Detecting Bias in International Investment Arbitration[*]

Anton Strezhnev[†]

Draft

March 12, 2016

Abstract

Foreign direct investment is increasingly coming under the governance of a patchwork of bilateral investment agreements among states that grant investors rights to legal recourse and arbitration in the event of property rights violations by a host country. These investor-state arbitrations often take place in international legal fora such as the World Bank's International Centre for Settlement of Investment Disputes (ICSID). While meant as an impartial alternative to weak host country legal systems, these international investment fora have been criticized for favoring the rights of investors over those of states. However, uncovering empirical evidence of this bias is difficult because of strategic pre-award settlement by parties to a dispute. If panel composition affects claimants' win probabilities, it also likely affects the probability of settlement. As a result, analyzing only those cases that result in a decision generates a form of selection bias. This paper outlines a new a method for estimating panel composition effects in the presence of non-random settlement. I apply this estimator to examine whether arbitrator career background affects the likelihood of claimant victory in ICSID arbitration. I find conditional evidence of pro-claimant bias among arbitrators from advanced economies. I estimate that when the presiding member is a national of an advanced economy and has had experience working in government as opposed to purely private law/academia, claimants are about 25% more likely to receive an award of damages.

[*] Paper presented at the 57th Annual Convention of the International Studies Association - Atlanta, Georgia. March 16-19, 2016

[†] Harvard University, Department of Government astrezhnev@fas.harvard.edu.

Introduction

International relations is increasingly becoming judicialized (Alter, 2014). States are increasing their reliance on formal and semi-formal methods for resolving disputes - in particular methods of arbitration and adjudication by panels of judges. This trend towards judicialization has been particularly pronounced in the area of investor-state dispute resolution. The rapid growth over the past two decades in the use of bilateral investment treaties (BITs) as a means of regulating property rights across borders is particularly indicative of this trend (Elkins, Guzman and Simmons, 2006). BITs often permit foreign investors to pursue recourse against violations of property rights committed by host countries. This recourse takes the form of arbitration in international legal fora - outside of the host country's legal system. One of the most common forums for investor-state dispute resolution is the International Center for the Settlement of Investment Disputes (ICSID). ICSID disputes usually take the form of claims by private firms brought against states, alleging a breach of contract or treaty. As such, they represent a new trend in international dispute resolution - the growth of disputes between private and public entities.

Decisions rendered by the arbitral tribunals charged with hearing these disputes have meaningful consequences for both parties involved. Awards in favor of firms – the claimants – against states – almost always respondents – for expropriation of property can amount to sizeable portions of state budgets. One notable decision, *Occidental v. Republic of Ecuador*, awarded claimants \$1.76 billion USD plus interest in damages – the largest award in ICSID history (Sabahi and Duggal, 2013).[1] Moreover, near universal adoption of the 1958 New York Arbitration Convention which stipulates that parties recognize foreign arbitration awards means that when states refuse to pay, claimants can enforce decisions through domestic courts and potentially seize assets.[2] Investor-state arbitration tribunals certainly have teeth and have developed far beyond being the curiosities of a small group of legal academics.

The turn by states to judicial mechanisms of dispute settlement in the investment sphere raises a number of important questions regarding what non-legal factors determine the outcomes of

[1] For comparison, according to the World Bank's World Development Indicators, Ecuador's nominal GDP in \$USD in the year of the award – 2012 – was \$87.6 billion.

[2] A famous example is the case of German national Franz Sedelmayer, who, in response to an unpaid award by the Russian federation, has attempted to seize Russian assets held abroad. Most recently, Sedelmayer was successful in obtaining a judgement from the Swedish Supreme Court permitting seizure of property in Stockholm held by the Russian government (Wrange, 2012).

these disputes. Investor-state dispute settlement institutions have come under extensive criticism from both politicians and legal scholars for systematically favoring the interests of multinational corporations over those of states.[3] More generally, Trakman (2013) notes that "an underlying concern among some developing states...is that the ICSID was established by, and arguably in the interest of, wealthy countries and their investors abroad." (606). Arbitrators are often seen as "male, pale, and stale," (Giorgetti, 2013) with the vast majority coming from wealthy advanced industrial economies that are typically capital exporters as opposed to investment hosts. Indeed, a central research question for international relations scholars studying this emergent system of investor state dispute settlement is the extent to which the composition of this institution generates systematic bias towards the interests of a particular set of actors. A number of recent studies within law have attempted to address this question,[4] but the implications should be of significant interest to scholars of international institutions and international relations more broadly.

Inferring systematic bias in ISDS, however, from empirical analyses of decisions is complicated by the structure of arbitration. By construction, arbitral institutions are designed to encourage parties to reach mutually acceptable compromises in lieu of a costly binding award. This "bargaining in the shadow of the law" is an element of almost all adversarial legal contexts, including in other international institutions like the WTO dispute settlement resolution. Pre-trial bargaining between claimant and respondent weighs the value of a proposed settlement against the cost of subsequent litigation and probability of winning before a tribunal. What this means in practice is that not all arbitration disputes receive a final award from an arbitral tribunal – many settle or are discontinued before reaching that stage. Because settlement is in part a function of beliefs about the probability of victory, in theory, the only disputes that go to a final award are those that are "close" cases.

The challenge for empirical research is that observed win-rates tell scholars very little about whether an institution is systematically biased in favor of one side or the other. Franck (2007) note that in ICSID disputes, claimant and respondent win rates are relatively close to 50%. the

[3] For example, in a recent Op-Ed, U.S. Senator Elizabeth argued " ISDS...wouldnt employ independent judges. Instead, highly paid corporate lawyers would go back and forth between representing corporations one day and sitting in judgment the next...If youre a lawyer looking to maintain or attract high-paying corporate clients, how likely are you to rule against those corporations when its your turn in the judges seat?" *See:* Elizabeth Warren, February 25, 2015 "The Trans-Pacific Partnership clause everyone should oppose." The Washington Post.

[4] See, for example Van Harten (2016).

classic Priest-Klein model of pre-trial bargaining suggests, when parties can settle in lieu of costly litigation, the share of plaintiff victories will in general tend towards 50% (Priest and Klein, 1984). Indeed, as Shavell (1996) notes, this is only a limiting case – because settlement failure in such models is due to incomplete information about outcome rather than the outcome itself, nearly any plaintiff win rate is possible irrespective of the underlying "strictness" of the legal standard for or against a plaintiff.

This paper chooses to avoid the tricky normative discussion of what a fair claimant win-rate "ought" to be, noting that the question is fundamentally unanswerable given strategic dynamics regarding which cases ultimately get litigated. Rather, it argues that while data on dispute outcomes tells researchers little about overall levels of systematic bias, it *does* provide information on individual sources of bias. Do different types of arbitrators influence how different types of panels ultimately decide cases? That is, irrespective of the strictness of the underlying legal standards, how will different panel compositions affect the ultimate outcome of a dispute?

From the standpoint of social science, ICSID arbitration presents a novel case for examining strategic and attitudinal models of judicial behavior where judges have mixed incentives with respect to maintaining legal credibility and rendering overtly "biased" decisions. While in an ideal legal system, the final outcome should be invariant to the lawyer making the decision given the same facts, this is clearly not the case in practice. Scholars of domestic legal systems have noted that factors as wide-ranging as race (Kastellec, 2013), gender (Farhang and Wawro, 2004), whether judges have daughters (Glynn and Sen, 2015), and hunger or fatigue (Danziger, Levav and Avnaim-Pesso, 2011) all affect the way in which a judge rules. Judges are not ideal-type analytical machines but humans responsive to human concerns. The investment arbitration system accentuates one of these concerns in particular – reputation. Arbitration is unique in that it relies so heavily on informal rather than formal systems of organization. For example, arbitrators are not appointed to fixed terms as is the case in many permanent court systems – both international and domestic – and must compete for re-appointment in order to remain within the arbitral community. Given this combined system of party influence and broader social pressures, what effect will changes in tribunal composition and professional background have on the outcome? For scholars of international institutions looking to understand why international courts behave the way they do, this is a very important question to answer in order to explain the winners and losers from the investment arbitration system.

I analyze two dimensions of individual background with respect to the presiding members of ICSID arbitral tribunals – nationality and career history. I find that tribunals where the president is a national of an advanced economy and had previously served in a government position are about 25% more likely to favor the claimant relative to other tribunals. In general, there is a strong, positive, statistically significant joint effect of these two variables on claimants' win rates in arbitration. Where the presiding member comes from a primarily governmental as opposed to professional/legal background, the nationality of the presiding member plays an important role, with arbitrators from wealthier countries. Conversely, I find no effect of presiding member nationality when the presiding member does not have a background. Overall, these results have important implications for how scholars should understand the growing professionalization of the investment arbitration Gaillard (2015). While it is troubling for respondent countries (typically developing states) that tribunal composition can so heavily influence win rates, there is a silver lining to the development of a distinct "elite" class of arbitrators. Legal norms coupled with material incentives for re-appointment appear to check overt expressions of national bias. Overall, these results suggest that professional norms are one mechanism international courts can insulate themselves from principal-agent pressures and operate with some independence of their creators – consistent with the "trustee" model of Alter (2008).

The rest of this paper is structured as follows. Section I provides an overview of the international investment legal regime and how it emerged as a means of regulating foreign direct investment in lieu of a single, formal international institution like the WTO. Section II reviews the existing empirical literature on judicial decisionmaking and develops a theory of how and why arbitrator background might affect arbitration outcomes and lays out the main hypotheses to be tested. Section III discusses the central empirical challenge of this analysis – the problem of settlement – and describes a novel approach to estimating that avoids the biases induced by conditioning only on those cases that go to trial. Section IV discusses the dataset and the results. I examine roughly 258 total disputes completed in the International Centre for the Settlement of Investment Disputes (ICSID) prior to April 2015, of which 180 resulted in a final award rendered by the tribunal. Section V explores avenues for future research and summarizes the main contributions of this paper.

1 Background

Investment Arbitration

Disputes between home and host countries over the treatment of foreign direct investment (FDI) date back many centuries. Indeed, as Frieden (1994) describes, variation in patterns of colonial occupation and control can in part be explained by tensions between colonial powers and local governments over the expropriability of investments. For much of the history of FDI, such disputes would often be resolved through bilateral diplomacy between governments, and, when negotiations failed, military force. The history of "gunboat diplomacy" during the late nineteenth century highlights the extent to which capital exporters were willing to use military coercion to secure private actors' property rights abroad. For example, throughout the first third of the 20th century, the United States often employed military interventions in Latin American states for the purpose of collecting debts owed to private U.S. citizens Vandevelde (2005).

During the latter half of the twentieth century, the global international investment regime saw a substantial shift towards legalization. Contrary to trade, which was based primarily around multilateral legal instruments like the General Agreement on Tariffs and Trade (GATT), investment legalization was driven by *bilateral* processes. The instrument that would come to create the patchwork international investment legal architecture was the BIT – the Bilateral Investment Treaty. BITs represent a commitment by states to respect the investment and property rights of one another's nationals. Parties to BITs typically commit to, among other things, restrictions on expropriation without adequate compensation, non-discrimination, national treatment, and some restrictions on capital controls. Elkins, Guzman and Simmons (2006) note an explosion of BIT ratification throughout the latter half of the twentieth century and particularly after the end of the Cold War. They hypothesize that competition among developing states for capital flows from developed states incentivized states to make treaty commitments to improve their business climate. Notably, the evidence for *whether* BITs actually increase FDI for developing countries remains mixed (Egger and Pfaffermayr, 2004; Yackee, 2008).

What is guaranteed under BITs, however, is litigation. One crucial provision within almost every BIT permits private investors to arbitrate against states for alleged violations of the BIT. Such arbitration provisions are not unique to BITs – indeed, many firm-state contracts, particularly in the energy sector, include such clauses – but the explosion in investment arbitration over the last few decades can very likely be tied to the growth in adoption and use of BITs. Simmons (2014) finds that when states ratify new BITs, the number of investment arbitrations initiated against them increases dramatically. Importantly, investor-state arbitration borrows heavily from concepts and practices developed in private international law and firm-to-firm arbitration (Drahozal, 2009). Arbitral panels are ad-hoc and the arbitration system has at its core a "market" of highly trained arbitration professionals who specialize in adjudicating such disputes (Rogers, 2004). Arbitrators do not retain permanent positions on tribunals – each arbitration panel is constituted independently by the parties to a dispute and in accordance with the chosen set of arbitration rules. The choice of arbitration rules is not inconsequential and often BITs will specify a set of arbitration forums and rules that prospective litigants can choose from. Institutions and rules vary somewhat in the degree of privacy and specific procedures available to the parties. For example, arbitrations registered under the United Nations Commission on International Trade Law (UNCITRAL) arbitration rules can remain completely confidential, while registrations with the International Centre for Investment Disputes (ICSID) are made public, even if the parties choose to keep the content of the award ultimately confidential.

ICSID

For the purposes of this paper, I focus primarily on arbitrations conducted under the International Centre for the Settlement of Investment Disputes (ICSID) due to practical considerations (the universe of ICSID cases is known while many disputes under other arbitral institutions remain confidential). However, there is strong reason to believe that ICSID is not an outlier, both in procedure and in relevance to investment arbitration as a whole. ICSID is one of the main pillars of investment arbitration. The ICSID Convention enjoys extremely broad acceptance among states, with currently 151 parties as of April 18, 2015.[5] It is a component of the World Bank Group and has been in existence since 1966 (Reed, Paulsson and Blackaby, 2011). Moreover, even states that are not party to the ICSID Convention have included ICSID provisions in their BITs via the ICSID Additional Facility – most notably, Mexico (Gonzalez de Cossio, 2008). Based on both temporal and geographic coverage, it is unlikely that the set of ICSID disputes is particularly

[5] See `https://icsid.worldbank.org/apps/ICSIDWEB/icsiddocs/Pages/List-of-Member-States.aspx`

unrepresentative of investor-state arbitration at large. Moreover, Franck (2010) finds no significant difference between amounts claimed and awarded in ICSID awards compared to a sample of investor-state awards from other institutions.

The structure of ICSID rules regarding tribunal constitution also has much in common with procedures used in many other arbitration settings. According to the ICSID Convention Arbitration Rules[6], the default tribunal – unless agreed to by the parties – is comprised of three arbitrators.[7] In a three-arbitrator panel, each party appoints one arbitrator and the third member, who serves as President, is decided by mutual agreement. Alternatively, parties may delegate the selection of the presiding member to their appointed co-arbitrators. Notably, to guard against potential co-national favoritism, the arbitration rules do not permit nationals of either party to serve on a tribunal. If the parties cannot make the appointments within 90 days of dispute registration, then one party may request that the Chair of the Administrative Council appoint the President and any remaining, un-appointed, arbitrators.[8] Appointments made by the Chair must be made from the "Panel of Arbitrators" – a list of qualified arbitrators named by states that is maintained by ICSID.[9] Once constituted, the arbitration tribunal hears arguments from both parties and, absent a settlement by the parties or a suspension of proceedings, renders an award. In some cases, the tribunal will choose to bifurcate proceedings into separate jurisdiction and liability proceedings, issuing an decision on whether the tribunal has jurisdiction prior to ruling on the actual merits of the case. While arbitrators have the option of writing a dissenting opinion for an award, unanimous awards are the norm. Finally, while parties have choice over whom they appoint, all arbitrators must be, in theory, independent and free from conflicts of interest. The Article 14(1) of the ICSID Convention stipulates that all arbitrators must be "persons of high moral character and recognized competence in the fields of law, commerce, industry or finance, who may be relied upon to exercise independent judgment." Indeed, the Convention permits parties to propose the disqualification of arbitrators who show a "manifest lack," of the aforementioned characteristics – a provision that may constrain overt displays of favoritism for a party. Whether

[6] `https://icsid.worldbank.org/apps/ICSIDWEB/icsiddocs/Pages/ICSID-Convention-Arbitration-Rules.aspx`

[7] While there are some arbitrations involving solo arbitrators, the vast majority of ICSID tribunals are 3-person panels.

[8] While it is rare for the Chair to appoint a non-Presiding member, it has occurred in cases where either an arbitrator resigns or a party is non-responsive to the request for arbitration.

[9] This is not a requirement of party appointments.

such formal declarations truly constrain arbitrators remains unclear, particularly given the apparent contradiction between the incentives generated by the party appointment system and the idea of independent arbitrators (Paulsson, 2010).

2 What Affects Arbitrators' Decisionmaking?

Political scientists often see courts as strategic actors and judges as motivated not only by proper interpretation of texts and legal argument, but also by personal preferences over outcomes or strategic interests. Scholars of the United States Supreme Court point to preferences over policy as an important driver of how justices ultimately decide cases.Segal and Spaeth (2002). Indeed, a significant component of research on courts focuses primarily on unpacking the degree to which law or politics influences judicial decisionmaking.(Bailey and Maltzman, 2008, e.g.)

In the international context, scholars often focus on the degree to which judges are biased in favor of those who appointed them. Because judges have incentives to maintain their position, they have reason to not rule harshly against their principals. However, these incentives are also balanced against the judges' embededness in a broader legal community. Explicit biases in decisions are frowned upon and for professional reasons, judges must also tailor their behavior to professional norms. It is for this reason that many scholars have questioned the implications of the classic principal-agent model for judicial behavior. For example, Alter (2008) argues that international courts are not merely puppets of their constituent states and that international judges, by virtue of their position in professional and legal hierarchies can avoid recontracting threats from below. Voeten (2008) finds that judges in the European Court of Human Rights are not systematically more likely to rule in favor of their home countries. Rather, judges are policy-seekers with issue-specific preferences (in the case of the ECHR, these preferences relate to the application of European human rights law).

The behavior of judges and adjudicators in international investment dispute settlement is particularly likely to require a balance between pro-appointer bias and wider professional and social concerns because the investment law community is remarkably close-knit. Puig (2014) finds that the ICSID arbitration community is dominated by a handful of highly prominent and influential individuals. Arbitrators tend to be Europeans or Americans and a small minority of individuals receive most of the appointments. Indeed, Ginsburg (2003) argues that professional barriers to entry - notably the requirements of legal experience - keep the arbitration community very closed. Moreover, as arbitration becomes an increasingly popular dispute resolution mechanism, competition amongst legal actors (such as law firms) to define the rules of the game so to speak has intensified. Ginsburg argues that this has to some extent, resulted in a convergence of legal cultures within the space of international arbitration.

Nevertheless, anecdotal evidence from arbitrators suggests that arbitrators are uniquely conscious of the preferences of the party that appointed them. One particularly explicit example where a party's promise of extra-legal costs was acknowledged to have influenced an arbitrator comes from a story from Judge Abner Mikva who served as the United States' apointee on the *Loewen v. United States* arbitration described in Schneiderman (2010). The arbitration was one of the first to arise out of non-discrimination provisions contained in the North American Free Trade Agreement (NAFTA). Mikva recounted that after his appointment, he was told by United States Department of Justice (DOJ) officials that "You know, judge, if we lose this case, we could lose NAFTA." Mikva replied "Well, if you want to put pressure on me, then that does it." While Mikva's co-arbitrators were described as leaning in favor of the claimant, the case was ultimately decided in favor of the United States on jurisdictional grounds.

While the *Loewen* case illustrates a particularly egregious form of appointing-party influence, biases in favor of one's appointing party can arise out of purely career-oriented considerations. First, because arbitrators compete for re-appointment, parties can impose costs on arbitrators who fail to back their position by by refusing to support an arbitrator's appointment in future. This dynamic suggests repeat players are the ones most poised to impose costs, and since the respondents often face multiple disputes from single-shot claimant firms, it is likely that such costs are asymmetrically imposed on respondents' appointees versus claimants'. Second, the Loewen example suggests that policy-oriented judges may be concerned about additional compliance costs on the part of the respondent. If an adverse ruling against a claimant is unlikely to be complied with and may result in adverse policy outcomes, an otherwise indifferent arbitrator may err on the side of the claimant.

The presence of party-bias is one reason for symmetry in the appointment process – each side receives one appointee who will, to some extent, advocate for their side. Under these circumstances, however, the "pivotal" voter on the panel becomes the presiding member of the tribunal, who is appointed ostensibly by agreement of the parties. Precisely because this member is likely to be the "swing" vote on the tribunal, parties frequently fail to arrive on a mutually acceptable nominee, necessitating an institutional appointment. While the presiding member is at least nominally meant to be free from party-induced biases, there are other extra-legal factors that may sway his or her decision. Franck (2009) examines whether the presiding members' country of origin affects the outcome of a given decision, testing the hypothesis that arbitrators from lower-income countries are more sensitive to the demands of respondent countries (typically also developing countries) and are therefore more likely to support the respondent. While Franck (2009) does not find an effect on outcomes, the study does find some evidence that developing-country arbitrators are more likely to render smaller awards against developing countries, tailoring the damages even when liability is found. In addition, country-of-origin bias may go in the other direction, with arbitrators from advanced economies being more favorable to the interests of the multinationals who typically act as claimants.[10]

The evidence on country-of-origin biases has shown that such biases *do* exist in other international court contexts. Within the ICJ, Posner and de Figueiredo (2005) find that judges tend to render decisions that are more favorable to the states that appoint them and to states that have a similar level of economic development as their home state. Voeten (2007) shows, however, that in the context of the European Court of Human Rights, pro-state bias could be explained by an arbitrator's professional background. Former diplomats were more likely to favor states over petitioners compared to judges with professional legal or academic backgrounds. Voeten (2008) further shows that former diplomats were more likely to support their home governments, but that home-state bias was not universal among all ECtHR judges. In fact, the presence of effect heterogeneity between these two sets of studies is consistent with career background being a moderator of nationality bias. Judges in the ICJ are primarily drawn from the ranks of domestic civil service, and have spent a substantial amount of time either directly working for their home governments. Even among academics, those appointed to the ICJ frequently served as legal advisors for governments (Hernández, 2013, 138). As such, these judges have a strongly internalized sense of the preferences of their particular states and have career trajectories and futures that are very

[10] See, Franck (2009), pp. 452 for a more extended discussions of possible mechanisms for this affinity.

much connected with the civil service. Contrast this with judges who primarily operate in legal academia or in private practice, who face an entirely different set of reputational constraints and considerations.

This is the second element of arbitrator background that may explain bias in decision-making. Professional background may moderate the expression of overt biases by introducing an additional set of incentives particular to an arbitrator's career trajectory. Pauwelyn (2015) notes that investment arbitrators can be characterized as a highly specialized, elite club, where a small number of individuals amass a large share of the overall number of appointments. These "arbitration professionals" are drawn primarily from the ranks of private legal firms and legal academia, in contrast to panelists in another related institution, the WTO, who tend to have careers in government. While essentially all arbitrators have some sort of background with the law, there is clear variation in the overall career tracks of these arbitrators. These two career trajectories – private sector law/academia and government service characterize a highly salient split in the investment arbitration community. Costa (2011) finds that a little over a quarter of investment arbitrators have previously held positions in governmental institutions, a minority compared to the number of private law and academic arbitrators. The existence of such a split suggests that otherwise identical arbitrators may be motivated by different sets of incentives based on their career paths. For "elite" arbitrators, maintaining the goodwill of the arbitration community is essential to securing future appointments and retaining one's status. Arbitrators frequently adopt the role of both judge and litigator across many different tribunals. As Kapeliuk (2012) notes that because arbitrators operate in what could be considered "market" and must be re-appointed to each tribunal, they have incentives to not cultivate animosity among their peers who may be the ones choosing to appoint them in the future. While government officials called to arbitrate may have an interest in joining the arbitration club, they have more extensive career options outside of arbitration. On the converse, arbitrators that are more heavily invested in the persistence of the international investment arbitration regime may be more sensitive to overall reputational concerns, relative to arbitrators that have less attachment to the system. Ginsburg (2003) comments that "in arbitration, perhaps more than any other field of law, the line between scholar and practitioner is blurred so that many leading scholars are involved in arbitrations, and many leading arbitrators take the time to write academic articles and books." Noting the positive externalities generated by uniformity for the overall application and prestige of international arbitration, Ginsburg (2003) argues that arbitrators have strong incentives to maintain its credibility, reputation and uniformity as a way of sustaining its overall value and as such, may have a tendency to attenuate any biases in their decision-making.

Therefore, I hypothesize that bias by arbitrators from developed countries towards claimants and firms will be attenuated by their career background. Bias will be more pronounced among arbitrators with prior government experience as opposed to purely private-sector or academic arbitrators. Career and nationality have an interactive effect on win-rates. Building on Franck (2009), I therefore suggest that the absence of a detectable marginal effect is a result of effect heterogeneity in the population, rather than the absence of any bias. I test this hypothesis using data on ICSID arbitrations concluded prior to April, 2015. The next section discusses some of the challenges of this empirical strategy and explains why legal researchers need to account for factors that affect settlement when trying to estimate panel composition effects.

Estimating Treatment Effects under Post-treatment Attrition

If disputants are behaving strategically, then inferring the effect of panel composition on the outcome of a dispute becomes difficult due to the screening process of pre-awards settlement. Factors that affect the expected outcome of a dispute may also affect elements of pre-trial bargaining such that settlements become more or less likely.

This poses a challenge for causal inference. Intuitively, if treatment affects survival, then the types of units that are observed under treatment may be qualitatively different from those that are observed under control. Because all inferences about the outcome are implicitly conditioning on cases that fail to settle and ultimately go to "trial," the average treatment effect (ATE) does not exist and comparisons between different treatment arms conditional on the outcome being observed can give misleading conclusions regarding the actual effect of treatment. Even if treatment is randomly assigned, conditioning on survival (that is, outcome observability) can break that randomization if some types of subjects are more likely to survive.

The problem of treatment affecting which units have measurable outcomes has been studied

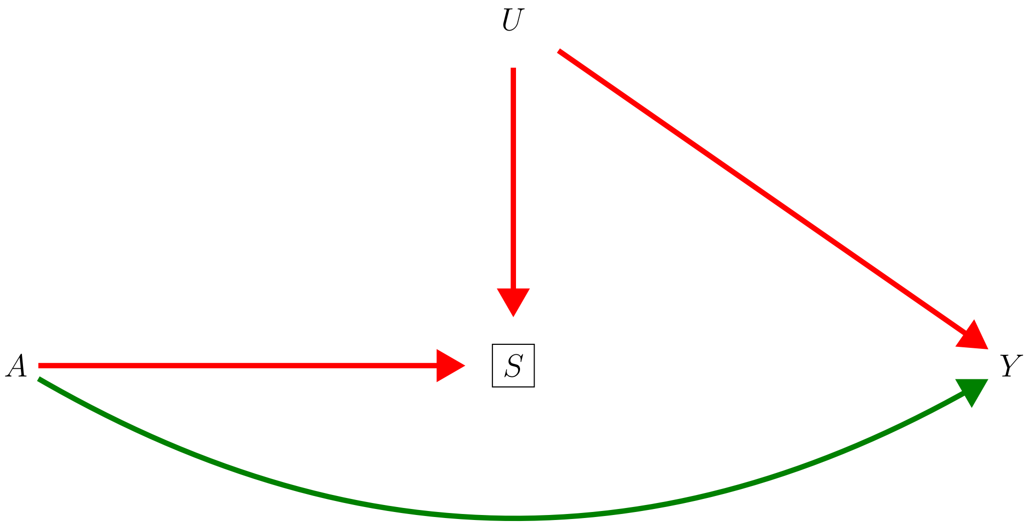


Figure 1: Directed Acyclic Graph illustrating Survival Bias.

in the statistics literature as "truncation-by-death" (Zhang and Rubin, 2003; Rubin, 2006; McConnell, Stuart and Devaney, 2008). The outcome of interest is undefined for subjects that "die." This problem often arises in medical studies with outcomes such as "quality of life" which cannot be measured for the deceased. However, this problem also appears in many social science settings – including the case of this paper. The party in whose favor an arbitration panel rules is undefined (in fact, it by definition does not exist) for cases that settle prior to a judgement. In non-legal political science contexts, one could envision an analogous truncation-by-death problem for studies of how congressional candidates' behavior in primary elections affects subsequent voting behavior – voting only exists for those candidates that win a general election.

This section discusses the necessary assumptions to estimate a valid causal quantity under attrition. I build on recent research in biostatistics to suggest a new causal quantity - the Average Survivor Controlled Effect which corresponds to a valid counterfactual comparison under a hypothetical intervention on attrition. This section will give an intuitive overview of the estimation procedure. For the full description of the method, see Appendix A.

The problem of analysing data where post-treatment attrition occurs is that it generates a form of selection bias when trying to estimate the effect of some treatment on an outcome. Consider a graph (Figure 1) describing a causal relationship between some treatment A and an outcome Y.

If treatment A also affects survival S, then naive analyses conditional on $S = 1$ will be biased due to a form of collider bias. Conditioning on S opens up an unblocked back-door path (the red path in the figure) into A (, through unobserved confounders U, violating the "adjustment criterion" (Shpitser and VanderWeele, 2011) for identifying the effect of an intervention on A holding constant S.

The intuitive reason for this is that when survival/settlement is affected by unobserved factors that also affect outcome, looking only at those cases where outcome was observed artificially induces a correlation between treatment and outcome. To use a medical analogy, if some treatment made individuals less likely to die, then paradoxically, the untreated individuals in a follow-up study would be healthier than those who received the treatment since treatment kept alive the less healthy individuals who were close to dying. We would conclude, incorrectly, that treatment made people less healthy. What happened though, was that treatment altered the underlying characteristics of the populations of the follow up study to make treatment vs. control incomparable. An analogous situation in the arbitration example for this paper would be if panels were more likely to settle when arbitrators had experience in government due to changes in beliefs over the probability of victory. This would mean that the cases that *didn't* settle were still too close to predict and therefore we would observe no change in voting rates even though the appointment had an effect on the underlying win rate.

Solving this problem is difficult and for the purposes of this paper, I can only rely on conventional "selection-on-observables" assumptions standard in observational causal inference designs. In order to estimate a valid treatment effect – what I term the "Average Survivor Controlled Effect" (ASCE) as it corresponds to the treatment effect under an additional intervention that guaranteed the case went to trial – I need to assume that not only are treatment and outcome unaffected by an unobserved common cause, *but also* that outcome and survival are unconfounded conditional on our covariates.

The second assumption can rule out a simple regression approach to estimation when some of the confounders of survival and outcome are post-treatment. One obvious such confounder is the time it takes for a dispute to resolve. Intuitively, the longer it takes a dispute to resolve, the more likely it is that the claimant will win (since the dispute is less likely to have been thrown out on a jurisdictional challenge if the parties have bifurcated proceedings). Estimation therefore uses a variation on the marginal structural model (MSM) inverse probability weighting

(IPW) approach described in VanderWeele (2009). While VanderWeele (2009) uses the MSM to estimate a controlled direct effect, an analogous method can be used to adjust for post-treatment confounders of survival and outcome.

I first estimate a regression model for the probability of survival using the full set of observations (cases where awards both were and were not rendered). This model includes treatment, post-treatment confounders and pre-treatment confounders. I then use this model to generate weights for each observation based on the inverse of the probability of observing that unit's "survival" status. I then run a regression of outcome on treatment and pre-treatment confounders that is weighted by the Inverse Probability of Survival Weights (IPSW) for the observations that did result in an award. Intuitively, this method up-weights cases that had a very low probability of resulting in an award (as we should expect more of them if cases were randomly decided) and down-weights cases with a very high probability of resulting in an award (as these are over-represented relative to others compared to if cases were randomly chosen). In practice, the models for the weights can be estimated using flexible semi-parametric regression methods such as generalized additive models in order to avoid making untestable linearity assumptions regarding the relationship between continuous covariates and treatment/survival. As VanderWeele (2009) notes, standard errors can be obtained using non-parametric bootstrap techniques.

Using this two-stage estimation technique I proceed to estimate the marginal and joint effects of arbitrator background and nationality on win rates.

Data and Methods

I obtained a dataset of all arbitration proceedings registered at the International Center for the Settlement of Investment Disputes (ICSID) between 1972 and April, 2015. For the purpose of this analysis, I focus on initial proceedings and not proceedings related to revisions or annulments of awards. The vast majority of these disputes involved a foreign investor (claimant) bringing a claim against a state (respondent). A handful of disputes involved two firms or a state bringing a claim against a firm, but these types of cases are very rare and because they are qualitatively distinct disputes are dropped from the dataset. Additionally, I remove all disputes that are still pending at the time of the collection of the dataset along with any tribunals composed of a solo arbitrator. Single arbitrator disputes fall outside of the scope of this paper as the theoretical argument depends on the dynamics of unilaterally-appointed arbitrators. However, these types of tribunals are relatively rare and the modal panel composition consists of two party-appointed arbitrators and a President appointed either by mutual agreement of the parties, selection by the co-arbitrators, or, if an agreement cannot be reached, the ICSID Administrative Council. I also exclude all disputes that were settled or discontinued prior to the constitution of a tribunal. Furthermore, because the quantity of interest depends on the president being the pivotal voter (between claimant and respondent appointees) I focus primarily on "competitive" tribunals where both claimant and respondent each appoint an arbitrator.[11] In total, the dataset consists of 261 arbitration proceedings, of which 180 had awards rendered. I also focus on estimating panel effects for the final panel composition of a dispute. Since arbitrators pass away or resign, some disputes see multiple assigned panels. While the question of "multi-shot" treatments (Blackwell, 2013) may be of interest, the data is too sparse in this particular case.

To code arbitrator backgrounds I used publicly available information, starting with arbitrators' online Curricula Vitae. If a CV could not be found (as is often the case for older arbitrators), I inferred background from articles written about the arbitrator or, in some cases, obituaries. I followed the definition in Costa (2011), coding an arbitrator as having a government background if they previously worked in an official capacity within a domestic executive, judicial or legislative branch (e.g. as a diplomat or a legislator). I exclude consultancy work with governments and focus only on formal/officially held positions in a government institution. Arbitrator nationality is directly obtained from the ICSID website. I operationalize the nationality variable as an advanced/developing economy binary. Arbitrators that are nationals of any country listed by the IMF's World Economic Outlook as an "advanced economy" (including dual-nationals) are coded as "advanced economy nationals."

For each dispute I also code the method by which each arbitrator was appointed - either by the claimant, respondent, by agreement of the parties, by agreement of the co-arbitrators, or by ICSID. This appointment data is obtained primarily from the texts of awards or other intermediate rulings of each panel. Where no texts are available, I look to secondary sources - in particular, the Investment Arbitration Reporter (IAReporter), a news service specializing in investor-state

[11] While infrequent, there are cases where respondents fail to appoint an arbitrator and the arbitrator is instead appointed by the institution.

arbitration.[12] I supplement this with additional data gathered on party appointments from Puig (2014). I am particularly interested in the binary indicator for whether the president is appointed by the institution or otherwise (by the parties' agreed-upon method).

The outcome – respondent victory – is coded as 0 if the final award finds that the respondent is liable for damages and awards a non-zero amount of awards to the claimant. Where the final award was kept confidential, the winner was inferred from secondary news reports from the Investment Arbitration Reporter. In nearly every case where final awards remained confidential, the general direction of the outcome could be obtained from reporting on the dispute. If parties agree to a settlement or allow the tribunal to lapse, this outcome is considered to be missing.

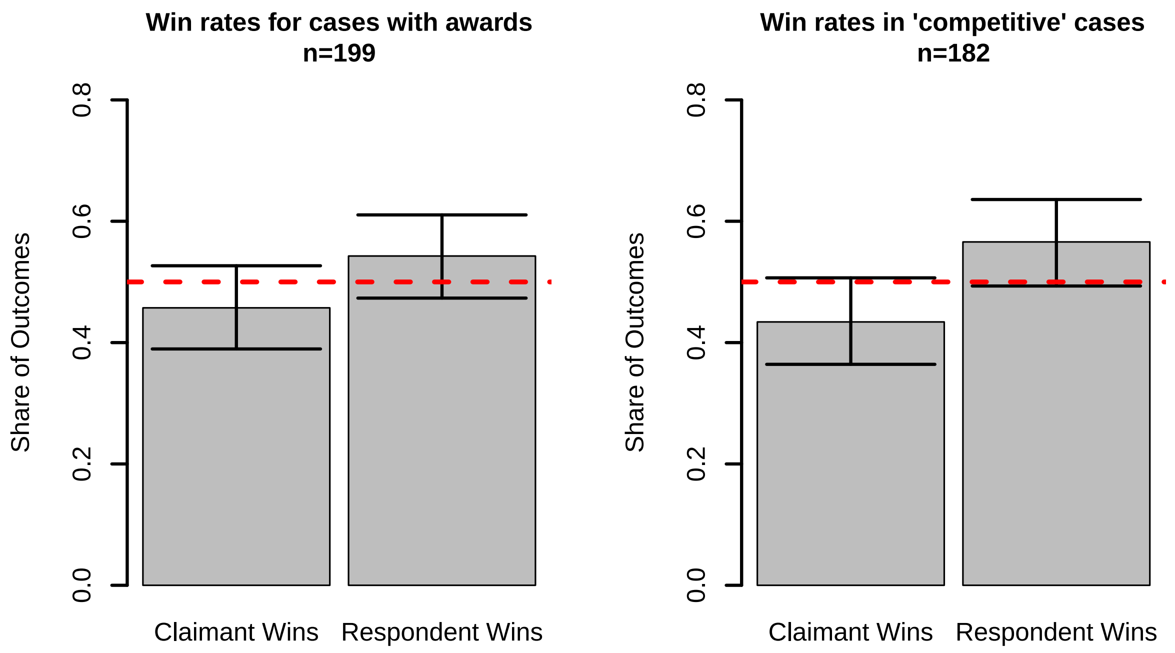
My set of pre-treatment covariates that affect either treatment/outcome or survival/outcome includes: legal basis (e.g. treaty or contract), whether a dispute is part of ICSID's Additional Facility, Economic sector of the dispute, respondent country income level, claimant origin country income level, the date of dispute registration, the length of time between registration and panel constitution, government background of claimant and respondent appointees, advanced economy nationality of claimant and respondent appointees, whether the president was appointed by the ICSID institution, and the number of past ICSID arbitration tribunals on which each arbitrator served (claimant's, respondent's and President). Post-treatment confounders of survival/outcome include time from panel constitution to outcome and the number of previous panels constituted.[13]

Most of the data on dispute characteristics, including dates, nationalities, and any intermediate outcomes, is obtained directly from ICSID's summary of each case.[14] For the legal basis variables, I code for whether the dispute arose out of a Bilateral Investment Treaty, a Contract between the firm and the host government, or the Energy Charter Treaty which regulates foreign direct investment related to electricity and comprises a sizeable number of ICSID cases. Economic sector is obtained from ICSID's own coding of the dispute and comprises 11 discrete categories: "Agriculture, Fishing & Forestry", "Construction", "Electric Power & Other Energy", "Finance", "Information & Communication", "Oil, Gas & Mining", "Other Industry", "Services & Trade", "Tourism", ""Transportation", and "Water, Sanitation & Flood Protection." Both respondent and claimant home country income are coded from the World Bank's 5-level classification of that

[12] http://www.iareporter.com/

[13] The latter is post-treatment since the president may have been appointed on an initial panel as opposed to a final panel.

[14] https://icsid.worldbank.org/apps/ICSIDWEB/cases/Pages/AdvancedSearch.aspx



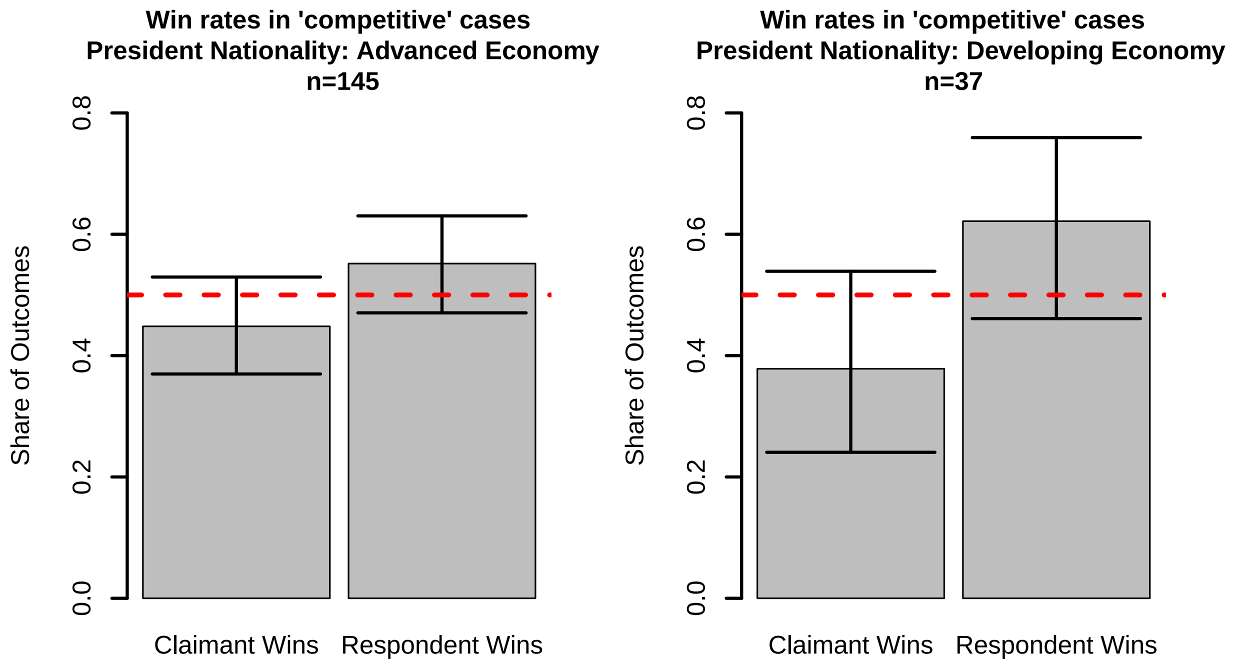
Lines denote 95% confidence intervals. Red horizontal line denotes the 50% win-rate threshold.

Figure 2: Baseline win rates in ICSID arbitrations

country in the year that the dispute was registered. The categories are: “High income: OECD”, “High income: non-OECD”, “Upper middle income”, “Lower middle income”, and “Low income.” Since many claimants are multi-nationals and may claim nationalities in different regions, I use the highest income level among claimant nationalities to estimate multinationals’ home country income.

For all treatment variables, to estimate the ASCE, I first fit a flexible generalized additive model of survival (award issued) on the pre- and post-treatment covariates. Continuous covariates enter the specification via an arbitrary smooth regression spline that is estimated from the data and makes no assumptions about the exact functional form relating the covariate to outcome. I then weight a linear probability model of respondent victory on treatment and pre-treatment covariates to obtain the estimated treatment effects. Standard errors are obtained using a non-parametric bootstrap procedure run for 1000 iterations.

Results



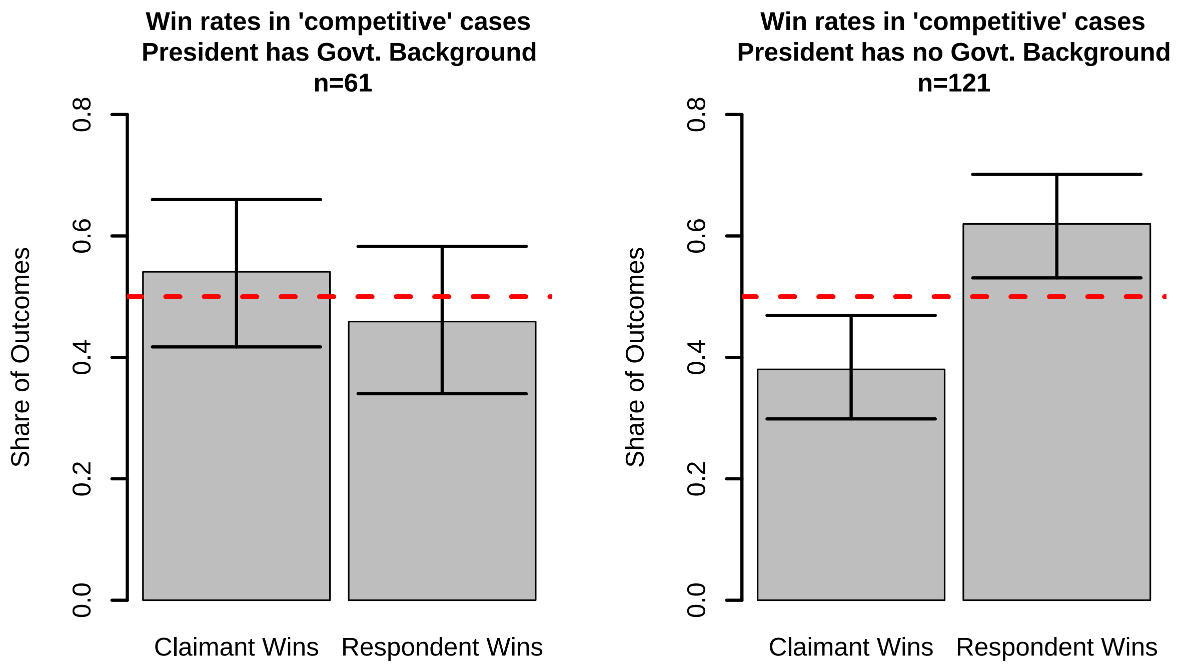
Lines denote 95% confidence intervals. Red horizontal line denotes the 50% win-rate threshold.

Figure 3: Win rates in ICSID arbitrations by presiding arbitrator nationality

Before going directly to the causal results, it is useful to simply look at the association between the treatment variables of interest and the outcome. Figure 2 plots the overall claimant and respondent win rates for all disputes in the dataset and all "competitive" disputes. As mentioned previously, the win-rate generally hovers close to 50%, with a slight edge towards the respondent, likely due to a selection effect due to truncation of the data at April, 2015 (disputes that finish early tend to be in the respondent's favor).

When looking at the marginal relationship between nationality and win rates (Figure 3, there does not appear to be much (if any) of an association between the two. While claimants appear to have a slightly higher win-rate among tribunal Presidents that are advanced economy nationals, the difference is well-attributable to sampling variation and random chance

However, plotting the outcome distribution by government/non-government background (Figure 4) reveals a potential effect. Among tribunals with presidents that have backgrounds working in government, claimants have a noticeably higher win rate. This difference becomes even more

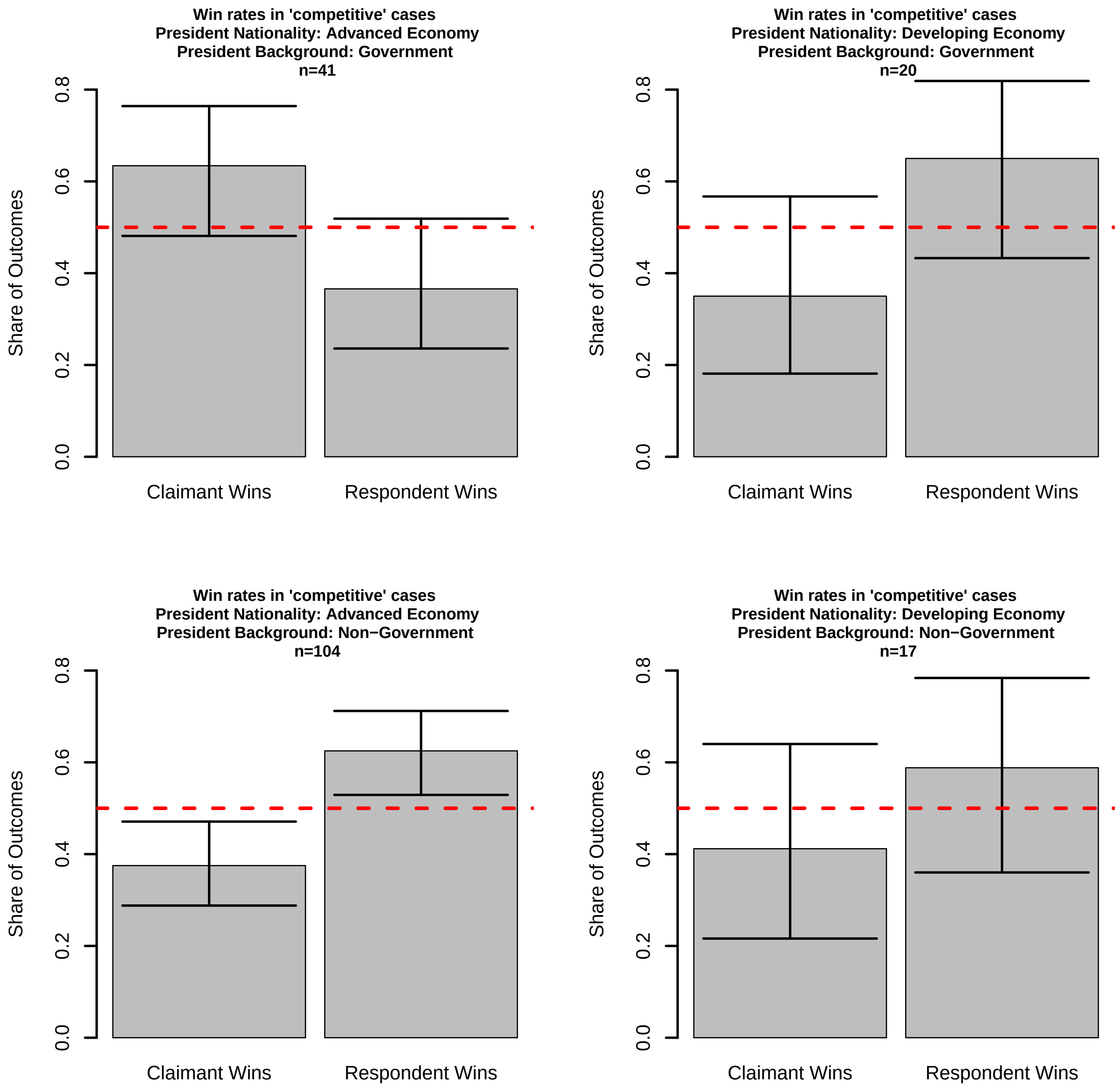


Lines denote 95% confidence intervals. Red horizontal line denotes the 50% win-rate threshold.

Figure 4: Win rates in ICSID arbitrations by presiding arbitrator background

striking when dividing the sample by both treatments (Figure 5). Out of the four possible combinations of nationality and government background, the only one where the claimant has a higher win rate than the respondent is the case of advanced economy presidents with a government background.

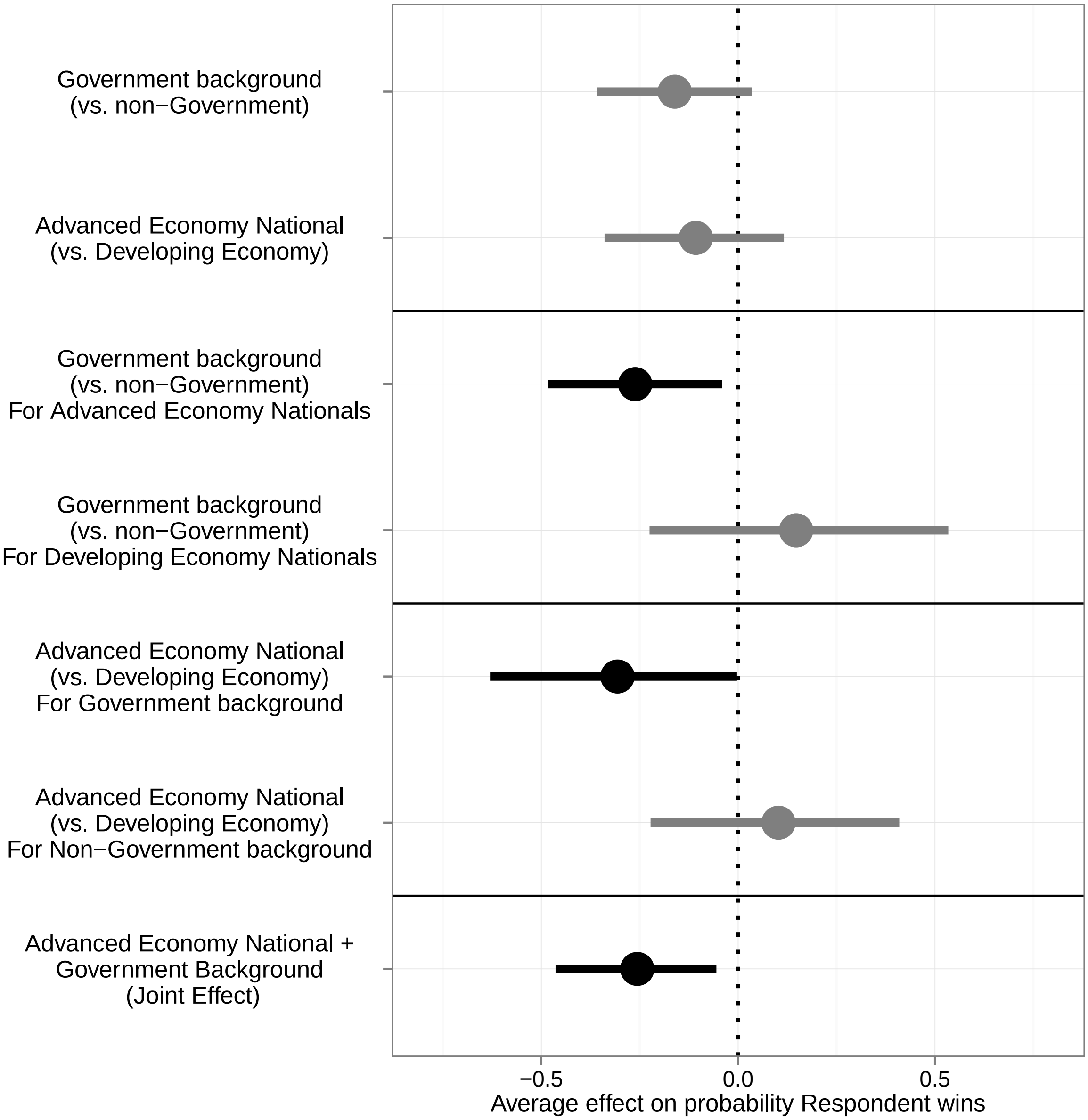
This effect persists after we further adjust for confounding variables. Figure 6 plots the estimated marginal effects of both of the treatments along with their combined interactive effect. While I fail to reject the null hypothesis of no-effect for the unconditional average treatment effects at the typical $p < .05$ threshold, I do find a statistically significant ($p < .05$) and negative effect of nationality *among presidents with government experience* on a respondent country's probability of winning. Conversely, government experience also has a negative effect on respondents' win probabilities *among nationals of advanced economies.* The average joint treatment effect is estimated at roughly 25 percentage points, meaning we can expect tribunals where an advanced economy government official is appointed president to be 25% more likely to render an award favorable to the claimant compared to all other possible tribunal presidents. Overall, the results strongly



Lines denote 95% confidence intervals. Red horizontal line denotes the 50% win-rate threshold.

Figure 5: Win rates in ICSID arbitrations by presiding arbitrator background and nationality

suggest that while nationality bias is certainly an issue within ICSID, its source does not come primarily from the private law community. Indeed, a greater reliance on elite legal professionals to constitute panels may have the positive effect of generating more favorable outcomes for states.



Lines denote 95% bootstrapped confidence intervals.

Figure 6: Average effects of arbitrator background and nationality on respondent win probability

Conclusion

Is Investor-State Dispute Settlement systematically biased against states? On its face, this question is unanswerable from looking at the outcome data alone. Claimants are strategic in when they choose to file disputes and both sides have incentives to reach a settlement when the outcome is evident. Win rates for both sides theoretically tend towards 50% irrespective of the legal standard.

A more tractable question is to examine individual rather than systemic sources of bias. How do different types of arbitrators decide cases? This paper examined the interaction of two elements of judicial identity that have been found to explain outcomes in other judicial contexts – nationality and government background. I found a strong *interactive* effect of these two variables on the relative success of claimants versus respondents in investor-state arbitration in ICSID. When tribunal presidents were nationals of advanced economies and had worked in government, claimants' win probabilities jumped significantly. Moreover, this effect is robust to the inclusion of a number of likely confounders of both outcome and treatment *and* settlement and treatment – the latter being a hidden source of bias in studies with post-treatment attrition. While any observational study relies heavily on the no-unobserved-confounding assumption to establish causality, the evidence presented here is certainly suggestive of a meaningful pattern.

These results have important limitations in that they are constrained to the set of publicly available and listed ICSID disputes. While ICSID disputes are somewhat representative of the overall universe of investment disputes they are just the tip of the international arbitration iceberg. Ad-hoc arbitrations not governed by ICSID rules tend to place more importance on the confidentiality of proceedings and in many cases, information about awards is simply unavailable. While disclosures of this information do happen from time to time, assembling a comprehensive and representative dataset of all arbitrations (as opposed to just ICSID ones), remains a challenging undertaking.

Nonetheless, these initial empirical results provide important information about which types of arbitration panels are more or less favorable to. This is not only useful information for practitioners, but it also point to a number of important policy implications for institutions seeking to improve the fairness and legitimacy of arbitration and minimize the impact of extra-legal biases on arbitrators' decisionmaking. First, it points to the value of expanding the overall pool of arbitrators beyond the small cohort of frequent and repeat arbitrators, to include more arbitrators from developing countries. Second, it points to the value of further professionalization of the arbitration community. As arbitration develops from a legal curiosity practiced by a handful of legal academics to a formal method of regulating and resolving disputes between states and firms over investments, the emergence of a set of professional norms governing arbitrator conduct may help to alleviate some sources of bias in how arbitrators make decisions.

Appendix A - More on the Average Survivor Controlled Effect

Let Y be the outcome of interest, A the treatment, and S an indicator variable denoting survival status. Y_i, A_i, and S_i denote the realized values of these variables for unit i. For simplicity assume binary treatment. Following the framework of Rubin (1974), we can write two sets of potential outcomes for each unit. $\{Y_i(1), Y_i(0)\}$ denote the potential outcome of Y for unit i under treatment ($A = 1$) and control ($A = 0$) respectively. Likewise $\{S_i(1), S_i(0)\}$ denote the potential survival outcome S for unit i under treatment and control.

Existing approaches to the problem of post-treatment attrition have focused on inferences within "principal strata" of the post-treatment variable. Frangakis and Rubin (2002) note that a post-treatment quantity S_i can be represented for each unit as a realization of a pre-treatment stratum defined by the joint potential outcomes for $S - \{S_i(1), S_i(0)\}$. They define principal strata causal effects (PSCEs) as the difference $Y_i(1) - Y_i(0)$ conditional on principal strata membership. That is, average PSCEs are treatment effects for the sub-population that would take on a common set of mediator values under treatment and control. In the case of binary treatments and where S denotes survival, there exist four principal strata: never survivors $\{S_i(1) = 0, S_i(0) = 0\}$, always survivors $\{S_i(1) = 1, S_i(0) = 1\}$, survivors under treatment $\{S_i(1) = 1, S_i(0) = 0\}$, and survivors under control $\{S_i(1) = 0, S_i(0) = 1\}$. Frangakis and Rubin (2002) and subsequent treatments of principal stratification for truncation-by-death consider estimation of average treatment effects for the "always survivor" strata – the Survivor Average Causal Effect (SACE) $E[Y_i(1) - Y_i(0)|S_i(1) = S_i(0) = 1]$ – as it is the only stratum for which both $Y_i(1)$ and $Y_i(0)$ are defined. That is, the average effect for units that would have survived regardless of treatment is the only well-defined causal estimand under the principal stratification framework for truncation-by-death.[15]

One challenge for identifying principal strata effects is that units' principal strata membership is only partially observed. For each unit, survival under the counterfactual treatment condition is unobserved. We do not know which units that are observed to have survived are part of the "always survival" stratum (that is, they would have also survived under the other treatment). As

[15] Other principal strata estimands can be targets in different settings. For example, in the partial compliance setting, instrumental variable designs identify a local average treatment effect (LATE) for the "complier" principal stratum (Angrist, Imbens and Rubin, 1996).

a consequence, VanderWeele (2011) notes that the SACE is often not point identified and inference strategies have to rely on bounds, strong assumptions regarding regarding possible strata coupled with sensitivity analyses for unknown parameters, or Bayesian approaches for inferring strata membership.

A common assumption made in the PS literature is that of "monotonicity" for the effect of treatment on survival. It is assumed that if a unit survived under control, then it would also survive under treatment – $S_i(1) \geq S_i(0)$. The assumption states a priori that treatment does not cause any individuals to not survive if they would survive under control and rules out the $\{S_i(1) = 0, S_i(0) = 1\}$ stratum. Under monotonicity, Chiba and VanderWeele (2011) show that the SACE is identified using the raw difference in means among treatment arms plus a user-specified sensitivity parameter. However, while monotonicity might be feasible in a medical context where the properties of a particular treatment drug are well-studied and known at an individual level, it is not feasible in most social science contexts, including the case of pre-trial settlement. We have strong reasons to suspect that effects on settlement are heterogeneous across units. As illustratedi n the toy model presented in the previous section, settlement probability decreases as the claimant's probability of victory approaches 0.5 but increases as the probability moves away from 0.5 and towards 1. If a hypothetical treatment increases claimant's win probability by .1, then the effect on settlement will be negative if the win probability changes from .4 to .5 and positive if the win probability changes by .5 to .6. Monotonicity is not a feasible assumption in this context.

In the absence of monotonicity assumptions, identification of principal strata effects becomes much more complex and infeasible in this particular context. One of the major limitations of the principal strata framework is its unwillingness to consider joint counterfactuals in which intervention occurs on both treatment and mediator - that is, $Y(a, s)$. In a discussion of the utility of PS for assessing mediation effects, Pearl (2011) comments that the implicit prohibition within principal stratification against counterfactual interventions on mediators is an unwarranted limitation on causal inquiry. Just as the principal strata framework forecloses consideration of "direct effects" (effects with pathways "deactiviated") in the context of mediation, it also limits consideration of interesting counterfactual quantities in the context of truncation by death. Joffe (2011) makes a similar criticism in that focus on the "always survival" stratum obscures other potentially interesting effects within other subsets of the population. An overly-strict interpretation of valid causal effects necessitating human-level interventions would invalidate nearly every observational study where direct experimenter intervention is replaced by assumptions regarding what Pearl (2011) calls "reasonable assumptions about how treatment variables are naturally chosen." Insofar as observational researchers are willing to accept "interventions" in terms of what nature assigns, then causal effects involving interventions on mediating variables – analogous to the direct and indirect effects – are valid in the truncation-by-death context.

The definition of potential outcomes for Y can be expanded to include $Y_i(a, s)$ which denotes the potential outcome for a unit i if treatment is set to a and survival status is set to s. Note that in the case of truncation by death, $Y_i(a, 0)$ is undefined, but $Y_i(a, 1)$ is defined for all units. To connect $Y_i(a)$ to $Y_i(a, s)$, I make the following assumption known as "composition" which allows the writing of total effects in terms of joint counterfactuals (VanderWeele and Vansteelandt, 2009).

Assumption 1. *Composition*

$$Y_i(a) = Y_i(a, S_i(a))$$

That is, the potential outcome for unit i when assigned treatment value a is equal to the potential outcome for i when it is assigned treatment value a and survival is set to the value that it would naturally take under treatment a ($S_i(a)$).

Under Assumption 1, we can write the Survivor Average Causal Effect for the survivor stratum as

Definition 1. *Survivor Average Causal Effect*

$$\begin{aligned} SACE &= E\left[Y_i(1) - Y_i(0) | S_i(1) = S_i(0) = 1\right] \\ &= E\left[Y_i(1, S_i(1)) - Y_i(0, S_i(0)) | S_i(1) = S_i(0) = 1\right] \end{aligned}$$

Again, because $Y_i(a, 0)$ is undefined, the SACE is only defined for units that have $S_i(1) = S_i(0) = 1$ – the survivor principal stratum. Using the joint counterfactual formulation of causal effects, however, allows for the consideration of different causal estimands that are defined not just for the survivor stratum.

I define the Average Survivor Controlled Effect (ASCE) as the expected difference between potential outcomes under treatment and control, *fixing* the survival intermediate to 1. Formally:

Definition 2. *Average Survivor Controlled Effect*

$$ASCE = E\left[Y_i(1,1) - Y_i(0,1)\right]$$

In contrast to the SACE, the ASCE is defined over the entire population of units. In addition, under Assumption 1, the ASCE and SACE are equivalent for the survivor principal stratum.

Proposition 1. *Equivalence of stratum-conditional ASCE and SACE*

Conditional on $S_i(1) = S_i(0) = 1$, the ASCE is equal to the SACE under assumption 1.

Proof. Start by writing the ASCE conditional on the survivor stratum.

$$E\left[Y_i(1,1) - Y_i(0,1) | S_i(1) = S_i(0) = 1\right]$$

Substituting what we know from the conditioning

$$= E\left[Y_i(1, S_i(1)) - Y_i(0, S_i(0)) | S_i(1) = S_i(0) = 1\right]$$

From assumption 1

$$= E\left[Y_i(1) - Y_i(0) | S_i(1) = S_i(0) = 1\right]$$

□

This suggests that under an additional no-interaction assumption, the SACE and ASCE are equivalent. Specifically,

Assumption 2. *No Principal Strata-ASCE Interaction*

$$E\left[Y_i(1,1) - Y_i(0,1) | S_i(1) = 1, S_i(0) = 1\right] = E\left[Y_i(1,1) - Y_i(0,1) | S_i(1) = s_1, S_i(0) = s_2\right] \forall s_1, s_2$$

Proposition 2. *ASCE - SACE equivalence under no-interaction*

Under Assumptions 1 and 2, the ASCE and SACE are equivalent.

Proof. By Law of Total Probability

$$E[Y_i(1,1) - Y_i(0,1)] = \sum_{\forall s_1, s_2} E\left[Y_i(1,1) - Y_i(0,1) | S_i(1) = s_1, S_i(0) = s_2\right] Pr(S_i(1) = s_1, S_i(0) = s_2)$$

From Assumption 2

$$E[Y_i(1,1) - Y_i(0,1)] = \sum_{\forall s_1, s_2} E\left[Y_i(1,1) - Y_i(0,1) | S_i(1) = S_i(0) = 1\right] Pr(S_i(1) = s_1, S_i(0) = s_2)$$

From Proposition 1

$$\begin{aligned} E[Y_i(1,1) - Y_i(0,1)] &= \sum_{\forall s_1, s_2} E\left[Y_i(1) - Y_i(1) | S_i(1) = S_i(0) = 1\right] Pr(S_i(1) = s_1, S_i(0) = s_2) \\ &= E\left[Y_i(1) - Y_i(1) | S_i(1) = S_i(0) = 1\right] \sum_{\forall s_1, s_2} Pr(S_i(1) = s_1, S_i(0) = s_2) \\ &= E\left[Y_i(1) - Y_i(1) | S_i(1) = S_i(0) = 1\right] \end{aligned}$$

□

Notably, the ASCE also appears analogous to a controlled direct effect fixing survival status at 1 (Pearl, 2001). Indeed, it can be interpreted as the controlled direct effect of treatment on outcome holding fixed survival. This type of effect – called "death blocking" was suggested by Joffe (2011) as a potential alternative to principal strata effects for truncation by death but was dismissed due to concerns that the counterfactuals involved are overly vague and implausible (e.g. it is unclear how units will be "forced" to survive and whether such an intervention would alter the potential outcomes). [16]. However, to some extent, concerns about the feasibility or even existence of an intervention on the mediator do not entirely diminish the usefulness of quantities like the controlled direct effect.

Even if it is difficult or impossible to conceive of a "non-invasive" intervention on survival (interventions that would fix S to 1 but nonetheless leave the outcome distributions unchanged), VanderWeele and Vansteelandt (2009) notes that controlled direct effects remain a theoretical

[16] Chaix et al. (2012) make similar critiques in response to a paper (Weuve et al. (2012)) that employs a method very similar to the estimator I describe but does not explicitly define the quantity being estimated as distinct from principal strata effects.

quantity of interest for mediation analysis. The CDE can be thought of as a way of ruling out alternative causal pathways – that is, it allows researchers to answer the question of whether there exists an effect if an entire causal pathway were deactivated. In a recent treatment of controlled direct effects for a political science audience Acharya, Blackwell and Sen (2015) note that if there exists a non-zero ACDE, then there must be some effect of treatment on outcome that operates outside of the mediator being controlled. A similar justification can be made for the ASCE in the context of truncation-by-death. Testing for whether the ASCE is non-zero permits researchers to answer the question of whether an observed relationship between treatment and outcome is driven *entirely by selection of which units have observable outcomes*. Overall, such issues of manipulability are, as VanderWeele and Vansteelandt (2009) notes, not unique to the question of interventions on mediators. For many observational studies, particularly in the social sciences, variables such as country development level or gender are ones for which counterfactual interventions are ill-defined. Yet if researchers are willing to accept that the effects of these variables are interesting in observational studies and that counterfactual thought-experiments provide a useful way of conceptualizing causality, then analogous interventions on mediators may be of theoretical interest even if an *actual* intervention is practically infeasable. If the thought experiment "what would be the treatment effect if all units in sample survived" is a useful one for researchers, then the ASCE is a useful estimand.

Identification and estimation assumptions follow directly from existing treatments of controlled direct effects. To begin, we make a standard consistency assumption (VanderWeele and Vansteelandt, 2009).

Assumption 3. *Consistency*

$$Y_i(a, 1) = Y_i | A_i = a, S_i = 1$$
$$S_i(a) = S_i | A_i = a$$

The consistency assumption states that the observed outcome Y_i for units with treatment level a and survival status 1 is equal to the joint potential outcome $Y_i(a, 1)$. Likewise, the observed survival status S_i for a unit with treatment level a is assumed to be equal to the potential outcome $S_i(a)$. Note the slight difference with the consistency assumption for controlled direct effects as

$Y_i(a, 0)$ is undefined.

Next, it is necessary to make two unconfoundedness assumptions conditional on covariates (VanderWeele, 2009). First,

Assumption 4. *Treatment-Outcome Unconfoundedness*

$$Y_i(a, 1) \perp\!\!\!\perp A|X$$

where X is the set of observed covariates that affect both treatment and outcome. Assumption 4 can be understood as a no-omitted-variables assumption for the relationship between the treatment and the outcome. Additionally, we need

Assumption 5. *Survival-Outcome Unconfoundedness*

$$Y_i(a, 1) \perp\!\!\!\perp S|A, X, L$$

where L is another set of variables that confound the survival-outcome relationship but which do not confound treatment and outcome. Crucially, variables in L can be post-treatment, which slightly complicates estimation as these variables cannot be conditioned on (the standard method for covariate adjustment).

Finally, we make a positivity assumption

Assumption 6. *Positivity*

For all treatment values a, survival levels s, covariate values x and l

$$Pr(A_i = a|X_i = x) > 0$$
$$Pr(S_i = s|A_i = a, X_i = x, Z_i = z)$$

This essentially amounts to a covariate overlap assumption in the distributions under treatment/control and survival/non-survival.

In the presence of post-treatment confounders, conventional regression estimators will be unable to consistently estimate the ASCE. This is because conditioning on a post-treatment variable blocks part of the pathway through which the treatment effect is transferred to the outcome. VanderWeele (2009) applies the technique of estimating Marginal Structural Models (MSMs) weighted by inverse probability weights (IPWs) developed by Robins, Hernan and Brumback (2000) to the task of estimating controlled direct effects. An analogous estimator is consistent for the ASCE.

Marginal Structural Models are models for the expectation of a counterfactual (Robins, Hernan and Brumback, 2000). The MSM of interest for the ASCE case is[17]

$$E[Y(a,1)] = \alpha_0 + \alpha_1 a$$

With binary treatment, we define two sets of weights for each observation

$$w_i^A = \frac{P(A = a_i)}{P(A = a_i | X = x_i)}$$

and

$$w_i^S = \frac{P(S = s_i | A = a_i)}{P(S = s_i | A = a_i, X = x_i, W = w_i)}$$

The denominators of the weights are, respectively, the probability that the unit received the treatment it did given pre-treatment covariates and the probability that the unit received the survival level it did given pre- and post-treatment covariates. The numerators are "stabilizing" probabilities that give less variable weights when probabilities in the denominator are small as recommended by Robins, Hernan and Brumback (2000).

Proposition 3. *Identifiability of ASCE*

Under Assumptions 1 - 6 plus an additional assumption that the weight models are correctly specified, a weighted regression of Y on A where each observation i is weighted by $w_i^A \times w_i^S$ is consistent for the parameters of the Marginal Structural Model

$$E[Y(a,1)] = \alpha_0 + \alpha_1 a$$

Proof. See Robins, Hernan and Brumback (2000), cited in VanderWeele (2009). □

[17] Note the slight difference with the MSM in VanderWeele (2009) which includes an interaction between a and s. Since $Y(a,0)$ is undefined, so is this interaction. Additionally, an s term is absent as it is unidentified along with the intercept (again, no variation in s).

It follows that the estimated coefficient on the treatment variable A in the weighted regression model is consistent for α_1 which is equal to the ASCE – $E[Y(1,1)] - E[Y(0,1)] = \alpha_0 + \alpha_1 - \alpha_0 = \alpha_1$.

It is worth noting that Weuve et al. (2012) and Tchetgen Tchetgen et al. (2012) propose a similar estimation strategy to account for attrition by death. However, these papers do not make clear the exact quantity that the IPSW (inverse probability of survival weighting) estimator estimates. The analysis here clarifies that, contrary to what is argued in Tchetgen Tchetgen et al. (2012), the IPSW estimator does *not* estimate a principal strata effect absent strong non-interaction assumptions. Rather, given the above assumptions which are analogous to those made by Weuve et al. (2012), the identified effect is what I call the ASCE – a treatment effect fixing survival (by some unspecified means) for all units in sample. Only if additional consistency and composition assumptions are made along with an assumption of no treatment effect heterogeneity across principal strata does the ASCE equate to the survivor causal effect within the always survivor principal stratum. However, even in the absence of some of these assumptions, the ASCE remains a valid quantity of interest for researchers interested in ascertaining whether an observed relationship in data denotes a meaningful effect or purely an artifact of post-treatment selection.

References

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

Alter, Karen. 2014. *The New Terrain of International Law: Courts, Politics, Rights*. Princeton University Press.

Alter, Karen J. 2008. "Agents or trustees? International courts in their political context." *European Journal of International Relations* 14(1):33–63.

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.

Bailey, Michael A and Forrest Maltzman. 2008. "Does legal doctrine matter? Unpacking law and policy preferences on the US Supreme Court." *American Political Science Review* 102(03):369–384.

Blackwell, Matthew. 2013. "A framework for dynamic causal inference in political science." *American Journal of Political Science* 57(2):504–520.

Chaix, Basile, David Evans, Juan Merlo and Etsuji Suzuki. 2012. "Commentary: weighing up the dead and missing: reflections on inverse-probability weighting and principal stratification to address truncation by death." *Epidemiology* 23(1):129–131.

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.

Costa, Jose Augusto Fontoura. 2011. "Comparing WTO Panelists and ICSID Arbitrators: The Creation of International Legal Fields." *Oati Socio-Legal Series* 1(4).

Danziger, Shai, Jonathan Levav and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences* 108(17):6889–6892.

Drahozal, Christopher R. 2009. "Private ordering and international commercial arbitration." *Penn State Law Review* 113(4).

Egger, Peter and Michael Pfaffermayr. 2004. "The impact of bilateral investment treaties on foreign direct investment." *Journal of comparative economics* 32(4):788–804.

Elkins, Zachary, Andrew T Guzman and Beth A Simmons. 2006. "Competing for capital: The diffusion of bilateral investment treaties, 1960–2000." *International Organization* 60(04):811–846.

Farhang, Sean and Gregory Wawro. 2004. "Institutional dynamics on the US court of appeals: Minority representation under panel decision making." *Journal of Law, Economics, and Organization* 20(2):299–330.

Franck, Susan D. 2007. "Empirically evaluating claims about investment treaty arbitration." *North Carolina Law Review* 86:1.

Franck, Susan D. 2009. "Development and Outcomes of Investment Treaty Arbitration." *Harvard International Law Journal* 50(2).

Franck, Susan D. 2010. "ICSID Effect-Considering Potential Variations in Arbitration Awards, The." *Va. J. Int'l L.* 51:825.

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

Frieden, Jeffry A. 1994. "International investment and colonial control: A new interpretation." *International Organization* 48(04):559–593.

Gaillard, Emmanuel. 2015. "Sociology of international arbitration." *Arbitration Journal* 31:1 – 17.

Ginsburg, Tom. 2003. "Culture of Arbitration, The." *Vand. J. Transnat'l L.* 36:1335.

Giorgetti, Chiara. 2013. "Who Decides Who Decides in International Investment Arbitration?" *University of Pennsylvania Journal of International Law* 35(431).

Glynn, Adam N and Maya Sen. 2015. "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?" *American Journal of Political Science* 59(1):37–54.

Gonzalez de Cossio, Francisco. 2008. "Mexico before ICSID: Rebel without a Cause." *J. World Investment & Trade* 9:371.

Hernández, Gleider. 2013. *The International Court of Justice and the judicial function.* Oxford University Press.

Joffe, Marshall. 2011. "Principal stratification and attribution prohibition: Good ideas taken too far." *The international journal of biostatistics* 7(1):1–22.

Kapeliuk, Daphna. 2012. "Collegial Games: Analyzing the Effect of Panel Composition on Outcome in Investment Arbitration." *Rev. Litig.* 31:267.

Kastellec, Jonathan P. 2013. "Racial diversity and judicial influence on appellate courts." *American Journal of Political Science* 57(1):167–183.

McConnell, Sheena, Elizabeth A Stuart and Barbara Devaney. 2008. "The truncation-by-death problem what to do in an experimental evaluation when the outcome is not always defined." *Evaluation Review* 32(2):157–186.

Paulsson, Jan. 2010. "Moral Hazard in International Dispute Resolution." *ICSID Review* 25(2):339–355.

Pauwelyn, Joost. 2015. "The Rule of Law Without the Rule of Lawyers? Why Investment Arbitrators are From Mars, Trade Panelists are From Venus." *CTEI Working Paper* .

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc. pp. 411–420.

Pearl, Judea. 2011. "Principal stratification – a goal or a tool?" *The International Journal of Biostatistics* 7(1).

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.

Priest, George L and Benjamin Klein. 1984. "The selection of disputes for litigation." *The Journal of Legal Studies* pp. 1–55.

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

Reed, Lucy, Jan Paulsson and Nigel Blackaby. 2011. *Guide to ICSID Arbitration: Version 2*. Kluwer Law International.

Robins, James M, Miguel Angel Hernan and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11(5):550–560.

Rogers, Catherine A. 2004. "Vocation of the International Arbitrator, The." *Am. U. Int'l L. Rev.* 20:957.

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

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.

Sabahi, Borzu and Kabir Duggal. 2013. "Occidental Petroleum v Ecuador (2012) Observations on Proportionality, Assessment of Damages and Contributory Fault." *ICSID Review* .

Schneiderman, David. 2010. "Judicial Politics and International Investment Arbitration: Seeking an Explanation for Conflicting Outcomes." *Nw. J. Int'l L. & Bus.* 30:383.

Segal, Jeffrey A and Harold J Spaeth. 2002. *The Supreme Court and the attitudinal model revisited*. Cambridge University Press.

Shavell, Steven. 1996. "Any frequency of plaintiff victory at trial is possible." *J. Legal Stud.* 25:493.

Shpitser, Ilya and Tyler J VanderWeele. 2011. "A complete graphical criterion for the adjustment formula in mediation analysis." *The international journal of biostatistics* 7(1):1–24.

Simmons, Beth A. 2014. "Bargaining over BITs, arbitrating awards: The regime for protection and promotion of international investment." *World Politics* 66(01):12–46.

Tchetgen Tchetgen, Eric J, M Maria Glymour, Ilya Shpitser and Jennifer Weuve. 2012. "Rejoinder: To Weight or Not to Weight?: On the Relation Between Inverse-probability Weighting and Principal Stratification for Truncation by Death." *Epidemiology* 23(1):132–137.

Trakman, Leon E. 2013. “ICSID Under Siege, The.” *Cornell Int’l LJ* 45:603.

Van Harten, Gus. 2016. “Arbitrator Behaviour in Asymmetrical Adjudication (Part Two): An Examination of Hypotheses of Bias in Investment Treaty Arbitration.” *Osgoode Hall Law Journal* 53.

VanderWeele, Tyler J. 2009. “Marginal structural models for the estimation of direct and indirect effects.” *Epidemiology* 20(1):18–26.

VanderWeele, Tyler J. 2011. “Principal Stratification–Uses and Limitations.” *The international journal of biostatistics* 7(1):1–14.

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

Vandevelde, Kenneth J. 2005. “A brief history of international investment agreements.” *UC Davis Journal of International Law & Policy* 12(1):157.

Voeten, Erik. 2007. “The politics of international judicial appointments: evidence from the European Court of Human Rights.” *International Organization* 61(04):669–701.

Voeten, Erik. 2008. “The impartiality of international judges: Evidence from the European Court of Human Rights.” *American Political Science Review* 102(04):417–433.

Weuve, Jennifer, Eric J Tchetgen Tchetgen, M Maria Glymour, Todd L Beck, Neelum T Aggarwal, Robert S Wilson, Denis A Evans and Carlos F Mendes de Leon. 2012. “Accounting for bias due to selective attrition: the example of smoking and cognitive decline.” *Epidemiology (Cambridge, Mass.)* 23(1):119.

Wrange, Pal. 2012. “Sedelmayer v. Russian Federation.” *Am. J. Int’l L.* 106:347.

Yackee, Jason Webb. 2008. “Bilateral Investment Treaties, Credible Commitment, and the Rule of (International) Law: Do BITs Promote Foreign Direct Investment?” *Law & Society Review* 42(4):805–832.

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.