Designing the Optimal International Climate Agreement with Variability in Commitments

Jordan H. McAllister and Keith E. Schnakenberg

Abstract We analyze the design of an international climate agreement. In particular, we consider two goals of such an agreement: overcoming free-rider problems and adjusting for differences in mitigation costs between countries. Previous work suggests that it is difficult to achieve both of these goals at once under asymmetric information because countries free ride by exaggerating their abatement costs. We argue that independent information collection (investigations) by an international organization can alleviate this problem. In fact, though the best implementable climate agreement without investigations fails to adjust for individual differences even with significant enforcement power, a mechanism with investigations allows adjustment and can enable implementation of the socially optimal agreement. Furthermore, when the organization has significant enforcement power, the optimal agreement is achievable even with minimal investigative resources (and vice versa). The results suggest that discussions about institutions for climate cooperation should focus on information collection as well as enforcement.

The international community has failed to significantly slow the pace of climate change, even though its causes are known and its effects are potentially devastating.[1] This failure is partially explained by incentives: individual governments are tempted to let other countries bear the costs of climate change abatement and so reap the benefits without taking on these costs themselves.[2] Thus scholars and practitioners have emphasized the importance of well-designed international institutions to help shape the incentives of national actors.[3] In so doing, an ideal institution would need to not only manage free riding but also reflect variability between countries, by asking countries that can reduce fossil fuel emissions at a lower cost to do more and requiring others to do less.[4] Agreements with nonuniform obligations are more cost-effective because member states have varying costs, as well as broader because variability ensures that each country's quota is appropriate given the country's relative capability.[5]

1. Stern 2007.
2. Barrett and Nitze 2007; Bernauer 2013; Clapp and Dauvergne 2011; Hovi, Sprinz, and Underdal 2009; Tavoni et al. 2011.
3. Martin 1992; Michaelowa and Michaelowa 2017; Ostrom 1990.
4. Keohane and Victor 2011; McGinty 2007.
5. Barrett 2003.

International Organization 76, Spring 2022, pp. 469–86
© The IO Foundation 2021
doi:10.1017/S0020818321000400

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

The importance of accounting for individual differences has been recognized in major efforts to create a global climate agreement. First, the precursor to modern climate agreements, the 1987 Montreal Protocol on the ozone layer, set the standard by assigning nonuniform commitments in which developing countries were given less strict emissions limits.[6] This was echoed in later climate agreements such as the 1992 United Nations Framework Convention on Climate Change (UNFCCC), which recognized the "common but differentiated responsibilities" of all countries.[7] In due course, the 1997 Kyoto Protocol set emissions limits for only developed countries and even set country-specific limits wherein economies in transition were allowed to use a different base year to compare their reductions to.[8] Most recently, the 2016 Paris Agreement simply let each country determine its own quota.[9]

Reflecting variability between countries is a desirable component of a climate agreement. First, a one-size-fits-all policy would be suboptimal because some countries are able to reduce fossil fuel emissions at a much lower cost than others. So in an agreement with uniform requirements, some countries bear disproportionate burdens while others that could do more are not required to do so.[10] Second, one-size-fits-all agreements suffer from a "broad versus deep" trade-off which limits their success. If international agreements require the same commitments of all members, then they can either have broad membership or require deep commitments, but not both.[11] By contrast, a variable agreement encourages both broader participation and deeper commitments for countries able to make them by setting each country's quota separately according to its capabilities. However, because countries' leaders are likely to be better informed about the individual circumstances in their own countries than are leaders of other countries or of international organizations (IOs), an agreement must rely to some extent on self-reports to implement a variable agreement—relatedly, it could be the case that an international agency (such as the United Nations Environment Programme) gathers information transmitted by countries before drafting an agreement (as in the Montreal Protocol), or that an agreement relies on each country to determine the level of commitment it is able to make (as in the Paris Agreement).[12]

Unfortunately, as Harrison and Lagunoff have argued, overcoming free riding while accounting for individual differences is difficult in the face of this asymmetric information.[13] To see why, consider an agreement involving only two countries and a

6. Barrett 2003.
7. Stern 2007.
8. Barrett 2003; Von Stein 2008.
9. Harstad 2018; Hovi et al. 2019.
10. Bernauer 2013. Though we are referring solely to efficiency when we discuss disproportionate burdens as suboptimal, one-size-fits-all agreements may be especially unacceptable to climate negotiators with preferences for equity. Dannenberg, Sturm, and Vogt 2010; Kennard 2021.
11. Barrett 1994, 2003.
12. The Montreal Protocol, UNFCCC, Kyoto Protocol, and Paris Agreement all rely on member states' self-reports for information collection. Barrett 2003; Hovi et al. 2019; Slechten 2020.
13. Harrison and Lagunoff 2017.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

mediator helping countries decide what commitments they must make under the agreement. Setting enforcement issues aside for the moment, assume the two countries have significant enough repeated interactions that they can make credible commitments to one another, knowing that they can mutually punish defection from their agreement.[14] The ideal agreement has a country commit to only a moderate reduction in emissions when it faces a hardship resulting from a change in economic or political conditions, and to a more significant reduction when it faces no such hardship.[15] Consistent with the free-riding problem, however, a country facing no hardship may still prefer to commit to only moderate reductions and let the other country take on more of the costs of climate action. This creates a problem for the mediator trying to decide what commitments each country should make under information asymmetry.[16] Each country is tempted to exaggerate its hardships so that it is allowed to make smaller commitments under the agreement.[17] Information asymmetry is significant enough that in some circumstances, the best agreement that can be implemented is one that has no variation in commitments.[18]

In this paper we analyze how to design an optimal climate agreement when countries' costs of climate action are private information. The problem the IO must solve is to maximize global welfare subject to the constraint that countries are willing to participate and also to submit accurate self-reports. To focus on the informational aspect of the problem, we assume that the IO has some (perhaps limited) ability to enforce the commitments member states make.[19] For instance, enforcement leverage may come from some positive linkage such as controlling access to a club good or other benefit that accrues to countries for being in good standing with the organization,[20] or even from some negative linkage such as punishments used to coerce countries into compliance.[21] Nonetheless, without any investigative resources, this leverage for inducing compliance does not help the organization design an optimal agreement.

Even as the IO has very high leverage for enforcement, it cannot prevent countries from exaggerating compliance costs. However, when the IO has the power to investigate the compliance costs of member states, it may be able to overcome misreporting while also allowing variability in commitments. Notably, the availability of

14. Hovi, Ward, and Grundig 2015; Milgrom, North, and Weingast 1990; Rosendorff and Milner 2001.
15. Koremenos 2005; Pelc 2009.
16. See also Slechten 2020.
17. See Herzing 2005, Karp and Zhao 2010, and Slechten 2020 for discussion of how the unobserved nature of economic shocks, abatement costs, or domestic opposition produces an incentive problem in which countries overstate the size of any such barriers to maintaining their commitments. This also mirrors the standard problem of mechanism design for public goods provision where agents have private information about their willingness to pay for the public good. Csapó and Müller 2013.
18. Harrison and Lagunoff 2017.
19. This simplifies the exposition of results, but we can easily provide microfoundations. In the online supplement, we discuss endogenous enforcement power through dynamic agreements or issue bundling. In both cases, punishments or rewards are self-enforcing and do not require direct policing power by the IO.
20. Keohane and Victor 2016; Nordhaus 2015; Stern 2007.
21. Barrett 2003; Martin 1992; Mitchell and Keilbach 2001.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

investigative resources means that the IO's leverage for encouraging compliance can also help it incentivize truth telling to make it so that the agreement is able to assign realistic commitments for each member state. This could be a benefit of major climate agreements with a verification procedure to check countries' reports, including the Montreal Protocol (for the ozone), UNFCCC, Kyoto Protocol, and Paris Agreement.[22] To increase their investigative capacity, IOs can also oversee investigations conducted by external bodies. Indeed, although most environmental organizations lack significant monitoring capabilities, they have historically been able to rely on nongovernmental organizations (NGOs) or other actors to conduct monitoring for them. For example, the 1979 Long-Range Transboundary Air Pollution Convention did not have the ability to monitor but was able to depend on environmental NGOs to do so.[23] Likewise, because free-riding behavior is so costly to all countries in an agreement, they will have an incentive to request or acquire extensive information on each other's true types.[24] So even if the organization itself lacks the ability to collect all of the information it needs, it can still expect entities such as NGOs and member states to conduct their own investigations.

We contribute to the literature in two main ways. First, we contribute to the literature on mechanism design and climate change cooperation. In general, we follow the literature in viewing the main function of climate agreements as attempting to overcome problems of free riding.[25] The closest papers are by Harrison and Lagunoff and Slechten.[26] Harrison and Lagunoff take a mechanism-design approach to constructing an international agreement to limit carbon consumption. In their model, the best implementable agreement subjects every country to the same requirements, even though the best possible policy would account for individual differences in costs and abilities. The problem is that if countries have private information about these differences, they may try to free ride by exaggerating their costs from reducing carbon consumption. We complement that paper by considering independent information gathering as a potential solution to this issue of misreporting.

Slechten also considers the possibility of verifying self-reports in a mechanism-design setting for environmental agreements.[27] In her model, states can voluntarily allow the IO to "certify" their private information, making it completely verifiable. By contrast, we consider probabilistic verification and allow the IO to freely allocate investigative resources determining the likelihood of verification. We also differ from Slechten by considering a multilateral agreement with no transfers, reflecting a weak IO that cannot freely move resources between countries.

22. See Barrett 2003, Hovi et al. 2019, and Slechten 2020 for descriptions of these procedures. Though these procedures are often used to check reports of compliance, we suggest that similar institutions could in theory be employed for wider information collection, as depicted in the model in Slechten 2020.
23. Dai 2007.
24. Martin 1992.
25. Hovi, Ward, and Grundig 2015; Nordhaus 2015; Ostrom 1990.
26. Harrison and Lagunoff 2017; Slechten 2020.
27. Slechten 2020.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Second, we contribute to the international relations literature on flexibility in the design of international institutions more broadly. Much of the work on flexibility comes from the literature on trade agreements such as the World Trade Organization.[28] In these papers, two factors driving the need for flexibility are changes over time that require renegotiation,[29] and unexpected situations within countries leading to a need for escape clauses[30] or compensatory payments.[31] We conceptualize flexibility in agreements differently than in these papers, and use the term *variability* instead to avoid conflating these ideas. Although these papers refer either to renegotiation or to forgiveness from punishment, we simply mean allowing the terms of the agreement to be different for each country. However, some of the basic insights about flexibility are similar, and we believe similar informational problems to those raised in this paper could apply to escape clauses, compensation, or renegotiation.[32] The flexible mechanism closest to the kind of variability we examine here is the escape clause, which enables countries to temporarily lower their commitments when their costs of abatement (or type) become higher due to an economic shock, military conflict, or another similar event.[33]

In particular, we address the climate agreement literature's commonly noted broad versus deep trade-off between the strictness of actions required under an agreement and the ability to gain more signatories.[34] The basic intuition is that an agreement that requires more significant action on the part of its members will be more costly for those countries, so all countries will be less inclined to sign the agreement initially.[35] In other words, an agreement with uniform commitments can either be small in membership but assign a difficult quota (in which case only the most committed countries would be members), or large in membership but assign an easy quota (in which case only the least committed countries would curb their emissions). This problem is particularly difficult when it comes to climate change because nonmembers of climate agreements are not only refraining from action but also making the problem worse because maintaining business as usual actually results in further deterioration of the climate. Therefore, it is not evident whether it is better to try to craft a deeply committed agreement of only the most willing countries and risk

28. Goldstein and Martin 2000; Herzing 2005; Pelc 2009; Rosendorff 2005; Rosendorff and Milner 2001; Sykes 1991. See also Gilligan and Johns 2012 for a useful review of some of this literature.

29. Barrett 2003; Harstad 2018; Koremenos 2001; Kucik and Reinhardt 2008.

30. Helfer 2012; Koremenos 2005; Koremenos, Lipson, and Snidal 2001; Rosendorff and Milner 2001; Sykes 1991.

31. Herzing 2005; Kucik and Reinhardt 2008.

32. Boockmann and Thurner 2006.

33. Pelc 2009. Regarding environmental agreements in particular, Karp and Zhao 2010 show how allowing a country that exceeds its emissions quota to use an escape clause enables an environmental agreement to simultaneously achieve efficient abatement and high participation. In fact, around one-fourth of environmental agreements contain an escape clause. Marcoux 2009.

34. Hovi, Ward, and Grundig 2015; Keohane and Victor 2016; Von Stein 2008.

35. Barrett 1994, 2003.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

letting less-willing countries pollute as normal, or to form a weak agreement that includes all countries but which does not require much of any of its members.[36]

Variability in commitments may increase the number of willing signatories by ensuring that countries that face higher compliance costs receive some relief. An agreement that allows variation in the requirements made of less-well-off countries or in less secure times will be able to both require more of countries and get more countries to join because these countries know they will not be asked to do more than they are able at any given time.[37] Here, our conception of variation is the extent to which the agreement's requirements account for individual differences. An optimally varying agreement softens the trade-off between breadth and depth but requires independent investigations of member states' claims to overcome the incentive of countries to exaggerate their costs of compliance. We now turn to the model.

The Game

In the benchmark model, the players comprise N countries, where $N = \{1, \ldots, n\}$. We assume that each country $i \in N$ chooses its own level of carbon emissions $c_i \geq 0$, and that all countries choose simultaneously.[38] However, we also assume that each country has a type θ_i which takes a value in $[\underline{\theta}, \bar{\theta}] \subset (0, 1)$. Higher types value carbon consumption more than other countries or would incur relatively large costs from reducing emissions, while lower types value abatement more than other countries or would incur relatively small costs from reducing emissions. For example, a high type of country could be one that will adapt easily to climate change or whose economic strategy entails continued investment in energy-intensive manufacturing sectors. A low type of country may be one invested in clean technologies or less energy-intensive industries. We may also think of types as related to internal political pressures within a country, given that the policy preferences of national leaders are influenced by the relative power of environmental interest groups, industry lobbyists, and so on. Types are private information but are commonly known to be independent and identically distributed along a continuous distribution F with density f and support in $[\underline{\theta}, \bar{\theta}]$.

36. Our benchmark model characterizes the optimal *fully inclusive* mechanism, meaning that all countries are willing to join, but in the online supplement we do include an example showing that the optimal compressed mechanism need not be fully inclusive and discuss the structure of noninclusive mechanisms with no investigations. Once we introduce investigations, the optimal mechanism is always fully inclusive.

37. Barrett 2003; Rosendorff 2005; Rosendorff and Milner 2001; Von Stein 2008.

38. In reality, countries may choose policies, such as taxes or subsidies, that indirectly regulate the aggregate amount of carbon emissions by private actors. For simplicity, we abstract away from these choices and act as if the country can simply choose the level of carbon emissions.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Based on these assumptions, this model then focuses on the possible strategies that could be employed by an IO to guide countries' compliance with global abatement goals. In particular, the IO is able to first collect information about countries' types by asking countries to voluntarily disclose this information through their announcements $\hat{\theta}_i$, and then it can recommend an emissions quota for each country. We denote this quota by $\tilde{c}(\hat{\theta}_i)$ to show that it may depend on type announcements. The IO in our model is a utilitarian planner that seeks to maximize the sum of all countries' payoffs.[39]

Each country's payoff is a function of how much carbon is emitted and its type, which can be written as

$$u_i(c, \theta_i) = \theta_i \log(c_i) + (1 - \theta_i) \log(\omega - C) \tag{1}$$

In this expression, $C = \sum_{i=1}^{n} c_i$ is the aggregate level of carbon emissions and $\omega > 0$ represents the carbon stock. If country i does not sign the agreement, it receives $u_i(\mathbf{c}, \theta_i)$ for its chosen level of emissions but pays a fixed penalty $K \geq 0$ for noncooperation. The first term in the utility function represents the countries' direct benefits from carbon consumption, while the second term represents the benefits from conservation. The type θ_i serves as a weight on these two components, so naturally countries value conservation less relative to consumption as θ_i increases.

This form captures the idea that high types' utility increases by more when they emit (because the utility function is written in a way such that emissions are valued relatively more and conservation relatively less when the type is high), while low types' utility increases by less when they emit (because the utility function is written so conservation increases in value and emissions decrease in value when the type is low). For a country that has not signed the agreement, it has this same utility function related to emissions and abatement but also pays a fixed penalty K for noncooperation. This fixed cost could be seen as the loss of any club benefits,[40] alliances, or simply prestige as a result of not being in the agreement. For the moment, we will treat K as exogenous. However, in the online supplement we consider an extension in which K arises endogenously from issue bundling. One thing we wish to make clear is that the IO need not have enforcement power in a formal sense (as these organizations often do not). Instead, fairly large values of K can be sustained by the IO just helping countries to coordinate on self-enforcing equilibria that punish climate defectors.

39. Our assumption of a utilitarian IO is primarily motivated by our research question. We want to know whether it is possible to implement a socially optimal agreement, so we take as given that the IO would want to implement such an agreement if it could. As a consequence, we assume away agency loss from delegating to the IO, even though this may be empirically relevant. For instance, if the IO sees its principal as primarily consisting of a small group of powerful countries (e.g., Copelovitch 2010), then the IO's actual objective may differ significantly from the one we use here. These considerations are interesting avenues for future research, but we put them aside to preserve clarity on the question of whether socially optimal agreements are possible.

40. Nordhaus 2015.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Overall, we analyze an institution in which each country makes an announcement to the IO and the IO recommends an emissions level. Although this seems restrictive if it is taken literally, the *revelation principle* implies that the outcome of any equilibrium from any institution can be implemented by a direct mechanism such as this one in which all players have an incentive to truthfully report their types. That is, suppose an outcome (i.e., emissions levels for each type) is implemented as an equilibrium to some potentially complicated mechanism. The revelation principle tells us that the same outcome can be implemented by a direct mechanism in which countries are asked to reveal their types and then are assigned emissions levels. Therefore, our setup encompasses a very broad set of institutions.[41]

The form of the utility functions and the choices of the IO in the benchmark model follow previous work by Harrison and Lagunoff.[42] The primary differences are that their model is dynamic in contrast to our static model, and that our model imposes a positive cost of noncooperation on countries that fall outside the agreement. These two choices are connected: the dynamic Harrison and Lagunoff model generates punishments for noncompliance endogenously, but our static model requires an exogenous punishment for noncooperation. Our setup enables simpler analysis and also lets us compare what happens when the IO can impose large costs for noncooperation versus when it cannot.[43]

Complete Information Benchmark

First, we solve for the quotas the IO would suggest assuming that there was perfect information about countries' types and that there was no need to deal with any perverse incentives held by the countries. This complete information benchmark has a socially optimal quota which is increasing in the global carbon stock and each country's type but decreasing in the number of countries, as suggested by the following lemma:

Lemma 1: Under complete information the socially optimal quota is $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$

Lemma 1 implies that those countries that value their emissions more are able to emit more under this quota (because the optimal quota is increasing in θ_i), but the

41. The fact that any implementable mechanism *can be* implemented by a direct mechanism does not imply that we should observe only direct mechanisms in the world. There may be political or practical reasons to use procedures that do not simply involve direct self-reports followed by quota recommendations. Our point is to analyze what can be achieved. The revelation principle suggests that analyzing truthful direct mechanisms is a useful way to do that because any equilibrium of any game with these preferences and types is outcome-equivalent to some truthful direct mechanism.

42. Harrison and Lagunoff 2017.

43. Also, in a dynamic model the IO may not benefit from imposing large costs for noncompliance because doing so reduces the ability of low types to credibly threaten to stop cooperating if other countries defect (Cirone and Urpelainen 2013). This suggests that a dynamic model with investigations is an interesting avenue for future research.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

quotas for all countries are also influenced by how many countries are a part of the agreement (because the extent of this allowance is decreasing in the total number of countries and increasing in the total available carbon stock).

Our complete information benchmark captures the idea that the optimal policy is variable in the sense of allowing for individual differences. In the next section, we introduce private information and reproduce the full compression result of Harrison and Lagunoff, which states that the best implementable quota is completely inflexible and thus requires the same actions of all countries.

Best Mechanism Under Constraints

To consider how an IO with some limited enforcement power but no investigative powers could attempt to implement an agreement, we examine the quotas the IO would suggest after imposing countries' incentive constraints. The first of these two kinds of constraints is the incentive compatibility constraint, which says that each country must find it in its best interest to reveal its true type given the quotas associated with each announced type:

$$\int_{\Theta_{-i}} u_i(\tilde{c}(\theta_i), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} u_i(\tilde{c}(\hat{\theta}), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \qquad (2)$$

for all $i \in N$ and $\hat{\theta} \in [0, 1]$, where $\tilde{\mathbf{c}}(\hat{\theta}_{-i})$ is the profile of all other emissions given that all other countries report their own types and comply with the recommended quota, and F_{-i} is the joint probability distribution of θ_{-i}. That is, given the quotas associated with each announced type, every type of every country weakly prefers to report its true type.

The second is the participation constraint, which states that each country must prefer joining the agreement under its recommended quota to leaving the agreement. This requires that the agreement must offer every type of country at least the utility it would get by not being in the agreement:

$$\int_{\Theta_{-i}} u_i(\tilde{c}(\theta_i), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} \max_{c_i \geq 0} u_i(c_i, \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) - K \qquad (3)$$

for all i and θ_i. That is, each country's payoff for going it alone is its payoff-maximizing emissions level minus the nonparticipation penalty K. The agreement must then offer every type of every country at least this payoff.

We solve for the best mechanism that satisfies the constraints (2) and (3). The incentive compatibility condition (2) represents the idea of the revelation principle: if an outcome is implementable at all, then it must be implementable by a direct mechanism in which all players are incentivized to reveal their true types. Thus, restricting ourselves to incentive-compatible mechanisms is natural in the sense that any achievable outcome can be implemented by one satisfying this constraint. The participation constraint (3) represents the idea that no country can be forced to sign an agreement. Thus, a feasible agreement must satisfy the property that every country would sign on to and comply with the agreement rather than going it alone. An agreement that nobody would join clearly cannot be effective. However, this constraint is limiting in a particular way: we look for the optimal mechanism among those that are fully

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

inclusive, that is, that *all* countries would sign. As we discuss in the next section and in the online supplement, without investigations it may not be the case that the optimal mechanism overall is fully inclusive.

The IO must maximize the sum of all these utilities based on type:

$$\Sigma_{i=1}^{n}[\theta_i \log(\check{c}(\theta_i)) + (1-\theta_i)\log(\omega - C)] \tag{4}$$

subject to (2) and (3). The full compression result is

Proposition 1: (Harrison and Lagunoff 2017) Assume $\underline{\theta} > \frac{\bar{\theta}}{n+\bar{\theta}-n\bar{\theta}}$. *The optimal mechanism under incomplete information in the baseline model is fully compressed: all countries must meet the same quota regardless of type announcements.*

Under the incentive compatibility constraint, the IO can assign all types of countries only the same quota. Such a result is referred to as "fully compressed," in that there is one quota for all countries despite the existence of different types. The reason this result is the same for every country is that all countries have the incentive to report a type higher than their true type to get a more lenient quota and therefore maximize their own utility. In other words, if quotas depend on countries' private information, then each country can indirectly free ride by reporting a type higher than its true type. This agreement is inefficient compared to the optimal agreement because it treats all countries as equal even though their costs of carbon abatement differ.[44] Notably, Proposition 1 makes use of the truth-telling constraint only. We have not yet analyzed how the participation constraint affects the optimal quota.

Clearly, a fully compressed quota meets the incentive compatibility constraint because no country can have a strict incentive to lie. To find the optimal quota, we can therefore identify the best compressed quota that satisfies all countries' participation constraints. This calculation boils down to finding the strictest quota that the highest type will be willing to accept. Given a compressed quota c^*, the highest type's utility if it opts out of the treaty is

$$\max_{c_H} \bar{\theta}\log(c_H) + (1-\bar{\theta})\log(\omega - c_H - (n-1)c^*) \tag{5}$$

Solving this maximization problem for c_H gives $c_H(c^*) = \bar{\theta}(\omega - c^*(n-1))$.

To get some intuition about the optimal compressed quota, it is useful to consider two extremes. The highest type incurs no cost from violating the agreement when K (the cost of not being in the agreement) is zero, and so the compressed quota c^* cannot be lower than the value at which the highest type receives equal utility from remaining in or leaving the agreement. This indifference condition is met only when the quota is set to the amount the highest type of country would choose

44. Though we model a situation in which countries are symmetric ex ante for simplicity, another issue with compressed agreements may be exacerbating inequality because policy and nonpolicy determinants of the costs of investment in clean technology are associated with national income. Baldwin et al. 2016.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

to emit were it not in the agreement:

$$c^* = \frac{\bar{\theta}\omega}{1 + \bar{\theta}(n-1)} \tag{6}$$

This is the lower bound on the optimal quota of allowed emissions. In the literature, this issue of agreements' commitment levels being dragged down by the least committed country is known as the law of the least ambitious program.[45] Essentially, in agreements that try to achieve broad membership but require the same quota of all countries, it is possible to assign every country only a "shallow" quota that is permissive enough for even the least committed country to meet (or else risk losing members).

If we assume that K is so large that the participation constraint no longer has an effect, then the compressed quota is simply the average quota under full information conditions:

$$\sum_{i=1}^{n} [\theta_i \log(c^*) + (1 - \theta_i) log(\omega - c^* n)] \tag{7}$$

Taking first-order conditions and solving for c^* (again using $T = \sum_{i=1}^{n} \theta_i$) yields $c^* = \frac{\left(\frac{T}{n}\right)\omega}{n}$. Notice that this is a particularly intuitive mechanism: it simply asks every country to follow the average quota under the full information solution. Note also that this is generally a lower (more demanding) quota than the one when K is zero. This illustrates the following result.

Proposition 2: The optimal quota is decreasing in K with $\lim_{K \to 0} c^* = \frac{\bar{\theta}\omega}{1+\bar{\theta}(n-1)}$ *and* $\lim_{K \to \infty} c^* = \frac{\left(\frac{T}{n}\right)\omega}{n}$.

Proposition 2 gives us some insights about how an organization must design an agreement when it has no independent investigative power. Recall that the parameter K is meant to represent the IO's enforcement leverage. For instance, if the IO controls access to a club good as Nordhaus suggested,[46] or if recognition by the IO is very valuable to the countries, then K may be large in the sense that the IO has leverage to make countries comply with an agreement. An insight from Proposition 2 is that this power is useful but not good enough to implement an optimal agreement. The reason is that this enforcement leverage enters countries' participation constraints but not their incentive compatibility constraints. In other words, an IO with a lot of leverage can induce countries to comply with the recommended quotas but cannot make them tell the truth about their types. As we show in the next section, introducing investigative powers helps the IO convert its leverage for compliance into leverage for honest reporting.

45. Hovi, Ward, and Grundig 2015; Underdal 1980.
46. Nordhaus 2015.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Noninclusive Mechanisms

Proposition 2 characterizes the best mechanism among those that are incentive compatible and fully inclusive. By fully inclusive, we mean that the participation constraint is satisfied for every country, so the agreement is a global one. However, the optimum among all possible mechanisms need not be fully inclusive. The central mechanism at play is the well-known broad versus deep trade-off from the literature on climate agreements. Since the optimal mechanism without investigations is fully compressed, IOs without investigative capacity can only be either broad (including many members) or deep (requiring strong commitments), given that all countries will be asked to meet the same quota despite inherent differences in their capabilities.

In our model, this trade-off is evident in the fact that the quota for every country is primarily a function of the participation constraint of the highest type. That is, all countries are asked to do only as much as the least environmentally friendly country is willing to do. Alternatively, a mechanism that is not fully inclusive lets signatories make deeper commitments but carries the disadvantage of leaving the countries with the highest carbon emissions completely unconstrained.

As we show by example in the online supplement, a noninclusive mechanism may be optimal in some cases. The structure of the optimal agreement is still similar to that in Proposition 2 in the sense that the mechanism should be fully compressed among those that sign the agreement to ensure incentive compatibility, but leaves nonsignatories (all countries above some type cutoff) free to choose their emissions levels outside of the agreement. This primarily depends on how much leverage the international community has over enforcement of an agreement (represented in our model by the parameter K). If K is near zero, meaning there is very little impetus to comply with any agreement, the participation constraints of the heaviest polluters become very salient, and it is likely that fully inclusive mechanisms will not be optimal. By contrast, as K becomes large, the optimal mechanism will be fully inclusive because all participation constraints are easily met. This also explains why the optimal mechanism in Harrison and Lagunoff is fully inclusive, because the dynamic model in that paper allows unbounded punishments, which closely corresponds to our model as K goes to infinity.[47]

Mechanisms with Limited Investigations

We now extend the benchmark model in the following way: the IO has a fixed amount R of investigative resources available and can distribute these resources however it chooses after seeing the type announcements of each country. Let $r_i(\hat{\theta}, \hat{\theta}_{-i}) \in [0, 1]$ denote the amount of investigative resources devoted to country

47. Harrison and Lagunoff 2017. See Finnemore and Jurkovich 2014 for an empirical discussion of inclusivity in international agreements.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

i when country i reports that its type is $\hat{\theta}$ and the other countries report $\hat{\theta}_i$.[48] We assume that i's type is verified with probability $r_i(\hat{\theta}, \hat{\theta}_{-i})$. The mechanism we consider is one where, if a country's verified type is different from its reported type, that country is removed from the agreement and incurs the cost K of noncompliance but is also free to choose its carbon emissions levels from then on.

The IO's budget constraint is that $\sum_{i=1}^{n} r_i(\hat{\theta}_i, \hat{\theta}_{-i}) = R$ for any $\hat{\theta} \in [\underline{\theta}, \bar{\theta}]^n$. Clearly if $R \geq n$, this corresponds to a complete-information case where the first-best quota is implementable as long as K is large enough. We are interested in the conditions under which the first-best quota is implementable for smaller values of R and in characterizing the optimal allocation of investigative resources in this case. Proposition 3 states our result.

Proposition 3: Assume $\underline{\theta} > \frac{\bar{\theta}}{n+\bar{\theta}-n\bar{\theta}}$. *For any* $R > 0$, *there exists* $K^*(R)$ *such that the optimal quota under complete information is implementable by a mechanism with limited investigations when* $K \geq K^*(R)$.

This result shows that the optimal quota may be implementable when the IO can conduct investigations. It also shows that enforcement leverage magnifies the effect of investigative resources. If K is large, then the optimal quota can be implemented with minimal investigative resources. More specifically, given *any* level of investigative resources, there exists some value of K such that the optimal rule can be implemented with that amount of resources. This illustrates the intuition given before: the effect of investigations is to convert leverage over compliance into leverage over honest reporting of types.

In Proposition 3, unlike the result in Proposition 2, the restriction to fully inclusive mechanisms is without loss of generality. For a variable agreement, the quota can be tailored to each individual country and type. An interpretation of this is that the ability to implement an agreement with variability solves the broad versus deep trade-off because we need not assign shallower commitments to lower types just to accommodate more difficult cases. In fact, Proposition 3 involves implementing a quota that is optimal among all possible quotas.

Note that, in contrast to many regulatory models with investigative capacity, our IO is using resources to directly investigate countries' costs of carbon abatement rather than to monitor emissions levels. So, like the model illustrating the broad versus deep trade-off in Barrett, our model assumes that emissions are perfectly observable.[49] Determining the right allocation of informational capacity between

48. Our inclusion of investigative resources places our model in line with regulation and crime deterrence models such as Becker 1968 and Mookherjee and Png 1994, where regulators must allocate investigative resources in addition to penalties. A key difference is that investigations reveal information directly about types rather than about violations themselves (which we assume are observable). Still, the case with large K but low investigative resources relates to the low-probability and high-penalty solutions in Becker 1968.

49. Barrett 1994.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

monitoring compliance and learning types may be an interesting area for future research in a version of this model in which emissions levels are not always observed. However, we expect that some investigative capacity must always be devoted to monitoring types. This is because, in contrast to the crime deterrence models such as that of Becker, the IO does not directly control the magnitude of the punishment.[50]

Discussion and Conclusion

Initially, our model allows us to see that the ideal agreement involves an optimal quota which is increasing in countries' true types, increasing in the total available carbon stock, and decreasing in the total number of countries. However, without investigations, the best incentive-compatible agreement is fully compressed, meaning all countries must meet the same quota regardless of type announcements. This is inefficient because those countries with high and low types are made to take the same actions, and thus those with low types are not contributing as much as they could, while those with high types are made to take on an unsustainable burden. The fully compressed agreement also gives rise to the familiar broad versus deep trade-off—because every country commits to the same quota, the behavior of all members of a fully inclusive agreement is determined by whether the least environmentalist country is willing to participate.

Only in the model with institutional oversight is variability possible. In this version of the model, we allow the institution limited resources to investigate countries' claims about their types. As a result, countries become increasingly likely to tell the truth about their types as the amount of the institution's investigative resources (related to the likelihood of being caught) or the cost of not belonging to the agreement (related to the punishment from being caught) increases. This "truth-telling" equilibrium possible under institutional oversight then allows an optimal assignment of emissions quotas, which ultimately enables the agreement to achieve both greater membership and greater compliance because countries are more willing to join and make deeper commitments in agreements with variability in commitments. Thus an optimal agreement supported by sufficient investigations and punishments can be both broad and deep.

Ultimately, these results suggest that an IO without the ability to investigate countries' claims can only inefficiently request the same quota of all countries, regardless of type. If the cost of not remaining in the agreement is very high, then this compressed quota is set at the average quota given all countries' types. However, if the cost of not belonging to the agreement is close to zero, then the quota must be set at the least committed country's willingness to abate or else risk losing members. By contrast, an IO with some ability to investigate countries' claims can incentivize

50. Becker 1968. This is related to the fact that, in Harrison and Lagunoff 2017, optimal mechanisms need not be fully compressed when transfers are allowed. However, we agree with those authors' assessment that an IO typically cannot freely choose transfers between countries.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

countries to reveal their true types in an effort to maximize global welfare by assigning appropriate quotas for each country, through increasing either its investigative resources or the costs of not being part of the agreement. Furthermore, if countries are able to punish defectors very effectively, as is the case in the dynamic setting of Harrison and Lagunoff, then the optimal agreement can be implemented with minimal investigative resources.

Our analysis focuses on simple mechanisms in which the IO elicits reports about countries' types and then suggests emissions quotas, but the revelation principle suggests that our analysis helps us understand what is possible in other, more familiar environments. In particular, a flexible interpretation of our model could motivate several examples from the literature of variability mechanisms paired with information collection by the international community. First, as Slechten pointed out, climate negotiations are often preceded by an umbrella convention which sets up institutions to gather countries' information related to climate change efforts.[51] In this environment, we can interpret the investigations in our model in a similar way to the certifications in her paper; one role of the IO in this stage of the negotiation may be to help verify the information submitted by the parties to reduce information asymmetries. Second, escape clauses in international agreements introduce a type of variability in line with that emphasized here by permitting parties to be subject to different requirements based on individual circumstances. These are often paired with procedures by which countries must submit verification of reasons for exceptions from usual rules. For instance, Article 8 of the London Dumping Convention permits exceptions to some articles in certain circumstances but requires contracting parties to consult with other parties and submit documentation of reasons for the exception.[52] Our model suggests that credible implementation of situation-specific variability in obligations may depend in part on these reporting procedures.

We leave several issues for future research. For example, one promising avenue for research might involve further probing the bargaining-over-institutions problem. As part of this effort, researchers could investigate under what conditions powerful states would form institutions (on the climate or other issues) with both investigative and enforcement abilities. Though the revelation principle suggests that *some* game should implement agreements that are outcome-equivalent to the direct mechanisms analyzed here, this could help political scientists predict whether such institutions are likely to be adopted given a particular existing set of bargaining protocols. A second promising avenue for future research may be to consider alternative objective functions for the IO. For instance, some IOs appear to primarily seek the approval of a small set of powerful states.[53] This may limit the ability to implement good agree-

51. Slechten 2020.
52. Marcoux 2009.
53. See Copelovitch 2010, which develops and tests this theory in the case of the International Monetary Fund.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

ments to the extent that it makes the IO reluctant to investigate the powerful countries. If the IO has sufficient difficulty committing to investigate powerful countries, then it may not be possible to implement agreements that account for the private information of those countries.

Supplementary Material

Supplementary material for this research note is available at <https://doi.org/10.1017/S0020818321000400>.

References

Baldwin, Elizabeth, Sanya Carley, Jennifer Brass, and Lauren MacLean. 2016. Global Renewable Electricity Policy: A Comparative Policy Analysis of Countries by Income Status. *Journal of Comparative Policy Analysis* 19 (3):277–98.

Barrett, Scott. 1994. Self-Enforcing International Environmental Agreements. *Oxford Economic Papers* 46:878–94.

Barrett, Scott. 2003. *Environment and Statecraft: The Strategy of Environmental Treaty-Making*. Oxford University Press.

Barrett, Scott, and Paul H. Nitze. 2007. *Why Cooperate? The Incentive to Supply Global Public Goods*. Oxford University Press.

Becker, Gary S. 1968. Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76 (2):169–217.

Bernauer, Thomas. 2013. Climate Change Politics. *Annual Review of Political Science* 16 (1):421–48.

Boockmann, Bernhard, and Paul W. Thurner. 2006. Flexibility Provisions in Multilateral Environmental Treaties. *International Environmental Agreements: Politics, Law and Economics* 6 (2):113–35.

Cirone, Alexandra E., and Johannes Urpelainen. 2013. Trade Sanctions in International Environmental Policy: Deterring or Encouraging Free Riding? *Conflict Management and Peace Science* 30 (4):309–34.

Clapp, Jennifer, and Peter Dauvergne. 2011. *Paths to a Green World: The Political Economy of the Global Environment*. 2nd ed. MIT Press.

Copelovitch, Mark S. 2010. Master or Servant? Common Agency and the Political Economy of IMF Lending. *International Studies Quarterly* 54 (1):49–77.

Csapó, Gergely, and Rudolf Müller. 2013. Optimal Mechanism Design for the Private Supply of a Public Good. *Games and Economic Behavior* 80:229–42.

Dai, Xinyuan. 2007. *International Institutions and National Policies*. Cambridge University Press.

Dannenberg, Astrid, Bodo Sturm, and Carsten Vogt. 2010. Do Equity Preferences Matter for Climate Negotiators? An Experimental Investigation. *Environmental and Resource Economics* 47 (1):91–109.

Finnemore, Martha, and Michelle Jurkovich. 2014. Getting a Seat at the Table: The Origins of Universal Participation and Modern Multilateral Conferences. *Global Governance: A Review of Multilateralism and International Organizations* 20 (3):361–73.

Gilligan, Michael J., and Leslie Johns. 2012. Formal Models of International Institutions. *Annual Review of Political Science* 15 (1):221–43.

Goldstein, Judith, and Lisa L. Martin. 2000. Legalization, Trade Liberalization, and Domestic Politics: A Cautionary Note. *International Organization* 54 (3):603–32.

Harrison, Rodrigo, and Roger Lagunoff. 2017. Dynamic Mechanism Design for a Global Commons. *International Economic Review* 58 (3):751–82.

Harstad, Bård. 2018. Pledge-and-Review Bargaining. CESifo Working Paper 7296, Center for Economic Studies, Munich.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Helfer, Laurence R. 2012. Flexibility in International Agreements. In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 175–96. Cambridge University Press.
Herzing, Mathias. 2005. Essays on Uncertainty and Escape in Trade Agreements. PhD diss., Stockholm University.
Hovi, Jon, Detlef F. Sprinz, Håkon Sælen, and Arild Underdal. 2019. The Club Approach: A Gateway to Effective Climate Cooperation? *British Journal of Political Science* 49 (3):1071–96.
Hovi, Jon, Detlef F. Sprinz, and Arild Underdal. 2009. Implementing Long-Term Climate Policy: Time Inconsistency, Domestic Politics, International Anarchy. *Global Environmental Politics* 9 (3):20–39.
Hovi, Jon, Hugh Ward, and Frank Grundig. 2015. Hope or Despair? Formal Models of Climate Cooperation. *Environmental and Resource Economics* 62 (4):665–88.
Karp, Larry, and Jinhua Zhao. 2010. International Environmental Agreements: Emissions Trade, Safety Valves and Escape Clauses. *Revue économique* 61 (1):153–82.
Kennard, Amanda. 2021. My Brother's Keeper: Other-Regarding Preferences and Concern for Global Climate Change. *Review of International Organizations* 16:345–76.
Keohane, Robert O., and David G. Victor. 2011. The Regime Complex for Climate Change. *Perspectives on Politics* 9 (1):7–23.
Keohane, Robert O., and David G. Victor. 2016. Cooperation and Discord in Global Climate Policy. *Nature Climate Change* 6 (6):570–75.
Koremenos, Barbara. 2001. Loosening the Ties that Bind: A Learning Model of Agreement Flexibility. *International Organization* 55 (2):289–325.
Koremenos, Barbara. 2005. Contracting Around International Uncertainty. *American Political Science Review* 99 (4):549–65.
Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001. The Rational Design of International Institutions. *International Organization* 55 (4):761–99.
Kucik, Jeffrey, and Eric Reinhardt. 2008. Does Flexibility Promote Cooperation? An Application to the Global Trade Regime. *International Organization* 62 (3):477–505.
Marcoux, Christopher. 2009. Institutional Flexibility in the Design of Multilateral Environmental Agreements. *Conflict Management and Peace Science* 26 (2):209–28.
Martin, Lisa L. 1992. Interests, Power, and Multilateralism. *International Organization* 46 (4):765–92.
McGinty, Matthew. 2007. International Environmental Agreements Among Asymmetric Nations. *Oxford Economic Papers* 59 (1):45–62.
Michaelowa, Katharina, and Axel Michaelowa. 2017. Transnational Climate Governance Initiatives: Designed for Effective Climate Change Mitigation? *International Interactions* 43 (1):129–55.
Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs. *Economics and Politics* 2 (1):1–23.
Mitchell, Ronald B., and Patricia M. Keilbach. 2001. Situation Structure and Institutional Design: Reciprocity, Coercion, and Exchange. *International Organization* 55 (4):891–917.
Mookherjee, Dilip, and Ivan P.L. Png. 1994. Marginal Deterrence in Enforcement of Law. *Journal of Political Economy* 102 (5):1039–66.
Nordhaus, William. 2015. Climate Clubs: Overcoming Free-Riding in International Climate Policy. *American Economic Review* 105 (4):1339–70.
Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
Pelc, Krzysztof J. 2009. Seeking Escape: The Use of Escape Clauses in International Trade Agreements. *International Studies Quarterly* 53 (2):349–68.
Rosendorff, B. Peter. 2005. Stability and Rigidity: Politics and Design of the WTO's Dispute Settlement Procedure. *American Political Science Review* 99 (3):389–400.
Rosendorff, B. Peter, and Helen V. Milner. 2001. The Optimal Design of International Trade Institutions: Uncertainty and Escape. *International Organization* 55 (4):829–57.
Slechten, Aurélie. 2020. Environmental Agreements Under Asymmetric Information. *Journal of the Association of Environmental and Resource Economists* 7 (3):455–81.

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge University Press.
Sykes, Alan O. 1991. Protectionism As a "Safeguard": A Positive Analysis of the GATT "Escape Clause" with Normative Speculations. *University of Chicago Law Review* 58 (1):255–305.
Tavoni, Alessandro, Astrid Dannenberg, Giorgos Kallis, and Andreas Löschel. 2011. Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game. *Proceedings of the National Academy of Sciences* 108 (29):11825–29.
Underdal, Arild. 1980. *The Politics of International Fisheries Management: The Case of the Northeast Atlantic*. Universitetsforlaget.
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.

Authors

Jordan H. McAllister is a PhD candidate in Political Science at Washington University in St. Louis. She can be reached at jordan.mcallister@wustl.edu.
Keith E. Schnakenberg is Associate Professor of Political Science at Washington University in St. Louis. He can be reached at keschnak@wustl.edu.

Acknowledgments

We thank Amanda Kennard and David Victor as well as participants in the Washington University formal theory seminar for valuable comments.

Key Words

Climate change; international institutions; game theory; mechanism design

Date received: June 16, 2020; Date accepted: May 18, 2021

https://doi.org/10.1017/S0020818321000400 Published online by Cambridge University Press