Domestic Politics and the Accession of Authoritarian Regimes to Human Rights Treaties[*]

James R. Hollyer[†]
New York University

B. Peter Rosendorff
New York University

Current Version: September, 2009

Abstract

A common claim in international relations theory holds that states will only join those international institutions with whose regulations they intend to comply (eg. Downs, Rocke and Barsoom 1996). In this paper, we offer a demonstration of when this claim might not hold. We construct a model of an authoritarian government's decision to accede to the UN Convention Against Torture (CAT). We demonstrate that authoritarian governments use the signing of this treaty - followed by the willful violation of its provisions - as a costly signal to domestic opposition groups of their willingness to employ repressive tactics to remain in power. In equilibrium, we find that authoritarian governments that torture more are more likely to sign the treaty than those that torture less, as is consistent with Hathaway's (2007) empirical findings regarding the CAT. We show that signatory regimes survive longer in office than non-signatories - and we provide empirical support for this prediction. The model further suggests that while CAT accession reduces levels of torture in signatory states, it also delays democratization in those states.

1 Introduction

Sovereign states that accede to international treaties, we are told, intend to comply with the obligations imposed by these treaties. The reasons for this claim are varied: International law (the Vienna Convention on the Law of Treaties in particular) declares that "every treaty ... is binding

[*] We would like to thank James Vreeland and Jennifer Gandhi for their generosity in providing access to their data. We would also like to thank Bruce Bueno de Mesquita, David Stasavage, the participants in the NYU IPE seminar series and the MPSA Panel on International Human Rights Agreements, and two anonymous reviewers for helpful comments and suggestions.

[†] Corresponding author. New York University, Department of Politics. 19 W. 4th St., 2nd Floor. New York, NY 10012. james.hollyer@nyu.edu

upon the parties." This declaration follows from the basic principle of international law *pacta sunt servanda* - treaties are to be obeyed. Downs, Rocke & Barsoom (1996) establish that those countries that are most likely to abide by the rules promulgated by an international institution are also those countries that are most likely to join in the first place. Failure to comply with treaty provisions is described as a "managerial problem," [1] or as a temporary aberration to be remedied by (re-)negotiation. [2] Tolerated temporary escape, [3] exchanges of information, and dispute resolution mechanisms are designed to complete gaps in treaty language or to generate better information about signatories' behavior [4] and to thus bring about treaty compliance. Where international treaties address issues of international externalities - such as trade, security, or the environment – the intent to comply is strengthened by the mutual gains associated with a predictable, stable, and cooperative international order. Similarly, states facing collective action problems may be inclined to forgo the temporary benefits of defection in order to remain within the society of cooperative nations, especially when future (relative to current) consumption is highly valued [5].

Recent scholarship has explored if these general findings also apply to the specific case of human rights treaties. Simmons (Forthcoming, 2009) argues that the major human rights treaties have been successful in reducing the prevalence of torture worldwide. She claims that countries accede to and ratify these treaties because they intend to comply with treaty provisions.[6] She acknowledges that there are some states that ratify, but do not greatly adjust their behavior. She describes these states as the "false positives,"the countries that sign human rights treaties and continue to torture. And she observes that these states tend to have relatively authoritarian regimes.

Simmons' observation reinforces the conclusions of Hathaway (2007), who finds a positive association between the practice of torture and accession to human rights treaties amongst highly authoritarian regimes. Hafner-Burton & Tsutsui (2005, 2007) confirm that indeed, accession to human rights treaties has little or no effect on the behavior of the world's worst repressors. As they put it, there is a "rising gap between states' propensity to join the international human rights

[1] Chayes & Chayes (1993)

[2] Koremenos (2005)

[3] Bagwell & Staiger (2005),Rosendorff & Milner (2001)

[4] Rosendorff (2005)

[5] Downs & Rocke (1995)

[6] chapter 3, p.42

regime and to bring their human rights practice into compliance" and this gap brings the efficacy of international law into fundamental question.

We thus have a puzzle: if states accede to and ratify agreements because they intend to comply with them, why do some states, particularly authoritarian states, ratify and fail to comply with human rights treaties?

We argue that authoritarian states ratify human rights treaties explicitly because they do *not* intend to comply. And it is important to those signatories that all observers understand that they have no intention of complying at the time of accession. The logic, while counterintuitive, is straightforward: an elite facing threats from a domestic opposition can mitigate these threats by engaging in torture. If there is any additional cost to the elite of signing and then being found to torture, the act of signing the agreement signals to the opposition the strength of the elite's commitment to remaining in power. Accession is a signal to the opposition of the very high value the elite places on holding onto power and its willingness to use torture if necessary. On observing the government's accession, the opposition - now better informed about the value the elite places on holding power - will rationally reduce its anti-regime activities. The government continues to torture, but will torture less. On the other hand a regime that doesn't sign shows itself to be vulnerable to the added costs associated with the use of torture. Thus, the opposition will increase efforts to remove the regime on seeing that the government does not sign.

This logic leads to two conclusions: First, more repressive regimes (regimes with elites more willing to use force to hold onto power) will sign and torture more frequently than less (or non-) repressive governments. Second, opposition political action falls in signatory states - yielding to reductions in the likelihood of regime collapse or transition. In the non-signatory states, opposition response actually rises, leading to more frequent regime failure.

The first finding is consistent with Hathaway's (2007) empirical results, and offers a theoretical explanation for the puzzle above. In order to check the veracity of the model, we test the second prediction: authoritarian regimes that accede to the treaty will enjoy longer tenures in office than those that do not. This is true for two reasons: (1) A selection effect implies that those regimes that will fight most strongly to remain in power are the same regimes that sign the treaty. (2) An information effect implies that domestic opposition groups will engage in fewer activities designed to overthrow a signatory government. We test this claim using data on accession to the UN Convention Against Torture (CAT) and find that it enjoys robust empirical support. Signatory regimes face a lower hazard rate than observationally similar non-signatories across a wide variety of empirical specifications.

Key to the causal logic of the argument is the notion that the CAT affects the costs to a member-state's elite of engaging in torture. We will argue that, aside from international opprobrium and withdrawals of concessions (or active sanctions) along other dimensions (such as international trade) by the international community [7], signatories of the CAT must consider the role of "universal jurisdiction"and the extradition clauses of the CAT when determining whether or not to employ torture. These additional considerations serve to make torture more costly given accession to the CAT than not. However, these costs do not directly translate into higher levels of compliance by signatory states. Rather, they allow accession to act as a costly signaling mechanism, such that the states that sign are those that are most likely to defy their treaty obligations.

This paper makes contributions to two literatures. It first speaks to the literature on selection effects and international institutions. While it may generally be the case that governments join treaties by whose provisions they intend to abide; there may exist circumstances in which governments benefit by acceding to treaties whose provisions they intend to defy. Our model offers one instance in which this may take place.

This paper also contributes to the literature on human rights law. We provide a theory of when and why authoritarian governments are likely to join human rights treaties, and provide empirical evidence in support of this theory. We also explore an "unintended consequence" of increased legalization of the human rights agenda - these legal instruments provide signaling opportunities to domestic oppositions of the elite's intent *not* to abide by its obligations. The human rights regime may in fact be counterproductive for reducing torture (at least in the long run) in that authoritarian states that accede to human rights treaties may enjoy longer tenures in office as a result.

[7] Hafner-Burton (2005)

2 Autocracies and the CAT

The United Nations Convention Against Torture and Other Cruel, Inhuman and Degrading Treatment or Punishment (CAT) was adopted in December 1984, went into effect in June 1987. It has been ratified by 139 states. It forbids

> "any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or intimidating or coercing him or a third person, or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity."[8]

The CAT requires that each member-state passes appropriate domestic laws making torture a crime, and requires that each state asserts jurisdiction when the crime occurs within its own territory, or the offender or the victim is a national of that state, or if the offender is present in its territory (if the member-state does not for some reason extradite the offender).

The emergence of a set of human rights treaties has been heralded as a major shift in the international system[9], and a measure of the success and efficacy of international law. While these agreements have been ratified by most states in the world; repressive state behavior has continued to rise over time. Hafner-Burton & Tsutsui (2005) report that in 2000, while the average state had ratified 80% of all available human rights treaties; 35% of states are reported as having violated these agreements. States then are clearly willing to sign human rights agreements and continue to violate their treaty commitments.

[8] CAT Article 1

[9] The major human rights treaties (in addition to the CAT) are the International Convention on the Elimination of All Forms of Racial Discrimination (adopted 1965), the Internatonal Convention on Economic, Social, and Cultural Rights (1966), The International Convention on Civil and Political Rights (1966), the Convention on the the Elimination of All Forms of Discrimination Against Women (1979) and the Convention on the Rights of the Child (1989). In addition other treaties, such as many Preferential Trading Agreements have both soft and hard prohibitions against human rights abuses (Hafner-Burton 2005).

The lack of compliance with human rights treaties is often viewed as stemming from a failure of enforcement. Enforcement of an international obligation has a number of prerequisites. First, failure to abide by the agreement must be observable. If violations are obscure, mixed in with noise or are otherwise difficult to observe or prove, there is little expectation of compliance. Second, there must exist a system of punishments to be imposed on a state or its elite in the event of a treaty violation to deter non-compliance. And third, there must be some mechanism or process by with these costs are actually applied or incurred. At the international level, this may be the withdrawal of concessions by a trading partner, or the application of sanctions. At the domestic level, failure to abide by a ratified and implemented international agreement is likely to be a violation of domestic law and subject to sanction by domestic authorities currently or in the future.

Human rights treaties are viewed as being weak on all three dimensions. Violations are difficult to observe.[10] There are low costs to non-compliance. And there are few mechanisms for enforcing the agreement[11]. If non-compliance costs are in fact quite low, following Downs, Rocke & Barsoom (1996), we would expect most or all states to sign the CAT, and that state behavior on signing would be little (or un-) changed. The pattern of accession and compliance is somewhat different however. Many governments do not accede to the CAT (in our sample of 129 authoritarian regimes between 1985 and 1996, 74 regimes were never signatories); others sign and reduce torture levels; while still others sign and continue to torture.[12]

[10] In 2002, Jay S. Bybee, Assistant Attorney General at the US Justice Department, for instance, asserted that the definition of torture in the CAT and its implementing US legislation was so extreme that only "serious physical injury, such as organ failure, impairment of bodily function, or even death" constituted torture under the CAT. Moreover even if any "interrogation method" might violate US law or its international obligations, a claim of necessity or self-defense "could provide justifications that would eliminate any criminal liability". See `http://news.findlaw.com/nytimes/docs/doj/bybee80102mem.pdf` accessed May 11, 2009.

[11] A number of scholars have argued that even weak enforcement regimes can influence state behavior - by the socialization into norms of appropriateness (Finnemore 1996) or cascades, where states feel pressured to conform (Keck & Sikkink 1998). Others have argued that the international regimes create openings for non-governmental actors (NGOs) to engage in information gathering, political action, legal manoeuvring etc. that influence state behavior (Neumayer 2005, Simmons Forthcoming, 2009). Moravcsik (2000) suggests that unstable democracies can "lock-in" human rights norms by treaty accession. Gilligan & Nesbitt (2007) argue that these norm-based arguments for the adoption of the CAT have not had any noticeable effect on torture levels.

[12] See Vreeland (2008) for a description of the variation in torture levels among dictatorships. Neumayer (2005) shows that torture levels fall in democracies and in other polities with richer civil society. Simmons (Forthcoming, 2009) argues that the CAT reduces torture in all but the most stable democracies and autocracies, also due to the presence of NGOs and other civil society actors. Powell & Stanton (2009) demonstrate that an average of 83 percent of CAT signatories violate some CAT provision each year, and 42 percent of signatories systematically violate CAT provisions.

Scholars have focused on domestic enforcement mechanisms to explain the observed variation in accession patterns and torture behavior. Hathaway contends that since domestic mechanisms to enforce compliance - such as an independent judiciary or an opposition party - are absent in autocracies, they find accession to human rights treaties essentially costless. In democracies, on the other hand, treaty violations are likely to impose costs on the incumbent government in the form of legal penalties or opposition attacks. Therefore, democracies are only likely to accede to human rights treaties if they are in compliance with these treaties' provisions *before* signing. Autocracies, however, should be willing to enter such treaties regardless of prior compliance. Both autocratic torturers and non-torturers will accede to the CAT. However, Hathaway's (2007) empirical findings indicate that, amongst autocracies, there is a positive association between torture and accession to the CAT.

Vreeland (2008) explores the domestic political and institutional dynamics of autocracies, and offers an explanation for Hathaway's puzzling finding. He contends that the positive association between levels of torture and accession to the CAT stems from omitted variable bias. More precisely, Vreeland argues that the presence of domestic opposition parties both causes autocrats to torture more and forces the these governments to sign human rights treaties. When opposition parties exist, there must be some freedom to engage in speech and activities that contradict the will of the incumbent government. In such a situation, opposition activists are likely to "cross the line" in their criticisms, leading the government to employ torture to maintain its control. Moreover, these opposition parties will pressure the government to enter into human rights agreements. Since Hathaway's regressions do not control for the presence of opposition parties, she finds a spurious association between torture and accession. When the presence of such parties is controlled for, the association between torture and the signing of the CAT drops to insignificance.

While Vreeland adopts a domestic politics argument to explain the pattern of accession; his theory appears to rely on out-of-equilibrium behavior. If, as Hathaway claims, human rights treaties do not constrain autocratic governments, what is motivating the domestic opposition to push for treaty accession in the first place? If, on the other hand, human rights treaties do constrain autocratic governments, Vreeland does not articulate how these treaties do so. Nor is it clear why, if autocratic governments are willing to so tie their hands, a treaty is necessary to enforce cooperation between the government and opposition.[13]

Hafner-Burton & Tsutsui (2007) also explore the link between autocratic accession to the CAT and torture levels. They argue (as in Hathaway 2007) that while there are vague political benefits from CAT membership ("window-dressing"); these treaties lack coercion and enforcement mechanisms, fail to make states internalize or acculturize international norms, and do not cause a domestic human rights institutional capacity to emerge. Autocracies are therefore unlikely to show any evidence of improvement in repressive behavior after accession. In fact, the effect on torture of CAT accession is least in the states that "need it most" - the autocratic torturers - as is consistent with Hathaway. This result is robust conditioning on measures of civil society.

Yet explanations that stress the CAT's lack of enforcement would seemingly suggest that *all* authoritarian regimes will sign. As mentioned above, this is not empirically the case. Moreover, it is unclear from either Hafner-Burton & Tsutsui (2007) or Hathaway (2007) why there exists a positive association between torture levels and CAT accession by authoritarian regimes. It therefore remains an open question as to why those states with the worst human rights records sign these agreements with the greatest frequency and then ignore their obligations.

In the theory developed below, we concur with Vreeland's focus on the role of the interaction between autocratic governments and their domestic opposition as it affects accession to human rights treaties. But we view this interaction quite differently. We assume a game is played between an office-seeking government and an opposition party. Their interaction is characterized by attempts to maintain (seize) power: the government can undertake costly measures to repress the opposition even as the opposition can take costly actions to remove the government. We assume that the opposition is imperfectly informed as to the costs repressive measures impose on the government.

[13] Empirically, the inclusion of a control for the presence of opposition parties causes the association between torture and the signing of the CAT to drop to insignificance only when a broad spectrum of other controls are also included in the Vreeland regressions. When only 'opposition parties' and 'torture levels' are used to predict signing, both are significant. Moreover, the inclusion of additional variables does not significantly reduce the magnitude of the coefficient on torture. Since there is a substantial amount of multi-collinearity between these 'torture' and 'opposition' measures, one cannot determine whether the newfound insignificance of 'torture' is simply due to problems of estimation. Without a more convincing theory of why the presence of opposition parties leads an autocrat to sign human rights treaties, there seems little reason to suppose otherwise.

We demonstrate that, in such a game, the governments may use accession to human rights treaties as a signal to the domestic opposition that they can repress at low cost. In such an equilibrium, those governments that sign the treaty would torture more heavily *ex ante* than those that do not. Moreover, we find that signatory governments are likely to survive longer in office than non-signatories.

3 Theory

Article 4 of the CAT states that “Each State Party shall ensure that all acts of torture are offences under its criminal law.” Moreover “[e]ach State Party shall make these offences punishable by appropriate penalties.” Article 5 requires that any State Party to the CAT take into custody any alleged offender that is present in its territory. And Article 6 requires that, if requested to do so, any State Party must extradite the alleged offender to any state with jurisdiction over the case, which may be defined by the nationality of the perpetrator or the victim. If no such extradition occurs, the State Party must try the offender domestically.[14] Finally Article 8 further requires signatories to treat violations of the prohibition on torture as extraditable offenses.[15]

The CAT does, therefore, make torture after accession a more serious offense.[16] Consider an autocrat inclined to torture in order extract information from or to punish a domestic political opponent. Should the autocrat, at some point in the future, find himself (and always its himself, not herself) out of power, deposed or otherwise overthrown, the consequences will differ depending on whether the state was a signatory to the CAT. The usual act of a falling autocrat is to abscond to another country, if he manages to remain alive or out of jail. Assume that the autocrat’s country were a Party to the CAT. If the country to which he escaped were also a CAT signatory, the autocrat’s successor can demand the autocrat’s extradition for trial for human rights violations. No such obligation would necessarily exist of the country is not a signatory to the CAT. On this

[14] This requirement is often referred to as establishing ‘universal jurisdiction’ for human rights offenses.

[15] United Nations Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment. `http://www.hrweb.org/legal/cat.html`

[16] Of the seven “core” human rights treaties, it may be argued that the CAT possesses the most serious enforcement mechanism. Goodliffe & Hawkins (2006) argue the CAT was the first treaty to apply the principle of universal jurisdiction to human rights law. As such, they suggest its enforcement mechanisms are more coercive than those of other human rights treaties.

basis, we argue that signing the CAT will at least weakly increase the penalties an autocrat would suffer after being evicted from office.

If an autocrat flees into exile - and if his state is unable or unwilling to try the him domestically - the now host nation, if it is a CAT signatory, has an obligation to try the ex-dictator for human rights offenses. It is reasonable to think that, if the state of the offending dictator had signed the CAT too, the pressures for arrest and indictment would be higher than if it were not a CAT signatory.

These provisions may increase the expected costs of torture substantially. In the event an autocrat is removed from office, the danger of extradition may substantially limit his possible destinations for exile. The long term costs of this restriction on his movement would be considerable. Clearly therefore, and contrary to much of the scholarship on the CAT, there are post-tenure liabilities associated with engaging in torture. While these might be perceived to be unlikely to occur, or might happen only the distant future; the costs from treaty violation are non-trivial in expected value.[17] We model these punitive mechanisms as increasing of the marginal cost of engaging in torture or repression.

The costs imposed by the CAT have most vividly been illustrated in the extradition proceedings in the British House of Lords against Augusto Pinochet in 1998. Famously, the Law Lords ruled that Pinochet may be extradited to face criminal charges in Spain. Offenses after 1988 were ruled as extraditable, as 1988 marked the year that the UK ratified the CAT and passed domestic implementing legislation [18]. This finding allowed the prosecution of Pinochet to proceed despite a negotiated amnesty with his successor regime. [19] [20]

[17] In the costly signaling model developed below, as the costs of treaty violation go to zero, all governments pool on signing. As the costs increase, only more repressive governments are likely to sign. As noted above, we do not witness all governments pooling on signing the CAT. Moreover, Hathaway's (2007) empirical findings are consistent with punitive costs that exceed the minimum threshold for separation. To the extent that Goodliffe & Hawkins (2006) are correct regarding the relatively punitive enforcement mechanisms of the CAT, we would be more likely to see this pattern of behavior in CAT accession than in the accession to other human rights treaties.

[18] Roht-Arriaza (2001)

[19] Jonas (2004), Roht-Arriaza (2001)

[20] It should be noted however that these provision of the CAT are not universally acted upon. For instance, the case against former Chadian dictator Hissène Habré has stalled in Senegal. Despite findings from both the UN Commission on Torture and the African Union that Senegal is obliged under CAT provisions to either extradite or try Habré for torture that took place while he was in office; the case remains stalled. (see *Press Release on Habré Trial*, Human Rights Watch, `http://www.hrw.org/en/news/2009/01/28/african-union-press-senegal-habr-trial`.) However, Habré's movements are, no doubt, quite limited. Even without a trial, the costs he faces, relative to a world with no

Paradoxically, the increased cost signing the CAT places on repression ensures that those countries that torture heavily are most likely to sign. Assume that autocrats vary in the costs they face from engaging in repression and further assume that opposition groups are unable to observe these costs. Those governments that can repress cheaply will be willing to engage in more torture than those that face higher costs. Opposition groups would prefer to engage in fewer costly efforts to remove the such governments. However, since all governments would like to intimidate the opposition, no government can effectively communicate whether it is truly a 'strong' or 'weak' type.

Since signing a human rights treaty imposes a cost on autocrats who torture - and only sufficiently 'strong' types would be willing to bear such a cost - accession to such a treaty may act as a credible signal to the domestic opposition of a government's type. If this is true, it is those governments that can repress at low cost who sign the treaty and continue to torture. High-cost governments do not sign. Such behavior would seem consistent with existing empirical findings.

Moreover, those autocrats who sign such treaties should survive in office longer than non-signatories. A *selection effect* implies that those regimes that will fight most strongly to remain in power are the same regimes that sign the treaty. And an *information effect* implies that opposition groups - on learning that the state has signed the treaty and is therefore a strong state - will engage in fewer activities designed to overthrow a signatory government.

3.1 Model

We model the signing of a human rights treaty as the outcome of an interaction between an autocratic government G and its domestic opposition D. Both are assumed to be office-seeking: i.e. each derives some value from holding power $R > 0$. In the contest for power, the government may - at some positive cost - engage in repressive measures entailing human rights violations against the opposition. Similarly, the opposition may undertake costly efforts to remove the government. The outcome of the contest for power will be determined, in part, by each party's respective choice of repression and effort level.

CAT, are quite considerable.

The sequence of the game is as follows: First, nature chooses the type of government $\theta \in [0, 1]$, where θ represents the cost of repression.[21] This variable is observed by the government, but not by the opposition. Second, the government chooses whether or not to sign a human rights treaty $s \in \{0, 1\}$. Third, the government and opposition simultaneously choose t - the level of repression - and e - the level of effort put into deposing the government. It is assumed that the choice of t is made at the constant marginal cost θ if $s = 0$ and $k\theta$ if $s = 1,\ k > 1$. Opposition effort e is chosen at cost $c(e),$ with $c'(e) > 0,$ and $c''(e) \geq 0$. Fourth, nature determines whether the government survives - with probability $\pi(t, e)$ - or not. We assume $\pi_t > 0,\ \pi_e < 0,$ and $\pi_{tt} < 0,\ \pi_{ee} > 0$. All payoffs are realized and the game ends.

For simplicity, we assume the opposition cost function is linear: define $c(e) = be,\ c' = b,\ c'' = 0$. We also assume a standard contest success function [22]: $\pi(t, e) = \frac{t}{t+e}$. We further let the distribution of government types be defined by the uniform distribution $f(\cdot)$ with support over the unit interval. This distribution is common knowledge.

Player utilities are defined by their expectation of holding office and the choice of $s,\ t$ and e by the autocrat and the opposition respectively. The autocrat's expected utility function is:

$$U_G(t, e, s; \theta) = \pi(t, e)R - [sk + (1 - s)]\theta t;$$

while the opposition's is defined as

$$U_D(t, e, s) = [1 - \pi(t, e)]R - be.$$

The government enjoys the rents from office R with probability $\pi(t, e)$ and pays a cost for repression equal to $k\theta t$ if $s = 1$ and θt otherwise. The opposition, on the other hand, obtains R with probability $1 - \pi(t, e)$ and pays a cost for its anti-regime efforts of $c(e) = be$.

[21] All results would be preserved if the government's type determined the value it places on office.

[22] Hirschleifer (1991) Skaperdas (1996)

3.1.1 Equilibrium

The game is solved through generalized backwards induction using the perfect Bayesian equilibrium concept. In the appendix we define an invertible function $\Psi(x)$ for any $x > 1$. Then we can characterize the unique semi-separating equilibrium in the following proposition:

Proposition 1. *If $k > \frac{3}{2}$ and $b > \frac{k}{3}\Psi^{-1}(k)$ then there exists a unique semi-separating equilibrium where for $\widetilde{\theta} = \Psi^{-1}(k)$*

- *If $\theta < \widetilde{\theta}$, $s(\theta) = 1$ and $e(1) = \frac{Rk\tilde{\theta}}{9b^2}$, $t(1,\theta) = \frac{R}{3b}\left[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}\right]$*
- *If $\theta > \widetilde{\theta}$, $s = 0$ and $e(0) = \frac{R}{9b^2}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)^2}{\left(1-\widetilde{\theta}\right)^2}, t(0,\theta) = \frac{R}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\left(\sqrt{\frac{1}{\theta}} - \frac{1}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\right)$*

Proof*: see appendix.*

In words, this equilibrium implies that, for any human rights treaty that makes repression sufficiently costly ($k > \frac{3}{2}$), and if the opposition faces sufficiently high marginal costs to anti-government efforts ($b > \frac{k}{3}\Psi^{-1}(k)$), relatively low-cost autocratic regimes will sign while relatively high-cost regimes will not. It is simple to see that these low-cost repressors would torture more heavily than high-cost governments absent the treaty. [23] *Contra* standard selection arguments, this equilibrium posits that it is precisely those regimes that are least likely follow the treaty's provisions absent any agreement that choose to accede to the human rights treaty. It is, however, consistent with the empirical evidence on the entry of authoritarian regimes into the CAT - those authoritarian regimes that torture more *ex ante* are more likely to sign.

The logic for this finding is straightforward. All autocratic governments seek to convince their opposition that they face a low cost to repression, as this will serve to reduce the level of effort the opposition will put into removing the autocrat. Signing a human rights treaty acts as a costly signal to the opposition of the government's low cost to repression. In equilibrium, the opposition will reduce the its anti-regime effort e if it observes that the government signs the treaty. This both benefits the government directly - as it faces a lower probability of removal from office $\pi(t, e)$

[23] We show in the appendix that the equilibrium level of torture absent a treaty, with some minor abuse of notation, is as follows $t(no\ treaty) = (\frac{R}{3b})(\frac{3b-\sqrt{\theta}}{3b\sqrt{\theta}})$ which is strictly declining in θ.

- and indirectly, as it can reduce its level of (costly) repression. However, whatever repression it continues to practice has become more costly for the government ($k > 1$).

The government must, therefore, weigh the costs of treaty accession against the benefits of lower opposition effort. If the penalty the treaty imposes on human rights violations is low ($k < \frac{3}{2}$), all autocrats pool on signing. If k goes to infinity, no government will sign (the threshold $\tilde{\theta}$ goes to zero making all types non-signers). For values of k between $\frac{3}{2}$ and infinity however, some governments choose to sign and others do not. Low-cost repressors benefit more from any reduction in opposition effort than high-cost repressors. A low marginal cost of repression θ implies that the government is highly responsive to any change in opposition effort levels. Thus, for any decline in e, the a low cost government reduces t more if θ is low than if it is high. Therefore, for any value of $k > \frac{3}{2}$, it is the low-cost governments ($\theta < \tilde{\theta}$) that benefit more from signing the treaty than high-cost ($\theta > \tilde{\theta}$) ones.

Using the equilibrium levels of repression practiced by the autocrat and effort exerted by the opposition, we can determine the probability of regime survival in equilibrium. We state these probabilities in the following Lemma 1:

Lemma 1. *In the semi-separating equilibrium, survival probabilities are given by the expressions* $\pi(t(1,\theta), e(1)) = 1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta}$ *for signatories and* $\pi(t(0,\theta), e(0)) = 1 - \frac{\sqrt{\theta}(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}$ *for non-signatories.*

Proof*: see appendix.*

This leads directly to the following result:

Proposition 2. *In the semi-separating equilibrium, signatories will survive (weakly) longer in office than non-signatories.*

Proof: $\pi(t(1,\underline{\theta}), e(1)) = 1 - \frac{k}{3b}\sqrt{\tilde{\theta}\underline{\theta}}\ \forall\ \underline{\theta} \in (0,\tilde{\theta})$; $\pi(t(0,\bar{\theta}), e(0)) = 1 - \frac{\bar{\theta}(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}\ \forall\ \bar{\theta} \in (\tilde{\theta}, 1)$. By contradiction: Consider $\underline{\theta} \in (0,\tilde{\theta})$ and $\bar{\theta} \in (\tilde{\theta}, 1)$, a signer and non-signer respectively. Assume the contrary: $\pi(t(1,\underline{\theta}), e(1)) < \pi(t(0,\bar{\theta}), e(0))$. Then $k\frac{\sqrt{\theta}}{\sqrt{\bar{\theta}}} > \frac{1-\tilde{\theta}^{\frac{3}{2}}}{\sqrt{\tilde{\theta}}(1-\tilde{\theta})} \Rightarrow F(\tilde{\theta})\frac{\sqrt{\theta}}{\sqrt{\bar{\theta}}} > F(\tilde{\theta}) \Rightarrow \sqrt{\underline{\theta}} > \sqrt{\bar{\theta}}$ which contradicts our initial assumption that $\underline{\theta} < \bar{\theta}$. □

The survival effect stems from two straightforward causes. First, it is evident from Proposition 1 that there exists a selection effect - autocrats who can repress more readily are more likely to sign the treaty than those for whom repression is costly. Second, the signing of the treaty conveys information to the domestic opposition, altering its optimal behavior. Signatory governments face less unrest than non-signatories[24].

The model predicts that signatories will reduce the level of torture they practice relative to a world where no treaty exists. It does not, however, yield clear predictions about the effect of the treaty on the level of torture practiced by non-signatories. Nor is it clear whether signatories will torture more or less than non-signatories when the option of signing a treaty is available. The intuition here is as follows: the treaty raises the cost of repression and, since it lowers the level of domestic opposition, renders such behavior less necessary. However, due to the selection effect discussed above, it is the low-cost repressors who sign the treaty, who are naturally prone to employ more draconian tactics. Thus, while signatories torture less than they would absent the treaty; comparisons between signatories and non-signatories yield ambiguous results. Note that the selection problem produced in measuring the effect of the treaty is precisely the opposite of that discussed by Downs, Rocke & Barsoom (1996) in regards to most international institutions.

Proposition 3. *Signatories of the treaty reduce torture levels, relative to a world without a treaty. It is ambiguous whether a given signatory will torture more or less than a given non-signatory.*

Proof: see appendix.

We should note that pooling equilibria exist in which no government signs the treaty, and pooling equilibria exist in which all states sign the treaty. However, when the costs of signing are large enough ($k > \frac{3}{2}$), the pooling equilibrium in which all states sign does not survive the intuitive criterion refinement (Cho and Kreps 1986). The equilibrium in which all states do not sign survives the intuitive criterion refinement.

Proposition 4. *When $k > \frac{3}{2}$ the pooling equilibrium in which all states sign the agreement does not survive the intuitive criterion refinement.*

[24] Equilibrium levels of effort are given by $e(1) = \frac{Rk\tilde{\theta}}{9b^2}$ and $e(0) = \frac{R(1-\tilde{\theta}^{\frac{3}{2}})^2}{9b^2(1-\tilde{\theta})^2}$ (see appendix). Since $\forall\, \tilde{\theta} \in [0,1]$, $\sqrt{\tilde{\theta}} < \frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}}$, it follows that the opposition always exerts less effort to remove signatory than non-signatory governments.

Proof: see appendix.

Hence when the expected costs of accession are large enough, there are two equilibria: one in which no states sign the agreement, and the semi-separating equilibrium in which the low cost torturers sign the agreement and the high cost torturers do not. When the costs of accession are low, the semi-separating equilibrium disappears, and we are left with the two pooling equilibria.[25]

3.1.2 Model robustness

In the Appendix, we demonstrate that semi-separating equilibria with similar properties exist for a number of alternative model specifications. Perhaps most significantly, we demonstrate that an analogous equilibrium exists when signing the treaty only results in punishment in the event the government steps down from office - i.e. punishments are strictly post-tenure. As is true when signing the treaty increases the government's marginal cost of repression, there exists a semi-separating equilibrium wherein only those governments that can repress at low cost sign (Proposition 5).

The application of post-tenure punishments does produce an additional effect not present in the baseline model. As the level of post-tenure punishment rises, the signatory governments grow increasingly willing to employ repression to remain in office and allocate more resources to torture in equilibrium (Proposition 6). We term this result the *commitment effect.* Knowing this, the opposition will devote less effort to removing signatory governments when post-tenure punishments are large. The difference in survival times between signatories and non-signatories is thus increasing in the level of post-tenure punishments.

In model extensions where post-tenure punishments are allowed, we also find that semi-separating equilibria will exist regardless of whether governments vary in their cost to repression (as in the baseline model) or in the value they place on office. In the appendix, we prove the existence of a semi-separating equilibrium in which governments vary in the value they place on office (and

[25] Note that as in the standard Spence (1973) signaling game, if there is perfect information about the type of the sender, no sender would incur the costs of signalling. In the Spence game, neither the high nor the low productivity workers would bother to spend money on education. Both types would pool on zero education. Here if there is perfect information about the type of regime, no regime would sign - all regimes would pool on not signing. Yet here, as in Spence, the presence of imperfect information makes incurring a cost to signal their type (for some types) worthwhile.

this is private information to the government). Only those governments that benefit greatly from remaining in power are willing to sign the treaty; those who benefit less are not so-willing to sign (Proposition 7). As is true in the baseline model, it is those governments willing to fight hardest to remain in power - and who thus practice the greatest levels of repression absent the treaty - who are most likely to sign.

3.1.3 Examples

In the empirical analysis that follows, we examine the empirical implications of the semi-separating equilibrium - noting that the equilibrium in which all states pool on not signing the CAT is clearly violated by the fact that some states actually do sign and others do not.

The semi-separating equilibrium predicts that (1) those authoritarian regimes that torture most heavily will be most likely to sign the CAT, (2) signatory regimes will survive longer in office than observationally similar non-signatories, and (3) CAT accession will result in a lowering in levels of torture.

These predictions - and the informational logic behind CAT accession - may seem counterintuitive. However, several examples of authoritarian regimes that sign the CAT fit this logic rather well. For instance, Chad became a CAT signatory on June 9, 1995. The Chadian regime - headed by Idriss Déby - faced extensive armed opposition at the time, which it repressed through the extensive use of torture.[26] The following year, the Déby regime unveiled a new constitution, which controversially granted sweeping powers to the presidency. This constitution was adopted on March 31, 1996 and presidential elections - in which there were reports of extensive irregularities - followed soon after.[27] According to our theory, Déby's decision to sign of the CAT acted as a signal to opposition forces of his intention to cling to power. Following Propositions 2 and 3, Déby would be predicted to survive in office and reduce torture levels, even as the opposition would be expected to reduce its efforts at bringing about his ouster.

In fact, Déby did remain in power following elections in 1996 and remains in power currently.

[26] In 1995, Chad had a value of 4 on Hathaway's (2007) 5 point torture scale, and a 3 on CIRI's (2007) 3 point scale. See also: James, Odhiambo. "Human Rights: Pattern Changes but Violations Continue. *Africa News.* August, 1995.

[27] 'Background Note: Chad.' US Department of State. Feb. 2009. `http://www.state.gov/r/pa/ei/bgn/37992.htm`

Torture levels in Chad declined in 1996 following accession to the CAT, as is in keeping with Proposition 3.[28] And, in 1997, several armed opposition groups ended their insurgency through negotiations with the government.[29]

Following a similar logic, many authoritarian regimes signed the CAT immediately following or preceding a transition of power.[30] For instance, the Museveni regime in Uganda and the signed the CAT in the year it assumed power.[31] Similarly, the Stevens government in Sierra Leone signed the CAT on March 18, 1985, immediately before handing power over to Stevens' chosen successor - Joseph Saidu Momoh - on November 28th of that year.[32] The theory predicts that CAT accession sends an informative signal of a regime's willingness to cling to office. Logically, the period immediately surrounding a change in the head of a regime would be a period of great uncertainty regarding the incoming elite's willingness and ability to so cling to power. This is also likely to be the period during which domestic opposition is particularly determined. As such, the informational value of signing the CAT is particularly great during transitional periods.

Of course, such results hardly constitute definitive evidence of the informational value of CAT accession, though they are suggestive. Indeed, case studies are unlikely to provide strong support for our theoretical claims. We, in essence, argue that authoritarian regimes sign the CAT as part of an effort to deter opposition groups from undertaking anti-regime activities. As is argued by Achen & Snidal (1989), the use of case study evidence is problematic for assessing the effectiveness of deterrence. Our preferred tests of our theory therefore consists of a large-N analysis examining the claim of Proposition 2 - that authoritarian CAT signatories survive longer in office than observationally similar non-signatories. We conduct such an analysis below.

[28] These ranked at a level of 2 on both the Hathaway and CIRI scales.

[29] 'Background Note: Chad.' US Department of State. Feb.2009.
`http://www.state.gov/r/pa/ei/bgn/37992.htm`

[30] In our sample, just under 15 percent of authoritarian regimes that joined the CAT did so in a year of transition. This was the most commonly observed time of CAT accession in our sample.

[31] Museveni assumed power on January 29, 1986 and the Museveni regime signed the CAT on November 3, 1986. Museveni currently remains in power.

[32] Momoh remained in power until April of 1992.

4 Empirics

We test the implications of Proposition 2 below. This proposition holds that authoritarian signatories will survive in office with higher probability than observationally similar non-signatories.[33] Empirically, therefore, one would expect CAT signatories to exhibit a lower hazard rate of regime failure than similar non-signatories. We use Cox proportional-hazards regressions on a cross-section of 129 authoritarian regimes during the 1985-1996 period to assess this claim.

The claim we test is *not* that signing the CAT causes an authoritarian regime to survive longer in office. Our theoretical prediction is that signatories face a lower probability of removal due, in part, to a selection effect. Signatories face a systematically lower cost to repression than non-signatories. Signatories also benefit from the revelation of information to opposition groups. Precisely because only low-cost repressors choose, in equilibrium, to sign the CAT, these regimes face a less restive domestic opposition. Domestic opposition groups are less willing to exert costly effort against regimes that reveal that they are low cost types by signing the CAT. Proposition 2 holds as a result of the cumulative force of these selection and information effects. Our empirics do not differentiate between the effect of selection and that of information.[34]

4.1 Data

To test Proposition 2, we use Vreeland's (2008) dataset on CAT accession, the Archigos database on political leaders and regime survival [35], and Gandhi and Przeworski's (2007) data on the longevity of authoritarian regimes. Since most of the variables in the theoretical model are characteristics of the regime and are relatively time-invariant, the unit of observation in our dataset is the regime.[36] In total, our dataset comprises 129 authoritarian regimes in the period 1985-1996.

- The dependent variable of the survival model is **sumten** - the sum total number of days

[33] Formally, $\pi(t(1,\underline{\theta}),e(1)) > \pi(t(0,\bar{\theta}),e(0)) \; \forall \; \underline{\theta} \in (0,\tilde{\theta}), \; \bar{\theta} \in (\tilde{\theta},1)$.

[34] It is, however, possible in principle to test the information effect separately from the selection effect. If our theory is correct, one would anticipate that signing the CAT should have a causal effect on levels of domestic unrest. Such a test is beyond the scope of the present work.

[35] Goemans (2006)

[36] Here 'regime' is defined as a single leader's tenure in the Archigos dataset. In this our analysis differs from both Hathaway's and Vreeland's. Both of these authors use country-year observations to predict the hazard rate for signing the CAT in a given country during a given year. We run an analogous model as a robustness check in Section 4.3.4

served in office - as taken from the Archigos dataset.

- The relevant explanatory variable is **eversign**, a binary indicator that takes a value of 1 if the regime in question was ever a signatory of the CAT.[37]

The control variables in the survival models we estimate are taken from Gandhi and Przeworski. They include the following:

- **resource** a binary indicator that takes the value 1 if the ratio of mineral exports to total exports exceeds 0.5. This export ratio likely affects the rents available to the autocrat from office and the potential rents that the opposition would gain from taking office. Greater ratios of mineral exports should be associated with greater rents. In the Przeworski & Gandhi dataset, this variable is time-invariant for each regime.

- **military** and **civilian** binary indicators that take the value 1 if the autocracy is a military regime or a civilian regime. The excluded category is a monarchy.

- **inherit** a binary indicator that takes the value 1 if the current autocrat inherited an organized opposition party.

- **acchead** the number of changes of regime experienced by the state under authoritarian control.

In addition to the above, we make use of the following controls:

- **tort** - torture levels. We use two measures of torture: one coded by Hathaway (2007) the other by Cingranelli and Richards (2007). Both variables are the same as used by Hathaway and Vreeland. The former is an ordinal index running from one to five, the latter an ordinal index from 1 to 3, with increasing values denoting a more widespread use of torture. We use the mean level of this variable for each regime in our dataset as a regressor.
- **gdpcap** - GDP *per capita*, in thousands of 1995 dollars, measured in PPP from the Penn World Tables 6.1. We use the mean level of this variable for each regime in our dataset as a regressor. We expect this variable to be associated with the returns to holding office.

- **pop** - population levels, in millions. We use the mean level of this variable for each regime in our dataset as a regressor. We expect this variable to be associated with the returns to holding office.

[37] This variable is coded based upon data made available by James Vreeland (2008) and initially coded by Oona Hathaway (2007)

- **cinc** - military capabilities index as constructed by the Correlates of War[38]. We use the mean value, by regime, of this variable over the 1985-1996 period. This variable is meant to proxy for the military capabilities of the regime.

- **communist** - an indicator variable for communist regimes taken from Vreeland.

4.2 Cox Estimates

To test the association between the signing of the CAT and the survival time of regimes, we first run a Cox proportional hazards model of the probability of regime failure. The Cox model provides an estimate of the hazard rate of a given regime i (the probability regime i collapses at time t given that it has survived until time t) conditional upon observed covariates: $h_i(t) = h_o(t)e^{X_i\beta}$, where $h_o(t)$ is the baseline hazard function. As our theory makes no predictions regarding the effect of time-in-office on regime survival, we use the semi-parametric Cox model that imposes no assumptions about the functional form of the hazard function.[39] Time is effectively treated as a nuisance parameter. The results of this regression are reported in Table 1.

The controls include the variables identified as significant to the survival of authoritarian regimes by Gandhi & Przeworski (2007), as well as an indicator of whether the regime was ever a signatory to the CAT, and variables relevant to the signing the CAT. As can be readily seen in Table 1, regimes that are signatories of the CAT have lower hazard rates than those that do not.[40][41] This difference is apparent even after controlling for other factors related to regime survival and for factors related to CAT accession.

The magnitude of this difference in hazard rates between signatories and non-signatories is difficult to interpret from the coefficients reported in Table 1. The Cox model is a non-linear specification, so the coefficient values can not be translated directly into marginal effects. The

[38] Correlates of War, National Material Capabilities v. 3.02 project (Singer 1987)

[39] The Cox model does assume hazard rates are proportional - that the shape of the hazard function does not differ across units. This assumption is discussed further below.

[40] The reported coefficients are not hazard ratios. A coefficient of zero implies that the variable in question has no effect on the hazard rate, negative coefficients imply that an increase in the variable reduces the hazard rate, and positive coefficients imply the reverse.

[41] We have also run models wherein the **eversign** variable is interacted with the indicators for regime-type. These estimates do not indicate any significant difference in the relationship between CAT accession and survival between military and non-military regimes.

difference between signatories and non-signatories is, therefore, best captured graphically and can be seen in Figure 1. This figure includes the hazard rates for both model specifications, with controls alternatively for Hathaway's and CIRI's measures of torture.

[Table 1 about here.]

[Figure 1 about here.]

These results are consistent with Proposition 2 above. Controlling for a number of observable covariates, signatory regimes face a lower hazard of removal than non-signatories.

Since we do not employ Heckman-type selection models to control for the governments' choice to sign - or not to sign - the CAT, these results cannot be interpreted causally. The negative relationship between regime hazard rates and CAT accession may be a product of selection effects. But, these effects are anticipated in our theory. Proposition 2 posits a simple binary relationship between the survival times of signatory and non-signatory regimes. These empirical specifications find such a relationship in the data.

4.2.1 The Treatment of Inherited Signatory Status

The above results pertain to *all* authoritarian signatories during the 1985-1996 period, including those regimes that inherited their signatory status from predecessor governments. Our theory, however, most directly pertains to those regimes that did not inherit signatory status. While governments who are signatories of the CAT as a result of their predecessors' decision to sign may choose to renounce the treaty; it may be plausibly argued that the signal sent by their decision to remain in or exit the CAT qualitatively differs from that of a regime that decides to accede of its own volition.[42] If this is true, the results above may be biased by the inclusion of such signatories. Indeed, if regimes that inherit their signatory status are particularly reticent to exit the treaty, the above results understate the association between CAT accession and regime survival.

To address this issue, we rerun the above analysis dropping all regimes that inherit their signatory status from the sample. This reduces the total size of our sample by 21 regimes, and drops

[42] We would like to thank an anonymous referee for raising this point.

the total number of signatories from 55 to 34. As can be seen in Table 2, when the empirical model above is run on the reduced sample, the coefficient on the **eversign** variable increases in magnitude and is significant at the 99 percent level across all specifications, despite the loss of statistical power that results from dropping observations. This finding is consistent with the suggestion that regimes that inherit their signatory status may be disinclined to exit the CAT. It also lends further support to Proposition 2.

[Table 2 about here.]

4.3 Robustness Checks

In addition to the empirical models discussed above, we have run a number of robustness checks to ensure that our findings are not spurious and that our models are not misspecified. We test the proportional hazards assumption of the Cox model to ensure that the shape of the baseline hazard function does not vary systematically by authoritarian regime-type.[43] We also preprocess our data using propensity score matching as advocated by Ho et al. (2007) to improve covariate balance. We restructure our data to estimate the effect CAT accession using regime-years as the unit of analysis. And we re-run our estimates after dropping particularly fragile regimes from our sample. These robustness checks are discussed in greater detail below. Results from these checks are reported in Tables 4 and 5.

4.3.1 Testing the Proportional Hazards Assumption

As discussed above, the Cox model assumes that the proportional hazards property holds. Any change in covariate values shifts the baseline hazard function $h_o(t)$ up or down. This assumption is quite strong for our sample. In particular, it seems quite plausible that shape of the hazard function may vary across authoritarian regime-types.[44] To ensure that our models are not misspecified,

[43] We have also run Weibull models as an alternative to the Cox. Unlike the Cox model, the Weibull makes parametric assumptions regarding the hazard function. It particularly assumes that this function is monotonic. Because we have little reason to support this assumption, we prefer the Cox estimates. Weibull estimates are of the same sign and of a similar magnitude to the Cox. However, they are not significant at conventional levels. Results are available from the authors on request.

[44] We would like to thank Bruce Bueno de Mesquita for raising this concern.

we therefore run Grambsch and Therneau and Harrell's rho tests of the proportional hazards assumption (for a discussion of these tests, see Box-Seffensmeier & Jones 2004).

To perform these tests, we extract the Schoenfeld residuals from the **Hath1** and **CIRI1** models above. These residuals are roughly equivalent to the observed minus the expected coefficient values at each failure time [45]. Systematic deviation of residual values from zero may indicate that the proportional hazards assumption is inappropriate. The Grambsch and Therneau test uses the absolute cumulative sum of these residuals to test the hypothesis that the proportional hazards property does not hold for the model as a whole against the null that it does. The Harrell's rho runs similar hypothesis tests for each covariate. It is based on the correlation between covariate and residual values. The results of the global tests and the covariate tests for regime-type are reported in Table 3. The tests do not reject the proportional hazards assumption.

[Table 3 about here.]

4.3.2 Matching Estimates

As an additional robustness check, we pre-process our data using propensity score matching. Matching helps to ensure covariate balance and overlap between treatment (signatory) and control (non-signatory) groups [46]. If signatories and non-signatories differ substantially on observed covariates, coefficient estimates may depend strongly on functional form assumptions of the empirical model. This danger is particularly great in this instance, as regimes select into signatory or non-signatory status and are thus likely to differ on observed covariates. Pre-processing matches signatories with non-signatories that are predicted to be similarly likely to sign. Any signatories or non-signatories for which a match does not exist are dropped from the dataset. Such a process helps to ensure covariate balance and limit the dependence of results on functional form assumptions [47].

To ensure covariate balance in the data, we first estimate the probability a given regime is a signatory of the CAT using a logit regression. (Results from this estimation are reported in Table 6 in the Appendix.) Regimes that signed the CAT are then matched with a control that had a similar

[45] Box-Seffensmeier & Jones (2004)

[46] Gelman & Hill (2006) Morgan & Winship (2007)

[47] Ho et al. (2007)

ex ante probability of accession. Matching is performed using Ho et al.'s MatchIt (2004) program, run from R 2.5.1. We use nearest neighbor matching (without replacement) using mahalanobis distance measures. Matching is one-to-one, so that each matched signatory is paired with one unique non-signatory.

The models labeled **Hath 1** and **CIRI 1** in Table 6 are used to produce the propensity scores on which matching is conducted. Scatter plots of these scores, differentiating between matched and unmatched treatment and control cases are presented in Figure 2 in the Appendix. The **Hath 1** model matches 55 treated cases to 55 controls; the **CIRI 1** model does the same. 19 control units are unmatched in each instance.

We run a Cox proportional hazards estimate of the effect of signing the CAT on regime survival using the resultant matched datasets. Coefficient estimates are reported in row 1 of Table 4.

[Table 4 about here.]

In both cases, the coefficient on the **eversign** variable is of the same sign and of a similar magnitude to those in the basic Cox models. When the propensity scores are generated using Hathaway's measure of torture, the **eversign** coefficient is negative and significant at the 95 percent level. When using CIRI's measure, it is significant at the 90 percent level.[48] Thus, in both the naive and matching estimates, signing the CAT is associated with lower regime hazard rates - supporting of the Proposition 2.

4.3.3 Dropping Fragile Regimes

It may be argued that our results are driven not by the causal mechanisms suggested in our model, but rather by the inclusion of particularly weak regimes in our dataset. Governments that struggle to survive for even a short time may avoid signing the CAT not because of the potential cost of punishment under its provisions, but rather because they lack the capacity to sign *any* treaty.

[48] Using nearest neighbor matching with a caliper of 0.5 standard deviations or the genetic matching method of Diamond and Sekhon (2006) generates similar results. The magnitude and sign of the **eversign** coefficient is similar in all models. These also resemble the results found in the basic Cox estimation. However, this coefficient is not significant at conventional levels when either caliper or genetic methods are used. This may be due to the fact that both caliper and genetic methods discard more observations than the one-to-one matching used here, leading to a loss of statistical power. Results using these alternate matching methods are available from the authors on request.

Such regimes may devote all their attention and resources to directly dealing with domestic threats. Since our pool of non-signatories may include such fragile regimes, we may find that non-signatories survive for shorter periods simply due to these outliers.[49]

To eliminate this possibility, we re-run our estimates after dropping all regimes that survive for less than 365 days and all regimes that survive less than 730 days from our sample. The results from these regressions are reported in rows 2 and 3, respectively, of Table 4. In both instances, the coefficient on **eversign** remains negative and has approximately the same magnitude as in our main results. These coefficient estimates are significant at the 10 percent level (as opposed to at the 5 percent level for our main results). However, one would expect the smaller sample size to lead to a loss of precision in our estimates.

4.3.4 Multiple Record Survival Analysis

In the empirical results so far, the unit of observation has been the regime. Our analyses compare the survival time of signatory regimes to observationally similar non-signatory regimes. These analyses are in keeping with the results of Proposition 2 and allow us to employ matching techniques to ensure covariate balance between signatory and non-signatory regimes.

However, our theory suggests that one should expect the hazard rate of a regime that signs the CAT at time t to fall in time $t+1$.[50][51] To test this hypothesis, we analyze the survival time of a panel of authoritarian regimes[52] between 1985 and 1996 where the unit of analysis is the regime-year. We employ the semi-parametric Cox proportional hazards model used above, adjusting for the presence of multiple (annual) observations for each regime. Our variable of interest is **cats_lag**, a binary indicator variable that takes the value of 1 in all years following that of CAT accession. We assess the conditional relationship between this variable and the regime hazard rate - the probability that the regime fails in year t given that it survived until year t.

Coefficient estimates for this model are reported in Table 5. The variable **growth** measures

[49] We would like to thank Jon Eguia for raising this issue.

[50] We would like to thank an anonymous reviewer for raising this issue.

[51] Strictly speaking, our model does not incorporate dynamics. However, this hypothesis is a reasonable extension of our claims.

[52] We only consider regimes that did not inherit signatory status from predecessor governments

the growth rate in *per capita* income in year t, **gdp_per_cap** measures GDP *per capita* using PPP dollars from the Penn World Tables 6.1, while **war** is a binary indicator that takes the value 1 if the regime is involved in a war. The **Hath** model controls for the Hathaway (2007) torture index, while the **CIRI** model controls for the CIRI (2007) torture measures.

As can readily be seen, the coefficient on the variable of interest **cats_lag** is negative in both models and significant, at the 90 percent level in the **Hath** model and at the 95 percent level in the **CIRI** model. This implies that governments that sign the CAT in year t experience lower hazard rates in year $t + 1$, in line with the predictions of the model above.

[Table 5 about here.]

5 Conclusion

The stylized facts on accession to the CAT are these: autocracies that torture are more likely to sign the CAT than those that do not; autocracies that sign the CAT continue to torture; overall the CAT appears to have reduced torture levels in signatory states; and autocratic torturers that sign the CAT survive longer in office than those that don't.

The model presented in Section 3 is consistent with and predicts all these findings. If authoritarian governments use the signing - and violation - of human rights agreements as a signal of their willingness to repress domestic opponents, our model predicts that those regimes that practice repression *ex ante* are most likely to sign. Moreover, in equilibrium, those states that sign continue to torture after accession. The informational effect acts as a threat: signing the treaty signals strength and a willingness to torture if necessary. A rational opposition reduces its political activity in response, and since torture and opposition effort are strategic complements, torture levels in these states are predicted to fall. So while the treaty is signed with an intent to defy it, torture levels do fall relative to what they would have been absent accession. They do not, however, go to zero.

The model is quiet as to whether torture rises or falls in non-signatory states before and after the CAT. But overall, the model predicts that aggregate torture levels will be no higher in signatory states after the CAT than they were before. This suggests an explanation for the finding that while signers continue to torture; overall human rights performance may in fact be improving over time.

The consequence of the effect of signing the CAT on domestic political action also has implications for leadership survival in autocracies. The signing by a strong state leads to less opposition and resistance, and a reduction in the likelihood of revolt or revolution. Hence autocratic signatories to the CAT are predicted to have longer tenures in office.

The CAT therefore presents an opportunity for authoritarian regimes that value office very highly to signal their intent to hold firmly onto that office. The signal associated with signing the agreement is informative in the sense that accession acts as a credible threat of a willingness to exert effort and incur high costs in order to remain in power. While the treaty designers (in the developed world, at least) probably had no intention for the treaty to play this role, hard-core authoritarian regimes appear to have taken advantage of this mechanism. This suggests, of course, that these autocrats have an incentive to try to find other mechanisms to signal their type in a credible fashion. While such signalling devices may exist - excessive crackdowns on opposition, detentions and disappearances of opponents, etc. - few are credible in their capacity to separate out types. All autocrats appear to pool (at least to an extent) on the use of repression. But, so long as some a degree of uncertainty over governments' willingness to employ repression exists, low-cost types will have an incentive to signal their status. Thus, they will resort to a variety of signaling mechanisms, including - it seems - accession to the CAT.

The mechanism we identify here is also consistent with the scholarship in international relations that stresses the useful informational role that can be played by international organizations (IOs). The usual view is that IOs generate information that helps the member states to engage in heightened cooperation and to coordinate common expectations, and that IOs provide focal points in environments with multiple equilibria. Here we identify an informational role that may be detrimental to the goal of alleviating torture - the aim of the treaty being signed. The conclusion we draw from this analysis is that if there is an incentive to credibly signal on the part of the autocrat, then the CAT offers a mechanism for such signaling. While other informational mechanisms may exist, it seems clear that the CAT plays this role.

What then to make of the CAT? While it may reduce torture in the most autocratic of states; those states sign precisely because they intend to continue torturing. Rather than failing to have any positive effect on those states that need the most help, it has countervailing effects: torture levels do fall, if only slightly, but those regimes that sign become more secure. The long-term effect of the CAT on levels of torture is, therefore, ambiguous. It is far from clear therefore whether the CAT improves societal welfare by reducing signatories' propensity to torture, or if it lowers welfare by increasing length in office of repressive autocratic regimes. The good intentions of the international community may have the unintended consequence of strengthening undemocratic regimes around the world.

We also have identified a new mechanism linking compliance costs to accession. Standard accounts suggest that the lower a country's cost of compliance, the greater its probability of treaty accession. In this model, the prediction is quite the opposite. The higher are the costs of compliance, the more likely an autocratic torturer is to sign the agreement. For the higher are the costs, the greater is the signaling value of the treaty in the contest between the elite and the domestic opposition. This result may have more general implications for theories that attempt to deal with the selection problem inherent in studies of international institutions. Compliance costs may have different effects than those widely postulated in the existing literature.

The findings of this research suggest an under-appreciated element of international institutional design. Agreements that focus on the nation state as a unitary actor, and ignore the effect of the institution on domestic politics - and in particular domestic conflict - may generate unanticipated and adverse effects. Policymakers, engaged in negotiations at the international level over the design of international institutions need to anticipate the effect of these agreements on the domestic polity. Agreements may come into effect exactly because they bolster the political survival of those leaders that sign them. When these leaders are autocratic, it is likely that they will use participation in international agreements to help prevent democratic reform.

How is possible that such a simple model yields such counterintuitive results? We believe this is a consequence of taking two aspects of international politics more seriously. First, domestic politics matter when it comes to a state's decision to accede or not to an international obligation. Second, international institutions generate information (if only by states' accession) that will affect the political calculus of the domestic groups engaged in political action. By combining the information generated by the international institution and an explicit political contest at the domestic level, we generate results that depart somewhat from the standard canon: countries may accede to treaties they intend to defy. Moreover, the international human rights regime may be extending the survival in office of the most autocratic torturers and delaying democratic reforms.

References

Achen, Christopher H. & Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41(2):143–169.

Bagwell, Kyle & Robert Staiger. 2005. "Enforcement, Private Political Pressure and the GATT/WTO Escape Clause." *The Journal of Legal Studies* 34(2):471–513.

Box-Seffensmeier, Janet M. & Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press.

Chayes, Abram & Antonia Handler Chayes. 1993. "On Compliance." *International Organization* 47(2):175–205.

Cingranelli, David L. & David L. Richards. 2007. "The Cingranelli-Richards (CIRI) Human Rights Dataset.".

Diamond, Alexis & Jasjeet S. Sekhon. 2006. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.".

Downs, George W. & David M. Rocke. 1995. *Optimal Imperfection?* Princeton University Press.

Downs, George W., David M. Rocke & Peter N. Barsoom. 1996. "Is the Good News About Compliance Good News About Cooperation?" *International Organization* 50(3):379–406.

Finnemore, Martha. 1996. *National Interests in International Society, Cornell Studies in Political Economy.* Cornell University Press.

Fudenberg, Drew & Jean Tirole. 1991. *Game Theory*. The MIT Press.

Gandhi, Jennifer & Adam Przeworski. 2007. "Authoritarian Institutions and the Survival of Autocrats." *Comparative Political Studies* 40:1279–1301.

Gelman, Andrew & Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods in Social Science Cambridge University Press.

Gilligan, Michael J. & Nathaniel H. Nesbitt. 2007. "Do Norms Reduce Torture?".

Goemans, Hein. 2006. Archigos: A Database on Political Leaders. Paper Presented at the Annual Meeting of the American Political Science Association.

Goodliffe, Jay & Darren G. Hawkins. 2006. "Explaining Commitment: State and the Convention Against Torture." *Journal of Politics* 68:358–371.

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

Hafner-Burton, Emilie M. & Kiyoteru Tsutsui. 2005. "Human Rights in a Globalizing World: The Paradox of Empty Promises." *American Journal of Sociology* 110(5):1373?1411.

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

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

Hirschleifer, Jack. 1991. "The Paradox of Power." *Economics and Politics* 3:177–200.

Ho, Daniel E., Kosuke Imai, Gary King & Elizabeth A. Stuart. 2007. "Matching as Preprossesing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* (forthcoming).

Ho, Daniel, Kosuke Imai, Gary King & Elizabeth Stuart. 2004. "MatchIt: Matching as Nonparametri Preprocessing for Parametric Causal Inference.". http://gking.harvard.edu/matchit/.

Jonas, Stacie. 2004. "The Ripple Effect of the Pinochet Case." *Human Rights Brief* 11(3):36–38.

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

Koremenos, Barbara. 2005. "Contracting Around International Uncertainty." *The American Political Science Review* 99:549–565.

Moravcsik, Andrew. 2000. "The Origins of Human Rights Regimes: Democratic Delegation in Postwar Europe." *International Organization* 54(2):217–252.

Morgan, Stephen L. & Christopher Winship. 2007. *Counterfactuals and Causal Inference.* Cambridge University Press.

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

Powell, Emilia Justyna & Jeffrey K. Stanton. 2009. "Domestic Judicial Institutions and Human Rights Treaty Violation." *International Studies Quarterly* 53(1):149–174.

Roht-Arriaza, Naomi. 2001. "The Pinochet Precedent and Universal Jurisdiction." *New England Law Review* 35(2):311–320.

Rosendorff, B. Peter. 2005. "Stability and Rigidity: Politics and the Design of the WTO's Dispute Resolution Procedure." *The American Political Science Review* 99(3):389–400.

Rosendorff, B. Peter & Helen V. Milner. 2001. "The Optimal Design of International Trade Institutions: Uncertainty and Escape." *International Organization* 55(4):829–857.

Simmons, Beth A. Forthcoming, 2009. *Mobilizing for Human Rights: International Law in Domestic Politics.* Princeton University Press.

Singer, J. David. 1987. “Reconstructing the Correlates of War Dataset on Material Capabilities of State, 1816-1985.” *International Interactions* 14:115–132.

Skaperdas, Stergios. 1996. “Contest Success Functions.” *Economic Theory* 7:283–290.

Vreeland, James Raymond. 2008. “Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture.” *International Organization* 62:65–101.

A Formal Proofs

A perfect Bayesian equilibrium of a signaling game consists of a strategy profile and a system of beliefs such that, (1) the sender chooses her strategy to maximize her utility subject to the receiver's strategy; (2) the receiver chooses her strategy to maximize her utility subject both to the sender's strategy and to her beliefs conditional upon the sender's message; and (3) the receiver's beliefs are updated according to Bayes' Rule, whenever possible [53].

Definition: Define a pair of strategies $\{(s,t),e\}$ where $s:[0,1]\rightarrow\{0,1\},\ t:\{0,1\}\times[0,1]\rightarrow\mathbb{R}_+,\ e:\{0,1\}\rightarrow\mathbb{R}_+$.

Definition: Define a function $\Psi(x)=\frac{1-x^{\frac{3}{2}}}{(1-x)\sqrt{x}}$. Note that since $\Psi(\cdot)$ is monotonic and decreasing for $x>0$, it is invertible. Denote the inverse of $\Psi(x)$ as $\Psi^{-1}(\cdot)$.

Proof of Proposition 1:

The proposition states that if $k>\frac{3}{2}$ and $b>\frac{k}{3}\Psi^{-1}(k)$ then there exists a unique semi-separating equilibrium where for $\widetilde{\theta}=\Psi^{-1}(k)$, where $s(\theta)=\begin{cases}1 & if\quad \theta<\widetilde{\theta}\\ 0 & if\quad \theta>\widetilde{\theta}\end{cases}$ and $e(s)=\begin{cases}\frac{Rk\tilde{\theta}}{9b^2} & if\quad s=1\\ \frac{R}{9b^2}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)^2}{\left(1-\widetilde{\theta}\right)^2} & if\quad s=0\end{cases}$ and $t(s,\theta)=\begin{cases}\frac{R}{3b}\left[\sqrt{\frac{\tilde{\theta}}{\theta}}-\frac{k\tilde{\theta}}{3b}\right] & if\quad \theta<\widetilde{\theta}\\ \frac{R}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\left(\sqrt{\frac{1}{\theta}}-\frac{1}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\right) & if\quad \theta>\widetilde{\theta}\end{cases}$. We prove this first by checking, given

[53] Fudenberg & Tirole (1991)

any signal, that each player is playing a best response. Then we check that given this behavior, no type has an incentive to send another signal. Thirdly, we specify the conditions for the threshold type $\widetilde{\theta}$ to be interior to the type space. Finally we establish uniqueness of the semi-separating equilibrium.

Firstly, suppose $s = 1$, $\frac{\partial U_G}{\partial t} = \pi_t(t,e)R - k\theta = 0$ yields a reaction function $t(e,1,\theta) = \sqrt{\frac{eR}{k\theta}} - e \Rightarrow t(\frac{Rk\theta}{9b^2}, 1, \theta) = \frac{R}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}]$. Therefore, autocratic signatory governments are playing a best response to their opposition when $t(1,\theta) = \frac{R}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}]$.

Suppose $s = 0$, $\frac{\partial U_G}{\partial t} = \pi_t(t,e)R - \theta = 0$ yields a reaction function $t(e,0,\theta) = \sqrt{\frac{eR}{\theta}} - e \Rightarrow t(\frac{R(1-\tilde{\theta}^{\frac{3}{2}})^2}{9b^2(1-\tilde{\theta})^2}, 0, \theta) = \frac{R(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}(\sqrt{\frac{1}{\theta}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})})$. Therefore, non-signatory autocratic governments are playing a best response to their opposition when $t\left(0,\theta\right) = \frac{R(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}(\sqrt{\frac{1}{\theta}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})})$.

The opposition's problem, if the opposition observes $s = 1$ is to maximize $EU_D = \int_0^{\tilde{\theta}}[(1 - \pi(t(e,1),e))R - be]f(1)d\theta$, where $f(1)$ is the posterior distribution of θ over updated support $\left[0, \tilde{\theta}\right]$, conditional on signal $s = 1$.

From the reaction function above, $\pi(t(e,1),e) = \frac{t(e,1,\theta)}{t(e,1,\theta)+e(1)} = 1 - \sqrt{ek\theta/R}$. Then:

$$\frac{\partial U_D}{\partial e} = \int_0^{\tilde{\theta}} [\frac{1}{2}\sqrt{\frac{k\theta R}{e}} - b]f(1)d\theta = 0$$
$$\Leftrightarrow e(1) = \frac{Rk\widetilde{\theta}}{9b^2}$$

Similarly, if the opposition observes $s = 0$, $EU_D = \int_{\tilde{\theta}}^{1}[(1 - \pi(t(e,0),e))R - be]f(0)d\theta$, where $f(0)$ is the posterior distribution of θ over updated support $\left[\tilde{\theta}, 1\right]$, conditional on signal $s = 0$. From the reaction function above, $\pi(t(e,0),e) = 1 - \sqrt{e\theta/R}$. Then:

$$\frac{\partial U_D}{\partial t} = \int_{\tilde{\theta}}^{1} [\frac{1}{2}\sqrt{\frac{R\theta}{e}} - b]f(0)d\theta$$
$$\Leftrightarrow e(0) = \frac{R(1-\tilde{\theta}^{\frac{3}{2}})^2}{(9b^2)(1-\tilde{\theta})^2}$$

In both cases the opposition is playing a best response to the government's action in the equilibrium specified.

Secondly, we must check whether autocratic governments of type $\theta < \tilde{\theta}$ have an incentive to deviate by setting $s = 0$, or if governments of type $\theta > \tilde{\theta}$ have an incentive to deviate by setting $s = 1$.

Consider a type $\theta \in (0, \tilde{\theta})$. If $s = 1$, $U_G = \pi(t(1,\theta), e(1))R - k\theta t(1,\theta) = (1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})R - \frac{Rk\theta}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}] = R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. If, on the other hand, $s = 0$, $U_G = \pi(t(0,\theta), e(0))R - \theta t(0,\theta) = R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. Defection then occurs iff $\pi(t(1,\theta), e(1))R - k\theta t(1,\theta) < \pi(t(0,\theta), e(0))R - \theta t(0,\theta) \Rightarrow R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2 < R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. This inequality holds iff $k > \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$or iff $\theta > \Psi^{-1}(k) = \tilde{\theta}$. But $\theta < \tilde{\theta}$, so there is no incentive to defect.

Consider a type $\theta \in (\tilde{\theta}, 1)$. If $s = 0$, $U_G = \pi(t(0,\theta), e(0))R - \theta t(0,\theta) = R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. If, on the other hand, $s = 1$, $U_G = \pi(t(1,\theta), e(1))R - k\theta t(1,\theta) = R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. Defection takes place iff $\pi(t(0,\theta), e(0))R - \theta t(0,\theta) < \pi(t(1,\theta), e(1))R - k\theta t(1,\theta) \Rightarrow R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})}^{2}) < R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. This inequality holds iff $k < \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$ or iff $\theta < \Psi^{-1}(k) = \tilde{\theta}$. But $\theta > \tilde{\theta}$, so there is no incentive to defect.

Thirdly, for $\tilde{\theta} = \Psi^{-1}(k)$ to be well defined, that is to lie in range $(0, 1)$, we require $k > \frac{3}{2}$ and for $1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta} > 0$ and $1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})} > 0$ we require $b \geq \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3(1-\tilde{\theta})}$ for all $\theta \in (\tilde{\theta}, 1)$ and $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$ for all $\theta \in (0, \tilde{\theta})$. Now, since $k = \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$, $\frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3(1-\tilde{\theta})} = \frac{k}{3}\sqrt{\tilde{\theta}\theta}$. Then $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$ for all θ. Then $b \geq \frac{k}{3}\sqrt{\tilde{\theta}, \theta}$ for all $\theta \in (0, \tilde{\theta})$ implies $b \geq \frac{k}{3}\tilde{\theta} = \frac{k}{3}\Psi^{-1}(k)$. Hence the conditions for the equilibrium are satisfied.

Finally note that since $\Psi(\cdot)$ is monotonic, there is a unique threshold type $\tilde{\theta}$, such that all types below $\tilde{\theta}$ sign the treaty and those above do not; moreover, we have fully specified the only possible out-of-equilibrium beliefs. Hence the semi-separating equilibrium is unique. □

Proof of Lemma 1:

$t(e, 1, \theta) = \sqrt{\frac{eR}{k\theta}} - e$; $\pi(t(e, 1, \theta), e(1)) = 1 - \sqrt{ek\theta/R} = 1 - \sqrt{\frac{Rk\tilde{\theta}}{9b^2}k\theta/R} = 1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta}$.
$t(e, 0, \theta) = \sqrt{\frac{eR}{k\theta}} - e$; $\pi(t(e, 0, \theta), e(0)) = 1 - \sqrt{e\theta/R} = 1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\tilde{\theta}}}{3b(1-\tilde{\theta})}$. □

Proof of Proposition 3:

$t(notreaty) = \frac{R}{3b}(\frac{3b-\sqrt{\theta}}{3b\sqrt{\theta}})$. Therefore, the existence of the treaty reduces torture levels amongst signatories iff $t(notreaty) > t(e, 1, \theta)$ for $\theta \in (0, \tilde{\theta})$. This implies $\frac{3b-\sqrt{\theta}}{3b\sqrt{\theta}} > \sqrt{\tilde{\theta}}[\sqrt{\frac{1}{\theta}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})}]$. Simple algebra yields: $3b(1 - \tilde{\theta}) - \sqrt{\theta}(1 - \tilde{\theta}) > \sqrt{\tilde{\theta}}[3b(1 - \tilde{\theta}) - \sqrt{\theta}(1 - \tilde{\theta}^{\frac{3}{2}})]$.

We know from proposition 1 that $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$, implying that when b is at its minimal value we have $k\sqrt{\tilde{\theta}\theta}(1-\tilde{\theta}) - \sqrt{\theta}(1-\tilde{\theta}) > k\tilde{\theta}\sqrt{\theta}(1-\tilde{\theta}) - \sqrt{\tilde{\theta}\theta}(1-\tilde{\theta}^{\frac{3}{2}})$. Substituting in $k = \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$, yields $(1-\tilde{\theta}^{\frac{3}{3}})\sqrt{\theta} - (1-\tilde{\theta})\sqrt{\theta} > (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\tilde{\theta}\theta} - (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\tilde{\theta}\theta} = 0$, which holds for all $\tilde{\theta} \in (0,1)$.

It now remains to examine the inequality for values of $b > \frac{k}{3}\sqrt{\tilde{\theta}\theta}$. Note that the derivative of the LHS of the inequality with respect to b is given by $3(1-\tilde{\theta})$. The derivative of the RHS is given by $3\sqrt{\tilde{\theta}}(1-\tilde{\theta})$. Now $3(1-\tilde{\theta}) > 3\sqrt{\tilde{\theta}}(1-\tilde{\theta})$ for all $\tilde{\theta} \in (0,1)$ implying that the inequality holds for all b. Thus, the treaty reduces torture levels amongst signatories.

Repeating the same process for non-signatories yields the following inequality $t(notreaty) > t(0,\theta;e(0))$ iff $\frac{3b-\sqrt{\theta}}{3b\sqrt{\theta}} > \frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}}[\frac{1}{\sqrt{\theta}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})}]$. This inequality reduces to $(1-\tilde{\theta})^2(3b-\sqrt{\theta}) > (1-\tilde{\theta}^{\frac{3}{2}})[3b(1-\tilde{\theta}) - (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}]$. We know from proposition 1 that $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$. Substituting in for the minimal value of b yields $(1-\tilde{\theta})^2(k\sqrt{\tilde{\theta}\theta} - \sqrt{\theta}) > 1-\tilde{\theta}^{\frac{3}{2}})[k\sqrt{\tilde{\theta}\theta}(1-\tilde{\theta}) - (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}]$. Also from proposition 1, we know $k = \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$. Substituting in for k yields $(1-\tilde{\theta})(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta} - (1-\tilde{\theta})^2\sqrt{\theta} > (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta} - (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta} = 0$ which always holds.

It remains to be seen whether the above inequality will hold for all values of b. Take the inequality $(1-\tilde{\theta})^2(3b-\sqrt{\theta}) > (1-\tilde{\theta}^{\frac{3}{2}})[3b(1-\tilde{\theta}) - (1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}]$ and take the derivative of each side with respect to b. The derivative of the left-hand side is $(1-\tilde{\theta})^2$; whereas, the derivative of the right-hand side is $(1-\tilde{\theta})(1-\tilde{\theta}^{\frac{3}{2}})$. Note that the derivative of the right-hand side is greater than that of the left-hand side for all $\tilde{\theta} \in (0,1)$. This implies that for low values of b, the existence of a treaty reduces torture amongst non-signatories. But that this effect may be reversed for high values of b.

It finally remains to compare signatories and non-signatories of the treaty. Take $\underline{\theta} \in (0,\tilde{\theta})$ and $\bar{\theta} \in (\tilde{\theta},1)$. $t(1,\underline{\theta};e(1)) > t(0,\bar{\theta};e(0))$ iff $\sqrt{\frac{\tilde{\theta}}{\underline{\theta}}} - \frac{k\tilde{\theta}}{3b} > (\frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}})[\frac{1}{\sqrt{\bar{\theta}}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})}]$. From proposition 1, $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\bar{\theta}}$. Substituting the minimal value of b yields $\sqrt{\frac{\tilde{\theta}}{\underline{\theta}}} - \frac{\sqrt{\tilde{\theta}}}{\sqrt{\bar{\theta}}} > (\frac{1-\tilde{\theta}^{\frac{3}{2}}}{)}1-\tilde{\theta})[\frac{1}{\sqrt{\bar{\theta}}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{k\sqrt{\tilde{\theta}\bar{\theta}(1-\tilde{\theta})}}]$. Substituting $\frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$ yields $\sqrt{\frac{\tilde{\theta}}{\underline{\theta}}} - \frac{\sqrt{\tilde{\theta}}}{\sqrt{\bar{\theta}}} > (\frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}})[\frac{1}{\sqrt{\bar{\theta}}} - \frac{1}{\sqrt{\bar{\theta}}}] = 0$. Since $\underline{\theta} < \bar{\theta}$ this inequality always holds.

It remains to be checked if signatories torture more than non-signatories for higher values of b. Take the inequality $\sqrt{\frac{\tilde{\theta}}{\underline{\theta}}} - \frac{k\tilde{\theta}}{3b} > (\frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}})[\frac{1}{\sqrt{\bar{\theta}}} - \frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})}]$ and take the derivative of each side with respect to b. The derivative of the left-hand side yields $\frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\tilde{\theta}}}{3b^2(1-\tilde{\theta})}$; whereas, the derivative of the right-hand side yields $(\frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}})(\frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b^2(1-\tilde{\theta})})$. Since $\frac{1-\tilde{\theta}^{\frac{3}{2}}}{1-\tilde{\theta}} > \sqrt{\tilde{\theta}}$ for all $\tilde{\theta} \in (0,1)$ the right-hand side is increasing faster than the left-hand side in b. This implies that signatories torture more than non-signatories for low levels of b, but that this relationship may be reversed as b rises. □

Proof of Proposition 4:

In the pooling equilibrium, in which all types sign the treaty, $U_G^{pooling}(\theta) = R\left(1 - \sqrt{\theta}\frac{k}{3b}\right)^2$. Consider now the out of equilibrium beliefs in which if a state does not sign the treaty, all posterior beliefs are that this is the state with the highest costs of torture, $\theta = 1$. Then $U_G^{defect}(1) = R\left(1 - \frac{1}{2b}\right)^2$. With these beliefs, is there an incentive for this type to defect? The answer is yes, iff $U_G^{defect}(1) > U_G^{pooling}(1)$ iff $R\left(1 - \frac{1}{2b}\right)^2 > R\left(1 - \frac{k}{3b}\right)^2$ iff $\frac{1}{2b} < \frac{k}{3b}$ iff $k > \frac{3}{2}$. Hence the pooling equilibrium in which all sign does not survive the intuitive criterion for $k > \frac{3}{2}$.

B Model Robustness

B.1 Post-Tenure Punishments

Assume that signing the CAT makes an autocratic leader vulnerable to post-tenure liability. That is, signing the CAT renders the leader open to punishment in the event that he is removed to office. Further assume that this additional punishment is a constant p. (In equilibrium, all autocratic governments repress, at least to some extent. The assumption that the punishment is a constant allows us to incorporate the possibility of post-tenure punishments without burdening the model with undue mathematical complexity.)

Denote the Government's (G's) utility as:

$$U_G(t, e, s; \theta) = \pi(t, e)[R + sp] - \theta t - sp$$

where θ denotes the marginal cost of repression, $s \in \{0, 1\}$ takes the value of 1 in the event G signs the CAT, and $\pi(t, e)$ is the contest success function $\pi(t, e) = \frac{t}{t+e}$. D's utility function is identical to that in the main model:

$$U_D(t, e, s) = [1 - \pi(t, e)]R - be$$

Proposition 5. *If $p > 0$, there exists semi-separating equilibrium where all governments with low costs to repression $\theta \leq \widehat{\theta}$ sign the treaty, and all governments with high costs $\theta > \widehat{\theta}$ do not sign for some $\widehat{\theta} \in (0,1)$.*

Proof: Firstly, the best responses are $t(e,1;\theta) = \sqrt{\frac{e(R+p)}{\theta}} - e$ and $t(e,0;\theta) = \sqrt{\frac{eR}{\theta}} - e$. The opposition's problem given these best responses are for $s = 1$ to maximize $\int_0^{\widehat{\theta}} (R\sqrt{\frac{e\theta}{R+p}} - be) f(1) d\theta \Leftrightarrow e(s=1) = \frac{R^2\widehat{\theta}}{9b^2(R+p)}$ and for $s = 0$: $\frac{\partial U_D}{\partial t} = \int_{\widehat{\theta}}^1 [\frac{1}{2}\sqrt{\frac{R\theta}{e}} - b] f(0) d\theta \Leftrightarrow e(0) = \frac{R(1-\widehat{\theta}^{\frac{3}{2}})^2}{(9b^2)(1-\widehat{\theta})^2}$. From the response functions of the government and opposition, we can derive the contest success function when $s = 1$, $\pi(t,e;s=1) = 1 - \frac{R\sqrt{\widehat{\theta}\theta}}{3b(R+p)}$. And when $s = 0$, $\pi(t,e;s=1) = 1 - (\frac{\sqrt{\theta}}{3b})[\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}]$. Governments will prefer to sign when $U_G(t,e,s=1;\theta) \geq U_G(t,e,s=0;\theta)$.

$$(R+p)\left[1 - \frac{R\sqrt{\widehat{\theta}\theta}}{3b(R+p)}\right] - \theta\left(\frac{R}{3b}\sqrt{\frac{\widehat{\theta}}{\theta}} - \frac{R^2\widehat{\theta}}{9b^2\,(R+p)}\right) - p \geq R\left[1 - \frac{\sqrt{\theta}}{3b}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right] - \theta\left[\frac{R}{3b}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\sqrt{\frac{1}{\theta}} - \frac{R}{9b^2}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right]$$

$$\frac{6b}{\sqrt{\theta}}\left[R\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right) - \sqrt{\widehat{\theta}}\right] \geq \left[R\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right]^2 - \frac{1}{(R+p)}\widehat{\theta}$$

The LHS is monotonic and decreasing in θ over the unit interval, and the right hand side is constant, so if $\widehat{\theta}$ is interior, it is the low cost (low θ) types that sign. To show $\widehat{\theta}$ is interior, it is necessary to show that there exists a value of $\theta \in (0,1)$ such that the above expression holds at equality when $\theta = \widehat{\theta}$. A fully analytic solution is elusive, but an interior solutions exist for a variety of parameterizations in simulations. For instance, at $(R,p,b) = \left(1, 100, \frac{1}{3}\right)$, $\widehat{\theta} = 0.497 \in (0,1)$.

Proposition 6. *As the level of post-tenure punishments grows increasingly large, signatory governments are willing to devote more effort to repression and the opposition devotes less effort to removing the government on witnessing treaty accession. Survival times of signatory governments will thus increase with p.*

Proof: From the response functions above, it follows that $t(e(1);\theta) = \frac{R\sqrt{\tilde{\theta}}}{3b\sqrt{\theta}} - \frac{R^2\tilde{\theta}}{9b^2(R+p)}$ and $e(s = 1) = \frac{R^2\tilde{\theta}}{9b^2(R+p)}$. Clearly the former is increasing and the latter decreasing in p. Moreover, survival is given by $\pi(t,e;s=1) = 1 - \frac{R\sqrt{\tilde{\theta}\theta}}{3b(R+p)}$ which is increasing in p.

B.2 Varying Values to Holding Office

Assume that governments vary in the value the place on retaining office, rather than in the costs they suffer from engaging in repression. Thus, the signing of the CAT will act as a signal that a government places a high value on office, rather than a signal of the government's low marginal cost to repression. Let θ be constant, and common knowledge. Denote the value the government attaches to office as $R_G \sim U[0,1]$. The government observes the realization of this variable. The opposition is only aware of the distribution from which it is drawn. Denote the value the opposition attaches to office as R_D. This value is a constant and is common knowledge. As in the above, allow for post-tenure punishments of value p. The government's utility function is thus given by $U_G(t,e,s;R_G) = \pi(t,e)[R_G + sp] - [sk + (1-s)]\theta t - sp$, while the opposition's is given by $U_D(t,e,s) = [1-\pi(t,e)]R_D - be$.

Proposition 7. *There exists $\tilde{R_G} \in (0,1)$ such that for $k < \frac{1-\tilde{R_G}}{\left(\sqrt{1+p}-\sqrt{\tilde{R_G}+p}\right)\sqrt{\tilde{R_G}}}$, a semi-separating equilibrium exists where all high-value governments $R_G \geq \tilde{R_G}$ sign the treaty; while all low-value governments $R_G < \tilde{R_G}$ do not.*

Proof: For $s = 1$, the government's response function is $t(e(1);R_G) = \sqrt{\frac{e(1)(R_G+p)}{k\theta}} - e(1)$; for $s = 0$, the government's response function is $t(e(0);R_G) = \sqrt{\frac{e(0)R_G}{\theta}} - e(0)$. The opposition's problem when $s = 1$ is given by $EU_D = \int_{\tilde{R_G}}^{1}[R_D\sqrt{\frac{e(1)k\theta}{R_G+p}} - be]f(1)dR_G$ where $f(1)$ represents the posterior distribution over $[\tilde{R_G},1]$ given that $s = 1$. Solving for this expression yields:

$$EU_D = R_D\sqrt{e(1)k\theta}\int_{\tilde{R_G}}^{1}\frac{1}{\sqrt{R_G+p}}f(1)dR_G - be(1)$$
$$\Leftrightarrow e(1) = \frac{R_D^2(\sqrt{1+p}-\sqrt{\tilde{R_G}+p})^2 k\theta}{(1-\tilde{R_G})^2 b^2}$$

Similarly, when $s = 0$, the opposition's problem is given by $EU_D = \int_0^{\tilde{R_G}}[R_D\sqrt{\frac{e\theta}{R_G}} - be]f(0)dR_G$ where $f(0)$ represents the posterior distribution over $[0,\tilde{R_G}]$ given that $s = 0$. Solving this expression yields:

$$EU_D = R_D\sqrt{e(0)\theta}\int_0^{\tilde{R_G}}\frac{1}{\sqrt{R_G}}f(0)dR_G - be(0)$$
$$\Leftrightarrow e(0) = \frac{R_D^2\theta}{b^2\tilde{R_G}}$$

A government will be willing to sign the treaty when $U_G(t(e(1);R_G),e(1);R_G) \geq U_G(t(e(0);R_G),e(0);R_G)$. This inequality then yields:

$$\left(\sqrt{R_G}-\frac{R_D}{b}k\theta\frac{1}{1-\tilde{R}_G}\left(\sqrt{1+p}-\sqrt{\tilde{R}_G+p}\right)\right)^2-\left(\sqrt{R_G}-\frac{R_D\theta}{b}\sqrt{\frac{1}{\tilde{R}_G}}\right)^2 \geq \left[2\frac{R_D}{b}k\theta\frac{1}{1-\tilde{R}_G}\left(\sqrt{1+p}-\sqrt{\tilde{R}_G+p}\right)\right]\left[\sqrt{R_G+p}-\sqrt{R_G}\right]$$

Note that the RHS is monotonic and decreasing in R_G. The LHS will be monotonic and increasing in R_G if:

$$\frac{R_D\left(\sqrt{1+p}-\sqrt{\tilde{R}_G+p}\right)k\theta}{(1-\tilde{R}_G)b} < \frac{R_D\theta}{b\sqrt{\tilde{R}_G}} \Leftrightarrow k < \frac{1-\tilde{R}_G}{\left(\sqrt{1+p}-\sqrt{\tilde{R}_G+p}\right)\sqrt{\tilde{R}_G}}$$

Note that if $\tilde{R}_G$ is interior, this implies that it will the high-value types that sign the treaty. Again a fully analytic solution is elusive, but an interior solutions exist for a variety of parameterizations in simulations. For instance, at $(p,k,\theta,b,R_D)=\left(50,\frac{1}{2},\frac{1}{2},\frac{1}{2},1\right)$, $\tilde{R}_G = 0.66576 \in (0,1)$.

C Matching Diagnostics

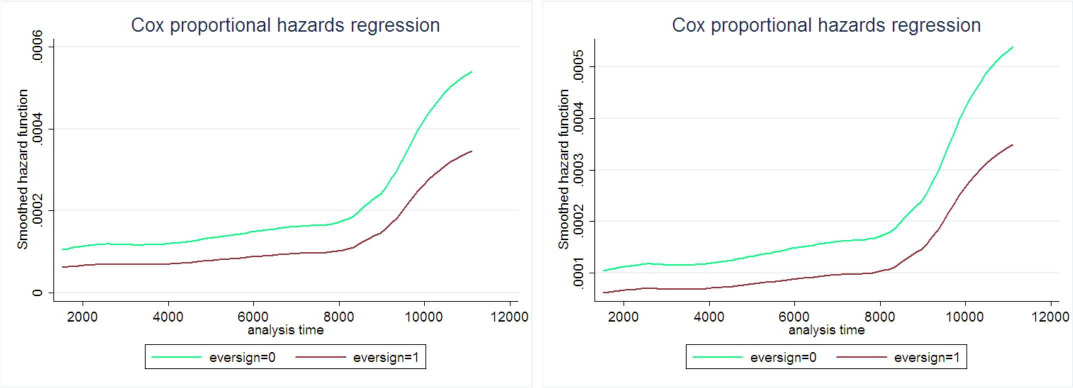
[Table 6 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

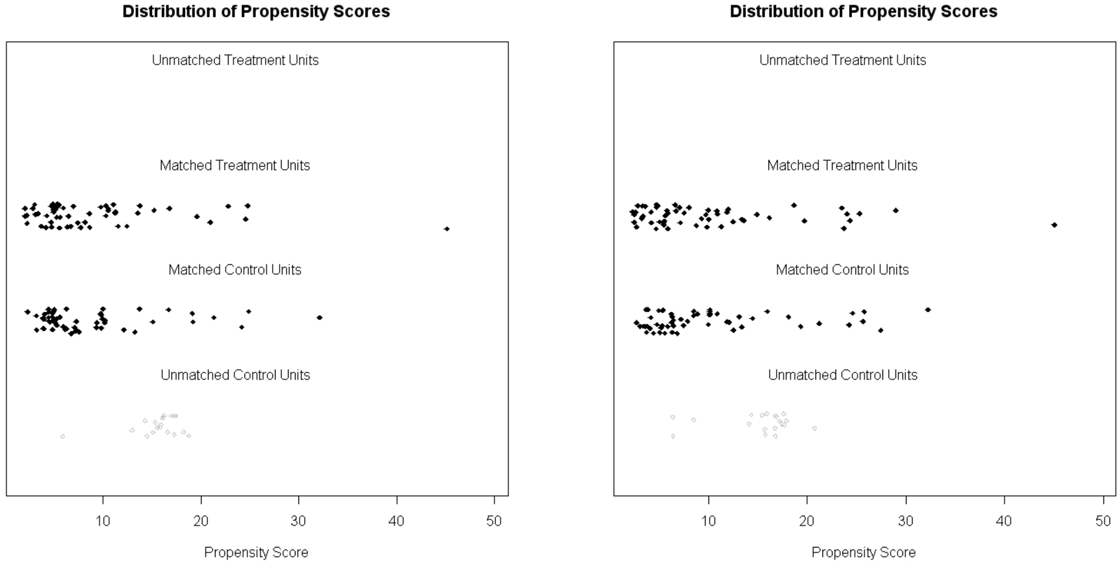
[Figure 4 about here.]

Figure 1: Hazard Plots for the Naive Survival Models



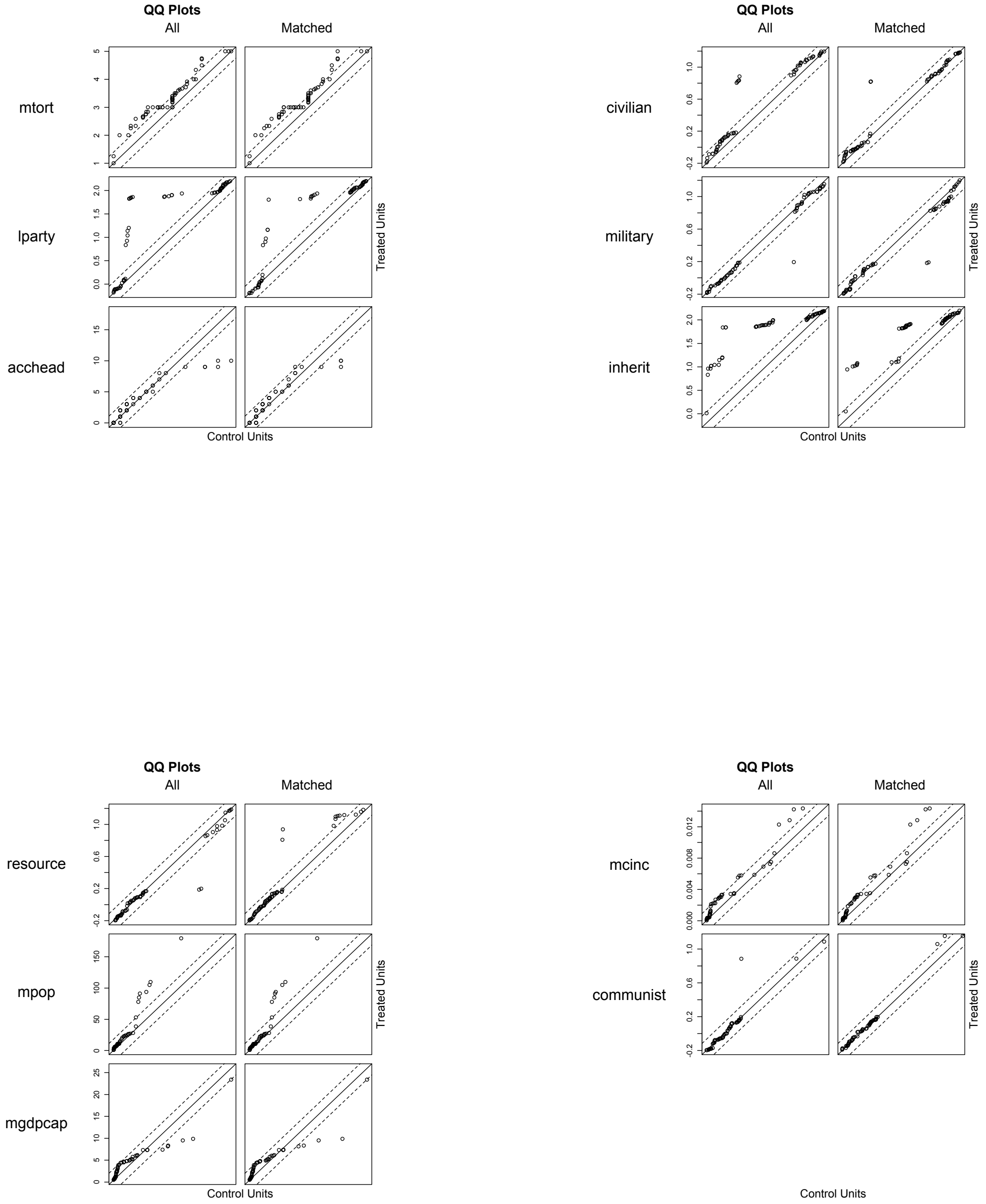
Hazard function estimates from the models **Hath 2** and **CIRI 2** described above. Estimates from the **Hath 2** model are to the left and those from the **CIRI 2** model are to the right. The plots depict the risk of regime failure at time t given that the regime survived until time t. Hazard rates are on the y-axis, regime tenure (measured in days) is on the x-axis.

Figure 2: Scatter Plots of Propensity Scores



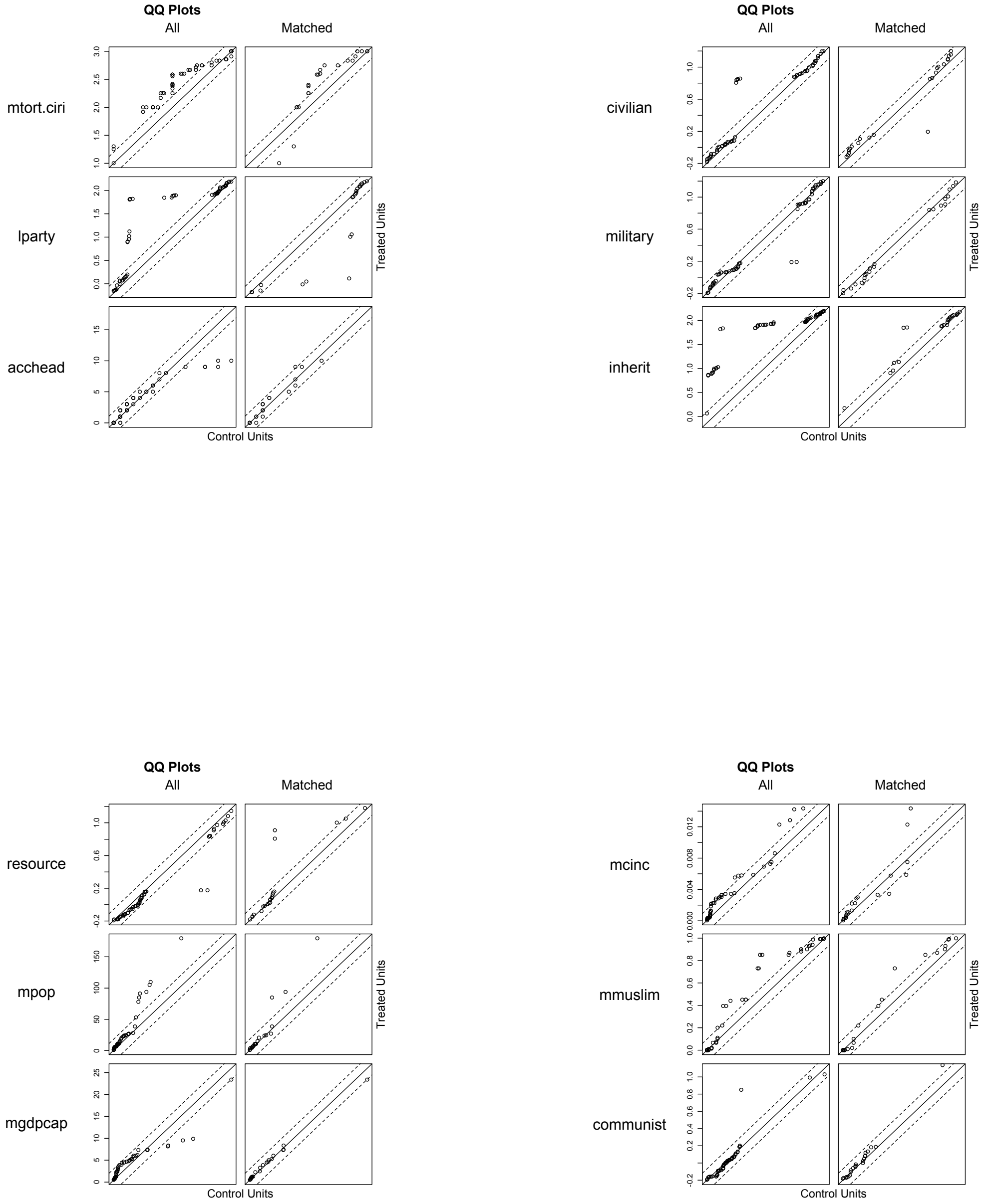
Scatter plots of propensity scores. Propensity scores estimated using the **Hath 1** model are to the left, those estimated using the **CIRI 1** model are to the right. Units labeled as treated signed the CAT, those labeled as controls did not. Both treated and control units enter into final estimates.

Figure 3: HathMatch Quantile-Quantile Plots



Q-Q plots for all covariates after matching using the **Hath 1** model. Identical distributions of each variable between treated and control groups would result in all points appearing on the 45° line. Plots before matching to the left, after to the right.

Figure 4: CIRIMatch Quantile-Quantile Plots



Q-Q plots for all covariates after matching using the **CIRI 1** model. Identical distributions of each variable between treated and control groups would result in all points appearing on the 45° line. Plots before matching to the left, after to the right.

Table 1: Coefficient Estimates from the Cox Survival Model

	Hath1	CIRI1	Hath2	CIRI2
eversign	-.559	-.544	-.539	-.526
	(.267)**	(.268)**	(.272)**	(.266)**
mtort	-.055	.042	.048	-.005
	(.176)	(.261)	(.164)	(.221)
acchead	.131	.135	.119	.122
	(.025)***	(.027)***	(.024)***	(.026)***
civilian	1.465	1.445	1.608	1.644
	(.582)**	(.591)**	(.569)***	(.567)***
military	1.215	1.18	1.416	1.44
	(.58)**	(.589)**	(.575)**	(.578)**
inherit	.214	.221	.217	.22
	(.177)	(.173)	(.155)	(.154)
lparty	-.698	-.736	-.693	-.686
	(.16)***	(.15)***	(.155)***	(.157)***
resource	.185	.138	.19	.234
	(.339)	(.37)	(.332)	(.299)
mpop	.006	.006	.	.
	(.007)	(.006)		
mcinc	18.424	20.947	.	.
	(57.465)	(59.282)		
mgdpcap	-.016	-.018	.	.
	(.036)	(.036)		
N	129	129	129	129

Results from a Cox Proportional Hazards estimate of regime tenure. Hazard rates estimates based on number of days in office as measured by the Archigos dataset. Estimates are constructed for a sample of 129 authoritarian regimes in the years 1985-1996. Coefficient estimates are of the form $h_i(t) = h_0(t)exp(\beta' X)$. **Hath** models use Hathaway's (2007) torture index; **CIRI** models use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 2: Coefficient Estimates Dropping Regimes that Inherited Signatory Status

	Hath1	**CIRI1**	**Hath2**	**CIRI2**
eversign	-.892	-.881	-.88	-.883
	(.33)***	(.323)***	(.325)***	(.316)***
mtort	.046	.114	.138	.142
	(.229)	(.326)	(.195)	(.221)
acchead	.121	.123	.106	.11
	(.031)***	(.032)***	(.03)***	(.031)***
civilian	1.299	1.294	1.51	1.567
	(.58)**	(.588)**	(.591)**	(.574)***
military	1.056	1.046	1.401	1.441
	(.581)*	(.586)*	(.593)**	(.587)**
inherit	.202	.207	.225	.214
	(.198)	(.194)	(.175)	(.178)
lparty	-.752	-.782	-.764	-.718
	(.186)***	(.177)***	(.178)***	(.181)***
resource	.11	.08	.045	.131
	(.389)	(.418)	(.393)	(.342)
mpop	.008	.008	.	.
	(.007)	(.007)		
mcinc	-5.619	-5.797	.	.
	(76.118)	(76.883)		
mgdpcap	-.034	-.035	.	.
	(.043)	(.042)		
N	108	108	108	108

Results from a Cox Proportional Hazards estimate of regime tenure. Hazard rates estimates based on number of days in office as measured by the Archigos dataset. All regimes that inherited their status as a signatory government from a previous regime are dropped from the sample. Estimates are constructed for a sample of 108 authoritarian regimes in the years 1985-1996. Coefficient estimates are of the form $h_i(t) = h_0(t)exp(\beta' X)$. **Hath** models use Hathaway's (2007) torture index; **CIRI** models use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 3: Proportional Hazards Tests

Test	**Variable**	**Hath1**	**CIRI1**
G & T	-	0.5200	0.6899
Harrell's rho	civilian	0.7932	0.7432
	military	0.9567	0.9447

Results from Grambsch and Therneau (G&T) and Harrell's rho tests of the proportional hazards assumption. Reports are p-values from a test of the hypothesis that the proportional hazards assumption is violated against the null that it is not. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 4: Robustness Checks

	Hath	CIRI	N
Pre-processing using	-.664	-.584	110
Propensity Scores	(.312)**	(.282)*	
Cox Estimates	-.474	-.471	110
1 year +	(.271)*	(.271)*	
Cox Estimates	-.591	-.596	104
2 years +	(.319)*	(.315)*	

Coefficients on **eversign** variable in robustness checks. In all models, hazard rates are estimated based on the number of days served in office, as measured by the Archigos dataset. Coefficient estimates in the first row are conducted using Cox proportional hazards regressions after pre-processing the data using propensity score matching. Matching is one-to-one using nearest neighbor estimates based on mahalanobis distances. Estimates in the second and third rows are from Cox proportional hazards models after all regimes surviving less than (respectively) one and two years are dropped from the sample. Estimates in the fourth row are from Weibull survival models. Estimates in the final two rows are from Weibull models where shape of the baseline hazard function is allowed to vary by authoritarian regime-type. Coefficient estimates for the **eversign** variable are reported in row 5, estimates on the military regime ancillary parameter are reported in row 6. Estimates in the column labeled 'Hath' control for Hathaway's torture measure; those in the column labeled 'CIRI' control for the CIRI measure. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 5: Coefficient Estimates Using Multiple Record Data

	Hath	**CIRI**
cats-lag	-.659	-.767
	(.394)*	(.39)**
torture	-.566	-.815
	(.265)**	(.357)**
acchead	.194	.152
	(.088)**	(.077)**
civilian	-.21	-.389
	(.86)	(.898)
inherit	-.126	-.079
	(.343)	(.351)
military	.113	.129
	(.949)	(.96)
resource	-.232	-.061
	(.502)	(.529)
growth	-.026	-.033
	(.008)***	(.009)***
gdp-per-cap	-.005	-.036
	(.065)	(.055)
cinc	183.853	176.481
	(91.283)**	(81.226)**
war	1.378	1.547
	(.468)***	(.486)***
population	-.014	-.015
	(.01)	(.01)
N	428	420

Results from a Cox Proportional Hazards estimation of regime tenure using a multiple record panel. The unit of estimation is the regime-year. The sample consists of a a panel of authoritarian regimes that did not inherit the status of signatory to the CAT between 1985 and 1996. Time-invariant covariates are characteristics of the regime that do not vary over time, time-varying covariates are characteristics that change from year to year. The **Hath** model use Hathaway's (2007) torture index; while **CIRI** model use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 6: Selection Model Coefficients

	Hath 1	**CIRI 1**	**Hath 2**	**CIRI 2**
tort	.546	1.517	.485	1.181
	(.355)	(.822)*	(.303)	(.679)*
lparty	.189	.418	.153	.313
	(.328)	(.424)	(.329)	(.383)
acchead	-.093	-.112	.	.
	(.081)	(.082)		
civilian	-.614	-.552	.	.
	(1.102)	(1.261)		
military	-.488	-.427	.	.
	(1.15)	(1.277)		
inherit	1.444	1.547	1.155	1.133
	(.453)***	(.502)***	(.428)***	(.468)**
resource	.337	.728	.	.
	(.808)	(.843)		
pop	.013	.017	.018	.021
	(.016)	(.015)	(.019)	(.019)
gdpcap	.065	.093	.095	.117
	(.091)	(.099)	(.076)	(.082)
cinc	-85.021	-135.3	-122.043	-149.68
	(154.581)	(145.245)	(163.268)	(161.531)
communist	1.546	2.243	1.255	1.596
	(.917)*	(1.071)**	(.834)	(.856)*
cons	-3.994	-6.539	-4.194	-5.792
	(1.4)***	(2.599)**	(1.1)***	(1.905)***
N	129	129	129	129

Estimates from logistic regressions predicting the signing of the CAT. The dependent variable is **eversign**, a binary indicator that takes the value 1 if a regime was ever a signatory of the CAT and zero otherwise. Models labeled **Hath** use Hathaway's (2007) definition of torture, those labeled **CIRI** use the CIRI measures. The full models in the two left-hand columns are used to generate propensity scores which are fed into a genetic matching algorithm. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.