, PCA Case No. 2012-17). Award. March 24, 2016.

[24] *Occidental Petroleum Corporation and Occidental Exploration and Production Company v. The Republic of Ecuador*. (ICSID Case No. ARB/06/11). Award. September 9, 2008. para 455.

on its bond contracts had an indirect expropriatory effect[25]. 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 avaialable 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 IAReporter 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

[25] *Abaclat and Others. v. The Argentine Republic.*(ICSID Case No. ARB/07/5). Decision on Jurisdiction and Admissibility. August 4, 2011.; *Giovanni Alemanni and Others. v. The Argentine Republic.* (ICSID Case No. ARB/07/8). Decision on Jurisdiction and Admissibility. November 17, 2014.; *Ambiente Ufficio and Others. v. The Argentine Republic.* (ICSID Case No. ARB/08/9). Decision on Jurisdiction and Admissibility. February 8, 2013.

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.[26] 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 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 firm-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.[27]

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,[28] health insurance market reform[29], price setting in energy markets[30], and zoning policy [31]. Some

[26] *Gami Investments, Inc. v. The Government of the United Mexican States* (UNCITRAL). Final Award. 15 November 2004.

[27] 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.

[28] *Chemtura Corporation v. Government of Canada.* (UNCITRAL). Award. August 2, 2010.

[29] *Achmea B.V. v. The Slovak Republic.* (UNCITRAL, PCA Case No. 2008-13). Award. December 7, 2012.

[30] *Iberdrola Energa S.A. v. Republic of Guatemala.* (ICSID Case No. ARB/09/5). Award. August 17, 2012.

[31] *MTD Equity Sdn. Bhd. and MTD Chile S.A. v. Chile.* (ICSID Case No. ARB/01/7). Award. May 25, 2004

disputes even allege damages due to a government's *failure* to regulate as in *Anderson v. Costa Rica.* [32]. Other policy challenges have been directed at broad taxation of particular industries [33] and even trade barriers like import bans, quotas, or tariffs that often have analogues in WTO disputes [34].

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

Table 2: Distribution of dispute type across 383 ISDS BIT claims

Table 2 summarizes the distribution of coded disputes in the sample of 383 BIT arbitrations analyzed in this article. 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 indeed 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 litigation aimed at chilling regulation. Rather, firms engaging in business with government agencies have found BITs a viable method for escalating what are essentially private disputes using 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,

[32] *Alasdair Ross Anderson and others v. Republic of Costa Rica.* (ICSID Case No. ARB(AF)/07/3). Award. May 19, 2010.

[33] *Murphy Exploration & Production Company International v. The Republic of Ecuador.* (UNCITRAL, PCA Case No. 2012-16). Final Award. February 10, 2017.

[34] *Pope & Talbot Inc. v. The Government of Canada.* (UNCITRAL). Interim Award. June 26, 2000.; *Tembec Inc. et al. v. United States of America.* (UNCITRAL). Decision on the Preliminary Question. June 6, 2006.; *The Canadian Cattlemen for Fair Trade v. United States of America.* (UNCITRAL). Award on Jurisdiction. January 28, 2008.

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 treates breaches of contract by states as a violation of international law, permitting investors to access international arbitration even in the case of a purely contractual dispute.

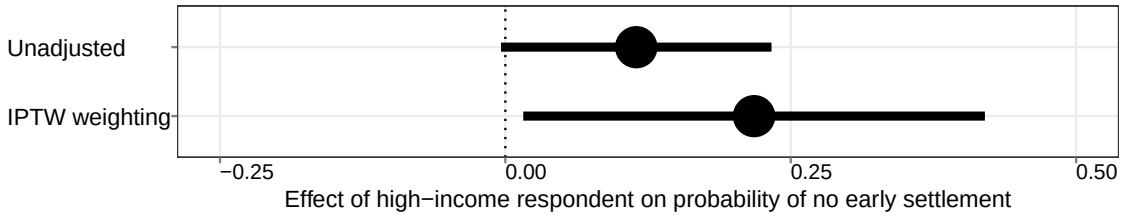
Second, BITs are also increasingly including provisions that limit investors claims in areas involving regulatory policy and other matters typically seen to be the purview of sovereign states. These exceptions either restrict the scope of the types of investments covered by the treaty or allow the government to permissly derogate from its BIT obligations under certain exceptional circumstances. UNCTAD's IIA mapping project identifies three types of substantive limitations: taxation, subsidies, government procurement policy along with four types of exemptions: security exceptions, public health and environmental exceptions, exceptions for prudential financial regulations, and other broad public policy 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 reviewability of awards, and, to some extent, the requirements for a tribunal to admit a claim (Jagusch and Sullivan, 2010).

Results

Figure 2 plots the estimated average treatment effect of having a high-income respondent in a dispute on the probability that the dispute will fail to settle.[35] All weighting models estimate robust 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.



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

Figure 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, 4, and 5 plot the difference in proportions between high- and low/middle- income respondent disputes prior to weighting and after. Imbalance on all covariates is

[35]Since there are comparatively fewer disputes with high-income respondents relative to low-income, I estimate 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 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. 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.

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 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.

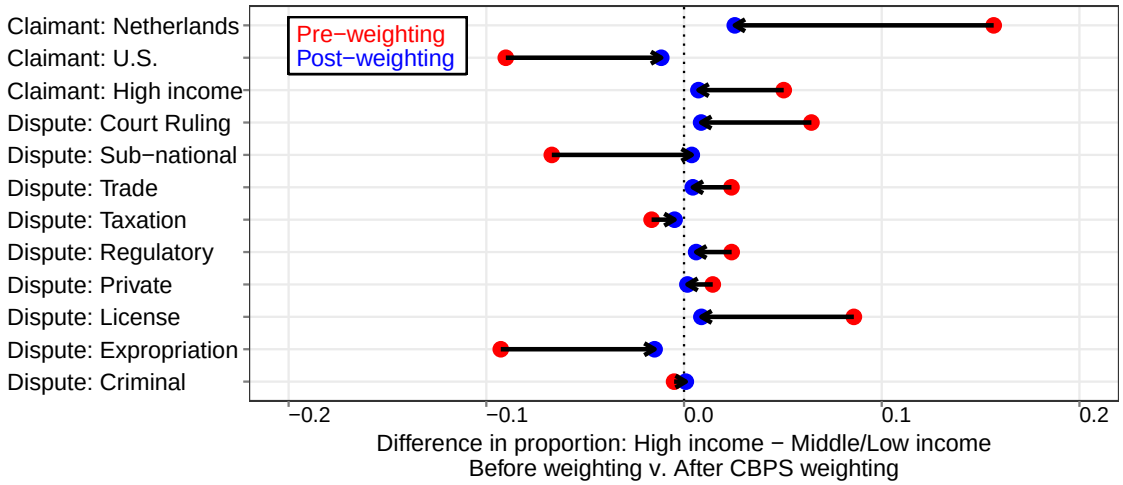


Figure 3: Covariate balance pre- and post- weighting – Claimant and dispute type

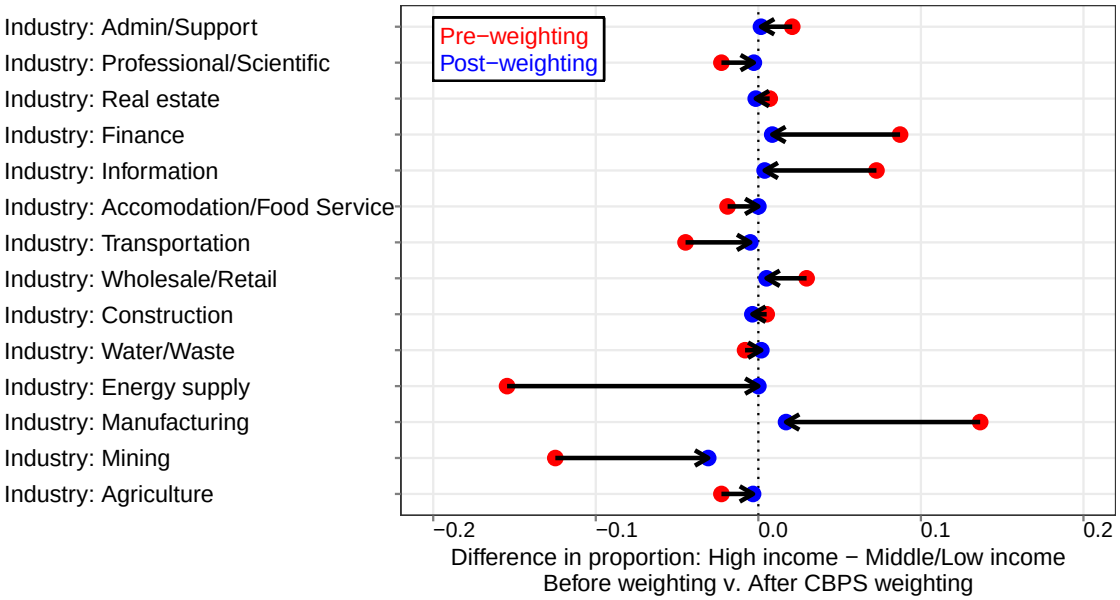


Figure 4: Covariate balance pre- and post- weighting – Industry of claimant

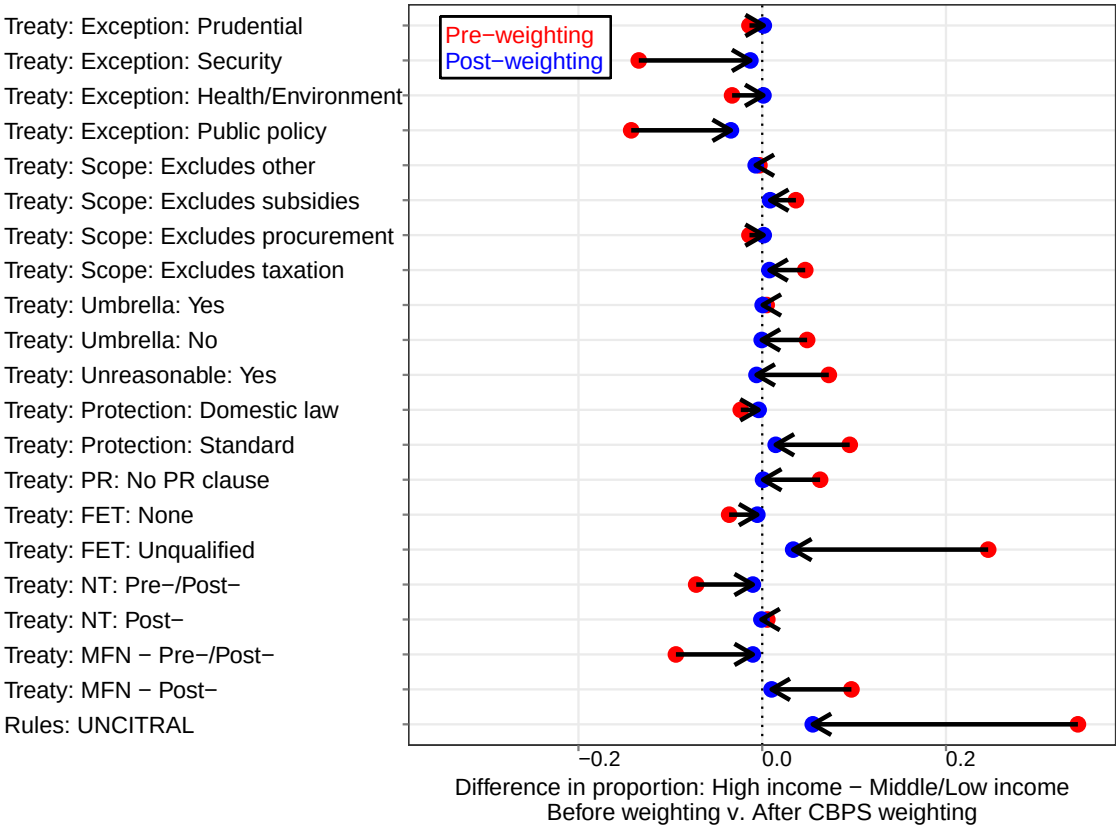


Figure 5: Covariate balance pre- and post- weighting – Applicable BIT provisions

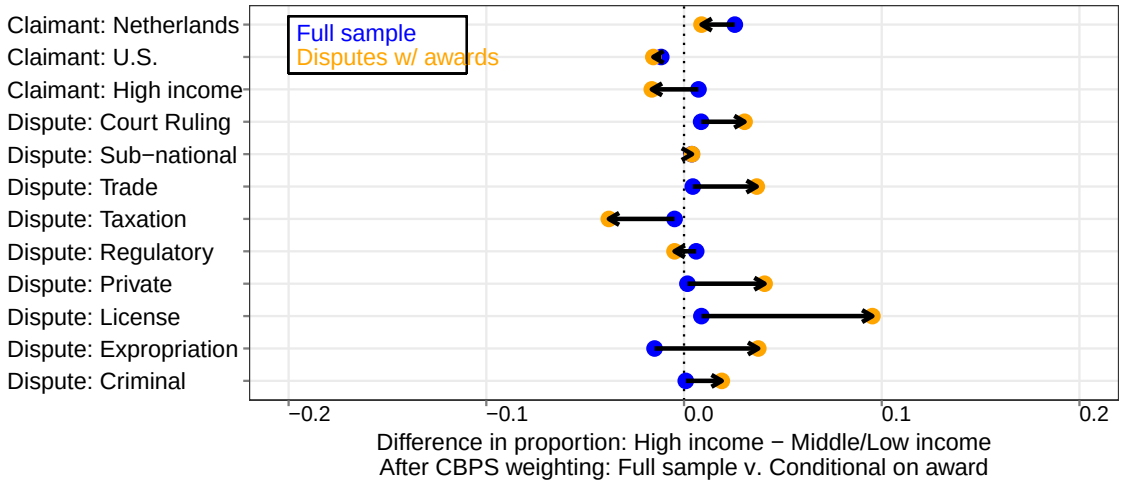


Figure 6: Conditioning on non-settlement breaks balance – Claimant and dispute type

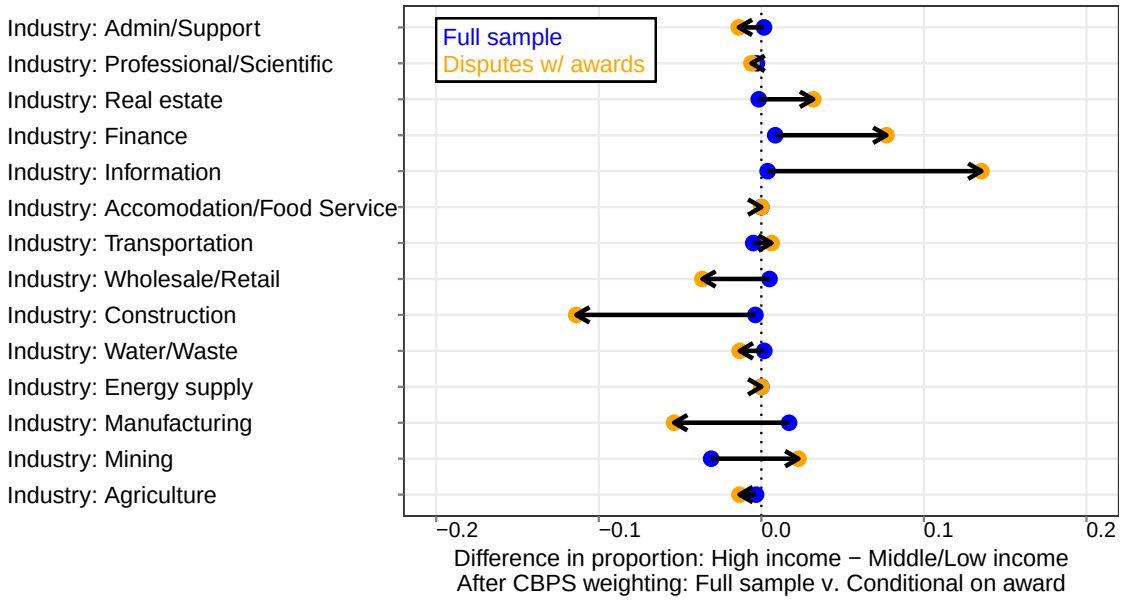


Figure 7: Conditioning on non-settlement breaks balance – Industry of claimant

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 balance. Figures 6, 7, and 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 is at work.

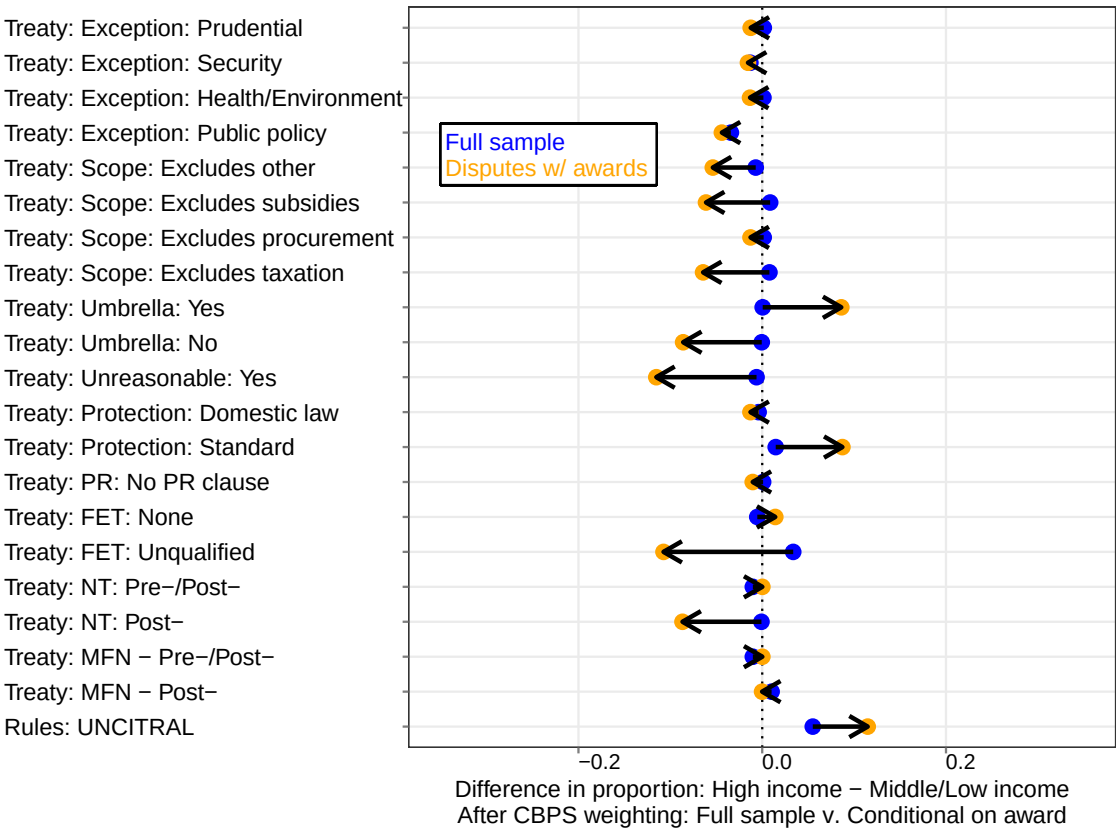


Figure 8: Conditioning on non-settlement breaks balance – Applicable BIT provisions

Estimating the weighting model for the principal scores is challenging given the large covariate space and the small sample. I started with an additive logistic model fit with the IPTW weights from the first stage. However, adding every single variable from the first stage very clearly exacerbated imbalance due to high estimation variance. Fitting a parsimonious model required selecting which variables exhibit the largest degree of imbalance and including them in the survivorship model, interacted with the treatment variable. Some variables with relatively small imbalance were omitted to keep the variability of the weights low. I iterated through a series of candidate regression models, and selected the model with the largest reduction in mean absolute imbalance across all covariates.

The resulting weights have the effect of reducing average absolute imbalance from 0.04 to 0.027, However, this improvement is not uniform across all covariates as figures 9, 10, and 11 illustrate. While most covariates see improvements in balance, there are a few on which imbalance is worsened after applying principal score weights, namely claimant nationality. While on average, this is off-set by improvements in balance for covariates like industry or dispute type, it points to some limitations in principal score adjustment with a high-dimensional covariate space. Optimal model selection remains an area for future research, but the application of weighting does appear to address most of the egregious cases of imbalance induced by selecting on those disputes with awards.

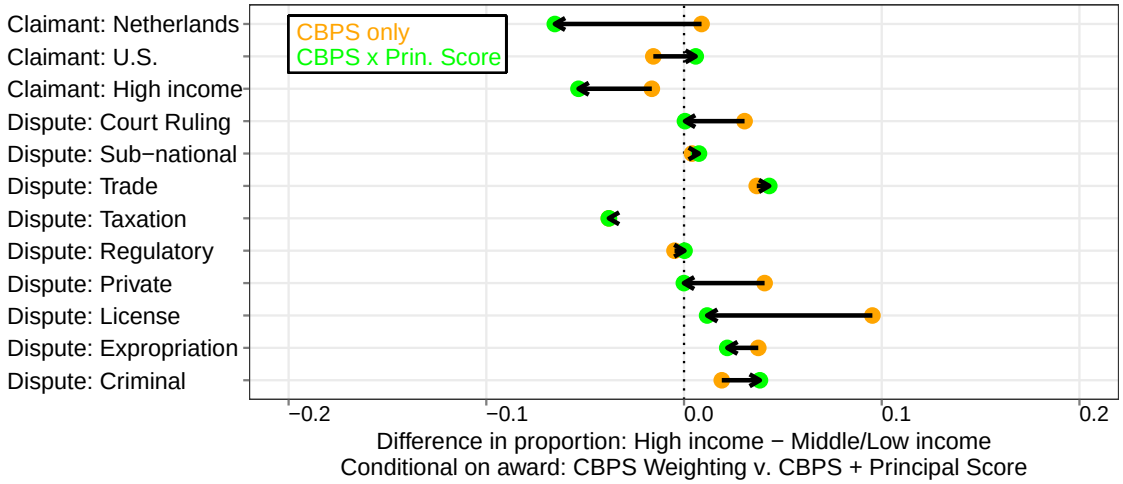


Figure 9: Conditioning on non-settlement breaks balance – Claimant and dispute type

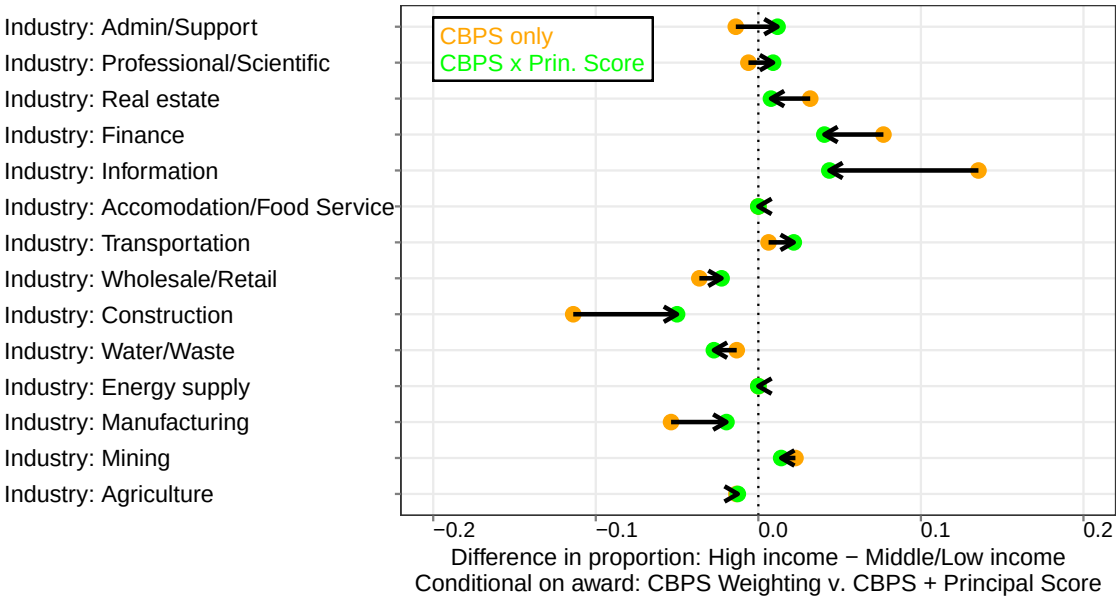


Figure 10: Conditioning on non-settlement breaks balance – Industry of claimant

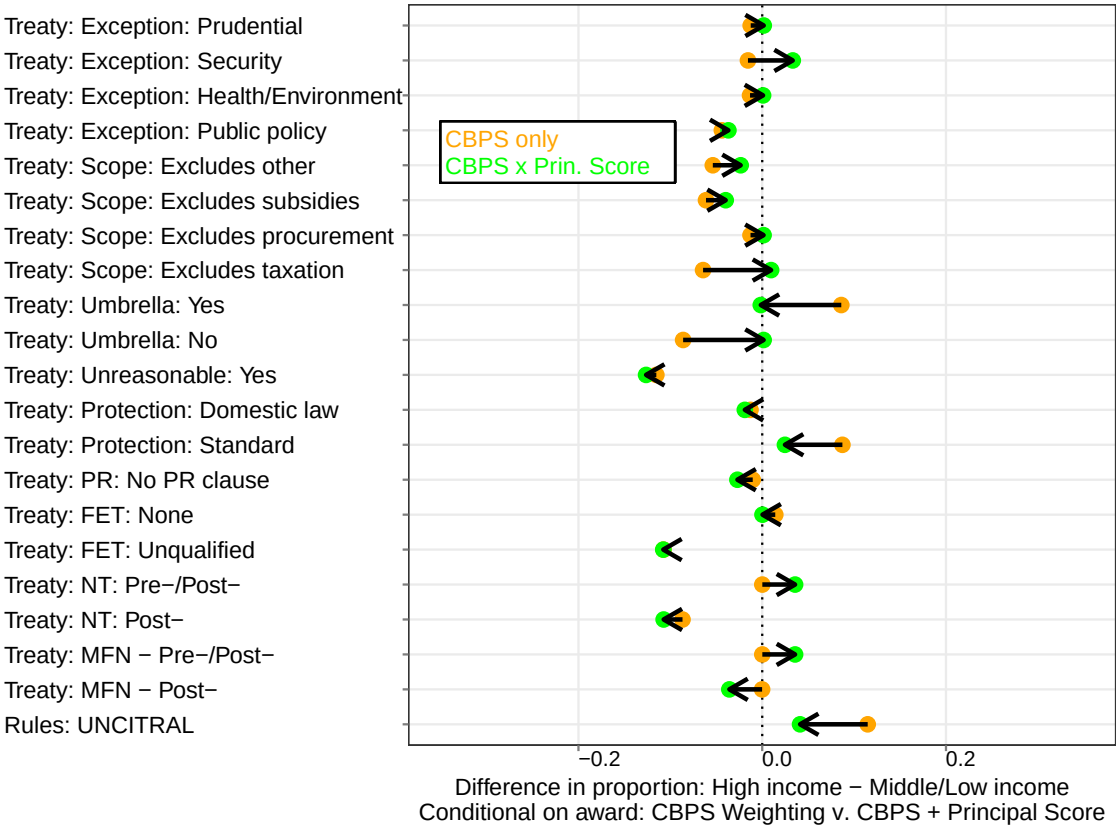
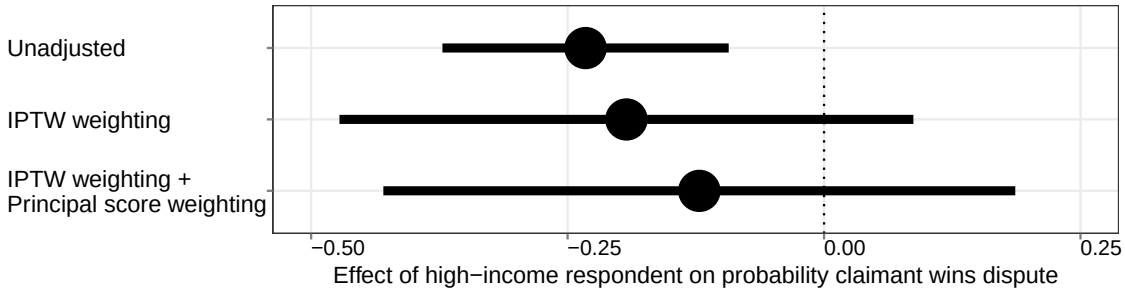


Figure 11: Conditioning on non-settlement breaks balance – Applicable BIT provisions



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

Figure 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 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. While the estimate with no covariate adjustment shows a strong negative effect of about 23 percentage points, incorporating covariates using IPT weighting reduces the point estimate slightly – to 19 percentage points – and substantially raises the uncertainty around the estimate. 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. Incorporating the principal score weights in addition to the propensity score weights further reduces the point estimate to about 12 percentage points. After adjustment, the effect of respondent income level on win probability is statisticall 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 an over-estimate given theoretical expectations over confounding. If there is a lurking confounder of principal strata 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 differentially fail to settle should be weaker than those that would settle under either condition. Overall, the results persuasively show evidence for a strong negative effect of respondent wealth on probability of settlement. The evidence for any advantage of wealth 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 betweeen

win-rates of high- and middle/low- income respondent governments can likely be attributed to attrition, rather than systematic bias by the tribunal.

Conclusion

For many policymakers and legal scholars, investor-state dispute settlement is at a cross-roads. While ubiquitous in Bbilateral investment treaties during the last few decades, ISDS provisions have become a stumbling block in negotiations for modern trade-agreements. Many critiques of ISDS allege that arbitration courts 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 article critically evaluates claims of bias against developing countries in investment arbitration proceedings. It emphasizes that analyses of win-rates are fundamentally meaningless without some 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 is 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, it argues 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 winning 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 in more instances. As a result, there may exist a set of claimants that would reach a settlement with a litigation-averse government that would not receive acceptable settlement offers from high-income governments.

Under this selection process, high-income governments are able to force some claimants with weaker cases to litigate instead of reaching a settlement.

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 article 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 dispute type and 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 compelled 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 leal 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.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2015. "Detecting Direct Effects and Assessing Alternative Mechanisms.".

Akerlof, George A. 1970. "The market for" lemons": Quality uncertainty and the market mechanism." *The quarterly journal of economics* pp. 488–500.

Allee, Todd and Andrew Lugg. 2016. "Who wrote the rules for the Trans-Pacific Partnership?" *Research & Politics* 3(3):2053168016658919.

Allee, Todd and Clint Peinhardt. 2014. "Evaluating three explanations for the design of bilateral investment treaties." *World Politics* 66(1):47–87.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91(434):444–455.

Aronow, Peter M and Allison Carnegie. 2013. "Beyond LATE: Estimation of the average treatment effect with an instrumental variable." *Political Analysis* pp. 492–506.

Aronow, Peter M and Cyrus Samii. 2016. "Does regression produce representative estimates of causal effects?" *American Journal of Political Science* 60(1):250–267.

Austin, Peter C. 2011. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research* 46(3):399–424.

Austin, Peter C. 2013. "The performance of different propensity score methods for estimating marginal hazard ratios." *Statistics in medicine* 32(16):2837–2849.

Bebchuk, Lucian Arye. 1984. "Litigation and settlement under imperfect information." *The RAND Journal of Economics* pp. 404–415.

Behn, D, M Langford and TL Berge. 2017. "Poor states or poor governance? Predicting outcomes in investment treaty arbitration." *Review of Law and Economics* .

Boyd, Christina L, Lee Epstein and Andrew D Martin. 2010. "Untangling the causal effects of sex on judging." *American journal of political science* 54(2):389–411.

Brower, Charles N and Sadie Blanchard. 2013. "What's in a Meme-The Truth about Investor-State Arbitration: Why It Need Not, and Must Not, Be Repossessed by States." *Colum. J. Transnat'l L.* 52:689.

Busch, Marc L and Eric Reinhardt. 2003. "Developing countries and general agreement on tariffs and trade/world trade organization dispute settlement." *J. World Trade* 37:719.

Busch, Marc L, Eric Reinhardt and Gregory Shaffer. 2009. "Does legal capacity matter? A survey of WTO Members." *World Trade Review* 8(4):559–577.

Chiba, Yasutaka and Tyler J VanderWeele. 2011. "A simple method for principal strata effects when the outcome has been truncated due to death." *American journal of epidemiology* p. kwq418.

Cole, Stephen R, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson and Charles Poole. 2009. "Illustrating bias due to conditioning on a collider." *International journal of epidemiology* 39(2):417–420.

Davis, Christina L and Sarah Blodgett Bermeo. 2009. "Who files? Developing country participation in GATT/WTO adjudication." *The Journal of Politics* 71(3):1033–1049.

Ding, Peng and Jiannan Lu. 2016. "Principal stratification analysis using principal scores." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Feller, Avi, Fabrizia Mealli and Luke Miratrix. 2017. "Principal Score Methods: Assumptions, Extensions, and Practical Considerations." *Journal of Educational and Behavioral Statistics* .

Franck, Susan D. 2009. "Development and outcomes of investment treaty arbitration." *Harv. Int'l LJ* 50:435.

Franck, Susan D. 2010. "Rationalizing costs in investment treaty arbitration." *Wash. UL Rev.* 88:769.

Franck, Susan D. 2014. "Conflating Politics and Development: Examining Investment Treaty Arbitration Outcomes." *Va. J. Int'l L.* 55:13.

Frangakis, Constantine E and Donald B Rubin. 2002. "Principal stratification in causal inference." *Biometrics* 58(1):21–29.

Frumento, Paolo, Fabrizia Mealli, Barbara Pacini and Donald B Rubin. 2012. "Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data." *Journal of the American Statistical Association* 107(498):450–466.

Gottwald, Eric. 2006. "Leveling the playing field: is it time for a legal assistance center for developing nations in investment treaty arbitration." *Am. U. Int'l L. Rev.* 22:237.

Greenland, Sander. 2003. "Quantifying biases in causal models: classical confounding vs collider-stratification bias." *Epidemiology* 14(3):300–306.

Greiner, D James and Donald B Rubin. 2011. "Causal effects of perceived immutable characteristics." *Review of Economics and Statistics* 93(3):775–785.

Guzman, Andrew T and Beth A Simmons. 2005. "Power plays and capacity constraints: The selection of defendants in World Trade Organization disputes." *The Journal of Legal Studies* 34(2):557–598.

Haftel, Yoram Z. 2010. "Ratification counts: US investment treaties and FDI flows into developing countries." *Review of International Political Economy* 17(2):348–377.

Ho, Daniel E, Kosuke Imai, Gary King and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15(3):199–236.

Hodgson, M. 2014. "Costs in Investment Treaty Arbitration: The Case for Reform." *Transnational Dispute Management (TDM)* 11(1).

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396):945–960.

Imai, Kosuke. 2008. "Sharp bounds on the causal effects in randomized experiments with truncation-by-death." *Statistics & probability letters* 78(2):144–149.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Imbens, Guido W. 2004. "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and statistics* 86(1):4–29.

Jagusch, Stephen and Jeffrey Sullivan. 2010. "A Comparison of ICSID and UNCITRAL Arbitration: Areas of Divergence and Concern." *The Backlash against Investment Arbitration: Perceptions and Reality* pp. 79–109.

Jo, Booil and Elizabeth A Stuart. 2009. "On the use of propensity scores in principal causal effect estimation." *Statistics in medicine* 28(23):2857–2875.

King, Gary and Langche Zeng. 2005. "The dangers of extreme counterfactuals." *Political Analysis* 14(2):131–159.

Kryvoi, Yaraslau. 2010. "Piercing the Corporate Veil in International Arbitration." *Global Bus. L. Rev.* 1:169.

Laird, Ian and Rebecca Askew. 2005. "Finality Versus Consistency: Does Investor-State Arbitration Need an Appellate System." *J. App. Prac. & Process* 7:285.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2016. How conditioning on post-treatment variables can ruin your experiment and what to do about it. In *Annual meeting of the Midwest Political Science Association, Chicago, IL, April*.

Nalebuff, Barry. 1987. "Credible pretrial negotiation." *The RAND Journal of Economics* pp. 198–210.

Neumayer, Eric and Laura Spess. 2005. "Do bilateral investment treaties increase foreign direct investment to developing countries?" *World development* 33(10):1567–1585.

Nunnenkamp, Peter. 2017. "Biased Arbitrators and Tribunal Decisions Against Developing Countries: Stylized Facts on Investor-State Dispute Settlement." *Journal of International Development* 29(6):851–854.

Pelc, Krzysztof J. 2017. "What Explains the Low Success Rate of Investor-State Disputes?" *International Organization* pp. 1–25.

Posner, Eric A and Miguel FP de Figueiredo. 2005. "Is the International Court of Justice Biased?" *The Journal of Legal Studies* 34(2):599–630.

Puig, Sergio. 2014. "Social Capital in the Arbitration Market." *European Journal of International Law* 25(02):387–424.

Rogers, Catherine A. 2005. "Transparency in international commercial arbitration." *U. Kan. L. Rev.* 54:1301.

Rosenbaum, Paul R. 1984. "The consquences of adjustment for a concomitant variable that has been affected by the treatment." *Journal of the Royal Statistical Society. Series A (General)* pp. 656–666.

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* pp. 41–55.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

Rubin, Donald B. 1986. "Comment: Which ifs have causal answers." *Journal of the American Statistical Association* 81(396):961–962.

Rubin, Donald B. 2001. "Using propensity scores to help design observational studies: application to the tobacco litigation." *Health Services and Outcomes Research Methodology* 2(3):169–188.

Rubin, Donald B. 2006. "Causal inference through potential outcomes and principal stratification: application to studies with" censoring" due to death." *Statistical Science* pp. 299–309.

Schultz, Thomas and Cédric Dupont. 2014. "Investment arbitration: promoting the rule of law or over-empowering investors? A quantitative empirical study." *European Journal of International Law* 25(4):1147–1168.

Simmons, Beth A. 2014. "Bargaining over BITs, arbitrating awards: The regime for protection and promotion of international investment." *World Politics* 66(1):12–46.

Stuart, Elizabeth A. 2010. "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1):1.

Tobin, Jennifer L and Susan Rose-Ackerman. 2011. "When BITs have some bite: The political-economic environment for bilateral investment treaties." *The Review of International Organizations* 6(1):1–32.

Trakman, Leon E. 2013. "ICSID Under Siege, The." *Cornell Int'l LJ* 45:603.

Van Harten, Gus. 2015. "Arbitrator Behaviour in Asymmetrical Adjudication (Part Two): An Examination of Hypotheses of Bias in Investment Treaty Arbitration." *Osgoode Hall LJ* 53:540.

VanderWeele, Tyler and Stijn Vansteelandt. 2009. "Conceptual issues concerning mediation, interventions and composition." *Statistics and its Interface* 2:457–468.

Vandevelde, Kenneth J. 1988. "The bilateral investment treaty program of the United States." *Cornell Int'l LJ* 21:201.

Waibel, Michael. 2010. *The backlash against investment arbitration: perceptions and reality*. Kluwer Law International.

Wellhausen, Rachel L. 2016. "Recent trends in investor–state dispute settlement." *Journal of International Dispute Settlement* 7(1):117–135.

Yackee, Jason Webb. 2016. "The First Investor-State Arbitration: The Suez Canal Company v Egypt (1864)." *The Journal of World Investment & Trade* 17(3):401–462.

Zhang, Junni L and Donald B Rubin. 2003. "Estimation of causal effects via principal stratification when some outcomes are truncated by death." *Journal of Educational and Behavioral Statistics* 28(4):353–368.