Unpacking Treaty Practice: The Differential Informative Power of Human Rights Monitoring Mechanisms

Sinh Nguyen*

December 15, 2015

Abstract

Monitoring mechanisms under the United Nations (UN) core human rights treaties include state reporting, inter-state communications, individual communications, inquiries, and country visits. The existing literature focuses almost exclusively on treaty effects at the aggregate level whereas, in practice, multiple human rights monitoring processes are often at work. This paper disaggregates treaty practice into simultaneous operation of multiple monitoring mechanisms and argues that variation in the legal design of monitoring mechanisms partially determines how effective they are in terms of changing state behavior. It then tests this argument by estimating the differential causal effects of the inquiries mechanism under Article 20 of the Convention against Torture (CAT), the individual communications mechanism under Article 22, and the country visits mechanism under the Optional Protocol to the Convention against Torture (OPCAT). To accomplish that, it, first, construct a structural model of the data generating process and formulates the causal quantities of interests; then, generates directed acyclic graphs to assess causal effect identifiability via the backdoor criterion; and finally, applies the targeted maximum likelihood estimation to produce effect estimates. The result indicates that the country visits mechanism under the OPCAT has a clear positive causal impact on human rights protection, but other less intrusive monitoring mechanisms do not. This finding suggests that international human rights institutions should become more legalized and more intrusive to have a significant causal impact on state behavior.

*Ph.D. student, Department of Political Science, Purdue University. Email: nguyens@purdue.edu

Introduction

In 1995 the Human Rights Committee (HRC), the treaty body under the International Covenant for Civil and Political Rights (ICCPR), in the case of *Arhuaco* v. *Colombia* (HRC, Communication No. 612/1995) found the Colombian government responsible for the arbitrary detention of Jose Vicente and Amando Vincente. As a result of the HRC's views in response to a submitted individual communication, the government under its own Law 288/96 rendered a favorable opinion to compliance and later let the case proceed to national courts (Ulfstein and Keller 2012, 365). Similarly, in 1998 the CAT Committee, the treaty body under the Convention against Torture (CAT), issued its decision in the case of *Ristic* v. *Yugoslavia* (CAT Committee, Communication No. 113/1998), finding that the government had violated its obligations under the CAT and ordering a remedy in the form of effective investigation and publication of the decision. Even though the CAT Committee's decision was not legally binding in terms of domestic enforcement, the country's Supreme Court nonetheless endorsed the CAT Committee's decision and ordered reparations for the victims (Ulfstein and Keller 2012, 369–370). These are just two examples that highlight the potential effect of optional monitoring mechanisms that go beyond the state reporting mechanisms under the United Nations (UN) human rights treaties. Whether this kind of effect is causal in nature and generalizable across international treaties is a key issue that has not been subject to any systematic investigation in the existing literature.

Since its establishment many decades ago, the UN human rights treaty system has proved crucial to the enterprise of international human rights protection. An expansive body of literature has examined the overall impact of particular human rights conventions (Hathaway 2002; Landman 2005; Neumayer 2005; Hafner-Burton and Tsutsui 2007; Simmons 2009; Hill 2010; Lupu 2013; Clark 2013). Yet, there have been very few systematic inquiries into the effectiveness of individual monitoring mechanisms under those conventions, including state reporting, state communications, individual communications, inquiries, and country visits. This paper takes on part of that challenge in the hopes of yielding more insights into the relationship between institutional design and empirical effectiveness of UN core human rights treaties at a more micro-level.

The motivating idea is to examine treaty effect from a new level of analysis by conceptualizing human rights treaty ratification not as a binary variable as in the existing literature, but

rather a varying collection of treatments. Each treatment would then correspond to ratification of one type of monitoring mechanisms. In the international law literature, Tyagi (2009, 127) has also proposed a similar conceptualization in the case of the ICCPR. I extend his suggestion to incorporate two other types of monitoring mechanisms, albeit under a different UN convention. This conceptualization is more in line with the actual practice of the UN human rights treaty system where multiple procedures are often employed simultaneously to monitor and hopefully improve human rights practices of member states. In addition to making a theoretical proposition, the key empirical question I seek to answer is, what are the disaggregated causal effects of the inquiries mechanism under Art. 20 of the CAT, the individual communications mechanism under Art. 22, and the country visits mechanism under the Optional Protocol to the CAT (OPCAT)? Thus, the *outcome* in this observational study is cross-national level of human rights protection. The *treatment* is state ratification of Art. 20 and Art. 22 of the CAT and the OPCAT. I use ratification as a generic term for a state's voluntary action to commit, or refrain from making a reservation, to an optional monitoring mechanism.

Monitoring mechanisms not covered in this study include state reporting, which is universally required in all human rights treaties, and state communications, which, unfortunately, have never been used. It is clear that this is only a first step to study the disaggregated effects of human rights monitoring mechanisms, focusing on three optional mechanisms to monitor the prevention of torture. I choose to study the effects of monitoring mechanisms under the CAT because it is one of the most prominent and well studied UN human rights conventions that deals with a crucial and non-derogable human right. As such, if institutional design, as I would argue, has any differential causal impact on human rights outcome, that should be reflected most clearly in the mechanisms that monitor the prevention of torture. While I expect similar patterns to hold in other human rights treaties, it remains a tentative belief pending on credible empirical evidence in future studies.

Table I summarizes the monitoring mechanisms available across nine UN core human rights treaties. Appendix A also provides the list of UN core human rights treaties and their current status of ratification. State reporting refers to the mechanism by which a state party submits self-reports on the measures it has taken and the progress made to fulfill its treaty obligations. Other monitoring mechanisms are optional and do not impose obligations on states parties unless

Table I: Monitoring mechanisms under UN core human rights treaties

	State reporting	Inter-state communications	Individual communications	Inquiries	Country visits
CERD	✓	✓	✓(optional)	✗	✗
ICESCR	✓	✓(OP)	✓(OP)	✓(OP opt in)	✓(OP)
ICCPR	✓	✓(optional)	✓(OP)	✗	✗
CEDAW	✓	✗	✓(OP)	✓(OP opt out)	✓(OP)
CAT	✓	✓(optional)	✓(optional)	✓(optional)	✓(OP)
CRC	✓	✓(OP)	✓(OP)	✓(OP opt out)	✗
CMW	✓	✓(optional)	✓(optional)	✗	✗
CRPD	✓	✗	✓(OP)	✓(OP opt out)	✓(OP)
CED	✓	✓(optional)	✓(optional)	✓	✓

OP: Optional Protocol.

they voluntarily commit to these mechanisms. When a state ratify these mechanisms, it is making a binding commitment to let international bodies hear complaints another state may bring against it (inter-state communications), allow treaty bodies to receive and adjudicate complaints against it by individuals within its jurisdictions (individual communications), answer inquiries and questions treaty bodies have regarding its treaty compliance (inquiries), and permit a third party to visit and investigate its human rights practices (country visits). In the next section, I present a theory about the potentially differential causal effects of monitoring mechanisms as a function of their legal design. I then detail and execute a research design and an estimation strategy to estimate the causal effects of the inquiries and individual communications mechanisms under the CAT and the country visits mechanism under the OPCAT. I conclude by describing the merits and implications of the findings.

Theoretical argument

My theoretical argument explains why human rights practices are likely to improve if states ratify a more intrusive mechanism than otherwise. In other words, intrusive mechanisms have a greater positive causal impact on state behavior. This argument takes the form of a causal chain with three interconnected components. The first component is the simple observation that monitoring mechanisms are procedures that create for states parties different kinds of obligations.

These obligations vary along three dimensions: monitoring extent, precision, and delegation to the monitoring bodies. As a result, mechanisms differ in terms of how intrusive they are to state sovereignty.

The second component is a logical reasoning about two implications of this intrusiveness. One is that ratifying an intrusive monitoring mechanism sends a credible signal about state intent, that is, states parties are serious about protecting human rights. The other is that intrusive monitoring by treaty bodies generates more concrete information about state compliance, that is, whether states parties are actually protecting human rights as they have promised. The third component is the proposition that signaling about state intent and information about state compliance feed into many empirically documented mechanisms of influence through which treaties affect state behavior.

In essence, my theoretical proposition about the differential causal effects of monitoring mechanisms is an informational argument. It builds up on previous findings in the literature about the causal process from treaty ratification through mechanisms of influence to human rights outcomes. Where it differs from, and contributes to, the literature is with regard to the first causal step from treaty ratification to mechanisms of influence. Targeting the first chain, this paper (a) disaggregates treaty practice into multiple monitoring mechanisms; (b) classifies them by their intrusiveness based on three dimensions of legalization; (c) introduces the signaling and informative power of monitoring mechanisms as a function of their intrusiveness; and (d) connects the signaling and informative role of monitoring mechanisms to the mechanisms of influence discussed in the literature. Once we have unpacked the causal chain, the observable implication is that ratifying a more intrusive mechanism causes an improvement in human rights practices. Figure 1 provides a simple flow chart that illustrates this proposed causal process. In what follows, I elaborate on each step of this causal process.

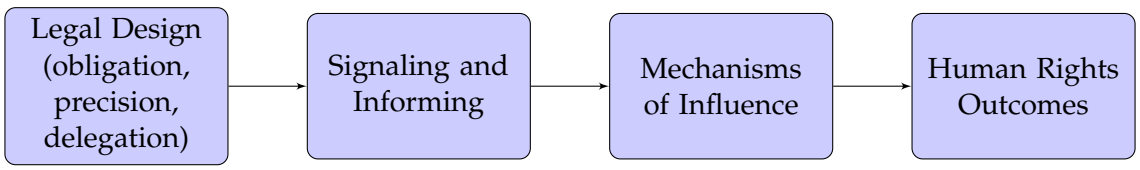

Figure 1: Causal process from monitoring mechanisms to human rights outcomes

Legal design of monitoring mechanisms

Among the key functions of monitoring treaty bodies under human rights treaties are to review application of the corresponding treaties, examine the progress that states parties have make in meeting their obligations, and report on state compliance. While the existing literature has many studies on the impact of particular human rights treaties, there is not a lot of empirical examination of monitoring mechanisms and how they differ in terms of intrusiveness. Here I draw on the legalization framework (Abbott et al. 2000; Goldstein et al. 2000) to form the basis for classifying monitoring mechanisms. This conceptual framework helps unpack the form and legal design of international agreements along three dimensions: obligation, precision, and delegation. Obligation concerns whether or not a particular rule imposed upon states parties is legally binding. Precision refers to the degree to which a particular obligation is unambiguously and clearly specified. The third dimension is delegation, which measures "the extent to which states and other actors delegate authority to designated third parties [...] to implement agreements" (Abbott et al. 2000, 415). One can classify monitoring mechanisms under human rights treaties along these three legalization dimensions as well. Note that, since legalization is defined "in terms of key characteristics of rules and procedures, not in terms of effects" (Abbott et al. 2000, 402), there is no risk of a logical circularity in my argument.

First, the legalization framework defines obligation as spanning a spectrum from being explicitly non-binding to simply hortatory to unconditionally binding. I modify this dimension for the context of human rights monitoring mechanisms to denote the *scope* or extent of procedural monitoring obligations. More extensive obligations created by a monitoring mechanism imply a more expansive scope of expected standards of behavior for its states parties. They also enable monitoring bodies to examine, discuss, and reveal information about human rights practices of its states parties to a greater extent. In effect, mechanisms that impose more extensive obligations intrude further into the domestic practices of member countries.

The active monitoring system that the Subcommittee on Prevention of Torture (SPT) operates under the OPCAT is particularly instructive. This system takes the form of regular and unrestricted visits to places of detention within any territories under the jurisdiction of OPCAT parties (Art. 4). Note that places of detention is more broadly defined under the OPCAT. Accord-

ing to its First Annual Report (p. 25), for example, the SPT visited police facilities, prisons, juvenile and shelter centers, children's homes, and drug rehabilitation and detoxification centers in Maldives in late 2007. Its Fifth Annual Report (p. 13) also states that the SPT "has made an effort to increase its activities in relation to non-traditional places of detention during 2011, including immigration facilities and medical rehabilitation centres." States parties to the OPCAT can only temporarily restrict the SPT's access under a limited number of conditions and they must grant access even in case of state emergency (Art. 14). Furthermore, the SPT is able to request any relevant information it needs and interview in private any persons deprived of liberty and practically anyone it believes can supply relevant information (Art. 14 and 15). States parties usually get notified and consulted about an upcoming visit, but the purpose is "to facilitate the visit, not to prevent it from occurring" and, in fact, "no State has objected to a visit proposed by the SPT" so far (Steinerte, Evans and de Wolf 2011, 98).

The second dimension of legalization is *precision*. The basic idea is that clearer and more determinative rules make it easier to detect state violations and produce information about compliance with greater clarity. Clear determination by treaty bodies "can more effectively alter the beliefs of domestic actors" (Wallace 2013, 4), among others. Even though they are usually confined to individual cases, judgments and recommendations by treaty bodies in response to submitted individual complaints about state violations almost always refer to specific human rights norms and treaty provisions. They, in effect, contribute to a more precise interpretation of international rules for national governments and domestic courts. This semi-judicial role of the monitoring treaty bodies, while not creating legally binding or domestically enforceable decisions, still represents a greater intrusion into the domestic governance of states parties than the universal, mandatory state reporting system.

The third dimension of legalization concerns *delegation*, which, in the context of UN human rights treaties, is about how much authority is delegated to treaty bodies for them to carry out monitoring functions, interpret rules, and promote implementation of treaty obligations. One way to gauge degrees of delegation across different types of monitoring mechanisms is by examining the procedures involved. One can argue that delegation monotonically increases from state reporting (treaty bodies only receive and make comments on self-reports submitted by states parties) to complaints mechanisms (treaty bodies receive and make judgments on submitted com-

plaints) to inquiries (treaty bodies actively seek and even publicize information about state violations and make recommendations concerning how to address and remedy for violations) to country visits (states are obligated to give treaty bodies access to any places within their jurisdiction for investigations and the SPT can go public with its reports by a simple majority decision in the CAT Committee).

In totality, roughly measured over three legalization dimensions, including obligatory monitoring scope, precision, and delegation, monitoring mechanisms like individual communications, inquiries, and particularly country visits are relatively more intrusive than state reporting, which is mandatory but imprecise and lowly delegated. This varying intrusiveness is certainly by design. It is not a coincidence that more intrusive mechanisms are usually presented in optional protocols or optional provisions in human rights treaties. How this intrusiveness by design shapes the informative power of monitoring mechanisms will be addressed next.

Signaling about intent and informing about compliance

Even when a state party is not targeted by a treaty monitoring mechanism, its ratification of this mechanism has already provided *ex ante* information about its intent to comply with treaty obligations. Just like ratifying a treaty, commitment to various optional mechanisms implies varying sovereignty costs, defined as constraints on the action and strategies of a member country. It should be noted that while some believe treaty commitment is essentially a costless signal (Hathaway 2002), others have argued it involves varying sovereignty costs (Goodman and Jinks 2003; Cole 2005, 2009; Simmons 2009), depending on the design of the treaties (Hafner-Burton, Mansfield and Pevehouse 2014).

In the case of the country visits mechanism under the OPCAT, this cost is certainly substantial. The history of the OPCAT is particularly illustrative of how concerns about potential sovereignty costs could delay or even stymie adoption of human rights monitoring mechanisms. The formulation of the OPCAT was modeled upon a proposal by the Swiss Committee against Torture, later renamed as the Association for the Prevention against Torture, at the time of the negotiations for the CAT in the late 1970s and early 1980s. Since this "Swiss model" was considered more intrusive to state sovereignty, especially compared to the "Swedish model" which the CAT

eventually adopted, it was shelved and subsequently revived only after the CAT went into effect in 1987 (Clark 2009; Evans 2011). It took 15 years for the "Swiss model" to be revived, modified into the OPCAT as we know today, and subsequently adopted.

Because of the variation in their implied sovereignty costs, ratifying monitoring mechanisms conveys signals of varying credibility to the signaled domestic and international audience. Compared to the monitoring mechanism under the OPCAT, monitoring under the CAT such as the state reporting system does not incur as much sovereignty cost. The smaller the costs, the less credible the commitment becomes, and the smaller the signaling value that a ratification generates. By this logic, CAT membership is less likely to separate genuine actors from insincere governments in the eyes of the signaled states. Therefore, the signaling value of ratification is directly proportional to the intrusiveness of a ratified monitoring mechanism. In short, the logic of credible commitment helps us establish the connection between the designed intrusiveness of a monitoring mechanism and the signaling value of its ratification.

Once ratified, the major work of optional monitoring mechanisms is to monitor, assess, and produce information about state compliance to treaty obligations. As a legal concept, compliance indicates "the degree of conformity between legal requirements and the actions of those subject to those requirements" (Martin 2013, 605). Compliance information, therefore, refers to the information that monitoring mechanisms generate regarding the potential discrepancy between what is legally required under the terms of a treaty and the observed behavior of a state party. It should be noted that my argument is not that ratification of (intrusive) monitoring mechanisms causes compliance. Instead, it is concerning the behavioral change of states, on average, as a function of their ratification of (intrusive) monitoring mechanisms. The outcome in my observational study, as stated previously, is human rights practices or behavior, which should be conceptually and operationally distinct from compliance. The former measures the distance between behavior and human rights standards whereas the latter evaluates the distance between behavior and legal obligations.

Furthermore, compliance is "orthogonal to the concept of causal effect" (Martin 2013, 605). Full compliance could mean zero causal effect of the treaty if states have already complied with their obligations prior to ratification. Conversely, noncompliance may still imply a marked improvement due to a substantial causal effect by human rights monitoring mechanisms. There is

also an additional measurement problem with using compliance as an outcome, which is that any measurement of compliance has to be legal in content and, absent an authoritative legal determination, is subject to different legal interpretations (Martin 2013, 592-597). Human rights practices, particularly in the area of physical integrity rights, on the other hand, have been measured continually across a large number of countries (Cingranelli, Richards and Clay 2013; Gibney et al. 2013; Fariss 2014).

The second logical implication of ratifying an intrusive monitoring mechanism is that they tend to have an edge in terms of producing more information about state compliance. The SPT's visit to Paraguay in 2009, for example, produces a 313-paragraph-long visit report on that country's torture record alone. Compared to the CAT Committee, the SPT has fewer restrictions when monitoring and exposing a government's torture practices or, for that matter, supporting a government's claim of compliant behavior. The SPT can push to publicize a state's bad records. Governments with a good record, by Art. 16(2) of the OPCAT, could also request the SPT to publish its positive reports about them, thus enhancing the ability of OPCAT to separate good and bad actors. According to the Fourth Annual Report of the SPT, for instance, five visit reports have been published following requests by Honduras, Maldives, Mexico, Paraguay, and Sweden. Individual communications mechanisms are also indicative of how a more intrusive mechanism that operates on precise rules is more likely to generate more information about state compliance. The reason is that decisions by treaty bodies with respect to individual complaints contribute to unambiguously defining the expectations of state conduct. Also, by examining submitted complaints and rendering judgments on their admissibility and merits, treaty bodies can publicly deliver concrete information and remove ambiguity about state compliance.

In short, how monitoring mechanisms are designed in terms of their intrusiveness and their implied sovereignty costs for states parties shapes their power. The power of optional monitoring mechanisms takes the form of generating signals about state intent and producing information about state compliance. The next subsection elaborates how this informative power feeds into various mechanisms of influence and ultimately results in behavioral change in states parties.

Mechanisms of influence

The channels through which human rights treaties affect changes in state behavior are often referred to in the literature as mechanisms of influence. Major mechanisms of influence that have been discussed involve international reputation (Guzman 2008; Brewster 2013), electoral accountability (Dai 2005), domestic judicial enforcement (Simmons 2009; Powell and Staton 2009), popular mobilization (Simmons 2009), acculturation (Goodman and Jinks 2013), and norms internalization (Keck and Sikkink 1998). Building on these theories about human rights causal mechanisms, I argue that intrusive monitoring mechanisms such as country visit under the OPCAT have a greater causal impact because the signaling value and the compliance information they generate would enable, activate and facilitate these mechanisms of influence.

First, ratifying monitoring mechanisms of varying signaling value, which is prior to and separate from compliance information that monitoring bodies subsequently produce, sets different baseline expectations of behavior for states parties. The logic is straightforward. Ratifying an intrusive monitoring mechanism that potentially imposes greater sovereignty costs will establish higher expectations of behavior. A country facing a higher public expectation of being a protective state is more likely to incur larger reputational costs for its public acts of treaty violation than a repressive country with a poor reputation to begin with (Brewster 2013, 528). This factors into reputational cost-benefit analyses that states parties conduct before engaging in behaviors a monitoring treaty body may detect and deem non-compliant (von Stein 2013). Maintaining a good reputation is important for states because it shapes the inferences that other actors make about their trustworthiness and potential for beneficial cooperation in other areas such as trade and foreign investment (Blanton and Blanton 2009, 2007; Hafner-Burton 2005). For this reputational mechanism to work, however, reputation for human rights practices should be linked to reputation in other areas that involve material costs and benefits (Brewster 2013; Guzman 2008, 533). Post-ratification compliance information certainly also feeds into reputation of states parties. Specifically, more intrusive monitoring mechanisms that are highly delegated are likely to convey more compliance information to legitimate international material sanctions against violators (Abbott et al. 2000, 418), create greater deterrence effects (Simmons and Danner 2010; Wallace 2013), and lead to better substantive outcomes.

Second, treaties can work their effects through the electoral accountability mechanism (Simmons 2009). Compliance with treaty obligations can be tied to the political accountability of state leaders, increasing their incentives to refrain from abusing rights so as not to hurt their chance to stay in power. This electoral mechanism often depends on several factors, including whether information about treaty compliance can be conveyed meaningfully to the electorate, whether a large enough segment of the electorate is invested in seeking treaty compliance by their leaders, and whether that segment of electorate can coordinate effectively among themselves to punish their leaders. Underlying these factors is the unique role of information, which is often provided by a free press and domestic and international human rights organizations (Ritter and Conrad 2012). In this regard, the informational advantage of intrusive monitoring mechanisms could come into focus and greatly facilitate the electorate to hold government officials accountable.

Third, human rights treaties can make their influence through domestic judicial litigations. This judicial mechanism relies on the presence of judiciary independence and effective domestic courts (Simmons 2009; Powell and Staton 2009). Unlike state reporting and even the two complaints procedures under the CAT, the OPCAT's country visits system is particularly useful in facilitating this domestic mechanism of influence. As Lupu (2013, 477-481) points out, even independent domestic courts can have enormous difficulties enforcing international commitments to protect physical integrity rights due to the high costs of producing legally admissible evidence. A major reason is that the government has considerable ability to control victims and hide or destroy evidence. The unprecedented system of visits by the SPT and its domestic counterparts, the National Preventive Mechanisms (NPMs), may target exactly this informational challenge. The regular, unannounced, and unrestricted nature of the SPT's visits, which includes confidential interviews with potential victims and relevant interlocutors, could produce the kind of evidence that domestic courts can use to constrain the behavior of state officials and hold them accountable. The SPT's Second Annual Report (p. 8), for example, mentions that the SPT "carried out unannounced visits to places of detention [and] had interviews in private with persons deprived of their liberty." Its Third Annual Report (p. 9) reiterates that "confidential face-to-face interviews with persons deprived of liberty are the chief means of verifying information and establishing the risk of torture." The same report also raises concern that many detainees whom the SPT spoke

with may suffer from reprisals and recommends their protection by the national preventive mechanisms (p. 11), which suggests that the SPT was able to gather information about state violations the way it was designed to do.

Fourth, human rights treaties can activate or facilitate mass mobilization by raising people's awareness of their rights and their chances of successful mobilization, thus increasing the likelihood that government leaders will make concessions and comply with treaty obligations (Simmons 2009). Because of its ability to provide better monitoring and greater scrutiny, an intrusive monitoring mechanism could supply citizens with more information to overcome their coordination problem (Ritter and Conrad 2012). An instructive example is the NPMs, which are mandated and encouraged by Art. 18(4) of the OPCAT to be modeled as an independent national human rights institution. The NPMs essentially serve as a domestic mirror of the SPT in terms of verifying a government's human rights practices, providing citizens with information about state abuses, and creating a focal point around which domestic groups and human rights nongovernmental organizations (NGOs) can mobilize against a repressive government. In addition, more intrusive mechanisms such as inquiries and country visits generally aim to address allegations of systemic, widespread, and serious cases of violations. The information they generate often goes beyond individual cases and concerns larger groups of victims. It, therefore, has the potential to motivate a larger number of people among the citizenry to mobilize around issues of state abuses.

Finally, mandated monitoring functions of treaty bodies enable international experts and transnational advocacy groups to leverage their information to pressure low-performing states to emulate better practices (Goodman and Jinks 2013) or even change the hearts and minds of government officials about human rights norms and human rights protection (Keck and Sikkink 1998). The more information they are able to produce and the more interaction with state actors they generate, the more likely it is to promote emulation and norm diffusion regarding expected human rights practices.

In summary, I advance a theoretical argument regarding the potentially differential causal effects of monitoring mechanisms as a function of their legal design. Varying degrees of intrusiveness are built in various monitoring procedures, which endows monitoring mechanisms with differential informative power to generate signaling value and produce compliance information.

This informative power enables and facilitates multiple mechanisms of influence to affect changes in state behavior. The hypothesis, therefore, is that ratification of more legalized and intrusive monitoring mechanisms, particularly the country visits system, would have a greater causal impact than less intrusive ones such as the universal state reporting system. The analysis that follows would provide an empirical test of that hypothesis.

Empirical analysis

Despite many attempts in the empirical literature, estimating the causal effects of human rights treaties remains a challenge in terms of producing compelling findings with a causal interpretation from observational data. In this section I estimate the causal effects of ratifying (i) the country visits mechanism under the OPCAT, (ii) the individual complaints mechanism under Art. 22 of the CAT, and (iii) the inquiries mechanism under Art. 20 of the CAT on human rights outcomes. I first formulate the causal quantities of interest within the structural causal modeling framework (Pearl 2009*b*,*a*, 2010). For identification, I generate a directed acyclic graph to encode background knowledge about the determinants of human rights outcomes and to inform covariate selection for adjustment via the backdoor criterion. I then employ the method of targeted maximum likelihood estimation to produce causal effect estimates. The results provide some evidence in support of my previously stated hypothesis.

Formulation

Our quantities of interest are the causal effects of the ratification of the OPCAT and the optional Art. 22 and Art. 20 of the CAT. Specifically, we want to know the change or difference in human rights outcomes on average for the population of states parties to the CAT that would have occurred if one of these three monitoring mechanisms had been ratified. An intuitive way to think about them is to imagine two identical worlds. In one world, all countries ratified the OPCAT (or Art. 22 or Art. 20); in the other, none did so. The difference between the averages of measured human rights outcomes in these two worlds is defined as the average total causal effect of ratification of a monitoring mechanism for a population. One obviously cannot run experiments on imaginary worlds and for each individual instance we only observe the actual outcome and ratifi-

cation status rather than the potential outcomes and the potential actions of ratification. Hence, the fundamental fact is that one can never obtain causal inferences from observational data without making assumptions about the data generating process.

I employ causal Bayesian networks (CBN) in the form of a directed acyclic graph (DAG) to explicitly represent these assumptions. A DAG comprises of nodes indicating random variables and edges denoting one variable's direct causal influence on another (Pearl 1988; Koller and Friedman 2009). Its algebraic equivalent is a set of functional equations that forms a non-parametric structural causal model. Each of the equations describes the dependence of one variable on its parents through a non-specified stochastic function as follows.

$$
\begin{aligned}
W &= f_W(U_W) \\
Z &= f_Z(W, U_Z) \\
A &= f_A(W, Z, U_A) \\
Y &= f_Y(W, Z, A, U_Y)
\end{aligned}
$$

This structural causal model describes the data generating process $O = (Y, A, W, Z) \sim P_O$ where Y is human rights outcome, A is ratification status of a monitoring mechanism, W is the set of time-independent baseline covariates, Z is the set of time-varying confounders, and P_O is the underlying joint distribution from which n country–year observations are assumed to be randomly sampled. We make no assumptions regarding the functional forms of all f's. In this structural causal model, the causal quantity of interest is $\tau = E(Y|do(A = 1)) - E(Y|do(A = 0))$ where the do-operator $do(A)$ stands for an active intervention to assign a treatment value to each observation. Specifically, $do(A = 1)$ indicates ratification of a monitoring mechanism while $do(A = 0)$ is for non-ratification (Pearl 2009*b*).

Identification

The key utility of DAGs is to represent compactly, and facilitate a factorization of, the joint distribution of all variables in our model of the data generating process. A factorized representation of the joint distribution enables us to model a causal relationship as an intervention to only change the treatment assignment mechanism (the specific equation $A = f_A(W, Z, U_A)$ that generates the

values of the treatment) while maintaining the rest of the model (the remaining equations) intact. The result is a modified, interventional model used to compute the new probability distribution of the outcome, which will yield the causal effect estimate of the treatment.

In the language of probability calculus, this procedure involves using a mathematical machinery called the *do*-calculus inference rules that Pearl (2009*b*) developed to translate an interventional expression (the *do*-expression) into standard conditional probabilities of observed data. This process of translation, premised on a graphical causal model, if successful, will result in causal effect identifiability. It also makes explicit all assumptions required for causal effect identification and makes it possible to use statistical techniques for parameter estimation and endow these estimates with a causal interpretation. Specifically, the inference rules allow $E(Y|do(A), W, Z) = E(Y|A, W, Z)$ if the outcome Y and the treatment A are d-separated (that is, if all non-causal paths from A to Y are blocked) by a conditioning set $\{W, Z\}$. This is the essence of identifiability by adjustment via the backdoor criterion (Pearl 2009*b*). In other words, if we can find a set of variables $\{W, Z\}$ that are non-descendants of A such that $\{W, Z\}$ d-separates Y from A, then the causal effect of A on Y is identifiable and a causal effect estimation problem becomes a statistical estimation. In that case, estimating the total causal effect $\tau = E(Y|do(A = 1)) - E(do(A = 0))$ becomes estimating the parameter ψ of the observed joint distribution

$$\psi(P_n) = E_{W,Z}\Big[E(Y|A = 1, W = w, Z = z) - E(Y|A = 0, W = w, Z = z)\Big]P(W = w, Z = z).$$

Note that, in addition to the *i.i.d.* assumption above, we also assume the conditioning set $\{W, Z\}$ is sufficient to make the counterfactual outcome Y conditionally independent from the treatment A. The plausibility of this assumption can only judged based on our substantive knowledge drawn from the existing literature. We further assume $0 < P(A = 1|W, Z) < 1$, that is, the probability of all observations receiving the treatment is positive at all $\{W, Z\}$. These assumptions are indispensable for causal inference regardless of the causal inference frameworks or estimation strategies one may use.

A recent software (Textor 2015) automates and facilitates the task of identifying a sufficient conditioning set $\{W, Z\}$ in a graphical model that approximates the data generating process (Koller and Friedman 2009; Pearl 2009*b*). Based on the background knowledge in the literature, I have derived a sufficient adjustment set that includes both time-independent and time-varying covariates to identify the causal effects of human rights monitoring mechanisms (Figure 2).

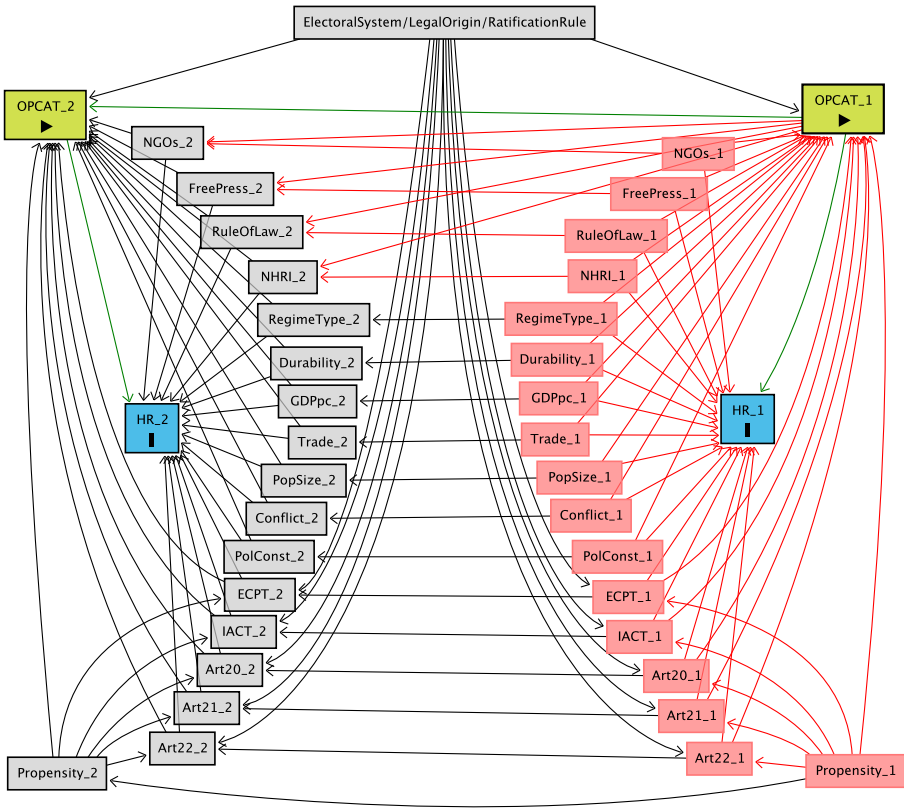

Figure 2: A graphical model of the underlying data generating process in two time periods. Red arrows are biasing paths while green arrows are causal paths. A sufficient adjustment set to identify the causal effect includes nodes in gray boxes. This sufficient adjustment set blocks all non-causal paths from the treatment $OPCAT_{t=2}$ to the outcome $HR_{t=2}$.

The treatment is OPCAT (or Art. 20 or Art. 22) ratification status and the outcome is human rights practices measured by the Human Rights Scores (Fariss 2014). Time-independent baseline covariates W include electoral system (Cingranelli and Filippov 2010), legal origin (Simmons 2009), and ratification rule (Simmons 2009). Time-varying confounders Z include measures of non-governmental organizations (Murdie and Davis 2012; Simmons 2009; Hafner-Burton 2008), freedom of the press (Conrad and Moore 2010), the rule of law (Powell and Staton 2009; Conrad 2013), national human rights institutions (NHRIs) (Goodman and Pegram 2012), regime types (Hathaway 2007; Chapman and Chaudoin 2013; Neumayer 2007), regime durability (Goodliffe and Hawkins 2006), gross domestic product (GDP) per capita, participation in international trade (Hafner-Burton 2013), population size, involvement in international or domestic conflicts (Chap-

man and Chaudoin 2013), political constraints (Hill 2014; Conrad and Moore 2010), and the overall propensity to ratify international treaties (Lupu 2014). Ratification status of other monitoring mechanisms under the CAT and the Inter-American Convention to Prevent and Punish Torture (IACT) and the European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment (ECPT) is also measured.

One more assumption implicit in my graphical representation of the data generating process is that any node only depends on its immediate parents (variables that emit an arrow directly to it) and not on its predecessors in the temporal ordering. This first-order Markovian assumption makes it possible to represent the joint distribution much more compactly and ease our parameter estimation. Based on the graphical causal model, a statistical model of treatment assignment (propensity score model) has, as its predictors, all variables that have a direct arrow pointing to the OPCAT node (electoral system, legal origin, ratification rule, human rights NGOs, freedom of the press, the rule of law, NHRIs, regime type, regime durability, GDP per capita, trade, population size, involvement in conflicts, political constraints, ECPT, IACT, Art. 20, Art. 21, Art. 22, and propensity to commit to treaties). Similarly, predictors in a statistical outcome model include the treatment variable (OPCAT ratification status) and all covariates in the treatment model except for the three time-independent covariates (electoral system, legal origin, and ratification rule) and treaty commitment propensity. Appendix C indicates the data sources of these measures and provides a variable description. Given the substantial rate of missing data during our coverage time periods for two key variables (human rights NGOs and treaty commitment propensity), I conduct multiple imputation ($m = 5$), using Amelia II program (Honaker et al. 2011), and combine estimates across imputed data sets. Appendix D provides summary statistics of the raw data and Appendix E provides information and diagnostic plots of multiple imputation.

Estimation

Once we have been able to determine identifiability of causal effects and translate them into parameters of the joint distribution of observed data, the next step is to select a good estimator. Propensity score-based estimators are some of the most popular methods for causal inference. The basic idea is to approximate an experiment by balancing the treatment group and the con-

trol group in an observational study on all relevant aspects (covariates) so that the two groups become as comparable as possible. Any resulting difference between the two groups in terms of the measured outcome is attributed to the treatment effect, assuming no unmeasured confounding factors. Rosenbaum and Rubin (1983) demonstrate that, instead of balancing on all measured covariates, one can just balance on the propensity score, which is the predicted probability of each unit receiving the treatment given the covariates. In other words, once adjusted for the propensity score via weighting, matching, stratification or covariate adjustment, the treatment group and the control group are similar on average in terms of the measured covariates, which helps address the treatment selection bias problem that was a subject of debate for some time in international relations literature (Downs, Rocke and Barsoom 1996; von Stein 2005; Simmons and Hopkins 2005) and enables an unbiased estimation of the treatment effect.

I, however, employ a different estimation technique called targeted maximum likelihood estimation (TMLE) that also involves estimating the propensity score, but uses it in a targeted manner to produce a more robust and potentially more efficient estimator (Van der Laan and Rose 2011). Its superior performance derives significantly from a machine learning approach called Super Learner (van der Laan, Polley and Hubbard 2007; Polley and van der Laan 2010). Super Learner uses an ensemble of parametric, semi-parametric, and non-parametric algorithms, each of them is weighted by its relative cross-validated predictive performance measured by a pre-specified loss function. These algorithms are then combined to produce a stronger hybrid predictive estimator (Pirracchio, Petersen and van der Laan 2015; Lendle, Fireman and van der Laan 2015). Super Learner, therefore, satisfies to a much greater extent the assumption of correct specification of models that describe the relationship between an outcome and its predictors. The collection of algorithms I use includes generalized linear models with and without interaction terms, generalized linear models with penalized maximum likelihood, generalized additive models, local polynomial regression model, spline regression model, generalized boosted regression model, and random forest.

In terms of implementation, TMLE, first, uses Super Learner ensemble prediction function to predict both $Q_0(A, W, Z) = E(Y|A, W, Z)$, which is the initial conditional mean outcome Y given the treatment A and covariates $\{W, Z\}$, and $g_0(A|W, Z) = P(A|W, Z)$, which is the propensity score or the probability of treatment assignment given the adjustment set.

Second, instead of using the propensity score g_0 to adjust the full likelihood function of the observed data like all propensity score-based methods, TMLE specifically targets the potential bias in Q_0 by computing the "clever covariate" $H_n(A_i, W, Z_i) = \left[\frac{I(A_i=1)}{g_n(A_i=1|W,Z_i)} - \frac{I(A_i=0)}{g_n(A_i=0|W,Z_i)}\right]$, which is a function of the treatment weighted by its probability of assignment. This covariate is subsequently used in a logit model to account for the remaining variation in the outcome Y by regressing Y on both Q_0 and H_n. The coefficient ϵ associated with the clever covariate H_n is obtained and used, together with the initial conditional mean outcome Q_0, to produce a modified, less biased conditional mean outcome Q_1 in a logit model:

$$\text{logit}\Big[Q_1(A, W, Z)\Big] = \text{logit}\Big[Q_0(A, W, Z)\Big] + \epsilon H_n(A, W, Z).$$

In this model $\epsilon H_n(A, W, Z)$ is the amount of bias reduction in the estimated outcome achieved by incorporating the information about treatment assignment mechanism.

Finally, the new predictive outcome function Q_1 is used to generate the predicted value of each unit under each treatment value ($Q_1(1, W, Z)$ and $Q_1(0, W, Z)$) and the causal effect estimate is the empirical mean difference of these predicted values. Inference about the causal effect estimate can be obtained via the influence curve (Van der Laan and Rose 2011) or the bootstrap method ($b = 1,000$ in this study) and pool them over multiple imputed datasets. Inferences obtained by both methods are reported.

While its implementation is more involved, TMLE performs better because of its double robustness (TMLE produces unbiased estimates if either the initial outcome model Q_0 or the treatment assignment model g_0 is consistent) and efficiency (if both Q_0 and g_0 are consistent). Each of these two models, in its own right, is already more robust to different functional form specifications thanks to the use of Super Learner. This feature of double robustness (Kang and Schafer 2007) is a clear advantage over propensity score methods, which mostly use main-term logistic regression models for propensity score estimation and, thus, are less likely to meet the assumption of correct model specification. It should be noted, however, that model specification of both Q_0 and g_0 in terms of covariate selection is justified by the previous DAG and, as a result, is only as good as the graphical causal model and its encoded assumptions. These assumptions may seem like strong limitations, but they are crucial if one wishes to make causal inference. Furthermore, while their plausibility is up for debate, their transparency in the form of a graphical model is an obvious plus.

Interpretation

Estimates of the causal effects of OPCAT, Art. 22, and Art. 20 ratification are presented in Table II and Figure 3. Statistical uncertainty for the effect estimates are obtained using the efficient influence curve and bootstrapping. Note that human rights outcome is measured on a rescaled 0–1 range with 0 as having the worst record and 1 as having the best record in terms of protecting physical integrity rights.

Table II: Estimates of ATE of monitoring mechanisms under the CAT and the OPCAT

Mechanism	ATE	SE	95% CI
OPCAT	0.045	0.017	(0.011, 0.080)
OPCAT (bootstrap)	0.045	0.016	(0.013, 0.077)
Art. 22	−0.116	0.036	(−0.188, −0.044)
Art. 22 (bootstrap)	−0.049	0.084	(−0.216, 0.118)
Art. 20	−0.060	0.128	(−0.317, 0.197)
Art. 20 (bootstrap)	−0.084	0.099	(−0.282, 0.114)

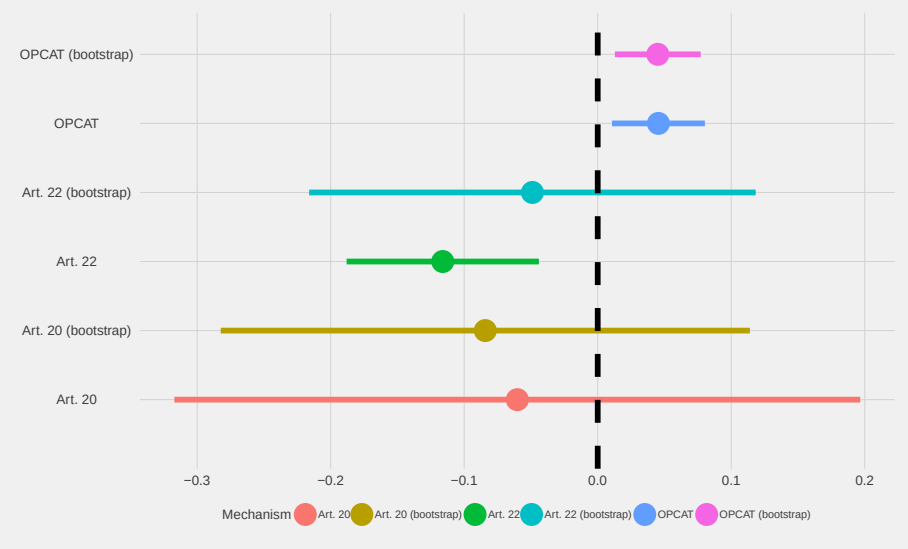

Figure 3: Average total causal effect of ratifying monitoring mechanisms on human rights outcome (measured in Human Rights Scores on 0 – 1 scale), 2006 – 2013

The results suggest that ratifying the OPCAT will, on average, improve human rights practices by 4.5% and this estimate is statistically significant, boosting our confidence in the positive causal impact of the country visits mechanism under the OPCAT. While this seems like a very small magnitude, given the large number of determinants of human rights outcome, some of

which are highly resistant to any meaningful change such as the nature of the political system or the absence of the rule of law, the finding that ratifying a single monitoring mechanism can lead to nearly 5% improvement is substantively significant and merits attention from human rights defenders and other stakeholders.

What is diverging from my theoretical expectation, however, is the causal effect estimates of two other optional monitoring mechanisms under the CAT. Making an optional declaration to recognize the competence of the CAT Committee to receive and consider individual complaints about government abuses does not seem to have any positive causal impact on state practices, if not an outright negative influence. Similarly, subscribing to the inquiries mechanism under Article 20 of the CAT actually worsens human rights performance, but this estimate is not statistically different from zero. A reasonable conclusion is probably that ratifying Art. 20 and Art. 22 under the CAT does not seem to exert any significant influence on human rights practices on average.

It should be remembered, however, that only states parties to the CAT can ratify these optional monitoring mechanisms. My empirical analysis reflects that legal restriction by only considering the observations that are CAT members since 2006 when the OPCAT went into effect. The effect estimates of ratifying these three mechanisms should be interpreted as their causal impact above and beyond that of the state reporting system under to the CAT. As more data become available, our findings could be revised in light of more evidence in the future. More generally, one may even speculate that monitoring mechanisms should be highly intrusive to have a positive causal impact on human rights outcomes whereas if they are designed to be just a little intrusive, they are unlikely to make meaningful improvements in human rights behavior. This speculation, however, dovetails with qualitative case studies in the international law literature that focus on regional human rights mechanisms (Phan 2012).

Finally, while a good case can be made that the country visits under the OPCAT is more intrusive, and there is solid evidence that it has a greater causal impact than other monitoring mechanisms, the positive relationship between designed intrusiveness and causal impact may not be linear. In other words, more intrusive and legalized monitoring mechanisms tend to have greater causal effects on average, but the intrusiveness might need to exceed a certain level before the causal impacts become meaningful and significant. This interpretation, albeit speculative at this point, might suggest an interesting direction for more future research by exploring, modeling,

and explaining the potential non-linearity of the relationship between legal design and causal impact of international human rights institutions.

Conclusion

In this paper, I aim to address an important question regarding the differential causal effects of monitoring mechanisms under the UN human rights treaty against torture. Answering this question has significant scholarly and policy implications. First, part of the ambitious research agenda of the legalization framework is to examine the causes and consequences of international legal design (Downs, Rocke and Barsoom 1998; Koremenos, Lipson and Snidal 2001; Koremenos 2005; Gilligan and Johns 2012; Abbott and Snidal 2013). This paper contributes to that agenda by examining the empirical implications of different legal designs of monitoring mechanisms under the CAT and the OPCAT. It answers in the affirmative the key question regarding whether legal rules that have stronger delegation, greater precision, and more extensive obligations lead to better human rights outcomes. More broadly, this research direction will hopefully advance our understanding about the potential causal impact of institutional design on substantive outcomes.

Second, another contribution of this paper involves examining human rights treaty effects from a new, disaggregated level of analysis. Contradictory findings unfortunately abound in the existing literature on human rights treaties. One of the reasons is probably that treaties have been so far examined only at the aggregate level. This paper reconceptualizes treaty ratification, unpacks treaty practice into multiple monitoring processes, and then estimates the causal effect of individual optional monitoring mechanisms. The results suggest that not all mechanisms of a human rights treaty are created equal or have the same impact on human rights practices of states parties. More intrusive mechanisms such as country visits may have the best chance to improve human rights outcomes. This is only the initial step in systematically evaluating the effectiveness of monitoring mechanisms and determining the kinds of institutional design that work, have no causal impact, or even backfire in protecting human rights, improving government accountability, and extending access to justice. Governments, international organizations, human rights activists and non-governmental organizations, and other stakeholders may benefit from further research in that direction.

Third, this paper develops a theory to explain why monitoring mechanisms may have differential effects on human rights outcomes. It advances an argument that highlights the legal design of monitoring mechanisms and its implications for their effectiveness in improving human rights practices of states parties. This differs from almost all existing theories, which mostly focus on external conditions such as the normative international environment, regime types, domestic attributes of states parties, the civil society, judicial independence, and executive job security, among others.

Finally, in terms of methodology, this paper draws on the structural causal framework and relies on graphical modeling to establish conditions for identifying the causal effects. It, therefore, attempts to set a study of causal inference in the area of international human rights research on a more transparent and sound footing. More studies should take advantage of recent advances in the causal inference literature and the structural causal modeling approach for clear and transparent causal effect identification. This paper also demonstrates an application of the TMLE method to political science data. TMLE promises superior performance in terms of estimation robustness and efficiency. More political science studies might consider using both the graphical modeling tool and machine learning-based estimation techniques so as to enhance their validity.

References

Abbott, Kenneth W and Duncan Snidal. 2013. Law, Legalization, and Politics: An Agenda for the Next Generation of IL/IR Scholars. In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, ed. Jeffrey L Dunoff and Mark A Pollack. Cambridge University Press pp. 33–57. 22

Abbott, Kenneth W, Robert O Keohane, Andrew Moravcsik, Anne-Marie Slaughter and Duncan Snidal. 2000. "The Concept of Legalization." *International Organization* 54(03):401–419. 5, 10

Blanton, Shannon Lindsey and Robert G Blanton. 2007. "What Attracts Foreign Investors? An Examination of Human Rights and Foreign Direct Investment." *Journal of Politics* 69(1):143–155. 10

Blanton, Shannon Lindsey and Robert G Blanton. 2009. "A Sectoral Analysis of Human Rights and FDI: Does Industry Type Matter?" *International Studies Quarterly* 53(2):469–493. 10

Brewster, Rachel. 2013. Reputation in International Relations and International Law Theory. In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, ed. Jeffrey L Dunoff and Mark A Pollack. Cambridge University Press pp. 524–543. 10

Chapman, Terrence L and Stephen Chaudoin. 2013. "Ratification Patterns and the International Criminal Court1." *International Studies Quarterly* 57(2):400–409. 16

Cingranelli, David L., David L. Richards and K. Chad Clay. 2013. "The Cingranelli-Richards (CIRI) Human Rights Dataset." *CIRI Human Rights Data Website:* `http://www.humanrightsdata.org`. 9

Cingranelli, David and Mikhail Filippov. 2010. "Electoral Rules and Incentives to Protect Human Rights." *Journal of Politics* 72(1):243–257. 16

Clark, Ann Marie. 2009. Human Rights NGOs at the United Nations: Developing an Optional Protocol to the Convention against Torture. In *Transnational Activism in the UN and EU: A Comparative Study*, ed. Jutta Joachim and Birgit Locher. Routledge. 8

Clark, Ann Marie. 2013. The Normative Context of Human Rights Criticism: Treaty Ratification and UN Mechanisms. In *From Commitment to Compliance: The Persistent Power of Human Rights*, ed. Thomas Risse, Stephen C. Ropp and Kathryn Sikkink. Cambridge: Cambridge University Press. 1

Cole, Wade M. 2005. "Sovereignty Relinquished? Explaining Commitment to the International Human Rights Covenants, 1966-1999." *American Sociological Review* 70(3):472–495. 7

Cole, Wade M. 2009. "Hard and Soft Commitments to Human Rights Treaties, 1966–2001." *Sociological Forum* 24(3):563–588. 7

Conrad, Courtenay R. 2013. "Divergent Incentives for Dictators: Domestic Institutions and (International Promises Not to) Torture." *Journal of Conflict Resolution* . 16

Conrad, Courtenay R. and Will H. Moore. 2010. "What Stops the Torture?" *American Journal of Political Science* 54(2):459–476. 16, 17

Dai, Xinyuan. 2005. "Why Comply? The Domestic Constituency Mechanism." *International Organization* 59(02):363–398. 10

Downs, George .W., David M. Rocke and Peter N. Barsoom. 1996. "Is the Good News about Compliance Good News about Cooperation?" *International Organization* 50:379–406. 18

Downs, George W., David M. Rocke and Peter N. Barsoom. 1998. "Managing the Evolution of Multilateralism." *International Organization* 52(2):397–419. 22

Evans, Malcolm. 2011. The OPCAT at 50. In *The Delivery of Human Rights: Essays in Honour of Professor Sir Nigel Rodley*, ed. Geoff Gilbert and Clara Sandoval Villalba. Taylor & Francis. 8

Fariss, Christopher J. 2014. "Respect for Human Rights Has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–318. 9, 16

Gibney, Mark, Linda Cornett, Reed Wood and Peter Haschke. 2013. "Political Terror Scale 1976-2012." *Political Terror Scale Website:* `http://www.politicalterrorscale.org` . 9

Gilligan, Michael J and Leslie Johns. 2012. "Formal Models of International Institutions." *Annual Review of Political Science* 15:221–243. 22

Goldstein, Judith, Miles Kahler, Robert O Keohane and Anne-Marie Slaughter. 2000. "Introduction: Legalization and World politics." *International Organization* 54(03):385–399. 5

Goodliffe, Jay and Darren G. Hawkins. 2006. "Explaining Commitment: States and the Convention against Torture." *Journal of Politics* 68(2):358–371. 16

Goodman, Ryan and Derek Jinks. 2003. "Measuring the Effects of Human Rights Treaties." *European Journal of International Law* 14(1):171–183. 7

Goodman, Ryan and Derek Jinks. 2013. *Socializing States: Promoting Human Rights Through International Law*. Oxford University Press. 10, 12

Goodman, Ryan and Thomas Pegram. 2012. National Human Rights Institutions, State Conformity, and Social Change. In *Human Rights, State Compliance, and Social Change: Assessing National Human Rights Institutions*, ed. Ryan Goodman and Thomas Pegram. Cambridge: Cambridge University Press. 16

Guzman, Andrew. 2008. "Reputation and International Law." *UC Berkeley Public Law Research Paper* . 10

Hafner-Burton, Emilie and Kiyoteru Tsutsui. 2007. "Justice Lost! The Failure of International Human Rights Law to Matter Where Needed Most." *Journal of Peace Research* 44(4):407–425. 1

Hafner-Burton, Emilie M. 2005. "Trading Human Rights: How Preferential Trade Agreements Influence Government Repression." *International Organization* 59(3):593–629. 10

Hafner-Burton, Emilie M. 2008. "Sticks and Stones: Naming and Shaming the Human Rights Enforcement Problem." *International Organization* 62(4):689–716. 16

Hafner-Burton, Emilie M. 2013. *Forced to Be Good: Why Trade Agreements Boost Human Rights*. Cornell University Press. 16

Hafner-Burton, Emilie M, Edward D Mansfield and Jon CW Pevehouse. 2014. "Human Rights Institutions, Sovereignty Costs and Democratization." *British Journal of Political Science* FirstView:1–27. 7

Hathaway, Oona. 2002. "Do Human Rights Treaties Make a Difference?" *Yale Law Journal* 111(8):1935–2042. 1, 7

Hathaway, Oona A. 2007. "Why Do Countries Commit to Human Rights Treaties?" *Journal of Conflict Resolution* 51(4):588–621. 16

Hill, Daniel W. 2010. "Estimating the Effects of Human Rights Treaties on State Behavior." *Journal of Politics* 72(4):1161–1174. 1

Hill, Daniel W. 2014. "Avoiding Obligation: Reservations to Human Rights Treaties." *Unpublished paper* . 17

Honaker, James, Gary King, Matthew Blackwell et al. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47. 17

Kang, Joseph DY and Joseph L Schafer. 2007. "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science* 22(4):523–539. 19

Keck, Margaret E. and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca: Cornell University Press. 10, 12

Koller, Daphne and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press. 14, 15

Koremenos, Barbara. 2005. "Contracting around International Uncertainty." *American Political Science Review* 99(4):549. 22

Koremenos, Barbara, Charles Lipson and Duncan Snidal. 2001. "The Rational Design of International Institutions." *International Organization* 55(4):761–799. 22

Landman, Todd. 2005. *Protecting Human Rights: A Comparative Study*. Georgetown University Press. 1

Lendle, Samuel David, Bruce Fireman and Mark J van der Laan. 2015. "Balancing Score Adjusted Targeted Minimum Loss-based Estimation." *Journal of Causal Inference* 3(2). 18

Lupu, Yonatan. 2013. "Best Evidence: The Role of Information in Domestic Judicial Enforcement of International Human Rights Agreements." *International Organization* 67(3):469–503. 1, 11

Lupu, Yonatan. 2014. "Why Do States Join Some Universal Treaties but Not Others? An Analysis of Treaty Commitment Preferences." *Journal of Conflict Resolution* p. 0022002714560344. 17, 31

Martin, Lisa. 2013. Against Compliance. In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*. Cambridge University Press pp. 591–610. 8, 9

Murdie, Amanda M. and David R. Davis. 2012. "Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs." *International Studies Quarterly* 56(1):1–16. 16, 31

Neumayer, Eric. 2005. "Do International Human Rights Treaties Improve Respect for Human Rights?" *Journal of Conflict Resolution* 49(6):925–953. 1

Neumayer, Eric. 2007. "Qualified Ratification: Explaining Reservations to International Human Rights Treaties." *The Journal of Legal Studies* 36(2):397–429. 16

Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 14

Pearl, Judea. 2009*a*. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3:96–146. 13

Pearl, Judea. 2009*b*. *Causality*. Cambridge University Press. 13, 14, 15

Pearl, Judea. 2010. "An introduction to Causal Inference." *The International Journal of Biostatistics* 6(2). 13

Phan, Hao D. 2012. *A Selective Approach to Establishing a Human Rights Mechanism in Southeast Asia: The Case for a Southeast Asian Court of Human Rights*. Procedural Aspects of International Law Leiden: Brill. 21

Pirracchio, Romain, Maya L Petersen and Mark van der Laan. 2015. "Improving Propensity Score Estimators' Robustness to Model Misspecification Using Super Learner." *American Journal of Epidemiology* 181(2):108–119. 18

Polley, Eric C and Mark J van der Laan. 2010. "Super Learner in Prediction." *Working Paper Series UC Berkeley Division of Biostatistics* . 18

Powell, Emilia J. and Jeffrey K. Staton. 2009. "Domestic Judicial Institutions and Human Rights Treaty Violation." *International Studies Quarterly* 53(1):149–174. 10, 11, 16

Ritter, Emily Hencken and Courtenay R Conrad. 2012. "International Human Rights Treaties and Mobilized Challenges against the State.". Unpublished paper. 11, 12

Rosenbaum, Paul R and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55. 18

Simmons, Beth A. 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge: Cambridge University Press. 1, 7, 10, 11, 12, 16, 31

Simmons, Beth A and Allison Danner. 2010. "Credible Commitments and the International Criminal Court." *International Organization* 64(02):225–256. 10

Simmons, Beth A and Daniel J Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *American Political Science Review* 99(04):623–631. 18

Steinerte, Elina, Malcolm David Evans and Antenor Hallo de Wolf. 2011. *The Optional Protocol to the UN Convention Against Torture*. Oxford University Press. 6

Textor, Johannes. 2015. "Drawing and Analyzing Causal DAGs with DAGitty." *arXiv preprint arXiv:1508.04633* . 15

Tyagi, Yogesh. 2009. "The Denunciation of Human Rights Treaties." *British Yearbook of International Law* 79(1):86–193. 2

Ulfstein, Geir and Helen Keller. 2012. *UN Human Rights Treaty Bodies*. Cambridge: Cambridge University Press. 1

van der Laan, Mark J, Eric C Polley and Alan E Hubbard. 2007. "Super learner." *Statistical Applications in Genetics and Molecular Biology* 6(1). 18

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media. 18, 19

von Stein, Jana. 2005. "Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance." *American Political Science Review* 99(4):611–622. 18

von Stein, Jana. 2013. The Engines of Compliance. In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, ed. Jeffrey Dunoff and Mark Pollack. Cambridge University Press pp. 477–501. 10

Wallace, Geoffrey PR. 2013. "Martial Law? Military Experience, International Law, and Support for Torture." *International Studies Quarterly* . 6, 10

A United Nations Human Rights Conventions and Status of Ratification

CAT Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, 1465 UNTS 85, adopted 10 December 1984, entered into force 26 June 1987, ratified by 158 states.

CED Convention for the Protection of All Persons from Enforced Disappearance, UNTS 2715 Doc.A/61/448, adopted 20 December 2006, entered into force 23 December 2010, ratified by 50 states.

CEDAW Convention on the Elimination of All Forms of Discrimination against Women, 1249 UNTS 13, adopted 18 December 1979, entered into force 3 September 1981, ratified by 189 states.

CERD Convention on the Elimination of All Forms of Racial Discrimination, GA Res. 2106 (XX), Annex, 20 UN GAOR Supp. (No. 14) at 47, UN Doc. A/6014 (1966), 660 UNTS 195, adopted 7 March 1965, entered into force 4 January 1969, ratified by 177 states.

CMW International Convention on the Protection of the Rights of All Migrant Workers and Members of their Families, GA Res. 45/158, Annex, 45 UN GAOR Supp. (No. 49A) at 262, UN Doc. A/45/49, adopted 18 December 1990, entered into force 1 July 2003, ratified by 48 states.

CRC Convention on the Rights of the Child, 1577 UNTS 3, adopted 20 November 1989, entered into force 2 September 1990, ratified by 194 states.

CRPD Convention on the Rights of Persons with Disabilities, UN Doc. A/61/611, adopted 13 December 2006, entered into force 3 May 2008, ratified by 157 states.

ICCPR International Covenant on Civil and Political Rights, 999 UNTS 171, adopted 16 December 16 1966, entered into force 23 March 1976, ratified by 168 states.

ICESCR International Covenant on Economic, Social and Cultural Rights, 993 UNTS 3, adopted 16 December 1966, entered into force 3 January 1976, ratified by 164 states.

B States Parties to Monitoring Mechanisms under the CAT and the OP-CAT, 2003 – 2013

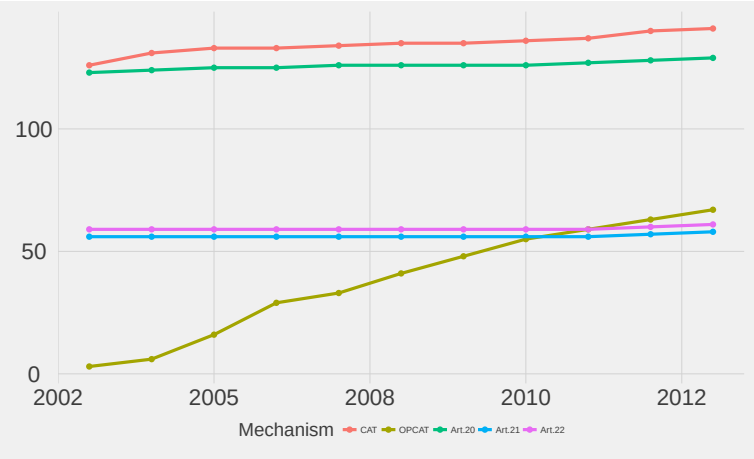

Figure 4: Number of states parties to human rights monitoring mechanisms under the CAT and the OPCAT (2003 – 2013), including state reporting (mandatory under the CAT), country visits (OPCAT), inter-state communications (Art. 21), individual communications (Art. 22), and inquiries (Art. 20)

C Variable Description

- **Human Rights Scores**: a country–year interval variable that measures respect for physical integrity human rights. Rescaled to a 0–1 range from the empirical range for ease of interpretation. The scores were generated using a dynamic ordinal item-response theory model that accounts for systematic change in the way human rights abuses have been monitored over time. The human rights scores model builds on data from the CIRI Human Rights Data Project, the Political Terror Scale, the Ill Treatment and Torture Data Collection, the Uppsala Conflict Data Program, and several other published sources.

  (http://humanrightsscores.org).

- **Ratification Status of Conventions and Monitoring Mechanisms**: A country–year binary variable coded 1 for ratification and 0 otherwise. Monitoring mechanisms coded include (i) Art. 20, (ii) Art. 21, and (iii) Art. 22 under the Convention against Torture, (iv) Optional Protocol to the Convention against Torture, (v) the Inter-American Convention to Prevent and Punish Torture, and (vi) the European Convention for the Prevention of Torture and Inhuman or Degrading Treatment or Punishment. Data are coded manually from the Office of the High Commissioner for Human Rights database.

  (http://www.ohchr.org/EN/HRBodies/Pages/HumanRightsBodies.aspx).

- **Political Terror Scale**: a country–year five-point ordinal variable measuring levels of political murders, torture, political imprisonment, and disappearances. Variable coded from 5 for worst level of abuses to 1 for least abuses. Its ordinal measurement restricts the application of many techniques; thus, it is used in this paper for multiple imputation purposes only.

  (http://www.politicalterrorscale.org).

- **CIRI Index**: a country–year ordinal variable measuring torture and other cruel, inhuman and degrading treatment or punishment; coded 0 if torture is practices frequently, 1 if torture is practiced occasionally, and 2 for no occurrences of torture. Its ordinal level of measurement and lack of temporal coverage (the project was halted after 2011) makes it difficult to measure the outcome variable. Thus, it is used in this paper for multiple imputation purposes only.

  (http://www.humanrightsdata.com).

- **Regime Type**: measured by the Polity Score. The Polity Score is a country–year interval variable measuring regime authority spectrum on a 21-point scale ranging from -10 (hereditary monarchy) to $+10$ (consolidated democracy).

  (http://www.systemicpeace.org/polityproject.html).

- **Regime Durability**: a country–year interval variable measuring the number of years since the most recent regime change (defined by a three-point change in the POLITY score over a period of three years or less) or the end of transition period defined by the lack of stable political institutions.

  (http://www.systemicpeace.org/polityproject.html).

- **GDP per capita**: a country–year interval variable measuring gross domestic product divided by midyear population measured in current US dollars.

  (http://data.worldbank.org/indicator/NY.GDP.PCAP.CD).

- **Trade**: a country–year interval variable measuring the sum of exports and imports of goods and services as a share of gross domestic product.

  (http://data.worldbank.org/indicator/NE.TRD.GNFS.ZS).

- **Population**: a country–year interval variable measuring the total number of residents in a country regardless of their legal status.

  (http://data.worldbank.org/indicator/SP.POP.TOTL).

- **Conflict Involvement**: a country–year binary variable from the UCDP–Prio Armed Conflict Dataset. It is recoded 1 to indicate a country's involvement in any side of an armed conflict anywhere and 0 otherwise. Armed conflicts are defined as "a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths."

  (https://www.prio.org/Data/Armed-Conflict/UCDP-PRIO/).

- **Ratification Rule**: a cross-sectional (country) five-point ordinal variable (1, 1.5, 2, 3, 4) measuring "the institutional "hurdle" that must be overcome in order to get a treaty ratified." Coding is based on descriptions of national constitution or basic rule (Simmons 2009).

  (http://scholar.harvard.edu/files/bsimmons/files/APP_3.2_Ratification_rules.pdf).

- **Legal Origin**: a cross-sectional (country) binary variable coded 1 for British legal orgin and 0 otherwise. Data are from the Global Development Network Growth Database.

  (http://www.nyudri.org/resources/global-development-network-growth-database/).

- **Electoral System**: a cross-sectional (country) nominal variable coded 0 for primarily presidential system, 2 for primarily parliamentary system, and 1 for hybrid or ambiguous system (Simmons 2009).

- **National Human Rights Institutions**: a country–year binary variable coded 1 for *de jure* organizational existence of an NHRI and 0 otherwise.

  (http://nhridata.weebly.com/data.html).

- **Freedom of the Press**: a country–year interval variable coded from 0 (the most free) to 100 (the least free).

  (https://freedomhouse.org/report-types/freedom-press).

- **The Rule of Law**: a country–year interval variable capturing "perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence." It ranges from -2.5 for the weakest to 2.5 for the strongest rule of law. Data come from the Worldwide Governance Indicators project.

  (http://info.worldbank.org/governance/wgi/index.aspx#home).

- **Treaty Commitment Propensity**: a country–year interval variable ranging from -1 to 1, measuring a country's commitment preference across a large number of treaties in many different domains. I use the first-dimension coordinates in Lupu (2014).

- **Human Rights NGOs**: a country-year interval variable measuring the number of human rights NGOs with locations in a certain country in a given year (Murdie and Davis 2012).

- **Political Constraints Index**: an expert-coded country–year interval variable on a scale from 0 (most hazardous - no checks and balances) to 1 (most constrained–extensive checks and balances).

  (https://whartonmgmt.wufoo.com/forms/political-constraint-index-polcon-dataset/).

D Summary Statistics

Table III: Summary Statistics of Raw Data

Statistic	N	Mean	Standard Deviation	Min	Max
Year	1,848	—	—	2,003	2,013
COW code	1,848	—	—	2	950
ID number	1,848	—	—	1	168
CAT ratification	1,848	0.801	0.399	0	1
Art. 20 ratification	1,848	0.749	0.433	0	1
Art. 21 ratification	1,848	0.335	0.472	0	1
Art. 22 ratification	1,848	0.353	0.478	0	1
OPCAT ratification	1,848	0.227	0.419	0	1
ECPT ratification	1,848	0.238	0.426	0	1
IACT ratification	1,848	0.096	0.295	0	1
PTS score	1,848	2.530	1.120	1	5
Human Rights Score (Fariss)	1,848	0.597	1.310	−2.640	4.710
Torture Index (CIRI)	1,499	0.614	0.664	0	2
Polity Score	1,745	3.670	6.370	−10	10
Regime Durability	1,782	26.900	31.200	0	204
GDP Per Capita	1,800	15,590	18,566	381	136,727
Trade	1,721	89.600	48.300	0.309	440
Population	1,848	39,726,899	139,837,190	102,369	1,357,380,000
Conflict Involvement	1,848	0.144	0.352	0	1
Ratification Rule	1,815	1.820	0.633	1	3
Legal Origin	1,837	0.287	0.453	0	1
System (electoral)	1,837	0.614	0.871	0	2
NHRI	1,848	0.640	0.480	0	1
Press Freedom	1,848	49.700	24	2	99
Rule of Law	1,848	−0.163	1.010	−2.5	2
Treaty Propensity	996	0.117	0.453	−0.985	0.993
Human Rights NGOs	651	3.340	9.430	0	71
Political Constraints	1,485	0.279	0.202	0	0.720

E Multiple Imputation of Missing Data

All variables in the data summary statistics are used in the multiple imputation stage to make the missing at random (MAR) assumption more plausible. Not all of them are included in the data analysis.

Table IV: Fractions of missing data by variables

Variables	**Fraction**
HRO locations	0.64773
Treaty Commitment Propensity	0.46104
Political Constraints	0.19643
Torture	0.18885
Disappearance	0.18994
Killing	0.18885
Political Imprisonment	0.18939
Trade	0.06872
Polity	0.05574
Regime Durability	0.03571
GDP per capita	0.02597
Ratification Rule	0.01786
Legal Origin	0.00595
Electoral System	0.00595

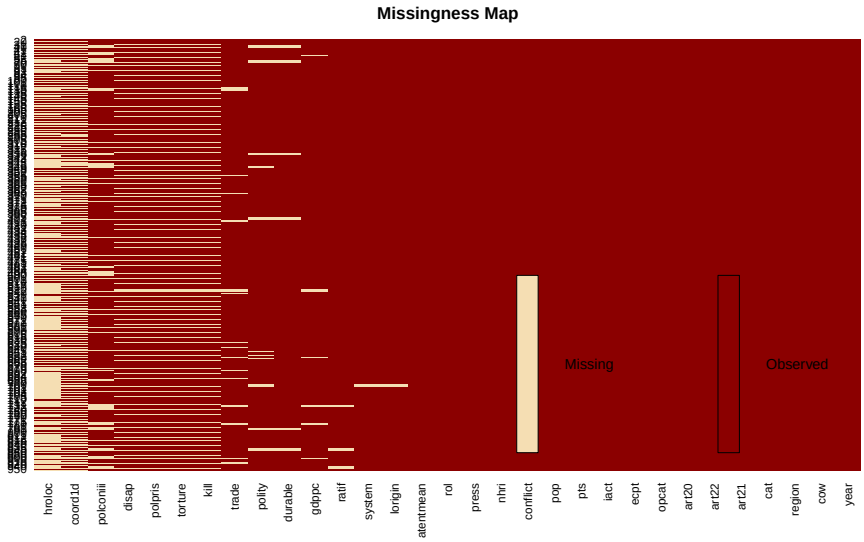

Figure 5: Map of missing data for multiple imputation

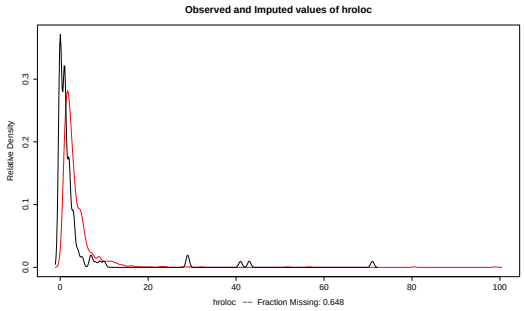

Figure 6: Comparing densities of observed and imputed data of human rights NGOs

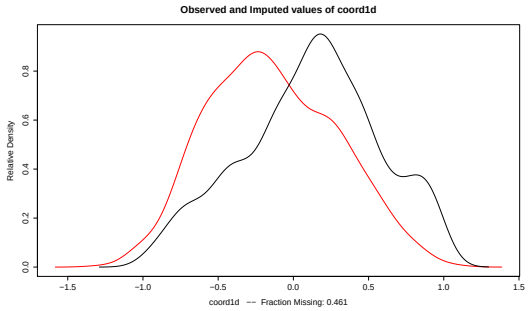

Figure 7: Comparing densities of observed and imputed data of treaty commitment propensity

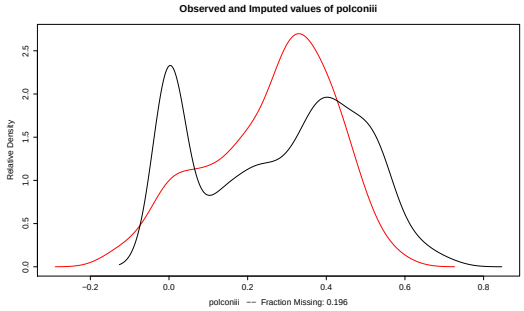

Figure 8: Comparing densities of observed and imputed data of political constraints

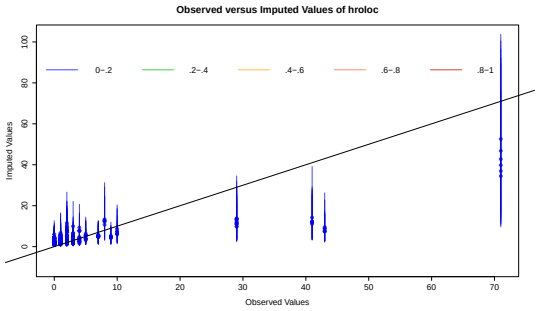

Figure 9: Overimputation diagnostic of observed and imputed data of human rights NGOs

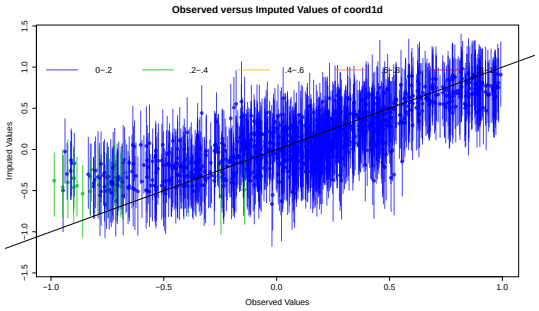

Figure 10: Overimputation diagnostic of observed and imputed data of treaty commitment propensity

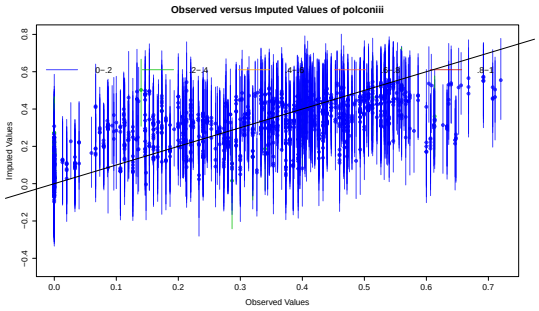

Figure 11: Overimputation diagnostic of observed and imputed data of political constraints

F OPCAT Propensity Scores across Imputed Datasets and by Ratification Status

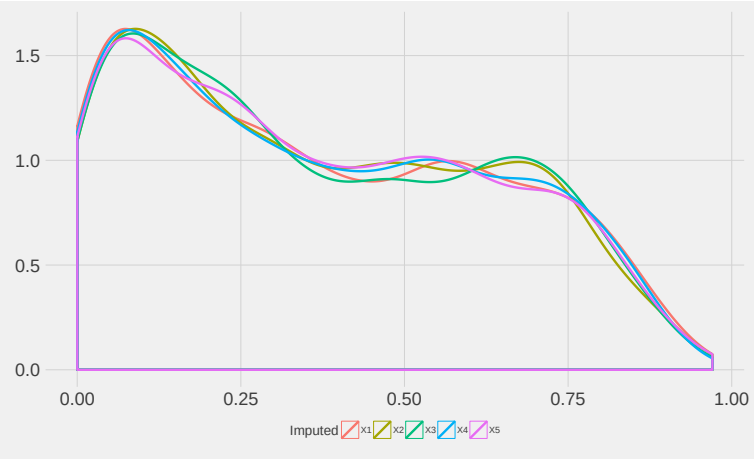

Figure 12: Distribution of predicted probabilities of OPCAT ratification across five imputed datasets

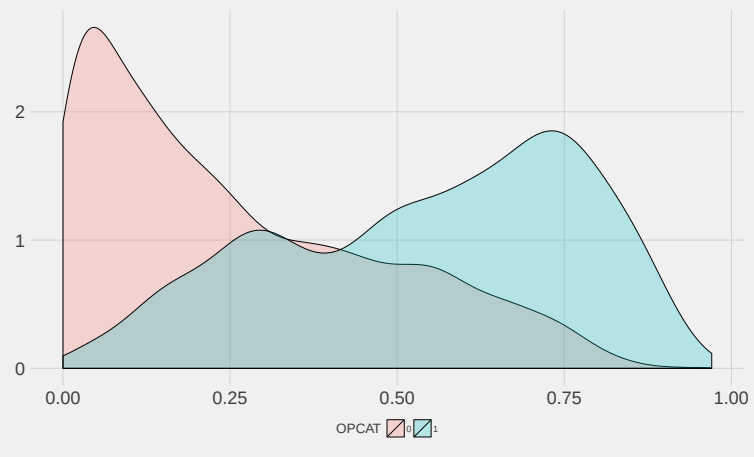

Figure 13: Distriution of predicted probabilities of OPCAT ratification between member group and non-member group summed over five imputed datasets

G Art. 22 Propensity Scores across Imputed Datasets and by Ratification Status

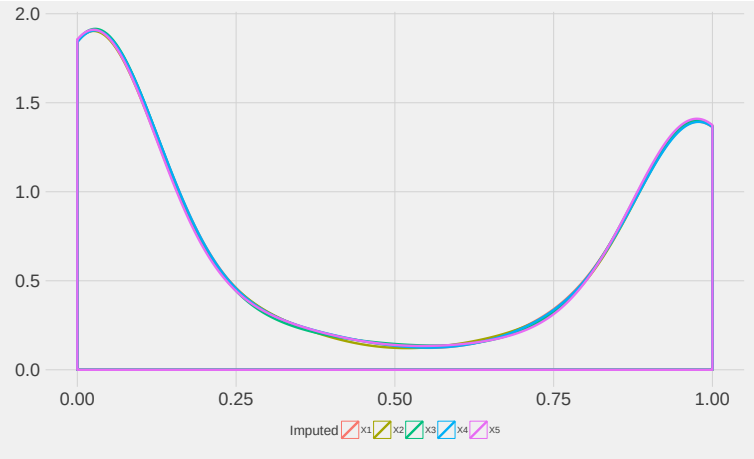

Figure 14: Distribution of predicted probabilities of Art. 22 ratification across five imputed datasets

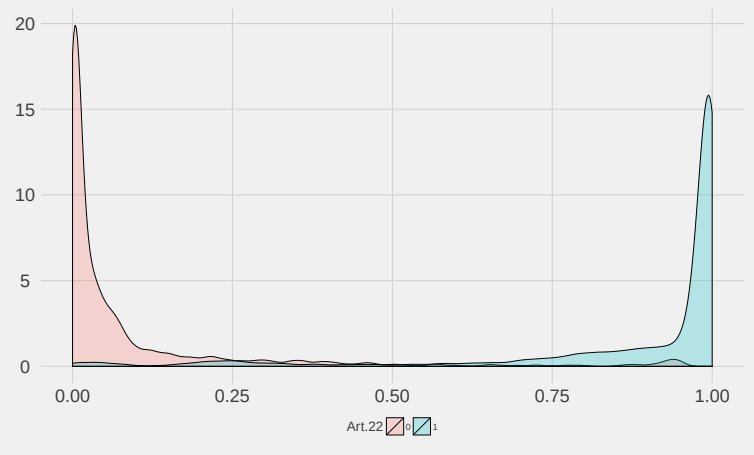

Figure 15: Distribution of predicted probabilities of Art. 22 ratification between member group and non-member group summed over five imputed datasets

H Art. 20 Propensity Scores across Imputed Datasets and by Ratification Status

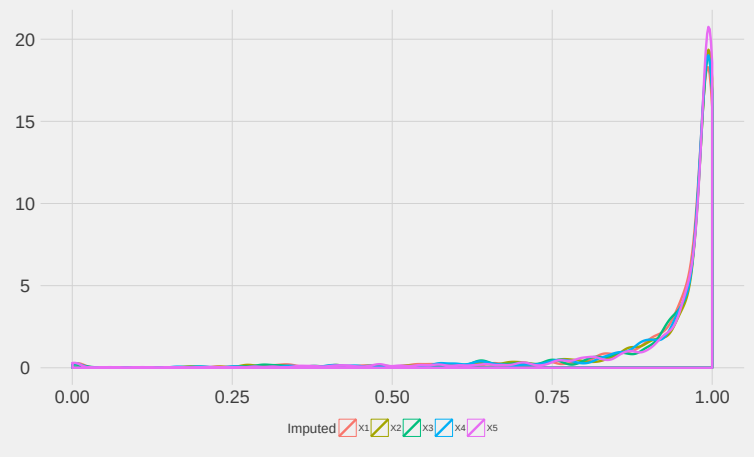

Figure 16: Distribution of predicted probabilities of Art. 20 ratification across five imputed datasets

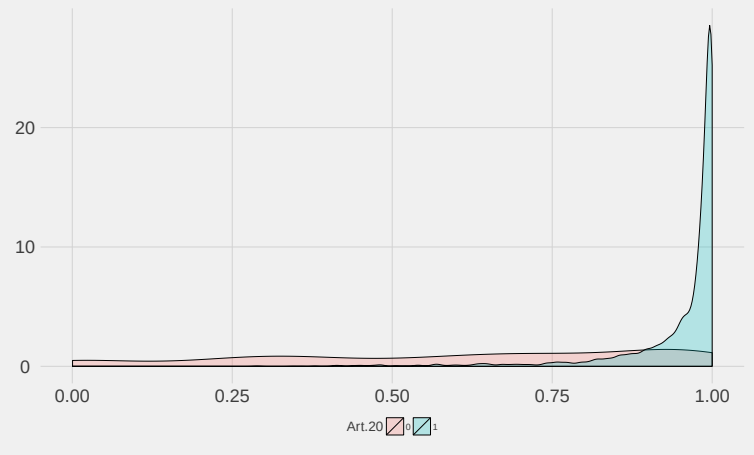

Figure 17: Distribution of predicted probabilities of Art. 20 ratification between member group and non-member group summed over five imputed datasets