Article

JOURNAL OF THEORETICAL POLITICS

A strategic theory of effective monitoring arrangements for international institutions

Journal of Theoretical Politics
2014, Vol. 26(4) 599–628

DOI:10.1177/0951629813511550
jtp.sagepub.com
SAGE

Thania Sanchez
Yale University, New Haven, CT, USA

Johannes Urpelainen
Columbia University, New York, NY, USA

Abstract

States often delegate compliance monitoring to international monitors. It is commomly assumed that these monitors will report accurately on the information they gather. However, the effectiveness of compliance monitors varies widely. Monitors may fail to collect information about non-compliance or even collude with non-compliant states and deliberately fail to report the information they gather. To explain this variation, we present a formal theory of how structural conditions and institutional designs lead to different levels of monitoring efficacy. We show that international institutions can improve monitoring by avoiding ruthless sanctions, and that intrinsically motivated monitors generally achieve better outcomes than neutral bureaucrats. Our theory contributes to the broad literature on international institutional design and treaty effectiveness.

Keywords

Compliance; enforcement; international institutions; monitoring; principal–agent models

1. Introduction

The Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal was created in part to halt the large amounts of exports of hazardous wastes by developed countries to developing countries that lacked the ability to properly dispose of the waste (Marcoux and Urpelainen, 2012). By limiting the amount and type of hazardous wastes countries could export, the treaty hoped that developed nations would produce less hazardous waste and dispose of it themselves in environmentally friendly

Corresponding author:
Johannes Urpelainen, Columbia University, 7th Floor, International Affairs Building, 420 West 118th Street, New York, NY 10027, USA.
Email: ju2178@columbia.edu

ways, instead of dumping it on developing nations. However, the shipment of hazardous wastes continues to occur even though it is illegal.

Italy has been a notorious non-complier with the treaty, as Italian toxic and hazardous waste has been dumped in countries such as Albania and Somalia, even though this is in violation of the treaty and EU regulations. According to Legambiente, a leading Italian environmental non-government organization (NGO), the Italian Customs Agency seized 7,400 tons of hazardous waste leaving the country illegally in 2011 alone (Ciafani, 2012). This is mostly accomplished as the criminal networks that export the waste illegally bribe or blackmail the monitors that are meant to check the transport of waste and procure the paperwork to authorize shipments (Massari and Monzini, 2006). Without reliable monitoring, the illegal transfer of hazardous waste continues.

The failure to effectively end the illegal shipment of hazardous waste demonstrates how sensitive international cooperation can be to monitoring failure. It is often difficult to obtain information about any given state's compliance record with international agreements. Indeed, states have incentives to hide non-compliant behavior from external observes, so monitoring is crucial for treaty effectiveness. As Barrett (2003, 37) writes, '[i]f non-compliance cannot be observed, then it cannot be punished. Effective monitoring is a prerequisite for enforcing compliance.' Often parties to a treaty seek to solve this problem by establishing monitors entrusted with the task of detecting non-compliance.

Scholars of international relations have paid scant attention to monitors' incentives to actually gather and report information.[1] Often theories assume that once monitors are in place they will provide the relevant information, although political economists have long emphasized that supervising effort is prone to collusion (Tirole, 1986). Here we present a theory that can explain why sometimes monitors may fail at their task. We present a strategic theory of treaty monitoring that pays special attention to issues of institutional design. We present a formal model that is developed around three key questions regarding the monitor's incentives. First, can the monitor profitably collude with non-compliant states? Second, if a monitor reveals information about non-compliance to other states, what are the consequences? Finally, if a monitor is negligent or engages in collusion, what is the probability of being caught later?

Our research finds that the best way to secure effective monitoring and compliance is to reward the monitor for successful information revelation while applying measures against collusion. Monitors should have incentives to collect information for the right reasons: to reveal this information to other states and publics rather than collude with the non-compliant state. When treaty monitors have incentives to collect and publicize information on non-compliance, high levels of treaty compliance by member states are attainable. However, often these incentives are missing, and a potential for monitoring failure is present.

With regard to institutional design, our analysis has two implications worth emphasizing. First, successful monitoring depends on the enforcement regime. While international cooperation theorists have emphasized the importance of strong sanctions for compliance (Carrubba, 2005; Downs et al., 1996), they have not considered the possibility that the enforcement regime needs to induce monitors to expend effort. When a monitor who is interested in cooperation expects the consequences of information revelation to be harmful to cooperation, she has an incentive not to reveal that information. In other words, when a monitor expects the punishment for lack of compliance to push actors

to leave a treaty or cooperative scheme altogether, she might think twice about revealing the non-compliant behavior. Consequently, a defector understands that she need not worry about monitoring. This result is particularly important given the debate between the 'enforcement' and 'managerialist' schools of international cooperation (Chayes and Chayes, 1995; Downs et al., 1996; Tallberg, 2002). While the enforcers have underscored the important of credible sanctions against defectors, the managerialists have warned that sanctions often do more harm than good. Our analysis provides a strategic rationale for limiting sanctions: if the monitor expects excessively ruthless enforcement, she will refuse to assist in carrying out sanctions. This result may explain limited enforcement 'under anarchy' (Axelrod and Keohane, 1985).

Second, we provide a strategic rationale for selecting monitors interested in compliance for intrinsic reasons (Mitchell, 1998). In addition to reducing the cost of monitoring, an intrinsic commitment to cooperation reduces the incentive to collude and ensures that the monitor will not have a perverse incentive to encourage non-compliance, so as to reap the reward for revealing it. Johns (2007) has previously shown that international bureaucrats may be more willing to truthfully reveal their information to bargaining states if they hold a bias in favor of the side with a strong outside option. Our model is different in that we focus specifically on monitoring, examining compliance incentives and the monitor's effort level, instead of just information revelation. We show that intrinsic interest in compliance allows effective monitoring without a risk of collusion. When intrinsically motivated monitors are not available, states should design institutions that offer generous rewards for revealing information about non-compliance, so as to reduce the attractiveness of collusion.

2. The state of the art on compliance monitoring

International cooperation produces collective benefits for multiple parties, so individual states are sometimes tempted to 'defect' or 'free ride' (Downs et al., 1996; Keohane, 1984). To enforce treaty compliance, the cooperating parties must not only credibly commit to punishing defectors, but also identify defections through monitoring. We define *monitors* as the actors entrusted with collecting and revealing information about compliance with an international rule or norm. Our definition encompasses official monitors, such as International Atomic Energy Agency (IAEA) inspectors or treaty secretariats, but it also applies to unofficial monitors, such as human rights activists (Dai, 2002). States need not even formally delegate authority to the monitor: as long as states hold the expectation that an actor will try to observe and report non-compliance, officially or unofficially, we consider the actor a monitor. Underlying this definition is the idea that states are motivated by the need to monitor compliance at the lowest possible cost (Dai, 2002). Thus, if an effective unofficial monitor, motivated by normative or material concern, is available to pull the 'fire alarm', states need not recruit an official monitor. However, if unofficial monitors are not available, an official 'police patrol' monitor may be necessary to secure compliance (McCubbins and Schwartz, 1984).

The conventional wisdom on international monitoring emphasizes that 'other-reporting' dominates 'self-reporting' because defectors have no incentive to reveal non-compliance to other states. Thus, while the originators of a treaty or norm are often also the subjects of monitoring, the monitors are usually third parties (Dai, 2002; Jo, 2008;

Mitchell, 1998). For example, the Chemical Weapons Convention entrusts the Organization for the Prohibition of Chemical Weapons with monitoring compliance. However, some monitors are not explicitly created by states. For example, the environmental NGO Greenpeace monitors compliance with whaling rules.[2]

We investigate the effectiveness of monitoring under various structural conditions and regime designs. We measure monitoring effectiveness as the ability of the monitoring regime to deter or remedy compliance failures. This effectiveness depends, of course, on the structural features of the issue at hand (Mitchell, 1994). However, the extant literature also offers insights into other strategic preconditions for effective monitoring. Mitchell (1998) recognizes that both self-reporting and other-reporting could provide states with limited or inaccurate information on compliance. In particular, he warns that some monitors, such as career bureaucrats, might not hold a strong normative commitment to the regime. Mitchell (1998, 124) also recognizes that increased transparency depends on 'enhancing incentives for reporting while reducing the incentives for not reporting'. However, he does not provide a detailed account of how exactly those incentives can be changed, notwithstanding his general recommendation to foster normative commitments to the regime.[3]

Dai (2002) also explores the conditions that impede monitoring in multilateral agreements. She argues that states prefer to delegate monitoring to victims of non-compliance if these victims can effectively detect non-compliance and state–victim preferences are closely aligned. If these conditions fail, states must either monitor themselves or delegate monitoring to international organizations (Dai, 2002, 416). Although Dai (2002) does not provide a formal analysis of monitoring, her analytical taxonomy and extensive empirical examples offer a useful basis for a comprehensive strategic theory of international monitoring.

In sum, the literature emphasizes that monitoring effectiveness depends on how the monitor relates to the states it is supposed to monitor. First, the dependence of a monitor on the state that it is supposed to monitor influences her effectiveness. This can be financial dependence or other factors that determine the monitor's existence. Second, if the monitor has a normative or material incentive to facilitate international cooperation, it is easier to induce monitoring efforts simply by structuring the international institution so that information leads to corrective measures, such as dispute resolution or sanctions that foster rather than undermine further cooperation.

The monitor's type will also influence her behavior. While the literature has discussed normatively motivated monitors, such as activist NGOs, these are not the only types of monitors that have incentives to be good monitors. For instance, employees at the Bureau of Democracy, Human Rights, and Labor of the US Department of State may be equally effective at getting information about other states' compliance. While these bureaucrats might be more invested in following the letter of the law rather than the spirit of it, for the purposes of our study they are well positioned to seek out compliance information, whereas other types of monitors may be less effective. In Table 1, we provide several examples of the two dimensions.[4]

In developing our theory, we build on ideas from contract theory (Aghion and Tirole, 1997; Tirole, 1986). This literature has recognized that collusion between agents and supervisors greatly complicates the optimal design of principal–agent contracts. In our case, the principal is the international institution and the agent is the target state, while

Table 1. A classification of treaty monitors.

	Independent	**Dependent**
Invested	Human Rights Watch Greenpeace International Amnesty International	UN human rights monitors Landmine Monitor Memorial
Not invested		Government agencies (Environmental Protection Agency) Private contractors (Saybolt)

the supervisor is the monitor. Our model is a stylized version of those used in the contracting literature, and we make assumptions about the nature of compliance monitoring and collusion based on substantive insights into international relations.

3. The argument

In this section, we present an informal summary of our formal analysis. In our model, a state that has undertaken an international obligation decides on a compliance level while a monitor decides on monitoring efforts. We will refer to the state as the *target state* because it is the state being monitored. If the monitor obtains information regarding non-compliance, it must decide whether to collude with the target state by not revealing information about non-compliance. If the monitor fails to obtain information, or obtains it but then colludes with the target state, this failure may be noticed later through external verification.

The model captures the two core incentives. First, monitors could have few incentives to collect or reveal information if their efforts do not lead to improved cooperation. Second, available information is often not used effectively. While Chayes and Chayes (1995) argue that information leads to compliance, the conclusion holds only if states actually use the information that they receive to improve compliance.

At worst, information about defections can even prompt states to abandon an international institution. For example, Australia suspended all relations with the UN human rights bodies after the release of a report critical of Australia's human rights record.[5] The problem looms particularly large if information revelation does not prompt sanctions or other actions to improve compliance. Why should a monitor reveal information if the most probable outcome is a decrease in cooperation by states? Thus, in some cases, a monitor might choose to work with the target state to improve compliance rather than reveal to the international community their poor record.

We emphasize both positive incentives (improving monitoring) and negative incentives (obstacles to effective monitoring). On the positive side, we recognize *rewards* and *external verification* as the main factors that facilitate monitoring. On the one hand, monitors must be rewarded for information collection and revelation. Some rewards are intrinsic, as a monitor might benefit from cooperation. For example, the World Trade Organization (WTO) allows states to litigate foreign discrimination. To the degree that actual litigation, or the threat thereof, prompts compensation or restores compliance, individual companies have an incentive to report abuse to their governments (Reinhardt, 2001). In addition, states could reward successful monitors financially. For instance, the

Landmine Monitor has received grants from several governments for procuring good information on compliance with the Ottawa Convention.[6]

Another important positive incentive is provided by external verification. If monitors understand that negligence might be detected and revealed by someone else, they have greater incentives to collect and reveal information. This incentive might be particularly relevant for such international organizations as the IAEA. If IAEA inspections fail to reveal a clandestine military nuclear problem, it is probable that the successful 'rogue state' will subsequently announce its new nuclear capabilities. At some point in the future the bad monitor will be held accountable for negligence, so the expected cost of 'slack' to the monitor is high. On the other hand, some other illicit behaviors simply cannot be detected without monitors. For instance, without monitors, it would be almost impossible to detect human trafficking.[7]

Conversely, negative incentives might stop monitors from revealing information. One previously unrecognized possibility that we emphasize is *collusion*. A monitor could receive side payments or concessions by the target state in exchange for withholding information. As an example, we discussed the case of Italian officers being bribed to allow hazardous waste shipments to go through.

Some monitors, such as morally driven activists, might be unwilling to collude. Often these are activist NGOs, such as Amnesty International. However, as stated earlier, other types of monitors may also have a strong interest in compliance. For example, a monitor such as the Preparatory Commission for the Comprehensive Nuclear Test Ban Treaty Organization could be interested in the survival of the monitoring organization.

Non-compliant states also use expulsion and coercion to prevent monitoring. For one, states can prevent monitoring by refusing monitors access to their territory or facilities. Mexico expelled various human rights observers from the country during the conflict in Chiapas in the late 1990s.[8] States also resort to coercion by threatening the infliction of physical harm on individuals working for the monitoring agencies or actually carrying out those threats. For example, there have been several reports that Mexican officials threatened various human rights activists, including members of the Centro de Derechos Humanos Miguel Agustin Pro Juarez. Human Rights Watch reports that members of this and other domestic monitors have received death threats.[9]

4. Formal analysis

We consider the following model of international monitoring. A *target state* has previously committed to abiding by the rules of an international institution, but compliance is imperfectly observable. The target state is subject to monitoring because it has incentives to defect. A *monitor* is responsible for monitoring compliance, but she is also a strategic player with her own preferences and incentives. We aim to uncover the conditions under which the monitor can induce the state to comply with the rules. For simplicity, we treat these rules as exogenous: the international institution is not a strategic player in the model.

4.1. Model

A simplified game tree is shown in Figure 1. The game begins as the target state chooses a level of *compliance*, $E \in [0, 1]$. The cost of compliance for the state is denoted by $e(E)$,

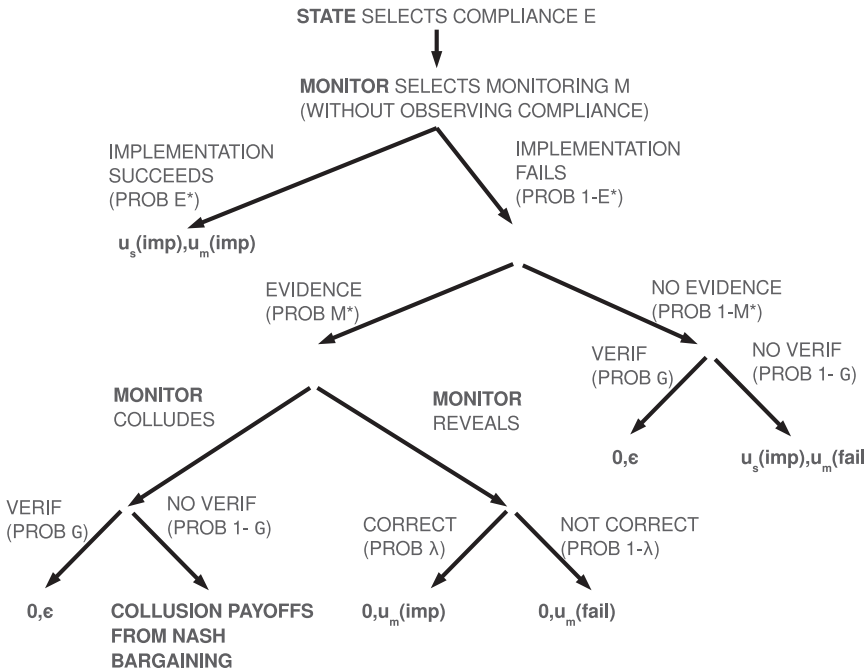

Figure 1. Game tree, including probabilistic moves by nature. For each outcome node, the first (second) payoff is for the state (monitor). Note that the implementation and monitoring costs, $e(E)$ and $m(M)$, are not shown because they do not vary across the final nodes.

an increasing and strictly convex function. To ensure that a unique interior equilibrium exists, let $e'(0) \to 0$ and $e'(1) \to \infty$.

The target state is implementing a project mandated by the international institution. Compliance comprises the efforts to do so. Following Aghion and Tirole (1997), we assume that the compliance efforts produce a binary outcome, 'success' or 'failure'.[10] Success is defined as the satisfactory implementation of treaty provisions, so that the target state has not violated the treaty. Failure is defined as failure to implement the treaty provisions.

Compliance efforts involve particular project implementation, so let E be the probability of success and $1-E$ the corresponding probability of failure. Both implementation and compliance success or failure are only observed by the state. The probabilistic formulation is plausible because compliance is rarely perfectly measurable (Dai, 2002; Chayes and Chayes, 1995). For example, an exogenous shock could prevent a state from implementing treaty obligations. Whenever such external shocks are present and imperfectly observable, compliance cannot be perfectly verified. If Chayes and Chayes (1995) are correct about the complexity and ambiguity of international law, such imperfect observability should be rather common.[11]

To illustrate, consider an example. Under the legal obligations of the 1989 Basel Convention on Hazardous Waste, the Netherlands is required to prevent exports of hazardous waste to Africa. The compliance effort E would be proportional to the resources that the

government allocates to the responsible implementation agency.[12] The failure outcome would be interpreted as at least one major illegal shipment of hazardous waste being sent to Africa. The success outcome would be interpreted as no such illegal shipment.

Next, the monitor selects a level of monitoring $M \in [0, 1]$.[13] The cost of monitoring is denoted by $m(M)$, an increasing and strictly convex function. Assume $m'(0) \to 0$ and $M'(1) \to \infty$ to ensure that a unique interior equilibrium exists.

Even though the monitor is officially supposed to carefully monitor the target state's behavior, in reality the monitor can also select a lower level of monitoring. To capture this fundamental incentive problem, we assume that the monitor directly selects the level of monitoring.[14]

Following monitoring and compliance decisions, the game has three possible subgames. First, implementation could succeed. In this case, non-compliance is not an issue so the game ends. Second, implementation could fail together with monitoring. Now the target state avoids punitive consequences for non-compliance. Finally, implementation fails but monitoring succeeds. In this case, the monitor must choose between collusion and non-collusion.[15]

4.1.1. Payoffs: implementation success If implementation has previously succeeded, monitoring is redundant.[16] The game ends. Intuitively, the monitor cannot reveal a defection if the state has honored the commitment. The payoff to the target state is denoted by $u_s(imp) > 0$ and the payoff to the monitor by $u_m(imp) > 0$. These two payoffs reflect the target state's willingness to comply, to avoid a scandal, and the monitor's interest in the state's compliance, respectively.

4.1.2. Payoffs: implementation failure, monitoring failure If implementation has previously failed, the monitor obtains hard and verifiable evidence on this failure with probability M. For example, if the Netherlands exports hazardous waste to Africa, Greenpeace International could document illegal shipments on the Atlantic. The existence of evidence is observed by both the monitor and the target state.

Given implementation failure, the game continues depending on whether the monitor obtains hard and verifiable evidence or not. First, suppose it does *not*. Two outcomes are possible. On the one hand, the target state may get away with implementation failure. For the target state, the payoff is again $u_s(imp)$, unchanged from successful implementation. Given that the implementation failure is never revealed, the state continues to enjoy a good reputation. The assumption that the payoff from successful implementation is unchanged suggests that the state has no intrinsic interest in compliance.[17] In this case, the monitor's payoff is $u_m(fail)$. We assume $0 < u_m(fail) < u_m(imp)$ because the monitor is at least marginally interested in the target state's compliance: the monitor prefers implementation ($u_m(imp)$) to failure ($u_m(fail)$).

On the other hand, external verification may occur: with probability G, nature reveals implementation failure and the monitor's failure. The target state obtains the lowest possible value from the game, namely zero 'scandal' payoff. Since external verification may also result in correction, we allow the monitor to obtain a small payoff $\varepsilon > 0$ upon external verification. The payoff is assumed to be 'small' because the scandal is assumed to be the dominant factor in shaping the monitor's payoff. The payoffs reflect the combination

of implementation failure *and* reputational loss. The state is shamed for non-compliance and the monitor is shamed for not reporting the non-compliance. For example, if the exportation of hazardous waste to Africa results in severe health problems among the local population, these consequences could later be attributed to hazardous waste that originated from the Netherlands. Consequently, the reputation of both the Dutch government and the treaty monitor would suffer. This move is observed by the monitor and the target state.

4.1.3. Payoffs: implementation failure, monitoring success Suppose now implementation fails but monitoring succeeds. The monitor can *bargain* with the non-compliant state. This bargaining is modeled as a simple Nash (1950) bargaining game and solved in the appendix. Intuitively, the monitor can either (i) reveal the target state's non-compliance or (ii) obtain an income transfer from the target state in exchange for not revealing non-compliance. In equilibrium, bargaining succeeds and *collusion* succeeds whenever the target state is willing to pay enough for non-revelation. Otherwise, the monitor reveals information to non-target states or international organizations.

The payoffs to the monitor and the target state from this bargaining game depend on the outcome. If the monitor does *not* collude, the target state's payoff is zero because a scandal ensues. This payoff reflects reputational loss from a scandal.

The monitor's payoff depends on whether the international institution or non-target states successfully correct the implementation failure in response to the monitor's information. By correction, we refer to measures that would fix the problem and foster future cooperation. For example, the state parties to the Basel Convention could develop a plan to compensate the victims of hazardous waste from the Netherlands.

This possibility of correction is much more important when the monitor reveals non-compliance than in the case of external verification. The idea is that if the state fails to implement and the monitor does not reveal information, both of them obtain low payoffs, respectively 0 and ε, reflecting a scandal. Even if the international institution were to correct the failure, the state and the monitor would have already lost their reputations, and so their payoffs would be low. The state would suffer a punishment for compliance failure, and the monitor would not benefit from a possible correction following the punishment, because the monitor itself would also be punished, perhaps by losing its monitoring rights.

Let $\lambda \in (0, 1)$ denote the probability of successful correction. For example, it could reflect the functioning of a dispute resolution mechanism or remedial action by non-target states. The monitor's expected payoff is $\lambda u_m(imp) + (1 - \lambda)u_m(fail)$. In other words, with probability λ the monitor is able to secure project implementation through information revelation. With probability $1 - \lambda$, this effort fails. Given that $u_m(imp) > u_m(fail)$, the monitor prefers a high correction probability λ.

Given these primitives, the 'disagreement payoffs' are zero and $\lambda u_m(imp) + (1 - \lambda)u_m(fail)$. If the monitor refuses to collude, the target state obtains the worst possible payoff while the non-target state secures the expected payoff from correction.

Assuming exogenous correction leaves the exact mechanism inside a black box. In practice, international institutions often allow states to restore compliance instead of a permanent suspension of cooperation (Chayes and Chayes, 1995; Rosendorff, 2005). For

Table 2. Notation.

Symbol	Interpretation
$E, e(E)$	Compliance by target state, cost of compliance effort
$M, m(M)$	Monitoring, cost of monitoring effort
$u_s(imp), u_m(imp)$	Successful implementation payoffs for target state and monitor
$u_s(imp), u_m(fail)$	Failed implementation payoffs without scandal for target state and monitor
λ	Probability of correcting the implementation failure upon revelation of non-compliance
G	Probability of external verification upon monitoring failure or collusion
0	Target state's scandal payoff
ε	Monitor's scandal payoff
U_s^{coll}, U_m^{coll}	Collusion payoffs endogenously derived from a Nash bargaining game

example, the probability λ could depend on the target state's ability and incentive to improve a tarnished reputation by publicly restoring compliance. In addition, λ could depend on the international institution's ability to impose sanctions on the target state (Downs et al., 1996). In the mathematical Appendix, we extend the model by endogenizing the target state's compliance decision and showing how the correction probability λ can be derived from strategic behavior without changing any of the conclusions of the simpler, exogenous model.

If the monitor colludes, external verification is again possible. With probability G, the target state and the monitor are caught in a scandal. In this case, they obtain the payoffs 0 and ε. With probability $1 - G$, they avoid a scandal: collusion is not revealed. In this case, the target state obtains the maximal payoff $u_s(imp)$ while the monitor obtains the failure payoff $u_m(fail)$. However, bargaining allows the target state to offer a side payment to the monitor. The expected surplus available for distribution is thus $(1 - G)(u_s(imp) + u_m(fail)) + G\varepsilon$.

In the Appendix, we show that collusion occurs in equilibrium whenever

$$(1 - G)(u_s(imp) + u_m(fail)) + G\varepsilon > \lambda u_m(imp) + (1 - \lambda)u_m(fail). \quad (1)$$

The left-hand side is the expected surplus available from collusion. The right-hand side is the sum of expected payoffs without collusion: the state faces a scandal because the monitor reveals information. All model notation is summarized in Table 2.

4.2. Equilibrium

We offer an informal summary of the equilibrium analysis in the main text. The technical details are found in the mathematical Appendix. Since the only source of uncertainty pertains to randomly drawn variables by nature, our solution concept is the subgame-perfect equilibrium.[18] An equilibrium of the game characterizes monitoring and compliance levels E^*, M^*, as well as bargaining strategies for the choice between collusion and information revelation.

Following the initial monitoring and compliance decisions, the game has three subgames of interest:

1. If the target state successfully implements the project, the game ends: there is nothing the monitor can do. **Probability**: E^*.
2. If the target state defects but the monitor does not acquire any evidence, the game again ends: the monitor cannot do anything. **Probability**: $(1 - E^*)(1 - M^*)$.
3. If the target state defects but the monitor obtains evidence, the monitor must choose: should it (i) collude with the target state or (ii) publicize the defection in view of corrective measures by the international institution or non-target state. The monitor selects collusion when (i) external verification is ineffective, so that collusion probably goes unnoticed and (ii) publicizing the defection probably does not prompt corrective measures by the international institution or non-target state. **Probability**: $(1 - E^*)M^*$.

To understand the equilibrium, the collusion decision is particularly important. Recall that collusion occurs whenever the target state is willing to pay enough to avoid a scandal. The target state's willingness to pay depends on the external verification probability G. When it is high, collusion is not profitable because external verification deprives the target state of the gains.

Suppose now collusion is mutually profitable. The target state and the monitor distribute the available bargaining surplus depending on the strength of their 'outside options,' namely payoffs upon information revelation by the monitor. If the monitor expects a large payoff from information revelation, it is in a strong bargaining position. If the monitor expects a small payoff from information revelation, it is in a weak bargaining position.

The expected payoffs from these games influence the strategically prior monitoring and compliance decisions. The best responses are characterized in Figure 2. For any parameters of the game, the monitor is only willing to invest heavily in monitoring if the probability of compliance failure is high. After all, if the target state complies, there is nothing that the monitor can do. Similarly, the target state is only willing to pay the compliance cost if monitoring is intrusive. If the probability of being caught is low, the target state can free ride without punitive consequences. Importantly, the figure also indicates that the game has a generically unique equilibrium, so our theory generates unequivocal predictions and is thus particularly suitable for systematic empirical applications.[19]

4.3. Empirical implications

Based on this analysis, we derive several empirical predictions concerning compliance and monitoring in an international institution. There are two relevant issues. First, is the monitor willing to collude? Second, given this decision, what are the equilibrium levels of compliance and monitoring?

Figure 3 represents the conditions under which the monitor colludes. The figure shows that when the monitor's revelation payoff is low, it *colludes*. But even then, an increase in the revelation payoff causes an increase in the monitor's equilibrium payoff due to a stronger outside option in bargaining. When the monitor exceeds the collusion threshold, above which collusion is no longer profitable, the monitor's equilibrium payoff begins to increase at a faster pace.

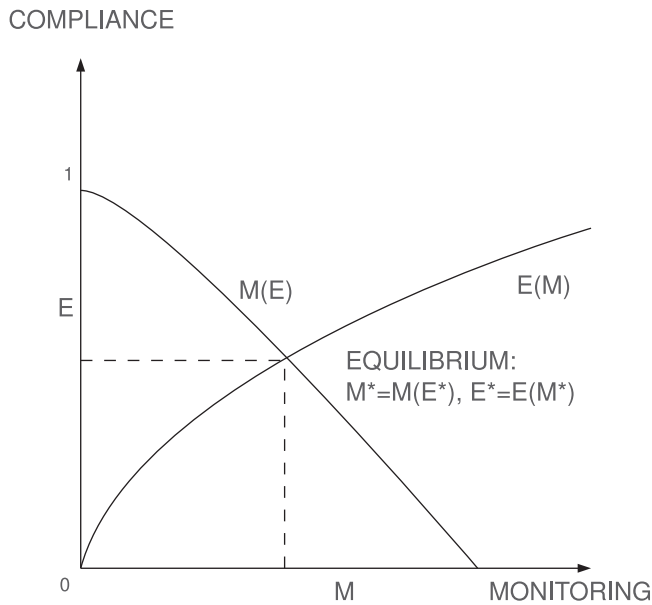

Figure 2. Best responses $M(E)$ and $E(M)$. The target state has an incentive to decrease compliance as monitoring decreases, while the treaty monitor selects a higher monitoring effort as compliance decreases.

4.3.1. Monitor's incentives Given this logic, how do the monitor's incentives to reveal information influence equilibrium compliance?

Proposition 1 (Monitor's incentives). *Increasing the correction probability, λ, or the monitor's valuation of treaty implementation, $u_m(imp)$, increases equilibrium compliance E^*. The relative size of this positive marginal effect slightly below and above the collusion threshold is ambiguous.*

This proposition examines how the monitor's belief in the international institution's ability to restore treaty implementation (λ) and valuation thereof ($u_m(imp)$) influence equilibrium compliance. As the value of either parameter increases, the monitor's payoff from revelation increases. This increases the monitor's payoff with (stronger outside option improves the bargaining position) and without (direct increase in payoff) collusion. The monitor's incentive to monitor increases, so the state has more incentives to comply.

Interestingly, it is unclear whether the positive marginal effect of increasing the values of λ or $u_m(imp)$ is higher or lower when the collusion threshold is reached. Intuitively, one would expect this marginal effect to be higher in the absence of collusion. However, this reasoning is invalid. In the presence of collusion, an increase in the monitor's revelation payoff has two effects: the monitor becomes more aggressive in monitoring *and* the target state's payoff from collusion decreases. Without collusion, only the monitor's behavior changes, but the change is larger. The relative size of these two effects is unclear.

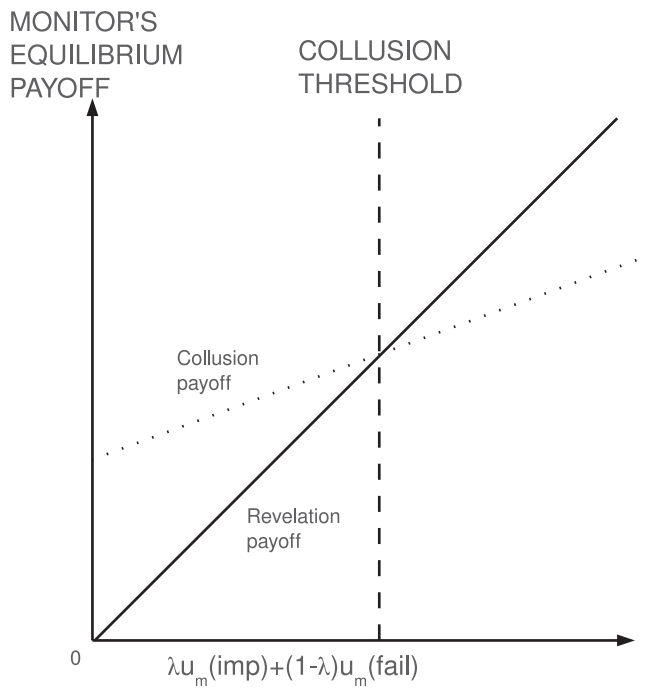

Figure 3. Collusion versus information revelation. The vertical dashed line is the collusion threshold, below (above) which the monitor colludes (not). The x-axis denotes the monitor's revelation payoff, while the y-axis denotes the monitor's equilibrium payoff. The dotted (solid) line denotes the monitor's expected payoff from collusion (information revelation). Note that the expected collusion payoff also increases with the revelation payoff on the x-axis.

A graphic illustration of this result is given in Figure 4. The left (right) panel illustrates the case in which the increase is larger in the revelation (collusion) region. This occurs when the monitor's behavior strongly reacts to changed incentives.

This proposition shows how monitors can be made more effective. If the monitors have a strong interest in successful international cooperation, it is important to ensure that the international institution *corrects* the failure with high probability λ. Similarly, it is important that the monitor is rewarded for detecting noncompliance, so that $u_m(imp)$ is high. In addition, the state should not be in a position to punish the monitor for revealing information. These predictions help us understand why some international institutions use certain monitors and what determines the probability of success and failure.

A relevant example of these virtues might be the WTO (Dai, 2002; Downs and Rocke, 1995; Rosendorff, 2005). A state's ability to litigate illegal foreign trade policies depends on information from exporters who suffer from discrimination. Since the WTO has a reasonably strong record in resolving trade disputes, exporters may expect monitoring to pay off. They can examine the characteristics of the situation, ranging from bargaining power between the trading states to the availability of precedents, to decide whether or not monitoring could lead to correction. In conducive circumstances, the WTO would

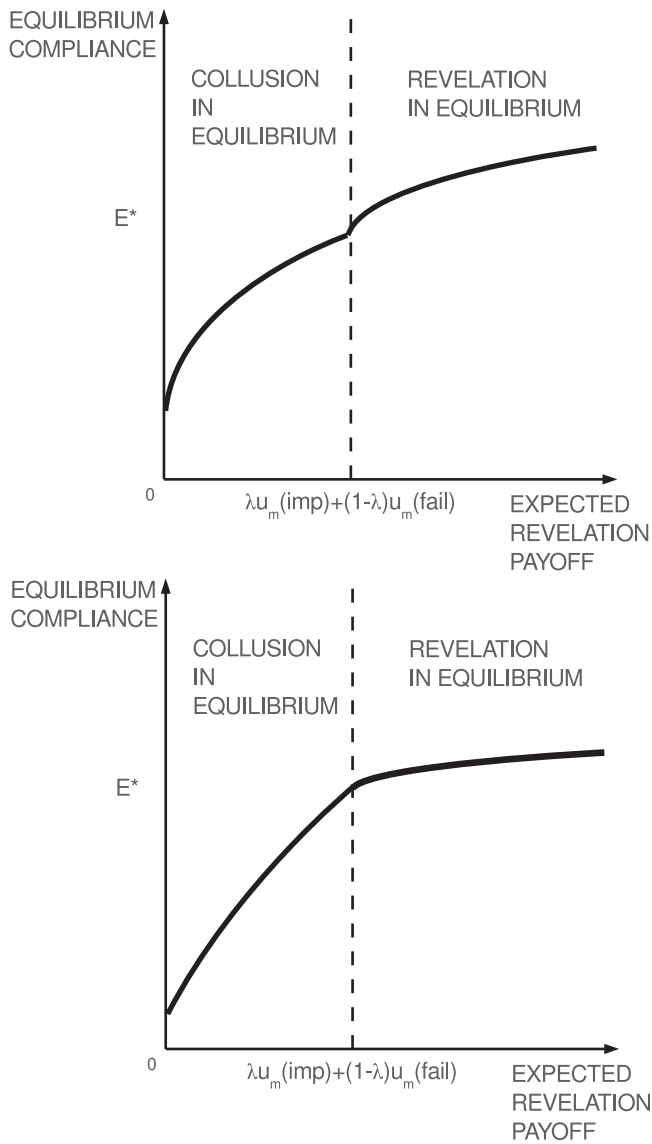

Figure 4. Equilibrium compliance as a function of the revelation payoff to the monitor. As information revelation begins to dominate collusion (dashed vertical line), the marginal effect of increasing the revelation payoff (x-axis) on equilibrium compliance E^* changes. The left (right) panel illustrates the case in which the marginal effect of changing monitoring incentives is larger (smaller) in the revelation region.

have a relatively high correction probability λ. This gives exporters strong incentives to monitor and report violations.

Importantly, the WTO also does not threaten defectors with ruthless punishments. Instead, it attempts to resolve the dispute and restore trade cooperation. Suppose that the WTO instead threatened defectors with ruthless punishments, so that λ would be low because defectors would immediately be driven into a costly and lengthy trade war (Rosendorff, 2005). In this case, an exporter might fear a trade war if non-compliance were made public and therefore choose not to reveal information. Indeed, Schwartz and Sykes (2002, 179) persuasively argue that the WTO dispute settlement mechanism has been successful at leading to peaceful conflict resolution precisely because it 'ensures that sanctions are not set too high.' Similarly, Busch and Reinhardt (2000) note that most WTO disputes never reach the panel stage, suggesting indeed the dispute settlement works best not because it can impose strong sanctions, but rather because it gives states incentives to settle their disputes.

4.3.2. State's incentives Consider now the state's incentives.

Proposition 2 (State's incentives). *Increasing the state's valuation of avoiding a scandal, $u_s(imp)$, increases compliance E^* and reduces monitoring M^*. However, it also increases the probability of collusion conditional on treaty implementation failure.*

This proposition shows that changes in the state's compliance incentives have a dual effect. On the one hand, if the state really wants to avoid a scandal it has a stronger incentive to comply. This is good news for compliance.

On the other hand, suppose that the state is indeed caught. In this case, the state has a stronger incentive to collude than previously because it wants to avoid the scandal even at a high cost. This creates incentives to collude; states that are inclined toward compliance are also inclined toward collusion.

4.3.3. External verification Let us finally examine how the probability of external revelation of implementation failure, G, influences equilibrium compliance. Intuition tells us that if the risk of being caught is high, the monitor and the state would have a strong common incentive to secure compliance. This reasoning fails.

Proposition 3 (External verification). *In the collusion region, increasing the probability of being caught G has an ambiguous effect on compliance and monitoring. In the information revelation region, increasing G causes an increase in compliance.*

In the collusion region, it is not clear whether increases in the probability of being caught increase or reduce compliance. On the one hand, the state has an incentive to increase compliance because it can thus reduce the risk of being caught. On the other hand, the monitor has already decided to collude, despite the risk of being caught, and this decision is not influenced by a small increase in the risk of being caught. However, the positive effect of monitoring on payoffs is *conditional* on not being caught, as seen in expression (11), so the monitor actually has *less* incentive to monitor if the probability of being caught increases.

When information revelation is possible, the effect of external verification is *amplified* by improved compliance. Not only does the state have stronger incentives to comply, but the monitor also increases monitoring for any given level of compliance. Why should

the monitor be more aggressive in monitoring if external verification is possible? Under external revelation, the monitor is implicated for negligence, so the monitor has a strong incentive to avoid this outcome by obtaining evidence. Consequently, improved external revelation produces a 'double dividend:' in addition to directly reducing the state's incentive to defect, it improves the performance of a monitor. Yet in equilibrium, actual monitoring effort might decrease because the deterrent effect induces so much more compliance that monitoring is no longer quite as profitable.

The International Monetary Fund (IMF) is a relevant example of these dynamics. While Dai (2002) correctly recognizes that states have few feasible alternatives to centralized monitoring of macroeconomic activity, she does not address the possibility that macroeconomic mismanagement is bound to become public as a crisis erupts. In the model, the IMF would have a high G because a crisis is not something the monitor or the target state has to reveal: financial markets observe and publicize the situation. When it fails to intervene before a crisis erupts, the IMF officials responsible for macroeconomic monitoring suffer a reputational loss.

If the IMF officials understand this, they also understand that they will be implicated unless they are able to warn the IMF Executive Board about future crises. Consequently, the logic of external verification allows the IMF to more effectively monitor macroeconomic policies. Not only are there few alternatives to centralized monitoring, but the incentive structure produced by a high external verification probability G also renders the IMF an unusually effective monitor.

For example, the IMF failed to foresee the financial crisis that Mexico fell into in 1994. As Woods (2006) notes, 'evidence suggests that the experts failed fully to recognize the risks faced by Mexico and failed to consider anything other than optimistic scenarios for the country.' The IMF was heavily criticized for its failure to predict or prevent the crisis. As one journalist noted, 'the fund has only recently begun beefing up its expertise in coping with weak banking systems and the dangers arising from "hot" money flows across national borders, which many analysts blame for Mexico's crisis in 1995 and the Asian crises of the past nine months'.[20] In response to the crisis, the IMF noted that they also needed to improve, so they could prevent future crises. In an address given by the Managing Director of the IMF the year following the Mexican crisis, he noted that 'the Fund also needs to make more use of financial market data in monitoring developments. We shall also be working toward the establishment of standards for the timely publication of economic data by members to enable markets to work more efficiently'.[21] By 1998, the IMF had done just that and the IMF announced that 'it might withhold assistance to countries that do not agree to give full and timely warnings about their financial difficulties'.[22]

5. Case study: The United Nations Oil for Food Program

To illustrate how the model fits with what can be empirically observed, we examine the case of Oil for Food Program implemented by the United Nations from 1995 to 2003. We chose this case to illustrate two key elements of our theoretical argument. First, Saddam Hussein's regime was able to export more oil than was allowed by colluding with monitors. Second, monitors lacked the intrinsic motivation to not collude, and thus fell

prey to the offers of collusion by Hussein's regime. These observations provide empirical insights into the strategic logic of state–monitor interactions.

5.1. Background

During the occupation of Kuwait by Iraq in August 1990, the United Nations Security Council issued Resolution 661, beginning an era of sanctions against Iraq by the international community. After the Gulf War ended, the Security Council reaffirmed the sanctions against Iraq with UNSC Resolution 687, with the exceptions of foodstuffs, medicine, and health supplies.[23] Hussein did not comply with many of the provisions of the resolution by not allowing weapons inspectors into the country, and thus the sanctions continued. The sanctions had a devastating effect on the Iraqi civilian population. As a report from a United Nations fact finding mission reported, Iraq was already in bad shape as 'the recent conflict has wrought near-apocalyptic results upon the economic infrastructure of what had been, until January 1991, a rather highly urbanized and mechanized society'.[24] In the summer of 1991, the United Nations proposed to the Hussein regime a 'oil for food' program, whereby Iraq would be allowed to sell oil, but would get food in return delivered to the civilian population, yet Hussein turned it down (Meyer and Califano, 2006, 3).

By 1995, the situation in Iraq had deteriorated drastically. A second attempt was made to propose the program. This new one had features that were intended to take into account Hussein's previous opposition to such a program, and this time the program gave more discretion over implementation to the Secretary General and did not propose more on the ground monitors (Meyer and Califano, 2006, 3–4). Resolution 986 was passed in 1995 creating the Oil-for-Food Program. The negotiation over implementation lasted through 1996 and in 1997 the program was up and running.

The Oil-for-Food Program allowed Iraq to sell a limited amount of oil, the proceeds of which were put in an account managed by the Secretary General, and the money was only to be used to buy medicine, health supplies, foodstuffs, and materials and supplies needed for the Iraqi people.[25] The program grew in scope over the years. By 1999, the Security Council had removed any limits to the amount of oil that could be sold, and the types of goods that could be purchased grew from basic food and medicine, to construction supplies and cultural funds.

5.2. Weak monitoring and corruption

Once the Oil-for-Food Program started, the Hussein regime sought in various ways to profit from the program and generate revenue for itself. As the Independent Inquiry Committee report on the Oil-For-Food Program (henceforth, the Volker report) concluded, by the end of the program it is estimated that the Hussein regime manipulated the program and generated over US$10 billion in illicit income (Volker et al., 2005, 1). In terms of our model, Hussein's efforts to generate illicit income can be thought of as minimal investments into compliance. Had Hussein instead forgone the income, this would have allowed a higher probability of complying with the regime. There were three main ways through which Iraq managed to generate revenue: surcharges on oil sales, smuggling oil,

and kickbacks on humanitarian goods. Each of these was possible due to weak or corrupt monitoring.

5.2.1. Illegal surcharges The Oil-for-Food Program allowed Iraq to choose the companies to which it would sell oil. This gave the Iraq leverage, as it could favor some companies and countries over others. In turn, this allowed Saddam's regime to generate money by charging an illegal surcharge per barrel that the companies had to pay directly to them. In other words, Hussein's regime looked for companies that were willing to give them a surcharge for the sale of oil, even though this was not authorized by the program (Volker et al., 2005, 3). Initially the surcharge was included in the contracts and companies deposited the surcharge to Iraq. This lasted until 2002 when the United Nations finally declared the surcharges to be illegal. Yet, why did the surcharge last as long at it did? After all, the Hussein regime managed to generate over US$228 million between September 2000 and August 2002 in surcharges (Volker et al., 2005, 21).

The monitoring of the entire program was under the control of the United Nations Security Council 661 Committee. This committee was created to oversee all aspects of the Iraq sanctions, and subsequently the implementation of the Oil-for-Food Program. It included all 15 members of the Security Council, and the body had a consensus voting rule. Any decisions had to be approved by all 15 members. The 661 Committee established a set of Oil Overseers in charge of keeping track of the oil sales in Iraq, including approving the contracts Iraq made with various companies. Although the 661 Committee was in charge of monitoring, they were not empowered or guided to do anything if they did detect a problem. As the Volker report noted, '[a]lthough established as a monitoring body, the 661 Committee's rules did not require it to take action of any kind in response to a report of a violation of the sanctions regime or a violation of the Programme' (Volker et al., 2005, 21). Thus, any response to a reported violation of the program would have to come from the permanent members, which included Russia and France, the two countries receiving the bulk of the oil contracts.

When the Oil Overseers did report to the Security Council the existence of the illegal surcharges, the Security Council did nothing (Volker et al., 2005, 29). This is consistent with the prediction of the model showing that when the probability of correction is low, monitors are less likely to gather or reveal information. Indeed, even after the Oil Overseers brought up the problem of the illegal surcharges, they were ignored, and even when they told companies the surcharges were illegal, everyone continued to pay them. The end of the surcharges happened not because the monitors detected the non-compliance, but because Iraq raised the price of the surcharges and buyers were not willing to pay it. In addition, the United States and the United Kingdom changed the setting of the price of oil to the market value at the time the oil was lifted, not at the time the 661 Committee set the price. Thus, by changing the price based on a retroactive rule, the buyers were less confident that a surcharge plus the market value would generate them a profit over buying oil elsewhere (Volker et al., 2005, 29–30).

5.2.2. Smuggling oil Another way that Hussein generated revenue was by selling oil outside of the program and getting a larger share of the revenue. Iraq's Oil Ministry was eager to sell more oil than it was allowed to sell, and keep as much of the money as

possible. Thus, the ministry proposed to two oil companies, Ibex Energy and Trafigura Beheer, the sale of illegal barrels. The proposed scheme would consist of these two companies buying legal oil from Iraq, proceeds to be paid to the United Nations escrow, but to add more barrels to their shipment, and the companies would sell that oil. Iraq would then receive 80% of the proceeds from those extra barrels (Volker et al., 2005, 176).

The key was to find a way to smuggle oil out of the country. The Ministry of Oil knew that Mina al-Bakr was a good port to use to smuggle the oil, since the port lacked oil metering. The only way to verify the quantities of oil loaded unto the ships was by the inspectors. The United Nations did not have their own inspection personnel, and instead had hired Saybolt Eastern Hemisphere BV, a private company that specializes in inspection, monitoring, and verification for oil, gas, and other chemicals. The Saybolt inspector on site was Armando Carlos Oliveira, and when the time came to smuggle the oil, Iraqi officials approached him and offered him a bribe in order to not report the extra barrels, and sign the appropriate paperwork authorizing the shipment (Volker et al., 2005, 179). Oliveira accepted the bribe and after the ship was loaded with the allowed amount, Oliveira and his team left the site, and then over 200,000 extra barrels of oil were loaded onto the ship. The Volker investigation found evidence that this happened on two occasions and this earned Iraq approximately US$10 million (Volker et al., 2005, 176).

This instance of smuggling would not have been detected were it not for the captain of the ship, Childakis Theofanis, who noticed the activity and wrote a letter to a United Nations Oil Overseer informing him of the extra barrels he had transported in his ship (Meyer and Califano, 2006, 98). And yet, even after his report, the United Nations did little to stop subsequent transports, did not increase monitoring, and simply asked Iraq to install oil meters in the port, which Iraq did not do. Although this was an instance where the smuggling and corruption was detected, a lot more smuggling occurred, as it is estimated that Iraq smuggled oil worth somewhere between US$5.7 and 8.4 billion.[26]

The bribery of the Saybolt officer, and likely many others that went undetected, is consistent with the predictions in the model. First, the monitors hired were private companies that had no investment or stake in the outcome. The companies were hired under a contract to do a job and get paid for it. The individual monitors on the ground were not invested in making sure Iraq complied with the program, and for this reason it made them easier to bribe. Second, the monitors knew that even if they did not accept bribes and reported non-compliance, there was little that was going to happen, and indeed in the one case where the captain of the ship was a whistleblower, nothing happened to either Iraq or the monitor. All of this is consistent with Proposition 1 of the model.

As Proposition 3 of the model shows, when external verification is unlikely, collusion will be common. This was the case here, as the only way to verify the shipments of oil were in compliance with the program rules was through the monitor. Of the estimated US$5–8 billion of profit generated by smuggled oil, there were only US$10 million that were detected, and this was due to one captain that notified the United Nations. Thus, there were no positive incentives for the monitor to do a good job, or negative incentives for failing to do a good job. Unsurprisingly, the monitors chose to collude.

5.2.3. Humanitarian kickbacks Iraq also raised funds from kickbacks on the contracts on the commodities that were to be distributed in Iraq, mainly the foods and medications. Just as it was allowed to choose the companies to which it sold oil, Iraq could also choose the companies that would receive contracts for the humanitarian goods. Hussein's regime began to charge a transportation fee that companies had to pay directly to Iraq. Subsequently Iraq began to demand a 10% kickback on all contracts, calling it an after-sales-service fee (Volker et al., 2005, 249). In order for the companies not to lose profits due to these surcharges, Iraq would inflate the cost of the goods in the contract, or have the companies distribute lower-quality goods than those provided and paid for in the contract. It is estimated that more than 2,200 companies went along with Iraq's demands and paid Iraq kickbacks amounting to US$1.5–3.5 billion (Volker et al., 2005, 249).

The kickback scheme was possible due to weak monitoring. The 661 Committee did not do a diligent job in reviewing the contracts and checking that there was no price inflation. In addition, the companies that were hired to monitor shipments of goods on the ground were not mandated to verify that the foods were of the quality that was contracted or determine the value of the shipment, nor were they to report if they detected smuggled goods. Their task was mainly to report that a shipment of a set quantity was delivered, so the UN could pay the contract (Volker et al., 2005, 26). Although, the final report of the investigation did not find that the goods monitors colluded with Iraq, it did find that monitors were badly managed and performed poorly (Volker et al., 2005, 475). Thus, Iraq was able to take advantage of the weakness of these monitors.

Proposition 1 states that the payoff to the monitor depends on the international institution's ability to restore treaty implementation. Once again, the monitors had no reason to believe that the United Nations would do anything about the reports of smuggled goods, and indeed seem frustrated that they were not mandated to do anything about the smuggling they observed (Volker et al., 2005, 474). Moreover, realizing that the United Nations was unresponsive and only cared if a shipment was received, the monitor did a weak job, not even meeting the industry standards for monitoring, as was reported by one of the former employees that testified before congress (Volker et al., 2005, 475). Uninvested and with no incentives to perform, weak monitoring emerged.

The theory we presented suggests that monitoring effectiveness depends on three key factors: the types of monitors, their incentives to gather information, and the possibility of external verification. This case illustrates the importance of these issues and shows how the model can help us understand compliance monitoring. With non-invested monitors with no incentives to do a good job, and little way to verify failures otherwise, the Iraq regime succeeded circumventing the goals of the program and the international sanctions against it.

6. Toward effective international monitoring

We summarize the implications of our study for international institutions in the form of two key lessons. For each lesson, we provide an application to the emerging global climate regime.

First, successful monitoring depends on the enforcement regime in surprising ways. While political economists, such as Carrubba (2005) and Downs et al. (1996), emphasize the importance of strong sanctions, overly ruthless sanctions could induce monitors to

hide information regarding non-compliance when cooperation is likely to break down as a result. If monitoring is a difficult challenge, strong sanctions could have negative consequences. Assuming ruthless sanctions reduce the correction probability λ because cooperation breaks down (Chayes and Chayes, 1995), a monitor has few incentives to reveal information rather than collude. Given this, target states need not worry about information revelation, and they can comply only to the level necessary to avoid external verification in the future.

This suggests a rationalist argument for limiting sanctions. However, being unresponsive to reports of non-compliance also increases the likelihood of weak monitoring, as monitors will see that their work is fruitless. There is an optimal, moderate level of response that maximizes compliance. Sanctions can be deemed too severe if they have a strong negative effect on the monitor's incentives to reveal information, tilting the balance in favor of collusion, and not severe enough to discourage the target state from non-compliance, given the opportunity to collude.

In the debate on a future climate regime, many have proposed that carbon tariffs help countries commit to climate cooperation (Barrett, 1997; Lessmann et al., 2009; Veel, 2009; Weber and Peters, 2009; Zhang, 2009). While the threat of trade sanctions could provide incentives for compliance, trade sanctions, by virtue of creating conflict among nations, especially in the North–South context, may increase the risk that climate cooperation is altogether suspended. If parties supposed to monitor climate cooperation want to avoid this outcome, they may have perverse incentives not to monitor national pledges or reveal instances of non-compliance. Our formal analysis provides a potential reason for being cautious about applying high carbon tariffs, especially since monitoring implementation on the ground in this area is particularly difficult. This insight is particularly important if the future climate regime will be built on variegated national commitments that are difficult to monitor.

Second, our analysis provides a rationalist foundation for selecting monitors with an intrinsic interest in treaty implementation and cooperation (Dai, 2002; Mitchell, 1998). Intrinsic motivation is difficult to identify, but one possible rule of thumb would be to see if a potential monitor, such as an NGO, would reap direct benefits (donations, new members, opportunities to profit) from compliance or alternatively be directly hurt by non-compliance. Using such monitors both reduces the cost of monitoring and helps avoid collusion. In contrast, disinterested bureaucrats must operate under effective external verification and generous rewards to resist the urge to collude. If intrinsically motivated monitors are not available, states can attain effective monitoring when one of two conditions is met: either external verification is available or the monitors can be generously compensated for revealing information about non-compliance. Similarly, our model also suggests that external verification might increase with the presence of multiple monitors. It is easier for target states to collude with one actor than with many.

In the case of the climate regime, these observations can help evaluate the ability of different monitors to improve compliance with treaty commitments. Given that the question of Measurement, Reporting, and Verification (MRV) was among the most contentious in the 2009 Copenhagen Conference, the policy importance of such results is pronounced. Somewhat surprisingly, however, strategic monitoring incentives are frequently overlooked in the policy debate. Breidenich and Bodansky (2009) discuss various

monitoring arrangements from national reporting to centralized international verification, but they overlook the fact that the efficacy of alternative monitoring arrangements depends on monitor incentives to find and reveal information. Busby (2010, 9) notes that China was a particularly strong proponent of having domestic monitors only: the Chinese delegation argued that outside monitors, whether other states or an international institution, would be excessively intrusive and violate state sovereignty. However, he does not offer an analysis of (i) the strategic reasons why China might be hostile to international monitoring or (ii) the effectiveness of domestic monitors for monitoring climate commitments. Michonski and Levi (2010) examine the degree to which extant international organizations may facilitate monitoring for the climate regime, but they focus on institutional capacity and omit the fact that strategic organizational incentives will also affect their monitoring efficacy.

Our theory can inform this debate. While the secretariat of the United Nations Framework Convention on Climate Change could, at least in principle, monitor treaty commitments in a centralized fashion, several problems exist. In addition to the heated sovereignty debate, especially in China and other rapidly industrializing countries, centralized monitoring may be problematic unless the international bureaucrats are genuinely committed to expending effort to reveal non-compliance. By contrast, interested monitors on the ground may be in a better position to reveal non-compliant behavior. In this case, however, it is integral to ensure that potential monitors are not themselves susceptible to influence by the target states. For these reasons, the ideal institutional design may well comprise (i) a network of voluntary monitors in different nations and (ii) a centralized verification agency that evaluates these monitors.

7. Conclusion

We have provided a theory that explains the variation in the effectiveness of monitoring of international agreements. Our central theoretical contribution is to highlight the conditions under which monitoring fails and succeeds. In particular, we have emphasized that successful monitoring depends on simultaneously (i) reducing the monitor's incentive to collude with non-compliant states and (ii) ensuring that information revelation elicits a favorable response. If both conditions hold, high levels of compliance, as well as successful corrective action, can be expected.

Based on this strategic logic, we derived two surprising yet important implications for treaty design. First, international institutions can improve monitoring incentives by refraining from ruthless sanctions, because monitors are unwilling to provide information that causes the regime to collapse. Second, normative monitors have a strong advantage over private companies or those not invested in compliance, because the latter are particularly susceptible to collusion.

The broader implications of our theory are significant both for scholars and practitioners. For scholarship, a particularly important lesson is the need for much greater precision in the analysis of the role that various non-state actors play in international cooperation. As our formal analysis demonstrates, the strategic interactions that influence international cooperation are highly complex and contingent. Future theoretical research should strive for models that extend and modify our core insights. Empirically, we foresee benefits from systematic hypothesis testing of our key comparative

statics. Even more broadly, our contribution demonstrates how scholars of international institutions can investigate design interactions in international institutions, thus reaching beyond the boundaries of 'rational design', as currently understood in the literature (Koremenos et al., 2001).

While we have not offered an empirical test, our model opens the door for testing in this area by providing concrete features of monitors and their environments that can be analyzed. First, the model has implications for the diverging incentives of various monitors. We expect that normatively committed and relatively independent monitors (i) induce greater compliance by states and (ii) collude with non-compliant states less frequently than other monitors. This expectation could be tested against quantitative data on the objectives of treaty monitors in different nations, as well as their dependence (financial, political) on the target state. For example, one may expect that domestic monitors are more (i) willing to collude with their home state and (ii) susceptible to state pressure than international monitors. Second, the theory could be tested by exploring how centralized treaty monitors, such as the Landmine Monitor, allocate their resources. In addition to monitoring potential defectors, we have found that treaty monitors expend effort when they expect the sanctions for defection to be effective but not overly costly. Since centralized treaty monitors allocate resources across multiple state parties, quantitative tests are feasible.

For practitioners, our incentive analysis also provides useful insights. By providing monitors with the right incentives, practitioners can improve treaty effectiveness and thus use international law to solve some of the difficult collective-action problems that the international community currently faces. However, the design of such incentives is bound to be a complex enterprize. Most importantly, our analysis emphasizes the importance of (i) avoiding overly ruthless sanctioning regimes and (ii) attracting treaty monitors with deep normative commitment. These observations could, for example, inform the design of a future climate regime, especially if it will be based on national pledges that allow variegated and flexible commitments to policies and measures.

Appendix

Nash bargaining for collusion

Suppose that implementation fails and monitoring succeeds. Now the state and the monitor must decide on collusion. This decision is modeled as a conventional Nash (1950) bargaining game. In this game, the monitor and the state simultaneously select demands $d_m, d_s \in [0, 1]$ for their share of the total value of collusion. If $d_m, d_s \leq 1$, then bargaining succeeds. Otherwise it fails.

Payoffs from bargaining failure reflect the monitor's decision to reveal information. A scandal ensues, so the state's payoff is zero. The monitor's decision depends on the success or failure of correction. The expected payoff is $\lambda \cdot u_m(imp) + (1 - \lambda) \cdot u_m(fail)$.

When bargaining succeeds, the divisible surplus is

$$(1 - G)\left[u_m(fail) + u_s(imp)\right] + G\varepsilon - \left[\lambda u_m(imp) + (1 - \lambda)u_m(fail)\right]. \qquad (2)$$

The multiplier $1 - G$ reflects the fact that with probability G, all surplus except the monitor's ε is lost because a scandal ensues.

The Nash bargaining solution requires that the equilibrium shares d_m^*, d_s^* maximize the Nash product. Given the disagreement points, the Nash product is U_s^{coll} multiplied by $U_s^{coll} - \lambda \cdot u_m(imp) + (1-\lambda) \cdot u_m(fail)$, where U_i^{coll} is player i's bargaining payoff. By standard arguments, each player obtains $\frac{1}{2}$ of the divisible surplus. In equilibrium,

$$U_s^{coll} = \frac{1}{2}(1-G)[u_m(fail) + u_s(imp) + G\varepsilon] - \frac{1}{2}[\lambda u_m(imp) + (1-\lambda)u_m(fail)]; \quad (3)$$

$$U_m^{coll} = \frac{1}{2}(1-G)[u_m(fail) + u_s(imp) + G\varepsilon] + \frac{1}{2}[\lambda u_m(imp) + (1-\lambda)u_m(fail)]. \quad (4)$$

This completes the analysis of the Nash bargaining subgame.

Equilibrium

The game has a generically unique equilibrium, as shown through backward induction in this section.

Successful implementation To begin with, if the state successfully implements the project, the payoffs to the state and the monitor are respectively $u_s(imp) - e(E^*)$ and $u_m(imp) - m(M^*)$. In equilibrium, the probability of this outcome is E^*. Both the state and the monitor have expended efforts, but since implementation succeeds, collusion and information revelation are irrelevant. Both players obtain relatively high payoffs.

Implementation and monitoring failure If implementation and monitoring fail, the payoff to the state is

$$(1-G) \cdot u_s(imp) - e(E^*). \quad (5)$$

With probability $1-G$, the state avoids punishment for implementation failure. For convenience, we write $V_s = (1-G) \cdot u_s(imp)$. The payoff to the monitor is

$$(1-G) \cdot u_m(fail) + G\varepsilon - m(M^*). \quad (6)$$

With probability $1-G$, the monitor is not being accused of negligence because the implementation failure is never observed by any third party. However, the monitor obviously cannot obtain the high implementation payoff $u_m(imp)$. We also write $V_m = (1-G) \cdot u_m(fail) + G\varepsilon$.

Implementation failure and monitoring success The most interesting subgame is when implementation fails but monitoring succeeds. In this instance, the monitor must decide on information revelation. If the monitor reveals the evidence, her expected payoff is

$$\lambda \cdot u_m(imp) + (1-\lambda) \cdot u_m(fail) - m(M^*), \quad (7)$$

as the international institution corrects the implementation failure with probability λ. If the monitor colludes, her expected payoff is

$$U_m^{coll} - m(M^*). \quad (8)$$

With probability $1 - G$, collusion succeeds and the implementation failure is never observed by third parties.

The monitor colludes if and only if

$$U_m^{coll} \geq \lambda \cdot u_m(imp) + (1 - \lambda) \cdot u_m(fail). \tag{9}$$

The left-hand side of this condition is the expected bargaining payoff from collusion, U_m^{coll}. The right side is the expected payoff from information revelation, as the international institution attempts to correct the implementation failure, perhaps by imposing sanctions. For future reference, we define $W_m(1) = U_m^{coll}$ and $W_m(0) = \lambda \cdot u_m(imp) + (1 - \lambda) \cdot u_m(fail)$. These are the expected monitoring payoffs from collusion and information revelation, respectively.

Upon implementation failure and successful monitoring, the expected payoff to the state is $-e(E^*)$ if (9) fails and $U_s^{coll} - e(E^*)$ otherwise. All else being constant, the state prefers collusion to being punished for compliance failure. We also define $W_s(1) = u_s^{coll}$ and $W_s(0) = 0$ for future reference. These are the expected payoffs from collusion and information revelation to the state, respectively.

Compliance and monitoring The analysis above characterizes all possible subgames of the game after compliance and monitoring decisions. What remains is to find the equilibrium levels of compliance and monitoring. Note in particular that since all subgames have a generically unique equilibrium, the uniqueness of compliance and monitoring best responses suffices for the existence of a generically unique equilibrium.

Given a compliance level E^*, the monitor selects a level of monitoring M to maximize

$$E^* \cdot u_m(imp) + (1 - E^*) \cdot (M \cdot W_m + (1 - M) \cdot V_m) - m(M). \tag{10}$$

The first term is the expected payoff given successful implementation. The second term is the expected payoff upon implementation failure, so that monitoring is potentially relevant. The first-order condition is

$$(1 - E^*) \cdot (W_m - V_m) = m'(M). \tag{11}$$

This condition defines a unique equilibrium because we have assumed that $m'(0) = 0, m'(1) \to \infty$. Again, unsurprisingly, the optimal level of monitoring is decreasing in the level of compliance E^* and increasing in the expected payoff from obtaining evidence upon implementation failure, $W_m - V_m$.

Given monitoring M^*, the state selects a level of compliance E to maximize

$$E \cdot u_s(imp) + (1 - E) \cdot (M^* \cdot W_s + (1 - M^*) \cdot V_s) - e(E). \tag{12}$$

Again, the first term captures the consequences of successful implementation and the second term failure. The first-order condition is

$$u_s(imp) - (M^* \cdot W_s + (1 - M^*) \cdot V_s) = e'(E). \tag{13}$$

With $e'(0) = 0, e'(1) \to \infty$, uniqueness is again guaranteed. Equally plausible, the optimal level of compliance is increasing in the level of monitoring M^* and decreasing in the expected payoff from successful monitoring, $M^* \cdot W_s + (1 - M^*) \cdot V_s$.

Proofs

Proof of Proposition 1 Suppose that λ or $u_m(imp)$ increases. This increases both the value of $W_m(1)$ and $W_m(0)$. Thus, the monitor's payoff from acquiring information increases *ceteris paribus* regardless of collusion. Consequently, for any fixed compliance effort E, the best-response monitoring $M(E)$ must increase. Use (11) and (13) to see that E^* must increase.

Consider now the marginal effect slightly below and above the collusion threshold. In the limit, with the change in λ or $u_m(imp)$ approaching zero from above, E^* is approximately the same both before the increase in λ or $u_m(imp)$. Below the collusion threshold, an increase in λ or $u_m(imp)$ increases the monitor's bargaining payoff from successful monitoring given implementation failure by $\frac{1}{2}$ and reduces the state's bargaining payoff by $\frac{1}{2}$, as shown in the analysis of the Nash bargaining subgame. The former (latter) effect shifts the monitor's best response $M(E)$ upward (the state's best response $E(M)$) downward. This causes an increase in equilibrium compliance E^* by expressions (11) and (13).

Consider now the case of being above the collusion threshold. Now the monitor's payoff increases by 1 given implementation failure. However, the state's payoff does not change at all. Thus, the change in equilibrium compliance depends on the relative curvatures of the monitor's and the state's best responses at this point, and for general functional forms these are indeterminate. ■

Proof of Proposition 2 By (11) and (13), increasing $u_s(imp)$ increases the monitor's best response $E(M)$ for any given M. In addition, M is strictly decreasing in E. Since the best response $M(E)$ does not depend on $u_s(imp)$ except through E, it must be that for any given E the best response $M(E)$ remains unchanged. Thus, the best response curve for E shifts upwards while the best response curve for M remains unchanged. With both best responses having a negative slope, the first part of the claim follows. For the second part, examine the state's equilibrium Nash bargaining payoff U_s^{coll}. ■

Proof of Proposition 3 In the collusion region, changes in G increase the best response $E(M)$ for a given M by (13) but decrease $M(E)$ for a given E by (11), so the effect on where the reaction curves meet is ambiguous. By (11) and (13), increases in G in the information revelation region must increase E^*. ■

Endogenous correction of non-compliance

Suppose that the monitor reveals information about non-compliance. In this case, the target state is allowed an attempt to restore compliance. The cost of this effort is denoted by $r(\lambda)$, where $r(\lambda)$ is a function similar to $e(E)$. It could even be the exact same function, $r = e$. The probability of success is λ. If the target state fails to comply, it incurs a punishment cost $Q > 0$, which can be thought of as a reputational loss and/or a sanction.

Given this, the state's objective function is

$$-(1 - \lambda)Q - r(\lambda). \tag{14}$$

The first-order condition is

$$Q - r'(\lambda^*) = \lambda^*. \tag{15}$$

Note that the optimal attempt λ^* is also the probability of successful restoring of compliance, as was the case in the original model with an exogenous λ . Note also that λ is strictly increasing in the severity of the punishment, $\lambda(Q)$.

The expected utility from the revelation of non-compliance is $-(1-\lambda)Q - r(\lambda^*)$, a strictly negative number. By the expected utility theorem, it can be safely normalized to zero by adding the constant $(1-\lambda)Q + r(\lambda^*)$ to all payoffs of the model.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

Notes

1. For rare exceptions, see Mitchell (1998) and Dai (2002).
2. Monitors can also be state bureaucracies, such as the Environmental Protection Agency. However, such monitors may have similar incentives to the state.
3. A related literature on conflict mediation explores whether biased mediators are better than neutral ones. For example, Kydd (2003) finds that biased mediators may be best at solving conflict, as they are more credible than unbiased mediators.
4. To be sure, not all treaty monitoring is sincere. As Beaulieu and Hyde (2009) argue, states often strategically invite monitors to raise their legitimacy, while expecting some non-compliance to go undetected. Similarly, Kelley (2008) argues that legitimacy, accompanied with a new norm of election monitoring, led states to invite monitors even when they continued to cheat. In such instances, the purpose of monitoring is to put a stamp of approval on state behavior.
5. See Human Rights Watch Report, *Australia Undermining Global Human Rights*, 31 August 2000, http://www.hrw.org/news/2000/08/30/australia-undermining-global-human-rights
6. See International Campaign to Ban Landmines Institutional Donor reports, http://www.icbl.org/index.php/icbl/About-Us/Donor (accessed February 2010).
7. See United Nations Office on Drugs and Crimes, *Global Report on Trafficking in Persons*, February 2009, https://www.unodc.org/documents/Global_Report_on_TIP.pdf.
8. Mexico expels 40 human rights observers. *Washington Post*, 12 May 1998.
9. See Human Rights Watch, *Human Rights Watch World Report 1997: Mexico*, 1 January 1997, http://www.refworld.org/docid/3ae6a8c110.html.
10. Of course, in reality varying degrees of compliance could exist. To simplify, we focus on a binary outcome variable.
11. If compliance were deterministic, the target state would basically choose whether or not to pay a fee for implementation. In failing to implement, the state would leave itself vulnerable to action by the monitor. Depending on the monitor's preferences, the state would either collude or accept revelation and a scandal. Thus, compliance would depend on its compliance, the monitor's incentive to collude, and the cost of colluding and a scandal to the target state.
12. Of course, the actual dumping would probably be done by private corporations. In this article, we focus on compliance by state parties.
13. Our assumption that monitoring follows compliance is innocuous. Since the monitor cannot observe compliance, equilibrium behavior would be unchanged if compliance and monitoring were simultaneous instead.

14. Our main results would also hold if it was mandated to select at least some minimal level $\underline{M} \in (0, 1)$.
15. The model could be complicated by assuming that, should implementation fail, the monitor's probability of obtaining evidence of non-compliance would decrease with the state's prior effort. We prefer our simpler specification, which follows Aghion and Tirole (1997), because it clearly illustrates the possibility that even sincere efforts to comply may fail, and this may carry negative reputational consequences for states (Chayes and Chayes, 1995).
16. In reality, it could be that the monitor threatens to falsely claim non-compliance even if the state complies. This would obviously reduce the state's incentive to comply. However, it seems safe to assume the monitor's bargaining power would be a lot weaker with than without compliance by the target state. Therefore, the possibility of extortion by the monitor would not compromise the target state's fundamental incentive to comply.
17. The main results would hold in a model extension with slightly different payoffs. However, we refrain from complicating the model by adding yet another payoff parameter.
18. Although the monitor cannot directly observe compliance, it is able to correctly infer the effort E^* in equilibrium.
19. Technically, the game may have multiple equilibria if the monitor is indifferent between collusion and information revelation. As usual, we omit this parameter set of measure zero throughout.
20. Blustein, P. IMF staff urges better economic monitoring, *Washington Post*, 31 March 1998.
21. Drawing lessons from the mexican crisis: Preventing and resolving financial crises– the role of the IMF. Address by Michel Camdessus, IMF Managing Director. Washington Conference of the Council of the Americas, 22 May 1995.
22. IMF to seek warnings from nations in need, *Washington Post*, 17 April 1998.
23. United Nations Security Council Resolution 687, 3 April 1991.
24. United Nations Security Council, Report to the Secretary General on humanitarian needs in Kuwait and Iraq in the immediate post-crisis environment by a mission to the area led by Mr Martti Ahtisaari, Under Secretary General for Administration and Management. Report S/22366, 20 March 1991.
25. United Nations Security Council Resolution 986, 14 April 1995.
26. United States Government Accountability Office, Report to Congressional Committees, United Nations: lessons learned from oil for food program indicate the need to strengthen UN internal controls and oversight activities. Report GAO-06-330, April 2006, p. 48.

References

Aghion P and Tirole J (1997) Formal and real authority in organizations. *Journal of Political Economy* 105(1): 1–29.

Axelrod R and Keohane RO (1985) Achieving cooperation under anarchy: Strategies and institutions. *World Politics* 38(1): 226–254.

Barrett S (1997) The strategy of trade sanctions in international environmental agreements. *Resource and Energy Economics* 19(4): 345–361.

Barrett S (2003) *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. Oxford: Oxford University Press.

Beaulieu E and Hyde SD (2009) In the shadow of democracy promotion: Strategic manipulation, international observers, and election boycotts. *Comparative Political Studies* 42(3): 392–415.

Breidenich C and Bodansky D (2009) Measurement, reporting and verification in a post-2012 climate agreement. Pew Center on Global Climate Change, April 2009.

Busby JW (2010) After Copenhagen: Climate governance and the road ahead. Council on Foreign Relations, August 2010.

Busch ML and Reinhardt E (2000) Bargaining in the shadow of the law: Early settlement in GATT/WTO disputes. *Fordham International Law Journal* 24: 158.
Carrubba CJ (2005) Courts and compliance in international regulatory regimes. *Journal of Politics* 67(3): 669–689.
Chayes A and Chayes AH (1995) *The New Sovereignty: Compliance with International Regulatory Agreements*. Cambridge: Harvard University Press.
Ciafani S (2012) *Ecomafia 2012: Environmental Organized Crime in Italy*. Legambiente.
Dai X (2002) Information systems in treaty regimes. *World Politics* 54(4): 405–436.
Downs GW and Rocke DM (1995) *Optimal Imperfection? Domestic Uncertainty and Institutions in International Relations*. Princeton, NJ: Princeton University Press.
Downs GW, Rocke DM and Barsoon PN (1996) Is the good news about compliance good news about cooperation? *International Organization* 50(3): 379–406.
Jo H (2008) Monitoring compliance: The design of monitoring institutions in international cooperation. PhD Dissertation, University of Michigan.
Johns L (2007) A servant of two masters: communication and the selection of international bureaucrats. *International Organization* 61(02): 245–275.
Kelley J (2008) Assessing the complex evolution of norms: the rise of international election monitoring. *International Organization* 62(02): 221–255.
Keohane RO (1984) *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton: Princeton University Press.
Koremenos B, Lipson C and Snidal D (2001) The rational design of international institutions. *International Organization* 554(4): 761–799.
Kydd A (2003) Which side are you on? Bias, credibility, and mediation. *American Journal of Political Science* 47(4): 597–611.
Lessmann K, Marschinski R and Edenhofer O (2009) The effects of tariffs on coalition formation in a dynamic global warming game. *Economic Modelling* 26(3): 641–649.
Marcoux C and Urpelainen J (2012) Capacity, not constraints: A theory of North–South regulatory cooperation. *Review of International Organizations* 7(4): 399–424.
Massari M and Monzini P (2006) Dirty business in Italy: A case-study of illegal trafficking in hazardous waste. *Global Crime* 6(3-4): 285–304.
McCubbins MD and Schwartz T (1984) Congressional oversight overlooked: Police patrols versus fire alarms. *American Journal of Political Science* 28(1): 165–179.
Meyer JA and Califano MG (2006) *Good Intentions Corrupted: The Oil for Food Scandal and the Threat to the UN*. New York: Public Affairs.
Michonski K and Levi MA (2010) Harnessing international institutions to address climate change. Council on Foreign Relations, March 2010.
Mitchell RB (1994) Regime design matters: Intentional oil pollution and treaty compliance. *International Organization* 48(3): 425–458.
Mitchell RB (1998) Sources of transparency: Information systems in international regimes. *International Studies Quarterly* 42(1): 109–130.
Nash JF (1950) The bargaining problem. *Econometrica* 18(2): 155–162.
Reinhardt E (2001) Adjudication without enforcement in GATT disputes. *Journal of Conflict Resolution* 45(2): 174–195.
Rosendorff BP (2005) Stability and rigidity: Politics and design of the WTO's dispute settlement procedure. *American Political Science Review* 99(3): 389–400.
Schwartz WF and Sykes AO (2002) The economic structure of renegotiation and dispute resolution in the world trade organization. *Journal of Legal Studies* 31(1): 179–204.
Tallberg J (2002) Paths to compliance: Enforcement, management, and the european union. *International Organization* 56(3): 609–643.
Tirole J (1986) Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, and Organization* 2(2): 181–214.

Veel PE (2009) Carbon tariffs and the WTO: An evaluation of feasible policies. *Journal of International Economic Law* 12(3): 749–800.

Volker PA, Goldstone RJ and Pieth M (2005) Independent inquiry committee into the United Nations oil-for-food programme. United Nations. http://www.iic-offp.org/index.html.

Weber CL and Peters GP (2009) Climate change policy and international trade: Policy considerations in the US. *Energy Policy* 37(1): 432–440.

Woods N (2006) *The Globalizers: The IMF, the World Bank, and their Borrowers*. Ithaca, NY: Cornell University Press.

Zhang Z (2009) Multilateral trade measures in a post-2012 climate change regime? What can be taken from the Montreal Protocol and the WTO? *Energy Policy* 37(12): 5105–5112.