International Theory (2017), 9:3, 353–380 © Cambridge University Press, 2017
doi:10.1017/S1752971917000082 First published online 31 August 2017

International institutions: weak commitments and costly signals

LISA L. MARTIN

Department of Political Science, University of Wisconsin-Madison, Madison, Wisconsin, United States

E-mail: lisa.martin@wisc.edu

As commitment devices, international institutions encourage cooperation by imposing costs on members who do not live up to their commitments. However, the costs that institutions can impose are limited, so that their commitment capacity is weak. Institutions can also impose costs as a condition of membership, allowing them to serve as costly signals. A model of weak commitment and costly signaling leads to a number of hypotheses about patterns of cooperation, institutional membership, and states' preferences over institutional design. For example, existing members of an institution should impose higher *ex ante* costs when a potential new member could either gain significant benefits from reneging on their commitments in the future, and when the new member expects to gain high benefits from future cooperation. These results are consistent with empirical work on institutions including peacekeeping and the World Trade Organization.

Keywords: international institutions; commitments; costly signals; international cooperation

In rationalist models of international institutions, they influence state behavior through creating commitments and serving as costly signals. As commitment devices, institutions impose costs on members that renege on agreements, thus encouraging cooperation. As signaling devices, they screen out potential members who will not reliably cooperate. Scholars typically understand commitment and signaling as alternative mechanisms, in contexts of bargaining or cooperation, and generally model them separately.[1]

While it is important, for purposes of conceptual clarity and theory development, to understand the differences between *ex ante* signals and *ex post* incentives that create commitments, in practice actual mechanisms might both sink costs (signal) and tie hands (commit). For example,

[1] See Fearon (1997) for a discussion of the distinction between commitment and signaling.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Slantchev (2005) observes that military mobilization is both costly and changes the chances of victory, so that it is simultaneously a signaling and commitment mechanism. As a practical example of commitment and signaling, consider China's entry to the World Trade Organization (WTO). The WTO has an elaborate dispute-resolution mechanism and the ability to authorize retaliatory tariffs against members who do not live up to their obligations, thus enhancing state commitments. Nevertheless, other WTO members have deep worries about whether China will, in fact, fulfill its WTO commitments. These worries led to a drawn-out accession process for China, during which it was required to implement numerous policy changes. These *ex ante* steps can be considered costly signals, so that the WTO serves as both a weak commitment and a signaling mechanism.

By creating costs associated with reneging on agreements, whether reputational or other costs, institutions create commitments that allow for mutually beneficial cooperation that could not be sustained outside an institution. However, in many cases institutions cannot change incentives so dramatically that they provide an absolute commitment. For example, military alliances create incentives to provide mutual assistance in cases of conflict. However, under duress some alliance members find that these incentives are in fact not high enough for them to come to an ally's assistance, so that they do not live up to their alliance obligations. The rate of reneging on alliance commitments is in fact quite high, with alliance members not living up to their commitments in times of war about 25% of the time (Leeds 2003). In general, while institutions such as alliances can create commitments, they usually do not have the resources or authority to deter all members from reneging under all circumstances.

I argue, therefore, that we should think of international institutions as 'weak commitment' devices. They change cost structures in such a way that they can allow cooperation to emerge that could not emerge in their absence. However, they cannot typically change cost structures sufficiently to deter all reneging on commitments. I develop a model of institutions as commitment devices in the presence of uncertainty about whether they are actually influential enough to prevent reneging on commitments. This model allows us to specify the conditions under which institutions as weak commitment devices can allow cooperation to emerge, and when we will observe states reneging on their commitments.[2]

I then suggest that institutions, beyond changing *ex post* incentives to enhance commitments, can also be *ex ante* costly to join and so

[2] Please note that the term 'weak' refers only to the commitment capacity of the institution, and is not intended to be pejorative. As the model will show, even institutions with only a weak commitment capacity do allow a degree of cooperation to emerge.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

simultaneously serve as signals, analogous to Slantchev's argument about military mobilization. I add this costly signaling function to the weak commitment model. This analysis gives rise to expectations about the conditions under which we will observe cooperation, institutions, and reneging; and it has implications for preferences over institutional design.

The first section of this paper provides a brief summary of the literature on institutions as signaling devices. The second develops the weak commitment model and then adds costly signaling to it. The third section focuses on the differences that emerge when signaling is added, discussing the costs and benefits associated with signaling and who, therefore, is likely to demand that institutions be costly to join. The final section draws out the implications of this analysis for institutional design and discusses empirical studies consistent with the model.

Institutions as signaling devices

The modern literature on international institutions tends to focus on their commitment properties. Going back to the original work on international regimes (Krasner 1982) and Keohane's (1984) seminal study of institutions, scholars have concentrated on the ability of institutions to either coordinate policy or to put in place monitoring and decentralized enforcement mechanisms that raise the costs to reneging on commitments. This tradition has continued in more recent work, as for example, in Guzman's (2008) study of why states comply with treaties. Guzman, writing from the perspective of a legal scholar, argues that the standard '3R' mechanisms of reputation, reciprocity, and retaliation explain why international law can be effective in committing states to live up to its terms. Other international legal scholars, such as Setear (2002), have also considered the use of institutions as signals, although Setear has focused more on domestic than international institutions in this capacity.

Fearon (1997) clarified the previously muddled distinction between commitment (tying hands) and signaling (sinking costs). A number of other authors have further developed the signaling side of the story. Morrow has modeled alliances as both commitment and signaling devices. Morrow (1994) argues that alliances enhance commitment by increasing the ability of allies to fight together, and act as signals because they involve sunk peacetime costs. He derives propositions about when alliances will form, when they will be credible, and when they will deter. Morrow (2000) focuses more directly on the signaling properties, asking why 'writing down' an alliance has any impact on its ability to deter attacks. Drawing on standard costly signaling models (as I do in this paper), he argues that alliances can only effectively deter if the signals that they send require their members to bear costs. Morrow points to the process of military coordination within alliances as the

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

major source of peacetime costs, as policy coordination itself is costly; and military coordination can leave an ally more vulnerable if it eventually must fight alone (Morrow 2000, 70). Kydd (2001) similarly considers North Atlantic Treaty Organization (NATO) expansion as a process of costly signaling, arguing that the reassurance that costly signals can provide is particularly valuable when uncertainty about allies' preferences is high.

Von Stein (2005, 2008) has also studied the signaling or screening properties of international agreements, directing her efforts toward careful statistical modeling of ratification and compliance decisions. Von Stein (2005) examines Article VIII of the International Monetary Fund (IMF) Treaty, which commits states that sign it to avoid certain currency practices such as interference with payments and discrimination among foreign currencies. Simmons (2000) had previously asked which states complied with their commitments under Article VIII, and found that the primary determinant was the ease with which they were able to comply.

Von Stein builds on Simmons' work by explicitly modeling the selection process by which states choose to enter Article VIII. She finds that Article VIII seems to act purely as a screening device. That is, it separates out those states that would find it difficult to comply with its terms, and adds no further commitment beyond the screening function. She concludes that 'the international legal commitment has little constraining power independent of the factors that lead states to sign' (von Stein 2005, 611). In other words, the only effect of Article VIII comes through the *ex ante* costs of entering it, which serve to screen out those states that will not be able to comply at low cost. Von Stein does not go into much detail about the nature of these costs or the source of their discriminatory power, but the evidence clearly points to a signaling function. The model in this paper allows for such an *ex ante* costly signaling function, while combining it with a weak commitment capacity. Von Stein (2008) considers ratification of international environmental agreements, in particular the UN Framework Convention and the Kyoto Protocol. Again, selection into these agreements is a prominent part of the causal mechanism, suggesting a signaling function.

Other authors have touched on the signaling properties of international institutions and agreements. Of note, Thompson (2006, 2009) discusses how the structure of the UN Security Council allows it to send effective signals when a state is attempting to coerce another. The use of the Security Council allows coercing states to strategically transmit information, so that the level of international support for the use of force becomes closely tied to Security Council approval. Haftel (2007) considers bilateral investment treaties (BITs), asking whether they serve primarily as credible commitments or signals. Developing countries sign BITs hoping that they will increase the flow of foreign direct investment (FDI). Haftel's statistical analysis suggests that the commitment function of BITs may be more

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

important than their signaling function, as FDI flows respond only to BITs that are mutually ratified, not to BITs that have merely been signed. Hyde (2011) analyzes international election monitoring as a costly signal, showing why it rapidly evolved into a widely observed norm. Walsh (2007) considers the general question of whether states are engaged in signaling games, focusing on the relationship between the United States and Soviet Union under Gorbachev. He argues that Gorbachev was sending costly signals to the United States, but that understanding the sources and consequences of these signals requires integrating domestic political conflict into the analysis. Gray (2013, 35) shows that entry into certain types of international organizations (IOs) sends a signal to international investors.

It is also of relevance to note discussions of institutions as signals in settings other than international politics. A number of authors have looked at democratic institutions as sources of costly signals (Schultz 1999; Mansfield and Pevehouse 2008; Hafner-Burton, Mansfield, and Pevehouse 2015). Moving more into the realm of economics, Bolle's work on corporate governance systematically studies governance institutions as signaling devices (Bolle 2002; Braham and Bolle 2006). Bolle emphasizes the potentially negative normative consequences of relying on corporate governance as a signal, as unless the signaling costs are precisely calibrated they can decrease social welfare. In the model in this paper, I likewise find that 'fine-tuning' the costs of signals is essential if they are to function effectively. Baglioni (2008) builds on Bolle's work by considering corporate governance as both a signal and a commitment device.

Model

In this section, I develop two simple games of incomplete information. Both involve a state, A, that decides whether to join an institution. A second state, B, then decides whether to cooperate with A. B is an existing member of the institution, while A is a prospective new member. A can be either a reliable or an unreliable type, distinguished by the fact that the reliable type receives a higher payoff from cooperation. For the unreliable type, the benefits of cooperation are not high enough to prevent reneging. An outcome where B cooperates but A does not imposes costs on B, as in a standard Prisoners' Dilemma. I first develop a model where the institution functions solely as an *ex post* commitment device, then add an *ex ante* signaling function to the institution.

Commitment model

If international institutions function solely as commitment devices, they impose a cost on members that do not live up to the terms of their

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

commitments – that is, that renege. Other members of an institution find such reneging costly. Consider, for example, members of the WTO deciding whether to admit a potential new member such as China. The WTO has in place monitoring and enforcement mechanisms that impose costs on members if they violate WTO rules. These mechanisms, if strong enough, should induce members to live up to their commitments. However, reneging nevertheless does sometimes occur, and it imposes costs on other members as their access to markets in the violating country is limited.

To model this situation in a simple manner, consider a country A that is deciding whether to join an institution. A is aware that the institution has the capacity to impose costs if A joins and then reneges on commitments. A second country, B, has to decide whether to cooperate with A, whether or not A joins the institution. B is uncertain of the benefits that A will reap from cooperation; in particular, whether the benefits are high enough to prevent A from reneging on its commitments. (I assume that A has complete information about its own payoffs.) With probability p, A is a 'reliable' type, meaning that its payoffs from cooperation are high. Figure 1 illustrates this game.

If A has joined the institution it receives a small payoff, ϵ, unless it reneges on its commitments.[3] We can think of this small payoff to A as representing the ancillary benefits of joining the institution even if B does not cooperate, such as reputational benefits. Table 1 summarizes the payoffs of this pure commitment game. State A receives benefits from cooperation with B; these benefits are higher if A is the reliable type. B also benefits from cooperation with A, but pays a cost if A reneges. A gets a reward if it suckers B into cooperating but then reneges; but this reward is reduced if A has joined the institution, as it will be punished for reneging. In order to illustrate the commitment effect, I assume that, in the absence of an institution that can punish reneging, A would prefer to sucker B rather than to cooperate, whether it is the reliable or the unreliable type.

The equilibrium of this game will depend, in part, on how large the cost c is that the institution can impose on members that renege on their commitments. I distinguish between two types of equilibria, a strong commitment equilibrium and a weak commitment equilibrium. Appendix 1 provides an equilibrium analysis of the pure commitment game.

Strong commitment equilibrium: If c is large, the institution would serve as a strong commitment device. If c is large enough that $b_u + \epsilon > a - c$, the prospect of punishment by the institution induces cooperation even from an unreliable A; both A types would choose, in the last stage, to cooperate

[3] Without this ancillary benefit, under some conditions an unreliable A would be indifferent between joining the institution and not, so that the model would not generate clear new insights.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

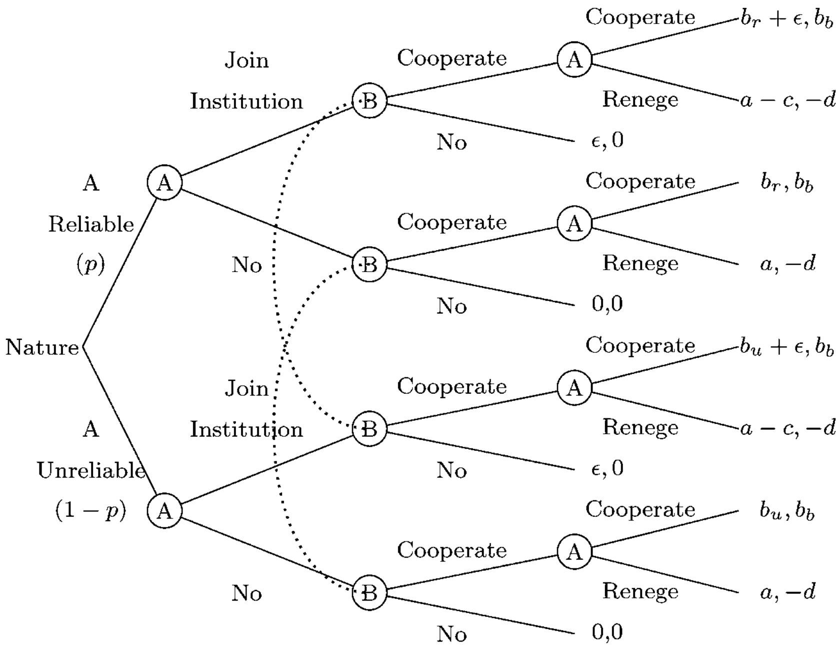

Figure 1 Pure commitment game.

Table 1. Payoffs of pure commitment game

	Reliable A	Unreliable A	B
No institution, no cooperation	0	0	0
No institution, A reneges	a	a	$-d$
Institution, A reneges	$a-c$	$a-c$	$-d$
Institution, no cooperation	ϵ	ϵ	0
Mutual cooperation	$b_r+\epsilon$	$b_u+\epsilon$	b_b

$a > b_r > b_u > \epsilon > 0.$

rather than to renege. Anticipating cooperation by A, B will cooperate if A joins the institution, knowing that institutional punishment will prevent A from reneging. Knowing that A will renege if it does not face institutional punishment, B will not cooperate if A does not join the institution.

Equilibrium strategies in the strong commitment equilibrium: Reliable A, join the institution; cooperate if B cooperates; if A chooses no institution and B cooperates, renege. Unreliable A, same as reliable A. B, cooperate if A joins the institution; do not cooperate if A does not join.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Weak commitment equilibrium: However, international institutions do not have adequate enforcement capacity to prevent all reneging.[4] It is difficult to conceive of states dedicating sufficient enforcement capacities to an institution that it was capable of deterring *all* reneging. Thus, to varying degrees, all international institutions are 'weak commitment' institutions, able to deter reneging from some members, but not from all. In a weak commitment equilibrium, which is of most interest for purposes of this paper as well as in practice, the punishment for reneging is sufficient to induce cooperation on the part of a reliable A but not an unreliable type. That is, a weak commitment equilibrium exists if $b_r + \epsilon > a - c > b_u + \epsilon > 0$. Under these conditions, it is an equilibrium for both reliable and unreliable A to join the institution. B will not be able to update its beliefs about A's type, and so will cooperate if $P > d/(b_b + d)$, meaning that the expected payoffs from cooperation are greater than 0.

Equilibrium strategies in the weak commitment equilibrium: Reliable A, join institution; cooperate if B cooperates; if A does not join the institution and B cooperates, renege. Unreliable A, join institution; renege if B cooperates; if A does not join the institution and B cooperates, renege. B, do not cooperate if A does not join the institution; if A joins, cooperate if $P > d/(b_b + d)$.

A reliable A benefits from a weak commitment institution because it gains cooperation with B, as long as B believes that A is likely reliable. However, an unreliable A also benefits from a weak commitment instituion, because it is able to induce B to cooperate and then will renege. State B's expected payoff from a weak commitment institution is also positive because it gains the possibility of mutually beneficial equilibrium. However, when B does cooperate, it runs the risk realizing the reneging payoff $-d$.

Overall, pure commitment institutions that provide only weak commitments do allow cooperation to emerge under some conditions, by overcoming the temptation to renege for some types of states. However, they also have some undesirable properties. In particular, they allow unreliable states to bluff, by joining the institution and then suckering other members into cooperation from which they will renege. Existing members

[4] Why members do not always endow institutions with sufficient enforcement capacities to induce a strong commitment equilibrium is an interesting question. The answer likely has to do with greater uncertainty in the environment than I assume in this model. In particular, the benefits of cooperation (for both A and B) are likely to be stochastic rather than fully predictable and constant, as for example in Downs and Rocke (1997). If in one period realized benefits of cooperation are subject to a negative shock, but the punishment for reneging remains high, states may find it best to opt out of the institution entirely, leading to the collapse of cooperation. Thus, as Downs and Rocke argue, we should see some 'imperfection' optimally built into institutions. That is, states do not create strong commitment institutions because they are concerned that those strong enforcement capacities could undermine cooperation over the long term.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

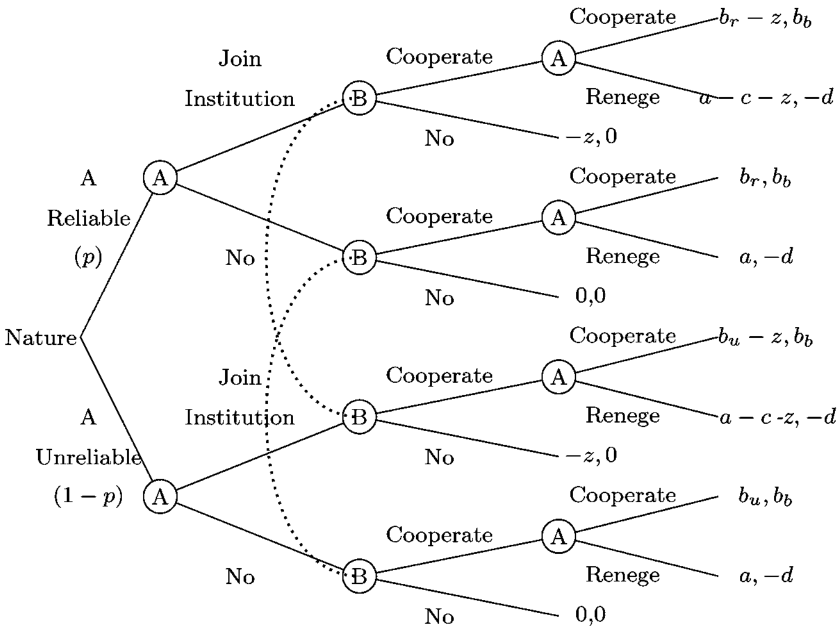

Figure 2 Signaling game.

of the institution can benefit, but only by exposing themselves to risk. In addition, because there is no *ex ante* cost to joining the institution, potential new members who are reliable cannot distinguish themselves from unreliable potential members, and so do not gain the cooperation of existing members if these states believe that A is not likely to be reliable.

Signaling model

The dynamics of this game change substantially when a new member has to pay an up-front cost to join the institution. In this case, the institution serves as a costly signal, and under some conditions reliable potential members can now effectively distinguish themselves from unreliable types, allowing more mutually beneficial cooperation to emerge. I will continue to focus on payoffs that lead to a weak commitment equilibrium, and Figure 2 illustrates the signaling game. However, now if A joins the institution it pays cost z on entry.[5] I use Perfect Bayesian Equilibrium as the solution concept for this game, and characterize the equilibria of this signaling and

[5] In the signaling game, I do not include the ϵ payoff to A if it joins the institution but cooperation does not materialize. This simplifies the notation, and we can assume that the z cost-of-entry parameter is net of these small ancillary benefits.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Table 2. Equilibria of signaling and commitment game

	Low reliability ($P < d/(b_b + d)$)	High reliability ($P > d/(b_b + d)$)
Low signaling cost ($b_r > a - c > z$)	Semi-separating: reliable A chooses institution. Unreliable A chooses institution with probability $pb_b/(d(1 - p))$. When B sees institution, cooperate with probability $z/(a - c)$	Pooling: all A's choose institution. B cooperates
Moderate signaling cost ($b_r > z > a - c$)	Separating: reliable A chooses institution. Unreliable A does not choose institution. B cooperates if it observes institution	Separating: reliable A chooses institution. Unreliable A does not choose institution. B cooperates if it observes institution
High signaling cost ($z > b_r > a - c$)	Pooling: no A's choose institution. B does not cooperate	Pooling: no A's choose institution. B does not cooperate

commitment game in Table 2. Appendix 2 provides an equilibrium analysis of the signaling game.

Two parameters describe the equilibria: B's prior belief that A is reliable (p) and the size of the signaling cost (z). A low reliability situation exists if $P < d(b_b + d)$ (left column in Table 2); otherwise we have a high reliability situation (right column). Signaling costs are low if $b_r > a - c > z$; they are moderate if $b_r > z > a - c$; they are high if $z > b_r > a - c$. The rows of Table 2 indicate different levels of signaling cost.

Consider first the situation where the costs of entering the institution are very high (bottom row). In this case, entry costs are so high that they outweigh the benefits of cooperation for even a reliable A. A pooling equilibrium then emerges, in which A does not join the institution and B does not cooperate. (Remember that, in the absence of an institution, even a reliable A succumbs to the temptation to renege.) Thus, if members of an institution miscalculate and set entry costs too high, they will deter even potential cooperating members from joining.

Consider next the situation of moderate entry costs, the middle row of the table. In this case, a separating equilibrium emerges. Reliable A's, with their higher benefits of cooperation, will choose to bear this cost, but unreliable types will not. State B can now fully update its beliefs about A's type, and will choose to cooperate with A if it joins the institution, but not otherwise. Signaling costs in this 'sweet spot' have many desirable properties. They allow the fullest extent of cooperation between reliable states, and eliminate any potential for unreliable states to bluff and sucker existing members. State B is not subject to risk in this situation, because it can fully distinguish reliable from unreliable types.

Table 3. Unreliable A payoffs, comparing signaling to pure weak commitment

	Low reliability	High reliability
Low cost to join	Indifferent: no cooperation	Small loss: bears signaling cost z, still gets high reneging payoff
Moderate cost to join	Indifferent: no cooperation	Large loss: does not induce B to cooperate
High cost to join	Indifferent: no cooperation	Large loss: does not induce B to cooperate

Finally, consider the top row of Table 2, where signaling costs are low. In this case, the equilibrium that emerges depends on B's prior belief about A's type. If B believes that A is likely reliable (high p), a pooling equilibrium will emerge. Because B believes that A is reliable, B will cooperate whenever A joins the institution. And because the cost of joining the institution is low, even unreliable A's will be willing to pay it. In this situation, an unreliable A will bluff and sucker B.

However, if B believes that the probability that A is reliable is low, we find a semi-separating equilibrium. In this case, B will not always cooperate because it faces a high chance of being suckered. Instead, if A chooses to join the institution, B's only equilibrium response is to cooperate probabilistically (as indicated in Table 2). If A does not join the institution, B does not cooperate. Reliable A's always join the institution; unreliable types join with some probability less than 1 (as indicated in the table).

Thus, setting entry costs low creates a number of inefficiencies. Reliable types can no longer fully distinguish themselves from unreliable types, so they cannot always induce B to cooperate, even when they pay entry costs. Unreliable types are also able to bluff, so that B sometimes has to bear the cost of A's reneging. States designing institutions therefore face a dilemma when they make *ex ante* demands of new members. If these demands are set too high, they will deter entry and lose out on potentially beneficial cooperation. On the other hand, if they set entry costs too low, they will allow unreliable new members into the institution and allow themselves to be suckered, at least probabilistically.

Comparing signaling and pure weak commitment

From states' perspectives, what difference does adding a signaling component to a pure commitment institution with weak enforcement capacities make? Tables 3–5 get at this question by considering the difference in

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Table 4. Reliable A payoffs, comparing signaling to pure weak commitment

	Low reliability	High reliability
Low cost to join	Gain: bears small cost of joining institution, but gains probabilistic cooperation	Small loss: gets cooperation, bears small cost z
Moderate cost to join	Gain: gets cooperation, bears moderate cost z	Moderate loss: gets cooperation, bears moderate cost z
High cost to join	Indifferent: no cooperation	Large loss: no cooperation

Table 5. Reliable B payoffs, comparing signaling to pure weak commitment

	Low reliability	High reliability
Low cost to A for institution	Small gain: gets some cooperation, but also bears chance of reneging	Indifferent (in expectation): cooperates, some reneging; bears risk
Moderate cost to A for institution	Large gain: gets cooperation with reliable A	Large gain: retains cooperation with reliable A, no risk of unreliable A reneging
High cost to A for institution	Indifferent: no cooperation	Loss: no cooperation

payoffs for our three types of states: unreliable potential members, reliable potential members, and existing members of the institution (state B). The cells in these tables indicate the difference in expected payoffs that occurs when an *ex ante* costly signal is added to a weak commitment institution.

Table 3 considers unreliable potential new members: states that would derive relatively little benefit from cooperation, and therefore are likely to renege on their commitments. Such states can never gain from demanding an entry cost; they can only lose or be indifferent. Indifference occurs when state B believes that A is likely unreliable. In this case, no cooperation would emerge under a weak commitment institution, and nothing changes when costly signals are introduced. On the other hand, when B believes there is a good chance that A is reliable, an unreliable A unambiguously loses when it has to send a costly signal. If the cost to join the institution is moderate or high, an unreliable A will not pay it and loses the opportunity to fool B. If the signaling cost is low, an unreliable A will be able to bluff, but is still slightly worse off than in the no-signaling situation because it has to pay the entry cost.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Table 4 looks at the situation from the prospective of a reliable potential member, a state A that will gain relatively high benefits from cooperation. The story is more complicated here, with a reliable A sometimes gaining from the opportunity to signal and sometimes losing. When B's prior belief is that A is probably unreliable, a reliable A gains from the opportunity to send a costly signal, because it can now differentiate itself unreliable types and gain B's cooperation. If the cost of the signal is set too high A will not gain these benefits, and is indifferent between a signaling institution and a pure weak commitment one.

However, when B believes that A is likely reliable, a reliable A can only lose from the addition of a signaling component, although the degree of loss varies depending on the cost of entry. When entry costs are low or moderate, so is a reliable A's loss. As in the weak commitment case, A cooperates with B; but it now has to 'burn money' by paying the *ex ante* cost. If entry costs are high, a large loss results for a reliable A, as it will not bear this cost and cooperation will not materialize. Thus, we would expect that potential new members who anticipate high benefits from cooperation would have varying attitudes toward demands that they pay an entry cost. If they see that this would allow them to differentiate themselves from unreliable states, they will happily pay these costs. But if existing members already believe that A is reliable, A will be opposed to any additional demands to ante up before entering the institution.

Finally, Table 5 considers the difference in payoffs for state B, an existing member of the institution. For empirical purposes, this table is likely the most useful, as existing institutional members will set the entry costs for new members; it is a reasonable simplification to assume that potential new members do not get to set their own entry costs. Table 5 therefore leads directly to hypotheses about institutional design. As indicated, existing members can gain from the introduction of costly signals, lose, or find themselves indifferent. The introduction of entry costs is in general helpful to existing members when they believe that there is a high probability that A is unreliable. As A more likely becomes reliable, entry costs could actually decrease B's welfare, unless B is able to hit the sweet spot that creates a separating equilibrium.

If B sets entry costs too high, no cooperation will emerge. When potential members are believed unreliable, this makes no difference to B, as it would not have cooperated anyway. But when potential members are more likely reliable, high entry costs hurt existing members of the institution by keeping out new members and making cooperation impossible. Likewise, if B sets entry costs too low, the results are mixed. This can lead to a small gain in welfare for B, as occasional cooperation emerges with reliable A's; but this effect is partially offset by the fact that B will now open itself to possible

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

exploitation by unreliable A's. Only when B sets entry costs at the appropriate intermediate level – not so high as to deter reliable A's from entering, but high enough that unreliable A's will not bear the cost – does B unambiguously gain from giving A the opportunity to send a costly signal.

This section has analyzed a pure commitment game, and a game in which the institution both includes a weak commitment capacity and demands an entry cost for new members. As long as the entry costs are set at an appropriate intermediate level, adding a costly signaling element provides substantial benefits for both reliable new entrants and existing members of the institution. By screening out unreliable types, the institution allows mutually beneficial cooperation to emerge that could not in the absence of the costly signal. This game therefore demonstrates that even if the primary effect of an institution is to screen out unreliable states, it can nevertheless have a substantial impact on the behavior of states that join, even in the face of weak commitment capacities.

Implications

This model gives rise to a number of empirical implications for the study of international institutions. Some implications regard states' decisions to join existing institutions, and others the behavior of states once they are in an institution. A number of implications also arise about different types of states' preferences over institutional design. In this section I will summarize some of the more prominent implications for empirical research, emphasizing those referring to institutional design.

First, consider pure weak commitment institutions – those that impose some cost for reneging, but that do not have sufficient enforcement capabilities to deter all reneging, and that do not impose an *ex ante* cost for joining the institution. When such institutions exist, there is no reason for states not to join them. The institution will then succeed in generating cooperation among reliable members (those who derive relatively high benefits from cooperation), but not among unreliable ones.

Weak commitment Hypothesis 1 (WC1): When institutions are costless to join, they will have a large membership.

Weak commitment Hypothesis 2 (WC2): When an institution is costless to join, its members will frequently renege on their commitments.

For purposes of this article, I am more interested in the properties of signaling institutions. In these institutions, because they sometimes screen out unreliable potential members, reneging should be less common than in pure commitment institutions. In addition, the frequency of reneging

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

should be a function of *ex ante* signaling costs. When such costs are low, reneging should be fairly common, although not quite as frequent as in pure commitment institutions. When signaling costs are moderate, they should fully screen out unreliable members, so that no reneging is observed. Finally, when entry costs are set too high, they are prohibitive and keep out even reliable members. We should not observe reneging, but we should also expect to see stagnant institutions with small membership.

Signaling Hypothesis 1 (S1): When entry costs are low, members should renege on their commitments fairly often.

Signaling Hypothesis 2 (S2): When entry costs are moderate, we should observe no reneging.

Signaling Hypothesis 3 (S3): When entry costs are very high, institutional membership should remain small, but we should not observe reneging.

The model also generates implications for states' preferences over institutional design, in particular over *ex ante* entry costs. First, consider the preferences of potential new members. Those who derive low benefits from cooperation but who could profit from suckering existing members into cooperation and then reneging will be opposed to any *ex ante* entry costs, as they can only lead to a decrease in utility.

Potential entrant Hypothesis 1 (PE1): Potential members who would derive low benefits from cooperation will oppose entry barriers.

In contrast, potential new members who would gain more from cooperation will have context-dependent preferences over entry costs. If the overall probability that a new entrant is reliable is low, those who would in fact be reliable will have a strong preference for the introduction of costly signaling. Such signaling costs will allow them to distinguish themselves from unreliable potential members. However, if the overall probability that a new member is reliable is relatively high, then reliable members have little to gain from the introduction of signaling costs, as they would have been able to gain the cooperation of existing members even in the absence of entry costs.

Potential entrant Hypothesis 2a (PE2a): When there are many potential new members in the population who would gain little from cooperation, those who anticipate large benefits from cooperation will favor the use of *ex ante* costs.

Potential entrant Hypothesis 2b (PE2b): When there are many potential new members in the population who would gain much from cooperation, they will be opposed to the use of *ex ante* costs.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Perhaps the most intriguing and immediately testable implications of the model involve the preferences of existing members of the institution over the introduction of signaling costs. Existing members have control over institutional design, so their preferences should be reflected in changes in the institution itself. The model implies that it is of great importance that existing members of an institution get the level of signaling costs 'just right'. If they set costs too low, they either leave existing members indifferent (if there are many reliable potential members in the population) or lead to only a small gain (if there are few reliable potential members in the population). If they set entry costs too high, existing members will either experience a loss (if there are many reliable potential members in the population) or be indifferent (few reliable potential members in the population). In contrast, moderate entry costs will serve as an effective screening device, separating reliable from unreliable potential members.

What does it mean to get signaling costs 'just right?' One important implication of the model is that the appropriate level of signaling costs is determined by the costs and benefits of cooperation for reliable *potential* members. That is, the benefits of cooperation for existing members and the costs to them if a new member reneges are *irrelevant* to existing members' preferences over signaling costs. Only the payoffs of potential members should enter into the optimal determination of entry costs. This logic implies that existing members should not have substantial disagreement among themselves about the 'right' cost of entry. Existing members must aim to set signaling costs so that they are lower than the benefits of cooperation for reliable new members (otherwise they would deter entry); but higher than the benefits of reneging on deals (otherwise they would not screen out unreliable new members). As either the benefits of cooperation or of reneging go up, existing members will calibrate by raising the costs of entry. These insights allow us to state the following hypotheses:

Existing member Hypothesis 1 (EM1): Existing members of an institution should express substantial concern that signaling costs be set at a level that is neither too high nor too low.

Existing member Hypothesis 2 (EM2): We should observe little conflict of interest among existing members about the appropriate level of entry costs.

Existing member Hypothesis 3 (EM3): As the benefits of cooperation for potential new members rise, signaling costs should go up.

Existing member Hypothesis 4 (EM4): As the benefits of reneging for potential members rise, signaling costs should go up.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

In the rest of this paper, I explore empirical applications of the weak commitment and signaling model, applying it to the institution of peacekeeping in civil conflicts and the expansion of the WTO.

Empirical applications

Peacekeeping

Peacekeeping is the quintessential ‘weak commitment’ institution. Peacekeepers enter a country when both sides in a civil war are willing to accept their intervention. The purpose of peacekeepers – as opposed to the occasional ‘peace enforcement’ mission – is to monitor a ceasefire and other terms of a peace agreement. They typically carry only light arms, and are not present in sufficient numbers to impose large costs on either side if they decide to being fighting anew. Thus, the question of why peacekeepers have any effect has been a puzzle.

Carter (2003) began developing an answer to this puzzle by arguing that the agreement between sides in a civil war to allow peacekeepers to enter itself serves as a signal to the UN about both sides’ intentions. Peace agreements are difficult to negotiate in civil wars, and the UN has a preference to see a clear agreement before sending in peacekeepers, so that the necessary conditions can be met for a separating equilibrium to exist.

More recently, Fortna and Martin (2009) have pushed further the idea that peacekeeping may serve primarily as a signal rather than as a commitment device.[6] They discuss the situation of fighting between a government and a rebel group. The rebel group is not sure whether the government is the type that will live up to the terms of a peace deal. The government can choose to negotiate a peace agreement with the rebel group; to invite peacekeepers in; or to continue fighting. The government pays a cost, in terms of violation of sovereignty and intrusive foreign troops monitoring its actions, if it allows peacekeepers to enter the country. Thus, peacekeeping has the potential to act as a costly signal of the government’s intention to comply with the terms of a peace accord.

The model allows for three possible outcomes: continued fighting; a peace accord (or truce) without the involvement of peacekeepers; or the intervention of peacekeepers. It leads to a series of predictions about when we are most likely to see each of these three outcomes. For example, the more a government values peace, the more likely it is to allow peacekeepers to enter. Similarly, the greater the sovereignty costs associated with peacekeeping, the less likely we are to observe that it occurs. Some results are

[6] Fortna and Martin (2009) model peacekeeping as a costly signal. In contrast to the model in this article, the institution has no commitment capacity. This article generalizes that model, adds the commitment dimension, and develops many additional observable implications.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

more counterintuitive. For example, consider the rebels' prior beliefs that the government is a reliable type (the analogue to p in the above model). As rebels become more certain that the government is reliable, the more *likely* we are to observe peacekeeping relative to continued fighting. However, an increased belief in reliability should also lead to a *decreased* chance of observing peacekeeping relative to peace agreements without peacekeeping, as rebels are more likely to be willing to accept agreements without peacekeepers. As in the model above, peacekeeping that involves a moderate *ex ante* cost will be the most effective at allowing rebels groups to distinguish between reliable and unreliable governments. A separating equilibrium will emerge when the up-front costs of peacekeeping are below the benefits of peace for a reliable government, but higher than the benefits of peace for an unreliable government.

Fortna and Martin subject the hypotheses derived from this model to a series of empirical tests, based on data from civil wars between 1989 and 1997 (64 cases). Because there are three possible outcomes, the appropriate method for analyzing these data is multinomial logit. They derive a series of proxies for the parameters of the model, for example, arguing that the duration of war is positively correlated with the benefits of peace and that democracy is positively correlated with rebels' prior beliefs about the reliability of the government.

They find substantial support for the predictions of the model. When considering the relative incidence of peacekeeping and continued fighting, the results are strong. The duration of war has a significantly positive coefficient, as predicted, while democracy also has a significantly positive coefficient. Other proxies also perform as expected and all but one meet standard tests of statistical significance. The multinomial logit specification also provides tests of the relative incidence of peacekeeping and peace agreements without peacekeeping. The results on this dimension are not quite as strong, but still promising. For example, the duration of war has the predicted positive effect, but democracy no longer has a statistically significant effect. Overall, they conclude that the signaling model provides substantial insight into the demand for peacekeeping, and helps us to understand its dynamics in a way that thinking of it purely as a commitment device could not.

The WTO

The GATT/WTO provides an excellent setting for testing the central implications of the signaling model. Over time, the demands made of countries that wish to join the GATT/WTO have changed substantially. The General Agreement on Tariffs and Trade (GATT) often had strikingly

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

low barriers to entry. For example, post-colonial states were guaranteed, under Article XXVI:5(c), accession to the GATT with essentially no bargaining or other costs involved (Copelovitch and Ohls 2012). This situation changed in 1995 with the creation of the WTO, as no similar provision was made for former colonies or dependencies. Since 1995, accession negotiations have tended to be drawn-out and contentious, especially in prominent cases such as the accession of Russia (Dyker 2004) and China.

A major turning point in the process of accession came during the Uruguay Round of negotiations that led to establishment of the WTO. The Uruguay Round Agreements themselves do not provide any explicit rules for the process of accession, stating only that accession is open to any country as long as that country agrees on terms with WTO member states (Kavass 2007, 455). In practice, the WTO secretariat has developed detailed accession procedures and has established an Accessions Division. The procedures laid out by the Secretariat are highly bureaucratic and 'labyrinthine' (Kavass 2007, 456). All of these procedures have created substantial costs for potential entrants, in particular countries transitioning away from communism. In fact, the costs of accession are high enough to be a major concern to developing and transition countries, and have prompted organizations such as UNCTAD and the World Bank to publish extensive guides to the accession process in an effort to assist states to overcome these hurdles (see United Nations Conference on Trade and Development 2001).

"Countries seeking to become Members of the WTO must be prepared to perform a hefty volume of highly demanding work. Not only do they need to submit a voluminous amount of documents and attend meetings to answer questions; they also may need to make extensive and substantial changes to their tariffs and taxes, as well as revise many of their existing laws and regulations in order to bring them into conformity with the WTO norms and standards" (Kavass 2007, 461).

Analysts express a consensus that the demands made of acceding countries have increased substantially over the GATT/WTO's history (Ognivtsev, Jounela, and Tang 2001, 173). The process of escalating demands actually predates the creation of the WTO, going back to the 1980s. Mexico's entry to the GATT provides a good example. It had reached an accession agreement during the Tokyo Round (1979), but decided in 1980 not to implement the agreement. In 1985 Mexico negotiated a new accession agreement that required a much higher entry cost than GATT members had imposed just 6 years earlier. The commitments to reduce trade barriers were far lengthier and more precise, with fewer loopholes to protect Mexican industries. These enhanced demands came at the behest of the United States. VanGrasstek finds the Mexican example typical: 'Many of the countries that acceded to the GATT during the 1980s found the process to be

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

more demanding, in large measure because of a change in policy on the part of the major trading countries' (2001, 127).

As the commitment and signaling model suggests, existing member states that worry about possible reneging of new members deliberately set barriers to entry high, even going beyond what the GATT/WTO agreements themselves specify (Butkeviciene *et al.* 2001, 230). The requirements that existing members impose on entrants, and the one-sided nature of the negotiating process, fit the assumptions of the model well. Analysts even argue that, beyond committing to comply with the WTO agreements, an acceding country must 'pay a "membership fee" in terms of specific concessions on tariff rates, commitments on agricultural subsidies and commitments on trade in services in return for its right to enjoy the benefits resulting from liberalization achieved in previous multilateral trade negotiations' (Ognivtsev, Jounela, and Tang 2001, 181). It is also worth noting that WTO procedures allow existing members a great deal of flexibility with respect to entry requirements. 'Paradoxically for a rules-based organization, the WTO has no clear rules for the "price" of membership' (Evenett and Primo Braga 2005, 2). The signaling model predicts that the entry barrier needs to be calibrated to each entrant's costs and benefits. Thus, the lack of clear rules is not a paradox, but exactly what we would expect to see.

The model predicts that we should see a higher entry cost imposed when a potential member could benefit substantially from joining and then reneging on its commitments (EM3), and we see widespread evidence of this dynamic in cases such as Mexico, Russia, and China. In their statistical analysis of GATT/WTO accession, Davis and Wilf (2014) find that the costs of entry decrease for states that are allies of existing WTO members, and for democracies. Ally status and democracy can both be interpreted as proxies for the benefits of reneging: close allies are likely to derive lower benefits from reneging that states who are not allies, and democracies are widely understood to be more rule-bound and thus to derive lower benefits from cheating. In addition, the transparency of democracies is likely to make reneging more difficult. Thus, Davis and Wilf's results support the model, in that states that are likely to gain less from reneging on their commitments do not have to pay as high an entry price.

The model also predicts that countries that would derive higher benefits from joining will be asked to pay a higher entry cost (EM4), and we also find widespread evidence supporting this implication in case studies. One analyst concludes that if existing members 'know that their interlocutor is under strong political pressure back home to secure accession at any cost, the negotiators in Geneva will feel even more secure in setting a high price' (VanGrasstek 2001, 136). The signaling model also suggests an interesting

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

twist on the results found by Allee and Scalera (2012) in their statistical analysis of the effects of the accession process. They argue that states that are subject to more stringent entry conditions undergo more liberalization, and thus obtain greater benefits on entry to the WTO. The signaling model suggests that the real story may be about reverse causation, in that states who expect to benefit more from entry will have to pay a higher price to join. Thus, the effect they identify may not be the direct causal result of policy changes during the accession process. Instead, these changes may be more about 'burning money', proving to existing members that the new entrant anticipates high benefits from cooperation and so is willing to pay this costly signal up-front.

This survey of evidence from case studies of GATT/WTO accession provides preliminary support for the signaling model. However, more systematic tests based on this experience should be possible. One research direction will involve looking more directly at negotiations over individual accessions, to determine whether the hypotheses summarized above about preferences over the terms of entry hold up. Another approach will include statistical analysis, attempting to explain variation in the entry fees demanded of different new members. The literature has identified a number of proxies for entry costs, ranging from the number of specific commitments made in particular sectors to the length of negotiations. While neither of these is a fully adequate indicator of *ex ante* costs, looking for robust results across a number of indicators should allow for more precise testing of the signaling model.

Conclusion

How do international institutions exert their effects on members? Typically, scholars have looked at the ability of institutions to enhance commitments by imposing costs on members if they renege. However, institutions rarely if ever have the capacity to fully commit all members to all of their commitments, all of the time. Instead, they operate as weak commitment devices, leading to enhanced but inconsistent cooperation. In this paper, I provide a model of institutions as weak commitment devices, then add to the model the potential for institutions to also perform a signaling function by requiring new members to pay an *ex ante* cost for joining the institution.

The model demonstrates that a weak commitment institution, on its own, has a number of undesirable properties. In particular, it allows unreliable new members under some conditions to join the institution and fails to ensure that cooperation emerges among reliable states. In contrast, when members must send costly signals to join the institution, and these signals

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

are appropriately calibrated, reliable members can consistently distinguish themselves from unreliable. In this case, the fullest extent of mutually beneficial cooperation emerges, and unreliable states are prevented from being able to sucker existing members of the institution.

One important lesson of this model is that institutions that 'only' screen can have a significant impact on levels of cooperation, contrary to claims in much of the literature. Even if an institution's main effect is to screen out those who are not very interested in cooperation, the institution can have a causal effect. This screening process reassures existing members of the institution that others are reliable, and allows new entrants who have a lot to gain from cooperation to differentiate themselves who do not. Thus, the process of screening via costly signaling, if the costs are set at appropriate levels, allows cooperation to emerge that would not be possible in the absence of the institution.

This model gives rise to a rich set of empirical implications. If we consider international peacekeeping in civil wars, conceiving of peacekeeping as a weak commitment and signaling institution fits the pattern of peace and conflict well. Work on the WTO suggests that the model's predictions about members' preferences for the costliness of signals hold up. The model could be extended to expansion of other important institutions, such as NATO, the EU, and regional trade agreements.

Acknowledgments

The author would like to thanks participants in seminars at the University of Wisconsin-Madison and the Ohio State University, as well as participants in the International Political Economy Society, for comments on drafts of this work. The author thanks to Andy Kydd for advice on elaborating the formal discussions of equilibria in the models. The author also thanks Inken von Borzyskowski, Patrick Kearney, Mark Toukan, and Kyle Marquardt for excellent research assistance. All errors are of course the author's responsibility.

References

Allee, Todd L., and Jamie E. Scalera. 2012. "The Divergent Effects of Joining International Organizations: Trade Gains and the Rigors of WTO Accession." *International Organization* 66(2):243–76.

Baglioni, Angelo. 2008. *Corporate Governance Institutions as Signalling and Commitment Devices*. Milano, Italy: Università Cattolica.

Bolle, Friedel. 2002. "Signals for Reliability: A Possibly Harmful Institution?" *CEJOR* 10: 217–27.

Braham, Matthew, and Friedel Bolle. 2006. "A Difficulty With Oaths: On Trust, Trustworthiness, and Signalling." *European Journal of Law and Economics* 22(3):219–32.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Butkeviciene, Jolita, Michiko Hayashi, Victor Ognivtsev, and Tokio Yamaoka *et al.* 2001. "Terms of WTO Accession." In UNCTAD 2001, 230–264.

Carter, Timothy A. 2003. "UN Peacekeeping: Treaties, Signaling, and Peace." Delivered at the American Political Science Association Annual Meeting, Washington, DC.

Copelovitch, Mark S., and David Ohls. 2012. "Trade, Institutions, and the Timing of GATT/WTO Accession in Post-Colonial States." *Review of International Organizations* 7(1):81–107.

Davis, Christina L., and Meredith Wilf. 2014. *Joining the Club: Accession to the GATT/WTO*. Princeton, NJ: Princeton University Press.

Downs, George W., and David M. Rocke. 1997. *Optimal Imperfection? Domestic Uncertainty and Institutions in International Relations*. Princeton, NJ: Princeton University Press.

Dyker, David A. 2004. "Russian Accession to the WTO – Why Such a Long and Difficult Road?" *Post-Communist Economies* 16(1):3–20.

Evenett, Simon J., and Carlos A. Primo Braga. 2005. "WTO Accession: Lessons from Experience." Trade Note 22, The World Bank Group, Washington, DC.

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands Versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

Fortna, Page V., and Lisa L. Martin. 2009. "Peacekeepers as Signals: The Demand for International Peacekeeping in Civil Wars." In *Power, Interdependence and Non-State Actors in World Politics: Research Frontiers*, edited by Helen V. Milner, 87–107. Princeton, NJ: Princeton University Press.

Gray, Julia. 2013. *The Company States Keep: International Economic Organizations and Investor Perceptions*. New York: Cambridge University Press.

Guzman, Andrew T. 2008. *How International Law Works: A Rational Choice Theory*. New York: Oxford University Press.

Hafner-Burton, Emilie M., Edward D. Mansfield, and Jon C.W. Pevehouse. 2015. "Human Rights Institutions, Sovereignty Costs, and Democratization." *British Journal of Political Science* 45(1):1–27.

Haftel, Yoram Z. 2007. "The Effect of BITs on FDI Inflows to Developing Countries: Signaling or Credible Commitment?" Paper Presented at the Ohio State University Workshop on Globalization, Institutions, and Economic Security, November, Columbus, Ohio.

Hyde, Susan Dayton. 2011. *The Pseudo-Democrat's Dilemma: Why Election Observation Became an International Norm*. Ithaca, NY: Cornell University Press.

Kavass, Igor I. 2007. "WTO Accession: Procedure, Requirements and Costs." *Journal of World Trade* 41(3):453–74.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy*. Princeton, NJ: Princeton University Press.

Krasner, Steven D. 1982. *International Regimes*. Ithaca, NY: Cornell University Press.

Kydd, Andrew. 2001. "Trust Building, Trust Breaking: The Dilemma of NATO Enlargement." *International Organization* 55(4):801–28.

Leeds, Brett Ashley. 2003. "Alliance Reliability in Times of War: Explaining State Decisions to Violate Treaties." *International Organization* 57(3):801–27.

Mansfield, Edward D., and Jon C. Pevehouse. 2008. "Democratization and the Varieties of International Organizations." *Journal of Conflict Resolution* 52(2):269–94.

Morrow, James D. 1994. "Alliances, Credibility, and Peacetime Costs." *Journal of Conflict Resolution* 38(2):270–97.

Morrow, James D. 2000. "Alliances: Why Write Them Down?" *Annual Review of Political Science* 3:63–83.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

Ognivtsev, Victor, Eila Jounela, and Xiaobing Tang. 2001. “Accession to the WTO: The Process and Selected Issues.” In UNCTAD 2001, 172–229.

Schultz, Kenneth A. 1999. “Do Democratic Institutions Constrain or Inform? Contrasting Two Institutional Perspectives on Democracy and War.” *International Organization* 53(2):233–66.

Setear, John K. 2002. “The President’s Rational Choice of a Treaty’s Preratification Pathway: Article II, Congressional-Executive Agreement, or Executive Agreement?” *Journal of Legal Studies* 31:S5–39.

Simmons, Beth A. 2000. “International Law and State Behavior: Commitment and Compliance in International Monetary Affairs.” *American Political Science Review* 94(4):819–25.

Slantchev, Branislav L. 2005. “Military Coercion in Interstate Crises.” *American Political Science Review* 99(4):533–47.

Thompson, Alexander. 2006. “Coercion Through IOs: The Security Council and the Logic of Information Transmission.” *International Organization* 60(1):1–34.

Thompson, Alexander. 2009. *Channels of Power: The UN Security Council and U.S. Statecraft in Iraq*. Ithaca, NY: Cornell University Press.

United Nations Conference on Trade and Development. 2001. *WTO Accessions and Development Policies*. New York: UNCTAD.

VanGrasstek, Craig. 2001. “Why Demands on Acceding Countries Increase Over Time: A Three-Dimensional Analysis of Multilateral Trade Diplomacy.” In UNCTAD 2001, 115–140.

von Stein, Jana. 2005. “Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance.” *American Political Science Review* 99:611–22.

von Stein, Jana. 2008. “The International Law and Politics of Climate Change: Ratification of the United Nations Framework Convention and the Kyoto Protocol.” *Journal of Conflict Resolution* 52(2):243–68.

Walsh, James Igoe. 2007. “Do States Play Signaling Games?” *Cooperation and Conflict: Journal of the Nordic International Studies Association* 42(4):441–59.

Appendix 1: Equilibrium analysis of pure commitment game

Strong commitment game ($b_u + \epsilon > a - c$): When the institution can impose a large punishment, even an unreliable A receives a higher payoff from cooperating than reneging. Thus, if A has joined the institution, both A types will choose to cooperate if B cooperates. Since all A’s that join the institution will cooperate, B will receive a higher payoff from cooperating than not, and so will always cooperate if A has joined the institution.

What if A chooses not to join the institution and B nevertheless chooses to cooperate? Then A will get a higher payoff from reneging than from cooperating, so A will renege. Anticipating this, B will not cooperate if A has not joined the institution. A thus receives a payoff of 0 if it does not join but gets the benefits of cooperation if it does; so in equilibrium, A will always join the institution.

Weak commitment game ($b_r + \epsilon > a - c$; $b_u + \epsilon < a - c$): As in the strong commitment game, if A has not joined the institution, B knows that A will renege and so will not cooperate.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

If A has joined the institution, it can impose a punishment cost that is sufficient to induce cooperation from a reliable A, but an unreliable A would choose to renege in the last stage because with the low punishment cost, the benefits of reneging are greater than the benefits of cooperation.

When B observes that A has joined the institution, if it chooses to cooperate it will receive the cooperative payoff, b_b, with probability p. With probability $(1-p)$, it will receive the sucker's payoff, $-d$. B will thus choose to cooperate if $p(b_b)+(1-p)(-d)>0$, or $P>d/(b_b+d)$.

Should a reliable A join the institution? If p is high, it gets the benefits of cooperation, so it will. If p is low, it still gets the ancillary benefits of institutional membership, ϵ, so it is still worthwhile to join.

Should the unreliable A join the institution? If p is high, it will be able to sucker B, so it will join. If p is low, like the reliable type it still gets the ancillary benefits of institutional membership, so it will join.

Because both A types join the institution, B is not able to update its beliefs about A's type, so as explained above its choice of whether to cooperate will be based on its prior beliefs p.

Appendix 2: Equilibrium analysis of signaling game

The signaling model presented in this paper is a standard one, and as usual I use the equilibrium concept of a Pure Bayesian Equilibrium.

High signaling cost ($z>b_r>a-c$): When the cost of entering the institution is higher than the benefits of cooperation for even the reliable type ($z>b_r$), both the reliable and unreliable types will not pay the cost. In this case, we find a pooling equilibrium in which A does not choose to enter the institution. Because even a reliable A will renege in the absence of the commitment effects of the institution (see Appendix 1), B will not cooperate ($0>-d$).

Could B benefit by deviating by cooperating? No: since A has not joined the institution, all A's will renege, so B could only lose by cooperating.

What if A deviated and paid the cost of entering the institution? Because $z>b_r$, even a reliable A would not benefit from this move. Off the equilibrium path, how would B respond if A were to join the institution? Since this is not an optimal move for either A type, it is reasonable to assume that B would not update beliefs about A's type in response to this off the equilibrium path move. Thus, B would cooperate if the expected payoff to cooperating is greater than the payoff for no cooperation: $p(b_b)+(1-p)>0$; or $P>d/(b_b+d)$. Knowing that B will cooperate off the equilibrium path when p is relatively large, should either A type deviate? No, because the cost of entering the institution (z) is greater than the benefits of cooperation for either type (b_r or b_u). Thus, beliefs and optimal

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

strategies off the equilibrium path support this pooling equilibrium where neither A type enters the institution when z is high.

Moderate signaling cost ($b_r > z > a - c$): A separating equilibrium exists when z is at a moderate level, so that a reliable A is willing to bear this cost but an unreliable A is not. In this equilibrium, B cooperates with A's that join the institution, but not those that refuse to join.

When $b_r > z$, the reliable A type will benefit from cooperation and is willing to pay the cost of entering the institution if B will respond by cooperating. However, when $z > a - c$, in equilibrium the unreliable A will not join the institution even if this move induces B to cooperate. As shown in Appendix 1, the unreliable type will always renege when B cooperates, because $a - c > b_u$ (the weak commitment condition). In the signaling game, the unreliable A's payoff if B cooperates is thus $a - c - z$; when $z > a - c$, the unreliable type is better off not joining the institution and getting the no-cooperate payoff of 0.

Because the unreliable type will not join the institution when the signaling costs are moderate, B will not cooperate if A does not join the institution. If A does join the institution, B can update beliefs and know with certainty that A is the reliable type, since the unreliable type is unwilling to pay the moderate signaling cost. Knowing that A is reliable, B will then cooperate ($b_b > 0$). Because B will cooperate if A joins the institution, the reliable type will pay the moderate signaling cost ($b_r - z > 0$).

Could a reliable A benefit by deviating and refusing to join the institution? No, because B would then not cooperate, and A's payoff would decrease from $b_r - z$ (which is positive) to 0. An unreliable A could not benefit by deviating and joining the institution, because its payoff would then be $a - c - z$, which is less than 0.

Could B benefit from cooperating even when A refuses to pay the signaling cost? Because this is a separating equilibrium, on observing that A does not join the institution, B knows with certainty that A is the unreliable type. Therefore, choosing to cooperate would reduce B's payoff to $-d$, so B will not deviate. B will also not deviate by refusing to cooperate when A joins the institution. Again, on observing that A joins the institution, B knows with certainty that A is the reliable type, so that B will get a payoff of b_b from cooperating, which is greater than 0, the payoff from no cooperation.

Low signaling cost: ($b_r > a - c > z$): When signaling costs are low, the equilibrium will depend on B's prior belief about A's reliability (p). A pooling equilibrium in which all A types join the institution and B always cooperates exists when p is high enough that B is willing to take the chance and cooperate. This occurs when B's expected payoff from cooperation is greater than the no-cooperation payoff, 0. Since all A's join the institution, B cannot update prior beliefs, and so calculates his expected payoff based on priors. The pooling equilibrium thus holds when B's expected payoff to

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

cooperating is greater than the no-cooperate payoff:

$$p(b_b)+(1-p)(-d)>0$$
$$p>d/(b_b+d). \tag{1}$$

When p is high, B cannot benefit by deviating and not cooperating, because this would reduce B's payoff to 0. Off the equilibrium path, how would B react if A were to refuse to join the institution? As established above, in the absence of the institution all A types will renege, so B will not cooperate if A does not join regardless of prior beliefs. Thus, neither A type can benefit from refusing to join the institution, because A's payoff would decrease to 0 $(a-c-z>0;\ b_r-z>0)$.

When signaling costs are low but p is below the threshold that allows B to cooperate $(P<d/(b_b+d))$, a semi-separating equilibrium emerges. A pooling equilibrium in which B cooperates cannot exist, because B's belief that A is reliable is too low to meet the condition in Equation (1). A separating equilibrium cannot exist in which only reliable A's join the institution, because the signaling cost z is so low that unreliable A's would bluff and pay the cost if this would induce B to cooperate. The only equilibrium in this instance is for unreliable A's and B's to both play a mixed strategy that leaves the other player indifferent between their pure strategies.

Let unreliable A choose the institution with probability x. A will choose x so that B is indifferent between cooperating and not cooperating. On observing that A joins the institution, B updates beliefs about A's type using Bayes' Rule, yielding the posterior belief that A is reliable with probability $p/(p+x(1-p))$. B's expected payoff from cooperating is then $(p/(p+x(1-p)))b_b+(x(1-p)/(p+x(1-p)))(-d)$. A will choose x so that B is indifferent between this expected payoff and the no-cooperation payoff:

$$(p/(p+x(1-p)))b_b+(x(1-p)/(p+x(1-p)))(-d)=0$$
$$x=pb_b/(d(1-p)). \tag{2}$$

On observing that A chooses the institution, B will cooperate with probability y that leaves unreliable A indifferent between choosing the institution and not. Unreliable A's payoff from choosing the institution is then $(y(a-c))+((1-y)0)-z$:

$$(y(a-c))+((1-y)0)-z=0$$
$$y=z/(a-c). \tag{3}$$

In the semi-separating equilibrium, reliable A will always choose the institution, gaining probabilistic cooperation from B. Reliable A's expected payoff from joining the institution in the semi-separating equilibrium is $(z/(a-c))(b_r-z)$. Could reliable A benefit by deviating and not choosing the

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press

institution? No, because then B would not cooperate (established above), reducing reliable A’s payoff to 0.

Could the unreliable type benefit from deviating from the mixed-strategy equilibrium? The unreliable type’s expected payoff in this equilibrium is $xy(a-c-z)+x(1-y)(-z)$, or $x(ya-yc-z)$, which is greater than 0. If the unreliable type were to deviate and not join the institution, the payoff would decrease to 0, since B never cooperates in the absence of the institution. If the unreliable type were to deviate and join the institution with certainty, B would not be able to update beliefs using Bayes’ Rule, as both A types would be pooling on the same strategy. Since p is relatively low, B would choose not to cooperate, reducing unreliable A’s payoff to $-z$. Thus, when signaling costs are low and B believes that A is likely unreliable (low p), the only the mixed-strategy equilibrium holds.

https://doi.org/10.1017/S1752971917000082 Published online by Cambridge University Press