

A proposal for the reform of the Kyoto Protocol: escape clauses under cost uncertainty

Larry Karp

Jinhua Zhao

University of California, Berkeley

Iowa State University, Ames

March 22, 2006

Abstract

A reform to the Kyoto Protocol that allows signatories to avoid achieving the target level of abatement upon payment of a fine, would achieve two goals. First, it would defuse one U.S. objection to the agreement: the concern that the cost of achieving the target might turn out to be extremely high. Second, unlike other cost-reducing measures (such as trade in pollution permits) it would increase the equilibrium number of signatories in a non-cooperative participation game. We study the participation game under an escape clause using both a Nash Equilibrium and the concept of a Stable Set when nations are “Farsighted”.

Keywords: Kyoto Protocol, escape clause, cost uncertainty, participation game, International Environmental Agreement, stable set with farsight, coalitional games.

JEL classification numbers

Preliminary and Incomplete Draft

1 Introduction

The control of greenhouse gasses (GHGs) requires international cooperation. The U.S. withdrawal from the Kyoto Protocol (“Kyoto”), and the questionable compliance of signatories, may render the agreement ineffective. The agreement may be weak because of design flaws and/or because of the inherent difficulty of inducing sovereign nations to provide a public good. The U.S. objected to (amongst other things) the Protocol’s imposition of an aggregate emissions ceiling, expressing the concern that the economic cost of achieving this target might be very large. If signatories discover that abatement costs are larger than anticipated, their compliance may erode. Since the treaty extends only to 2012, it is worth understanding how its successor should be designed. A reform to the Kyoto Protocol that allows signatories to avoid achieving the target level of abatement upon payment of a fine, would achieve two goals. First, it would defuse one U.S. objection to the agreement: the concern that the cost of achieving the target might turn out to be extremely high. Second, unlike other cost-reducing measures (such as trade in permits) it would increase the equilibrium number of signatories in a non-cooperative participation game. We study the participation game under an escape clause using both a Nash Equilibrium and the concept of a Stable Set when nations are “Farsighted”.

Although there is much more uncertainty about the benefit of GHG abatement than about the cost of abatement, arguably cost uncertainty is more important to the design of an International Environmental Agreement (IEA). Kyoto will be in force only four years; the duration of its successor is also likely to be fairly short. During this period, we will learn the costs of a particular level of abatement. Barring a catastrophic event, our information about the benefit of this abatement will probably change only slightly during this period. Therefore, for the purpose of designing a short-term IEA, it makes sense to treat abatement costs as a random variable whose value will be realized during the lifetime of the IEA, and the abatement benefit as a random variable that will be realized in the distant future.

Many papers (including Carraro and Siniscalco (1993), Barrett (1994), Bloch (1997), and Dixit and Olson (2000)) and several books (including Finus (2001), Batabyal (2000), and Barrett (2003)) study the formation of an IEA using the non-cooperative Nash Equilibrium (NE) to a participation game. Details vary across the models, but the basic structure of this “standard model” is that in the first stage (the participation game) homogeneous countries decide whether to join an IEA, and in a second stage (the abatement game) they choose the abatement level. The critical assumptions are that (i) the IEA does not prescribe specific actions, and (ii) the act of signing the IEA solves the collective action problem amongst members. Therefore, in the abatement stage members act to maximize IEA joint welfare. In this model, the IEA (typically) contains two or more members. If, contrary to the assumptions of this model, sig-

natories committed to a particular level of abatement, the IEA size in the NE would (typically) be 0. Commitment to a specific level of abatement increases the temptation to free-ride (by not signing the IEA), thus decreasing the incentive to join such an agreement.

An important result from this model is that the equilibrium size of the IEA tends to be small in circumstances when the potential benefits from cooperation are large. Holding fixed the benefit from abatement, the potential benefit from cooperation is large when abatement costs are small. Other things equal, an *increase* in the abatement cost increases the equilibrium size of the IEA, simultaneously decreasing the benefit from cooperation. The “membership cost” is the net cost to a nation of joining the IEA: its cost of abatement minus the benefit it obtains from the additional abatement. The standard model implies that a higher membership cost increases the equilibrium membership size. Although this result is somewhat counter-intuitive, its explanation (discussed below) is straightforward, given the assumptions of the game.

This result implies that efforts to reduce the costs of IEA compliance might backfire, by reducing equilibrium membership. There has been substantial interest in reforms to make Kyoto more efficient, thereby reducing membership costs. After initial resistance, Kyoto signatories accepted tradeable permits and “joint implementation”, and are discussing the use of “clean development mechanisms”.¹ These reforms reduce the membership cost. Another possibility is a hybrid policy that uses tradeable permits with a price ceiling; a regulator has the power to increase the allocation of permits to defend the price ceiling ((Pizer 2002), (Kopp, Morgenstern, Pizer, and Gherzi 2002), (Victor 2003)). The current carbon reduction agreement amongst northeastern U.S. states uses such a policy. This policy caps abatement costs and therefore reduces expected costs. If applied to Kyoto, this kind of policy would have removed, or at least weakened one of the U.S. objections to the agreement. The irony is that if the standard model of IEAs is a reliable description, the hybrid policy would have *decreased* nations’ incentive to join Kyoto.

The standard model implies that the central impediment to a successful IEA is the difficulty of inducing nations to forgo the temptation to free-ride, rather than design flaws such as the failure to deal efficiently with the possibility of unexpectedly high abatement costs. Design changes that increase efficiency might even be counterproductive. A corollary is that a successful IEA requires some kind of external punishment; Barrett (2003), chapter 15, makes this case persuasively. Some have proposed reforming the World Trade Organization to permit the use of trade sanctions against countries that do not abide by a climate change IEA [cites here].

¹Under joint implementation, a signatory obtains credit for abatement by investing in carbon abatement or sequestration activities in another member country. Under the proposed clean development mechanism, a signatory obtains abatement credit by investing in abatement or sequestration activities in a developing non-signatory country.

There appears to be little chance that the environmental tail will wag the trade dog, even if such an outcome were desirable. If an effective IEA really does require an external threat, the prospect of dealing successfully with the problem of climate change seems poor.

In our view, the pessimistic conclusion that effective IEAs require an external threat, and that they are unlikely to benefit from design changes, is exaggerated. That conclusion is a consequence of the assumption that the IEA stipulates only that it will maximize members' welfare. That assumption does not describe how IEAs actually behave. In one sense this assumption exaggerates IEAs' power, because clearly these are not capable of solving the collective action problem (although they might ameliorate it). In another sense, the assumption understates IEAs' power, because these are capable of prescribing contingent actions.

We modify the standard description of an IEA by replacing the assumption that signatories commit (only) to maximizing members' joint welfare, with the assumption that they can sign a simple contract. In our setting, nations face uncertain costs of abatement. The IEA consists of a contingent contract that contains an escape clause. This contract consists of two parameters, a prescribed level of abatement and a cost of exercising the escape clause (a "fine") that exempts the signatory from the requirement to abate. Receipts from the escape clause payments are returned equally to all signatories, except for a positive (possibly very small) amount that is lost as a transactions cost. In the first stage nations decide whether to join the IEA (the participation game). In the second stage (the abatement game) each signatory learns its abatement cost, which is private information, and decides whether to abate or to invoke the escape clause, taking as given other signatories' decision rules. Non-signatories have a dominant strategy. The Nash equilibrium to the abatement game depends on the outcome of the participation game (the number of signatories) and on the two parameters of the IEA (the prescribed level of abatement and the size of the fine).

Kyoto does have a prescribed level of abatement – a feature that the U.S. criticized – and in that respect it does not conform to the standard description of an IEA. Our proposal differs from Kyoto by including the escape clause. This modification has two desirable effects. First, it protects signatories from unexpectedly high abatement costs, and thus answers one of the U.S. objections. Second, it overcomes the "perverse" feature of the standard IEA, in which lowering the cost of membership also lowers the equilibrium membership size. In our proposal, a lower fine corresponds to a lower cost of membership, and it also leads to an increase in equilibrium membership size in most cases. For some levels of the fine, all countries sign the IEA.

In order to be able to study the effect of an escape clause in a simple setting, we ignore trade in emissions permits, an important feature of Kyoto. Trade in permits equalizes marginal abatement costs across countries, but *total* abatement costs still differ, so even with trade there is a role for the escape clause. Tradeable permits (with or without a price ceiling) "merely"

reduce expected membership costs. When IEA members behave non-cooperatively at the abatement stage, the escape clause can both reduce membership cost and increase equilibrium membership size.

It would be unreasonable to think that so simple a design change could solve the free-rider problem. We interpret the result as showing that a design change can substantially ameliorate the free-rider problem. This may appear to be a fairly non-controversial claim, but it is contrary to the IEA literature discussed above. That literature implies that design changes that reduce costs are futile, or even counterproductive. This conclusion emerges from a model that has become so widely used that it has taken on an air of inevitability. However, that standard model of an IEA is neither plausible nor descriptive.

The standard model uses a NE to a one-shot participation game. Several recent papers on IEAs (including Diamantoudi and Sartzetakis (2002) ?, and ?) are based on a more sophisticated interpretation of rationality, in which nations understand how their provisional decision to join a coalition would affect other nations' participation decisions; nations are farsighted. These papers are based on the theoretical developments in the study of games of coalition formation; important papers in this area include Chwe (1994), Mariotti (1997), Xue (1998) and Ray and Vohra (2001). We discuss the role of foresight in the model with an escape clause.

Section 2 sets out our model. Section 3 analyzes the one-shot NE, and Section 4 studies the non-cooperative participation game when nations are farsighted.

2 The Model

Each of N homogenous nations decides whether to join an IEA to reduce a global pollution. When nations make this decision the terms of the IEA are taken as given. The IEA specifies a target level of abatement, normalized to 1, and it contains an escape clause that allows a signatory not to abate if it pays a fine, F . Abatement is a global public good, with constant marginal expected benefit, $b > 0$. If m countries abate, all countries receive the expected benefit bm .

In the case GHGs and a short-lived IEA, it is reasonable to treat b as a constant. Potential environmental damages are caused by the stock, not the flow of GHGs. During a short period of time (less than a decade), the change in the stock of GHGs is small, relative to the stock. Provided that expected marginal benefit is a continuous function of the stock, a change in the number of countries that abate has a negligible effect on the marginal benefit of abatement. Hereafter we choose units of the value of abatement to normalize by setting $b = 1$.

When nations decide whether to join the IEA, they do not know their true abatement cost. At this stage nations are identical; they all face the same probability distribution for costs.

Nation i knows that its abatement cost θ_i , is a random variable drawn from $\Theta = \{\theta_L, \theta_H\}$, with $\theta_H > \theta_L$ and $p \in (0, 1)$ equal to the probability that $\theta_i = \theta_H$. The distributions of θ_i , $i = 1, \dots, N$, are independent.²

The IEA game has three stages. The fine F and the level of abatement (normalized to 1) are determined in stage 0. We do not model this choice, although we consider its welfare consequences. (The abatement level determines Θ and p .) The parameters F, Θ, p are common knowledge. In stage one, nations play a *participation game* in which they decide whether to join the IEA. We study two types of equilibria to the participation game, the NE (Section 3) and an equilibrium based on farsightedness (Section 4). The outcome of this game is a partition between signatories and non-signatories. Nations understand how their participation decision affects the final outcome in stage two. In stage two, each nation observes its own abatement cost θ , it knows whether it is a signatory, and it knows the total number of signatories. Based on this information, nations play a non-cooperative *abatement game*, each deciding whether to abate.

If M nations sign the IEA in the first stage and $M - m \geq 0$ of them invoke the escape clause in the second stage, revenue from the fine is $(M - m)F$. This revenue is shared equally among the M signatories (perhaps to provide a club good), except for a fraction $0 < 1 - \phi < 1$ that is lost to transactions costs. Each of the signatories receives a transfer of $\frac{M-m}{M}\phi F$.

2.1 Stage two: the abatement game

Without loss of generality, we define the level of abatement (normalized to 1) as the excess over the individually rational level of abatement in the low cost state. This definition implies that

$$\theta_L > 1. \quad (1)$$

This inequality implies that non-signatories' dominant strategy is not to abate.

A signatory must abide by the terms of the IEA. A signatory's abatement decision depends on its cost realization $\theta \in \Theta$ and the number of signatories $M \in \mathcal{N} \equiv \{0, 1, \dots, N\}$. The signatory's action set is $\{0, 1\}$ where 1 stands for abating and 0 stands for not abating. Its strategy is a mapping from $\Theta \times \mathcal{N}$ to $\{0, 1\}$. We consider only *symmetric pure strategy* Nash equilibria, hereafter referred to as simply NE.

There are three types of NE in the abatement game. In a type 0 NE, each signatory's strategy is not to abate for any cost realization; in a type 1 NE, each signatory's strategy is to abate only if its own cost is θ_L ; and in a type 2 NE, each signatory's strategy is to abate for either cost realization. A nation that is indifferent between two actions breaks the tie by abating.

²In future work we will consider correlated costs.

When a signatory decides on its strategy, it needs to assess the cost realizations (and thus the actions) of the other signatories. When there are M signatories, the probability that m of the $M - 1$ other signatories have cost θ_L is given by the binomial formula

$$p_{m,M-1} \equiv \frac{(M-1)!}{m!(M-1-m)!} (1-p)^m p^{M-1-m}. \quad (2)$$

2.2 Types of NE

We identify the combinations of F and M that support each of the three types of NE, beginning with the type 0 NE. Suppose the IEA has M members and consider a particular signatory i with cost θ_L . If it chooses not to abate, it pays the fine F . Since in a type 0 NE none of the other signatories abates, the total fine received is MF and each signatory receives a “rebate” of ϕF . Therefore, i ’s net benefit of not abating is $u_{na,0} = -(1-\phi)F$, where subscript na denotes “not abate” and 0 denotes a type 0 NE.

If i chooses to abate in a type 0 IEA, it incurs abatement cost θ_L and receives the benefit of $1 + \frac{M-1}{M}\phi F$, consisting of the benefit of pollution reduction and the rebate of other signatories’ fines. Thus, i ’s net benefit of abating is $u_{a,0}(\theta_L) = 1 + \frac{M-1}{M}\phi F - \theta_L$.

Nation i will not abate given cost θ_L if and only if $u_{na,0} > u_{a,0}(\theta_L)$, or $F < F_1(M)$, where

$$F_1(M) \equiv \frac{\theta_L - 1}{1 - \phi/M}. \quad (3)$$

Consider next the type 1 NE, in which the equilibrium action is to abate only in a low cost state. If signatory i has cost θ_L , and believes that each of the other $M - 1$ signatories abates only if its cost is θ_L , i ’s expected payoff if it abates is

$$u_{a,1}(\theta_L) = \sum_{m=0}^{M-1} p_{m,M-1} \left\{ m + 1 - \theta_L + \frac{M-1-m}{M}\phi F \right\}. \quad (4)$$

The term in the brackets measures the benefit of abating when m of the other $M - 1$ signatories are in state θ_L and thus choose to abate. Signatory i ’s expected benefit of not abating is

$$u_{na,1} = \sum_{m=0}^{M-1} p_{m,M-1} \left\{ m - F + \frac{M-m}{M}\phi F \right\}. \quad (5)$$

Signatory i will abate given cost θ_L if and only if $u_{a,1}(\theta_L) \geq u_{na,1}$, which is equivalent to $F \geq F_1(M)$. Similarly, we can show that in a type 1 equilibrium i will not abate given cost θ_H if and only if $F < F_2(M)$, where

$$F_2(M) \equiv \frac{\theta_H - 1}{1 - \phi/M}. \quad (6)$$

Since $F_2(M) > F_1(M)$, we know the abatement game has a unique type 0 NE if and only if $F < F_1(M)$. Similarly, the abatement game has a unique type 1 NE when $F_1(M) \leq F < F_2(M)$.

Consider now the type 2 NE. A signatory has the greatest incentive to not abate when its cost is θ_H . If the penalty F is so high that a signatory chooses to abate even given θ_H , it will also abate given cost θ_L . Suppose signatory i 's cost is θ_H . Given that all other signatories abate in both states, i 's expected payoff if it abates is $u_{a,2}(\theta_H) = M - \theta_H$. If it chooses not to abate, its payoff is $u_{na,2} = (M - 1) - F(1 - \phi/M)$. It abates at cost θ_H if and only if $u_{a,2}(\theta_H) \geq u_{na,2}$, or, after simplification, $F \geq F_2(M)$. This inequality is also the condition for the existence of the type 2 NE. In summary, we have

Proposition 1 *The abatement game in stage two has a unique type 0 NE if and only if $F < F_1(M)$, a unique type 1 NE if and only if $F_1(M) \leq F < F_2(M)$, and a unique type 2 NE if and only if $F \geq F_2(M)$.*

Figure 1 graphs $F_1(M)$ and $F_2(M)$, which are decreasing and approach $\theta_L - 1$ and $\theta_H - 1$ respectively. The figure shows the regions of the different types of NE, conditional on M, F . For brevity, we will sometimes refer to a “type of IEA” to mean an IEA that results in a certain type of NE at the abatement stage. The fact that the graphs of $F_1(M)$ and $F_2(M)$ are decreasing means that for a given level of F , an IEA member's incentive to abate (weakly) increases with the number of members. The reason for this relation is that each member's share of the revenue from the fine is $\frac{1}{M}$, so the net fine (after the rebate), $\frac{M-\phi}{M}F$, increases with the number of members. It is more expensive to exercise the escape clause in an IEA with more members. If there were no rebate ($\phi = 0$) membership costs would be independent of M .

2.3 Payoffs in NE

Let $\pi_{s,i}(M; F)$ and $\pi_{n,i}(M)$, $i = 0, 1, 2$, be, respectively, the expected payoffs of a signatory and a non-signatory in a type i IEA with M members and fine F . (s denotes “signatory” and n denotes “non-signatory”.) Non-signatories never abate.

Since no signatories abate in a type 0 NE, $\pi_{s,0} = u_{na,0} = -(1 - \phi)F$ and $\pi_{n,0} = 0 = \pi_{s,0} + G_0$, where $G_0 = (1 - \phi)F$. In a type 1 NE, a signatory abates only if its cost is θ_L , so its expected payoff is

$$\pi_{s,1}(M; F) = pu_{na,1} + (1 - p)u_{a,1}(\theta_L) = M(1 - p) - G_1(F) \quad (7)$$

where $u_{a,1}$ and $u_{na,1}$ are given in (4) and (5), and

$$G_1(F) = (1 - p)\theta_L + Fp(1 - \phi). \quad (8)$$

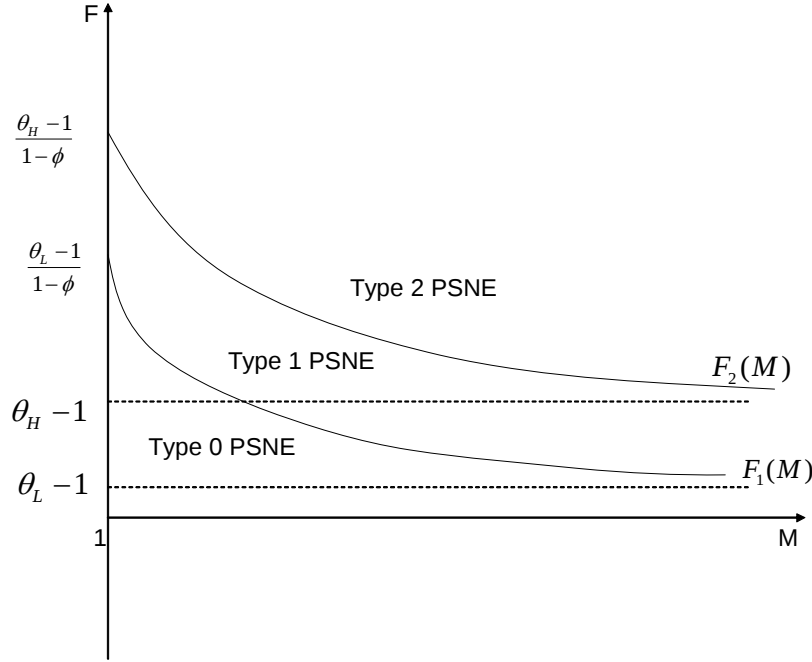


Figure 1: The Three Types of Nash Equilibria

(In a type 1 equilibrium, the expected fraction of nations that abate is $(1 - p)$; a signatory's expected abatement cost is $(1 - p)\theta_L$ and its expected fine payment, net of reimbursements is $Fp(1 - \phi)$.) The expected payoff of a non-signatory in a type 1 NE is

$$\pi_{n,1}(M) = \sum_{m=0}^M p_{m,M} m = M(1 - p) = \pi_{s,1}(M; F) + G_1(F, \phi). \quad (9)$$

For a type 2 IEA, the associated payoffs of signatories and non-signatories are

$$\pi_{s,2}(M) = M - \bar{\theta}, \quad (10)$$

and

$$\pi_{n,2}(M) = M = \pi_{s,2}(M) + G_2, \quad (11)$$

where $\bar{\theta} = p\theta_H + (1 - p)\theta_L$ is the expected value of θ , and $G_2 = \bar{\theta}$.

In all three types of NE, a non-signatory has a higher payoff than a signatory, and the additional benefit (or the gain) of the non-signatory, G_i , is independent of the IEA's size, M . The membership cost of the IEA is the reduction in expected value for a non-signatory that decides to join the IEA, under the assumption that other nations' actions (as distinct from their strategies) remain fixed. Since the increased benefit from abatement is the constant 1, the membership cost is $G_i - 1$. All nations enjoy the global public benefit of abatement, but the

members incur a cost of G_i . We have the following relation

$$(1 - \phi) F < \theta_L \implies G_0 < G_1 < G_2. \quad (12)$$

Provided that the fine and/or the transactions cost are sufficiently small, the membership cost is higher in IEAs that produce a higher level of abatement. A smaller fine or smaller transactions cost decrease membership costs in a type 1 equilibrium.

The payoff functions used above are indexed by i , the type of equilibrium, which is determined by M, F . We use $\pi_s(M, F)$ and $\pi_n(M, F)$ to denote the equilibrium payoff of a signatory and non-signatory, recognizing the endogeneity of the equilibrium type.

2.4 Relation to the “standard model”

The IEA in our model has less power (relative to the standard model), because it is not able to solve the collective action problem (i.e., to enable members to guarantee to maximize joint welfare). However, nations are able to commit to following a simple contingent contract, to either abate or to pay to exercise an escape clause. Thus, nations in this model have a different type of commitment ability, but it is not obviously greater or weaker relative to the standard model.

In a type 1 equilibrium, a nation that has a low abatement cost (and therefore chooses to abate) has a higher payoff than a high-cost nation (which decides to invoke the escape clause). The low-cost nation has an incentive to pressure non-abating signatories to abide by the IEA and pay the fine.

In a type 0 equilibrium, all signatories have negative payoff, $(1 - \phi) F$. We could modify the game by allowing the IEA to costlessly disband if all signatories agree to do so. That change causes $u_{na,0} = 0$, and the value of $F_1(M)$ in equation (3) is replaced by $\frac{\theta_L - b}{\phi - \phi/M}$. The change requires slightly more complicated notation, because it introduces an additional action (voting) at the abatement stage. The assumption that disbanding is costless is probably not reasonable. If disbanding is an option that costs more than $(1 - \phi) F$ (as we hereafter assume) the IEA would never disband.

3 Nash equilibrium of participation game

Here we describe the NE to the participation game when the size of the fine is taken as given in the participation game. We show that for an interval of F , a reduction in membership costs (smaller F) increases equilibrium membership size. We then compare the equilibrium membership size and global welfare in our setting with two versions of the standard model,

where a reduction in membership costs decreases equilibrium membership size. In both our setting and in the standard model, the equilibrium is subgame perfect. However, the second stage game is different in the two models. In the standard model the IEA members behave as a coalition and play a non-cooperative game with non-signatories. In our model all agents play a non-cooperative game, but IEA members are subject to different rules.

3.1 The participation game with an escape clause

Here we consider the NE to the participation game when the IEA uses the escape clause and chooses the fine *before* nations decide whether to join. Figure 2 reproduces Figure 1 and adds three curves; recall that F_1 and F_2 are the lower and upper boundaries, respectively, of the region where there is a type 1 equilibrium in the abatement game. (Table 1 in Appendix A collects the definitions of various curves and functions.) $\tilde{F}(M)$ is the locus of points at which $\pi_{s,1}(M; F) = 0$ (signatories in a type 1 equilibrium have 0 payoff); payoffs are negative to the left of that line. $F_0(M)$ is the locus of points at which $\pi_{s,1}(M; F) = \pi_{s,2}(M)$ (signatories' payoffs are equal in a type 1 and a type 2 equilibrium); below that line, signatories' payoffs are higher in a type 1 equilibrium. These functions are

$$\tilde{F} \equiv (M - \theta_L) \frac{p - 1}{p(\phi - 1)} \quad F_0 \equiv \frac{\theta_H - M}{1 - \phi}.$$

The inverse function of $F = F_i(M)$ is $M = F_i^{-1}(F)$.

In addition to N (the number of nations), this model contains four parameters: the cost parameters θ_i , the probability of a high cost, p , and the transactions cost parameter ϕ . Even with this simple model, there are many possible configurations of the various curves used to determine the equilibrium. A complete taxonomy would be uninteresting, so we consider a single case that holds for reasonable parameter values. In the text we use Figure 2 to discuss informally the relation between the fine, F , and the set of NE to the participation game. Appendix A provides a formal description of this correspondence, and the proof.

The equilibrium number of members must be an integer. Define $h(x)$ to be the smallest integer not less than x . The formal presentation in Appendix A respects the integer constraint, but here, in order to keep syntax manageable, we ignore this constraint. For example, if we state that a point x on the curve F_1 is a NE, we mean that for the value of F given by the vertical coordinate x , $h(F_1^{-1}(F))$ (i.e., the smallest integer greater than or equal to $F_1^{-1}(F)$) is a NE. With this understanding, the heavily shaded curves in Figure 2 show the NE correspondence (for all positive values of F). We discuss this figure, and then discuss the NE. We use subscripts to denote the coordinate of a point; e.g., F_b is the vertical coordinate of point b and M_b is the horizontal coordinate.

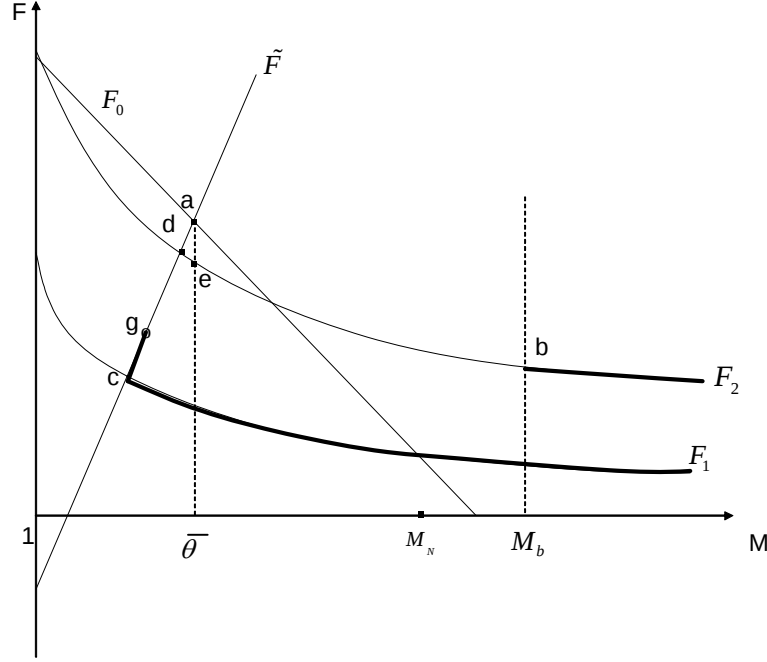


Figure 2: The Nash Equilibria in the Participation Game

Define $M_b \equiv \frac{\bar{\theta}-1+p}{p}$; denote point b as the intersection between F_2 and the vertical line at M_b .³ At point b , a signatory is indifferent between remaining in the IEA that results in a type 2 equilibrium, or leaving the IEA and being a non-signatory in a type 1 equilibrium. If the size of the IEA (on or above F_2) is less than M_b , a member would want to leave the IEA when this defection causes the abatement stage equilibrium to switch from type 2 to type 1. A simple calculation shows that point b always lies above F_0 as shown in Figure 2.

Point g (defined in the Appendix) is the highest point on the line $\tilde{F}$ for which there is a NE in the participation stage that leads to a type 1 equilibrium at the abatement stage. Define M_N as the value that satisfies $F_1(M) = F_0(M)$. Appendix A shows that $M_N < M_b$, as Figure 2 illustrates.

Figure 2 embodies several assumptions. The most important of these are that horizontal distance between curves F_2 and F_1 is greater than 2 at point ‘ e ’ and that $M_b - \bar{\theta} > 2$. The appendix contains necessary and sufficient conditions for these assumptions. Sufficient conditions are that θ_L is “moderately large” (e.g. at least 4), that there is a non-negligible difference between high and low costs, and that the probability of high abatement costs is moderate or

³[A later draft will fix a small error to Figure 2 and the analysis. Define the value $F_{b'} = \sup \{F \mid h(F_2^{-1}(F)) = h(F_2^{-1}(F_b))\}$ and define point b' as a point on F_2 with vertical coordinate $F_{b'}$. The heavy line on the curve F_2 (and all of the comments regarding that region) should be extended to point b' .]

small (e.g., less than 0.5). Given our normalization, $\theta_L = 4$ means that even if nations were certain that costs are low, an IEA would have to contain at least four members for the IEA to improve members' welfare.

These assumptions imply that if, at the participation stage, a signatory considers defecting from a candidate NE on F_2 below point e (i.e., for an IEA larger than $\bar{\theta}$), then the signatory knows that the resulting outcome at the abatement stage will be a type 1 NE rather than a type 0 NE. They also insure that for any candidate NE on $\tilde{F}$ consisting of fewer than $\bar{\theta}$ members, no non-signatory wants to defect by joining the IEA. (If it were to join the IEA, the membership would still be lower than M_b , the critical value below which a nation prefers being a non-signatory to an IEA that results in a type 1 NE in the abatement stage, rather than a signatory to an IEA that results in a type 2 equilibrium.)

Three facts about the equilibrium set are obvious. First, a membership size of 0 (not shown in Figure 2) is always a NE. Second, outcomes with negative payoffs for IEA members cannot be NE to the participation game. This fact means that outcomes to left of the upper envelope of curves F_1 and $\tilde{F}$, and below F_2 , and outcomes above F_2 with $M < \bar{\theta}$ are not NE. Third, in view of inequality (1), a member is deterred from leaving the IEA only if its defection would change the equilibrium at the abatement stage, e.g. from a type 2 to a type 1, or from a type 1 to a type 0. This fact means that only the “nearest integers” on or to the right of curves F_1 , F_2 , and $\tilde{F}$ are candidate NE.

At points on F_1 below point c , signatories have a positive payoff. These points are NE in the participation stage, leading to a type 1 equilibrium in the abatement stage. A signatory knows that its defection would lead to a type 0 equilibrium in the abatement stage, and a 0 payoff for non-signatories. Non-signatories do not want to join the IEA because the result would still be a type 1 equilibrium. (This and other conclusions follow from our parametric assumptions.) Other NE leading to type 1 abatement-stage equilibria consist of points between c and g on $\tilde{F}$. Here, a member's payoff is non-negative. Defection by a signatory leads to a type 0 abatement-stage equilibrium (and a 0 payoff for non-signatories). Defection by a non-signatory would not change the abatement-stage equilibrium from a type 1, leaving the defector with a lower payoff.

Points on F_2 below point b are NE that lead to type 2 equilibria in the abatement stage. At these points, members do not want to leave the IEA, since that would result in a type 1 equilibrium and a lower payoff. Points above b on the envelope aeb are not NE, even though members' payoffs are non-negative there. At these points, any member would want to defect, since it prefers to be a non-signatory in a type 1 equilibrium in the abatement stage. For $F > \max \{F_b, F_g\}$ there are no NE other than 0.

The full-information first best outcome would have all nations abate even when they have high costs if and only if $N > \theta_H$. If cost realizations are private information (as we assume)

and we impose incentive compatibility at the abatement stage, a type 2 equilibrium consisting of all N nations leads to higher expected welfare than a type 1 equilibrium if and only if $N \geq M_N$. It is straightforward to show that $M_N < \theta_H$. Thus, the incentive compatibility constraint reduces the parameter space for which it is optimal to abate in *only* the low cost state (e.g., it decreases the lower bound on N above which abatement in both states is socially optimal). In this respect, the incentive compatibility constraint (at the abatement stage) leads to excessive abatement. This result is due to the fact that the type 1 equilibrium requires that some nations incur a transactions cost, causing the type 1 NE welfare to be lower than in the first best outcome. In contrast, there are no transactions cost in a type 2 equilibrium.

However, our game includes not only the abatement stage incentive compatibility constraint, but also the constraint implied by non-cooperative behavior at the participation stage. The smallest NE in the participation stage that produces a type 2 equilibrium at the abatement stage is $M_b > \theta_H$. Thus, (assuming that F can be chosen in stage 0) it is possible to induce all nations to join the IEA and then abate in both states of nature if and only if $N \geq M_b > \theta_H$. Therefore, if it is feasible to induce a type 2 equilibrium (by choice of F) it is always optimal to do so. If $N < M_b$, it is not feasible to induce a type 2 equilibrium.

In summary, it is possible to induce all nations to sign the IEA by choosing a sufficiently small fine ($M = N$). If $N < \theta_H$ the optimal IEA leads to the first best pattern (abatement in only the low cost state) but involves expected transactions cost equal to $F_1(N)p(1 - \phi) = Np(1 - \phi)\frac{\theta_L - 1}{N - \phi}$. If $M_b < N$ the optimal IEA leads to the first best pattern (abatement in both states) without transactions costs. If $\theta_H < N < M_b$ the optimal IEA leads to too little abatement, since abatement occurs only in the low-cost state but it is optimal to abate in both states; in addition, there is the transactions cost, as above.

3.2 Review of “standard result”

The standard model assumes that the IEA maximizes members’ joint welfare. Equivalently, the IEA makes its decision *after* nations have made their participation decision. In order for this assumption to be consistent with our setting, we require that the abatement decision is a binary choice, e.g. because of some technological constraint. To demonstrate the fact that in the standard model a decrease in membership cost reduces membership size and global welfare, we first consider the simplest case, where the IEA is not able to observe agents’ costs, or to induce them to reveal those costs. Here, the expected cost of a unit of abatement in both stages of the game is $\bar{\theta}$.

We then consider the situation in which the IEA is able to induce members to reveal their costs by means of an escape clause. These two games differ only in the policy menu available

to the IEA (abatement choice vs. level of fine), not in the timing of the decisions. The ability to use a fine increases the IEA's set of feasible outcomes, relative to the case where the IEA directly chooses abatement. Thus, conditional on membership size, the IEA payoff is (weakly) higher when it is able to use a fine. However, the switch from choosing abatement to choosing a fine lowers equilibrium membership size and reduces global welfare. We then compare the outcome in these models with the outcome when the IEA uses an escape clause and the fine is taken as given in the participation game.

3.2.1 The IEA chooses output conditional on M

Suppose (for now) that after nations have decided whether to join the IEA, the IEA decides whether to abate. Conditional on M , the expected payoff to a signatory is $\Pi(M) = \max \{0, M - \bar{\theta}\}$. The IEA instructs (all) members to abate if and only if $M \geq \bar{\theta}$. The NE of the participation game is $h(\bar{\theta})$. To confirm this, note that if there are $h(\bar{\theta})$ members, each signatory's payoff is non-negative; no signatory would want to leave the IEA, because the resulting IEA would choose not to abate, leaving the defector with a 0 payoff. No non-signatory wants to join, since in view of inequality (1), $h(\bar{\theta}) > h(\bar{\theta}) + 1 - \bar{\theta}$. The non-signatory's payoff in the NE exceeds its payoff if it joins the IEA.

In this model, the membership cost equals $\bar{\theta} - 1$. The level of membership, $h(\bar{\theta})$, weakly increases with membership cost; $h(\bar{\theta})$ is constant between integers, and jumps up by one unit as $\bar{\theta}$ passes through an integer value. From equation (12), membership costs in the IEA with an escape clause are never greater, and sometimes are less than $\bar{\theta} - 1$.

The equilibrium global welfare of N nations is $(N - \bar{\theta}) h(\bar{\theta})$. As $\bar{\theta}$ increases between integers welfare falls, but welfare has an upward jump as $\bar{\theta}$ passes through an integer value. (When N is an even integer, welfare is maximized at $\bar{\theta} = \frac{N}{2}$.) The fraction of potential welfare achieved in equilibrium is $\frac{h(\bar{\theta})}{N}$. This example illustrates why the standard model leads to a rather pessimistic view of IEAs: they achieve a substantial portion of potential gains from cooperation only when those potential gains are small. IEAs are effective only when they are unimportant.

This model assumes that when nations decide whether to join the IEA, they anticipate that the IEA will maximize members' joint welfare in the abatement stage. Members understand that their participation decision might have an effect on the IEA's action. If instead, the IEA makes the abatement decision *before* agents decide whether to join (as is the case for Kyoto), the equilibrium membership size is 0. In this case, there is nothing to offset nations' temptation to free-ride.⁴ The assumption that the IEA abatement decision is conditioned on membership

⁴Kyoto solved this free-rider problem by stipulating that the agreement would not enter into force unless a

size therefore increases the equilibrium size from 0 to $h(\bar{\theta})$.

3.2.2 The IEA chooses a fine conditional on M

Suppose now that the IEA is able to use an escape clause with a fine. As above, the IEA decision is made after nations decide whether to join; the decision is conditioned on M . If it is optimal for the IEA to induce a type 1 equilibrium, it chooses the smallest fine that will achieve this, $F_1(M)$. Using equation (9), the expected payoff of a member is

$$\hat{\Pi}(M) = \max \{0, (1-p)M - G_1(F_1(M)), M - \bar{\theta}\}.$$

Clearly $\hat{\Pi}(M) \geq \Pi(M)$. It is straightforward to show that for sufficiently small M it is optimal to set $F = 0$ (so that no member abates), and for sufficiently high M it is optimal to set $F \geq F_2(M)$ (so that members abate for both high and low costs). For a range of intermediate values of M it is optimal to induce a type 1 equilibrium at the abatement stage. Thus, $\hat{M} < \bar{\theta}$, where $\hat{M} = \sup \{M \mid \hat{\Pi}(M) = 0\}$. It is also easy to see that in this model, the equilibrium membership size is $h(\hat{M})$. Allowing the IEA to use a fine and escape clause, rather than requiring it to directly chose abatement, makes the IEA more efficient, and reduces membership costs, conditional on M . Since $h(\hat{M}) \leq h(\bar{\theta})$, we have another example where a reduction in membership costs (weakly) reduces equilibrium membership size and may reduce equilibrium global welfare.⁵

3.2.3 A comparison

Table 1 compares the equilibrium IEA size for games with different timing and different policy instruments. The size of the IEA depends on both of these features. Allowing the IEA to choose the level of the policy variable before rather than after nations decide whether to join, increases the equilibrium size if the IEA uses a fine/escape clause; the timing change decreases the size if the IEA selects the abatement level directly.

minimum level of participation was achieved.

⁵If $h(\hat{M}) = h(\bar{\theta})$, membership size and global welfare is the same under the two policy instruments. A sufficient condition for $h(\hat{M}) < h(\bar{\theta})$ is that $\hat{M} + 1 < \bar{\theta}$. When $\phi \approx 1$, the necessary and sufficient condition for the latter inequality is that $\theta_L + 1 < \bar{\theta}$. The change in global welfare from switching to the more efficient instrument equals

$$\Delta \equiv (N - \bar{\theta})h(\bar{\theta}) - ((1-p)N - G_1(F_1(\hat{M})))h(\hat{M}).$$

For example, let $p = 0.1$, $\theta_L = 4$, $\theta_H = 8$, and $1 - \phi \approx 0$, so $G_1(F_1(\hat{M})) \approx 3.6$. In this case, $h(\hat{M}) = 4$ and $h(\bar{\theta}) = 5$, so $\Delta \approx (N - 4.4)5 - (.9N - 3.6)4 = 1.4N - 7.6$. Provided that $N \geq 6$ (i.e., assuming that the IEA that chooses abatement can not obtain global cooperation), the switch to the more efficient policy instrument reduces global welfare.

	IEA Policy instrument	
	Level of abatement	Level of fine with escape clause
Before nations make participation decision	$M^* = 0$	M^* can be large
After nations make participation decision	$M^* = h(\bar{\theta}) \geq 2$ (small)	$M^* = h(\hat{M}) \leq h(\bar{\theta})$ (even smaller)

Table 1: Equilibrium IEA size in four different games

Non-signatories have a dominant strategy in the abatement game; the outcome of the participation game therefore depends on how a nation's participation decision will affect the signatories' actions in the abatement game. The critical membership size, below which the IEA does not ever abate, or below which it abates in only the low-cost state, depends on the timing and on the policy instrument.

The fact that IEA members obtain a rebate, and that the rebate decreases with the number of IEA members (making it more expensive to exercise the escape clause) is critical to the ability of the escape clause to increase equilibrium membership. If $\phi = 0$, so that firms receive no rebate, then the graphs of F_1 and F_2 are flat lines. In this case, a member's equilibrium action in the abatement stage is independent of the number of members. In that situation, nations have no incentive to join the IEA in the participation stage (because their decision has no effect on member's behavior), and the equilibrium size of the IEA is 0.

4 Farsighted stability in the participation game

The Nash equilibria in the previous section assume that in the participation game nations believe that: (i) if one nation deviates from equilibrium by withdrawing from or joining the IEA, other nations do not respond by changing their participation decisions, and (ii) each nation acts on its own, i.e., nations do not join or withdraw together in a coalition. In this section, we study the IEA sizes when these assumptions are gradually relaxed.

Nations are likely to be more sophisticated, and more able to react to deviations, than the NE assumes. For example, a nation may perform a thought experiment to predict how its deviation from a particular equilibrium candidate would precipitate changes in other nations' actions. The nation would compare the status quo payoff with the payoff under the *eventual* equilibrium following its own deviation and other nations' response – not on the payoff that would result if no other nations responded. Such nations, using the terminology of Chwe (1994), are *farsighted*.

To understand the effects of farsightedness on the formation of IEAs, consider the IEA of size $M = h(\bar{\theta})$ when $F > F_a$ in Figure 2. In this IEA, each signatory's payoff is $\pi_{s,2}(h(\bar{\theta})) \geq 0$, where the inequality is strict unless $h(\bar{\theta}) = \bar{\theta}$. Suppose a signatory to the IEA withdraws. The immediate result of this deviation is an IEA with less than $\bar{\theta}$ members, and lying to the left of the curve $\tilde{F}$. In that outcome, each signatory's payoff is negative, so the remaining signatories would also withdraw, eventually leading to zero membership. Foreseeing the subsequent reactions of the other signatories, the first signatory will not withdraw, because doing so leads to a zero payoff instead of the non-negative payoff under the equilibrium $h(\bar{\theta})$. Thus, no signatory to the IEA of size $h(\bar{\theta})$ wants to withdraw. We show that when nations are farsighted, this IEA is stable; in the previous section we saw that it is not a NE.

The second implicit NE assumption may or may not be reasonable, depending on whether nations can credibly coordinate *in the negotiation process* before the IEA is signed. For example, if a group of nations agree to join the IEA together, can they sign a binding agreement to guarantee that they will act as a group and no members will act differently? Although the negotiation process eventually produces a binding agreement (the IEA), binding agreements within the negotiation process before the IEA is formed may be harder to justify. Unlike the IEA, these pre-IEA agreements are at best informal.

If binding agreements are not possible, nations will act alone in making their participation decisions. But if binding agreements are possible, coalitional deviations have to be considered in studying the participation game. Continuing with the above example, if, through some exogenous processes, the current proposal is the trivial IEA with zero members, nations acting alone will not be able to form the IEA of size $h(\bar{\theta})$. However, when binding agreements are possible, a group of $h(\bar{\theta})$ nations want to move the IEA from size zero to size $h(\bar{\theta})$. In both cases, the nations can be farsighted.

Thus, in addition to the assumptions of rationality associated with a Nash equilibrium (e.g., each nation optimizes given the other nations' actions), we assume in this section that the nations are farsighted. Depending on whether pre-IEA binding agreements are possible, we will analyze the negotiation game under two sets of assumptions:

Assumption 1 (Unilateral Farsight) *The nations are farsighted, but coalitional deviations are not possible: each nation acts on its own in deciding whether or not to join or to withdraw from the IEA.*

Assumption 2 (Coalitional Farsight) *The nations are farsighted, and coalitional deviations are allowed.*

We analyze the participation game under Assumption 1 and calculate the associated stable

set. We then show that a particular element of this set corresponds to the stable IEA under Assumption 2.

4.1 Unilateral farsighted stable set

When nations are farsighted, solution concepts different from Nash equilibrium are needed to predict the kinds of IEAs that will form in the participation game. One such concept is Chwe (1994)'s farsighted stable set (FSS). We describe a variation of Chwe's definitions using Assumption 1.

In the participation game, let an outcome be a partition of the nations into signatories and non-signatories, and let Z be the set of outcomes. Consider two outcomes $a, b \in Z$. Denote $a \rightarrow_i b$ if nation i can move the outcome from a to b . For example, if i is a non-signatory, it can change the outcome by joining the IEA, making it one member larger. The preference ordering of nation i between two outcomes is given by $\prec_i$: $a \prec_i b$ if i prefers b to a . We define a dominance relation between two outcomes that allows each nation to act only unilaterally, but not in coalitions.

Definition 1 (Chwe (1994)) *An outcome $a \in Z$ is (unilaterally) indirectly dominated by outcome $b \in Z$, denoted as $a \ll b$, if and only if there is a sequence of outcomes, $z_0, \dots, z_m$ with $z_0 = a$ and $z_m = b$ and nations $0, \dots, m-1$ such that $z_j \rightarrow_j z_{j+1}$ and $z_j \prec_j b$ for all $j = 0, \dots, m-1$.*

That is, starting with outcome a , m nations makes sequential unilateral changes, generating a sequence of intermediate outcomes, $z_1, \dots, z_{m-1}$. Each nation (j) in the sequence prefers the final outcome (b) to the interim outcome that it faces (z_{j-1}). Thus, if $a \ll b$, there is *some* sequence of deviations from a that takes the outcome to b , and it is rational for each agent in that sequence to make the deviation.

The farsighted stable set (FSS) is essentially von Neumann and Morgenstern (1953)'s stable set armed with the indirect dominance relation. Due to the restriction to unilateral deviations in Assumption 1, we define a *unilateral* FSS.

Definition 2 (Chwe (1994)) *Given the set Z of outcomes and relation $\ll$, set $V \subseteq Z$ is a unilateral farsighted stable set (UFSS) of $(Z, \ll)$ if and only if*

- (i) V is internally stable: $\nexists a, b \in V$ such that $b \ll a$, and
- (ii) V is externally stable: $\forall b \in Z \setminus V, \exists a \in V$ such that $b \ll a$.

We say that an IEA with M members is “unilaterally farsighted stable” (or simply “stable” when there is no ambiguity) if and only if $M \in V$, the UFSS.

To understand the two requirements, note that if $a \ll b$ then a and b cannot both be internally stable, otherwise some sequence of players would cause a defection from a to b . Further, if b is outside the FSS, then there must be an element $a \in V$ that indirectly dominates b : if no such element a exists, then b would be stable. The FSS thus contains all the outcomes that are not indirectly dominated by other outcomes, and excludes all the outcomes that are indirectly dominated by some other outcomes.

As Chwe (1994) showed using the Condorcet Paradox, the UFSS does not exist when circular decisions arise. In our setting, a nation might withdraw from an IEA anticipating that another nation would join in its place; the new member would have the same incentive to withdraw, leading to a cycle of one nation withdrawing and another joining. Circular decisions are typical of coalition formation problems with farsight, and as we show in Appendix ??, they also arise in our model for (M, F) such that a Nash equilibrium with a strictly positive number of signatories exists in the participation game. We assume that nations can find a way to “break the cycle;” for example, we can follow Mariotti (1997) and impose large negative payoffs when circular decisions arise.

Since the nations are assumed to be *ex ante* identical in the participation game, and since we have ruled out cyclical outcomes, we can identify each outcome by the size of its associated IEA, rather than by the identities of the nations. That is, $Z = \mathcal{N}$, and each nation is either a signatory or a non-signatory. This observation simplifies the determination of indirect dominance relation between two outcomes (or two IEAs).

Lemma 1 *Consider two IEAs of sizes $M, M' \in \mathcal{N}$ respectively.*

- (i) *Suppose $M > M'$. Then $M \ll M'$ if and only if $\pi_s(m; F) < \pi_n(M'; F)$ for all $m = M, M - 1, \dots, M' + 1$.*
- (ii) *Suppose $M < M'$. Then $M \ll M'$ if and only if $\pi_n(m; F) \leq \pi_s(M'; F)$ for all $m = M, M + 1, \dots, M' - 1$.*

The proof of the Lemma is a direct consequence of Definition (2) and is not presented. Since cyclical outcomes are ruled out, we only need to search “in one direction” in deciding the dominance relation. For example, when $M > M'$, M' indirectly dominates M if a signatory to the IEA of size M wants to withdraw, anticipating the subsequent withdrawal by other signatories until the IEA settles at size M' . In the process of moving from M to M' , no non-signatories have incentive to join the IEA, because otherwise circular decisions arise, resulting in large negative payoffs.

The difficulty in finding the UFSS is that determining the stability of one IEA requires knowing other stable IEAs. Unless we know at least one element of the UFSS, it is not possible to determine the other elements. However, if we have identified the smallest element of the

UFSS, a simple recursive procedure determines larger IEAs in the UFSS. This recursion uses the following:

Definition 3 *Given an IEA of size $M_0 < N$, the set $\mathcal{M}(M_0)$ generated by M_0 is a finite and strictly increasing sequence of integers, $\mathcal{M}(M_0) \equiv \{M^j(M_0), j = 1, 2, \dots, k\}$, such that*

$$\begin{aligned} M^1 &= M_0, \\ M^j &= h(m^j), \quad \text{where } m^j = \min\{m \in \mathcal{R} : \pi_s(m) = \pi_n(M^{j-1})\}, j \geq 2 \\ M^j &\leq N, \quad j = 2, \dots, k. \end{aligned} \quad (13)$$

The sequence $\mathcal{M}(M_0)$ depends on F , but we suppress that argument. We provide conditions under which $\mathcal{M}(M_0) = V$. Given an IEA of size M_0 , the set $\mathcal{M}(M_0)$ is generated by a simple sequence of comparisons. Starting with $M^1 = M_0$, the next element M^2 is the smallest IEA size such that a signatory's payoff in M^2 is no less than the non-signatory's payoff in M_0 . This condition guarantees that a signatory in the larger IEA M^2 does not want to withdraw, because doing so will lead other signatories to withdraw until M_0 is reached, resulting in a lower payoff. Once we identify M^2 , the next element M^3 is found through the same procedure. This process is to be repeated until the greatest possible element M^k is reached.⁶

We use Definition 3 to construct V by setting M_0 equal to the smallest element of V . No equilibrium outcome, including (M_0, F) , can lie in a region where signatories have negative profits, i.e. the open set consisting of $M > 0$ and (i) below the upper envelope of F_1 and $\tilde{F}$, and also below F_2 , and (ii) less than $\bar{\theta}$ and on or above F_2 . However, M_0 can lie outside or on the boundary of this region, or it can equal 0. Neither V or $\mathcal{M}(M_0)$ contain a type 0 IEA with positive members. We cannot exclude the possibility that both of these sets contain one or more type 1 IEAs followed by one or more type 2 IEAs. If this possibility occurs, then there is a single switch, because the smallest type 1 IEA (in either set) lies below F_2 and the smallest type 2 IEA lies on or above that curve. The following Proposition shows the relation between $\mathcal{M}(M_0)$ and V .

Proposition 2 *$\mathcal{M}(M_0) = V$ if and only if the following three conditions hold:*

- (i) *When $M_0 = 0$ then $M^2 - 1$ is not a type 0 IEA (in the abatement stage)*
- (ii) *Either all positive M^j are the same type IEA; or there is a switch from a type 1 to a type 2 IEA, and the smallest type 2 IEA occurs at $M^j < M_b$.*
- (iii) *When $M_0 > 0$ then $\pi_s(M_0) \geq \pi_n(M)$ for $M < M_0$.*

⁶The equation $\pi_s(m) = \pi_n(M^j)$ may have two solutions m_1 and m_2 , where m_i is associated with a type i IEA. That is, it may occur that $\pi_{s,1}(m_1) = \pi_{s,2}(m_2) = \pi_n(M^j)$. Since $m_1 < m_2$, the procedure picks $h(m_1)$ instead of $h(m_2)$.

The intuition for Proposition (2) is straightforward. Conditions (i) and (ii) of the proposition are used to demonstrate internal stability of $\mathcal{M}(M_0)$, and condition (iii) is used to show external stability of $\mathcal{M}(M_0)$. By construction of $\mathcal{M}(M_0)$, an IEA of size M^{j+1} is not indirectly dominated by M^j or by smaller IEAs. Showing that M^j is not indirectly dominated by M^{j+1} (or by larger IEAs) uses the first two conditions in the proposition. When $M_0 = 0$ conditions (i) and (ii) imply that the “last signatory” of the smallest positive stable IEA (M^2) would prefer to remain a nonsignatory. This requirement is needed because otherwise M^2 would indirectly dominate M_0 . The last signatory of the smallest positive IEA *does* want to join if the IEAs M^2 and $M^2 - 1$ are the same type, or if the IEA M^2 is type 2 and the IEA $M^2 - 1$ is type 1. (The second possibility uses condition (ii).)

If all positive IEAs are the same type, then inequality (1) guarantees that the “last signatory” to the IEA with $M^{j+1} > M_0$ members would rather remain a non-signatory to an IEA with $M^{j+1} - 1$ members. Therefore M^{j+1} does not dominate M^j . If the sequence of positive stable IEAs switches from a type 1 at M^j to type 2 at M^{j+1} , but the switch occurs below M_b , the “last signatory” to M^{j+1} again prefers to remain a non-signatory to the IEA with $M^{j+1} - 1$ members. Thus, under the first two conditions of the proposition, the set $\mathcal{M}(M_0)$ is internally stable.

To verify external stability of this set, we note that all IEAs strictly between M^{j+1} and M^j are indirectly dominated by M^j because of the monotonicity in M of payoffs. Similarly, IEAs larger than M^k (defined as the largest element of $\mathcal{M}(M_0)$) are indirectly dominated by M^k . When $M_0 = 0$ there are no IEAs smaller than $M^1 = M_0$. When $M_0 > 0$, condition (iii) implies that IEAs smaller than M_0 are indirectly dominated by M_0 .

Proposition (2) implies that $M^{j+1} - M^j > 1$.⁷ Not all (in fact, “very few”) integers are elements of V . Consequently, for arbitrary F it is not true in general that making nations farsighted enables them to achieve global cooperation. Ray and Vohra (2001) obtain a similar result.

Proposition (2) enables us to find the UFSS V , by identifying its smallest element, M_0 , and then applying the recursive relation in equation (13). The three conditions provide the information needed to identify M_0 ; that is, we *choose* M_0 in order to satisfy these conditions, so these conditions do not restrict the applicability of the algorithm. We can set $M_0 = 0$ and test whether conditions (i) and (ii) hold. If they do not hold, then $M_0 > 1$, and we identify its value using conditions (ii) and (iii). To verify that a candidate $M_0 > 0$ is the smallest element of V we only need to verify that it is not indirectly dominated by smaller IEAs.

We illustrate this process using the example in Figure 3. As in the previous section, the

⁷This claim follows directly from Definition 3 when successive elements of V result in the same type IEA. When successive elements result in different type IEAs, the claim follows from condition (ii). If $M^{j+1} - M^j = 1$ and M^{j+1}, M^j are type 2 and type 1 IEAs then condition (ii) is violated.

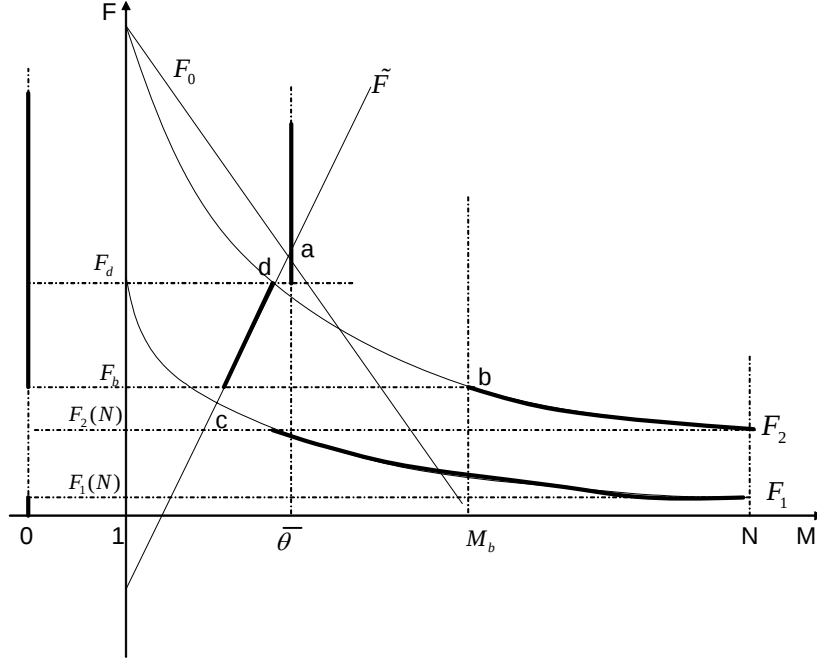


Figure 3: First Elements of the UFSS

equilibrium set (here, V) depends on the relative positions of the curves F_0 , F_1 , F_2 , $\tilde{F}$ and M_b . Figure 3 uses the definitions and assumptions implicit in Figure 2; in addition, it assumes that (i) point c is above the line $F_2(N)$,⁸ (ii) the horizontal distance between $\tilde{F}$ and F_1 at $F = F_b$ is not less than one, and (iii) the horizontal distance between $\bar{\theta}$ and F_1 at $F = F_d$ is not less than one.

For this configuration of curves, we have:

Example 1 (i) If $F < F_1(N)$, the UFSS contains the single element of an IEA with zero membership.

(ii) If $F \in [F_1(N), F_2(N))$, the UFSS is the set $\mathcal{M}(M_0)$, where $M_0 = h(F_1^{-1}(F))$.

(iii) If $F \in [F_2(N), F_b)$ where $F_b = F_2(M_b)$, the UFSS is the set $\mathcal{M}(M_0)$, where $M_0 = h(F_2^{-1}(F))$.

(iv) If $F \in [F_b, F_d)$, where F_d is the level of F where $\tilde{F}$ and F_2 cross, then the UFSS is the set $\mathcal{M}(M_0)$ where $M_0 = 0$. The second element of $\mathcal{M}(M_0)$ is $M^2 = h(\tilde{F}^{-1}(F))$.

(v) If $F \geq F_d$, the UFSS is the $\mathcal{M}(M_0)$ where $M_0 = 0$. The second element of $\mathcal{M}(M_0)$ is $M^2 = h(\bar{\theta})$.

⁸For $\phi \approx 1$, point c lies above $\theta_H - 1$ (a necessary condition for c to lie above $F_2(N)$) if and only if $\theta_H - \theta_L > 1$. This inequality is very likely to be satisfied for the problem of climate change, where there is a large difference between possible abatement costs.

Figure 3 graphs the first one or two elements of the UFSS, which are represented by the bold lines (ignoring the integer constraint). To see how Proposition (2) is used, consider first the case when $F \in [F_1(N), F_2(N))$. (The same reasoning applies to the case when $F \in [F_2(N), F_b)$.) Let M' be the IEA determined from the curve $F_1(\cdot)$: $M' = h(F_1^{-1}(F))$. We showed in the previous section that the IEA of size M' is a Nash equilibrium, i.e., $\pi_{s,1}(M') \geq \pi_{n,1}(M' - 1)$. The inequality implies that $\pi_{s,1}(M') \geq \pi_{n,1}(M)$ for all $M \leq M' - 1$ since $\pi_{n,1}(\cdot)$ is increasing; therefore, condition (iii) in Proposition 2 is satisfied. Since IEAs of sizes higher than or equal to M' are of the same type 1, Condition (ii) is satisfied when $M_0 = M'$. Thus, M' is the smallest element of the UFSS, or $M_0 = M'$.

When $F \in [F_b, F_d)$, there are two possible switches of IEA types: from type 0 to 1 (along curve F_1) and from 1 to 2 (along curve F_2). If $M_0 = 0$, the assumption that the horizontal distance between $\tilde{F}$ and F_1 is no less than one implies that condition (i) in Proposition 2 is satisfied. Further, condition (ii) is satisfied since the curve F_2 is to the left of M_b when $F > F_b$. Since 0 is the smallest possible IEA, we know $\mathcal{M}(0)$ is the UFSS. Similarly, when $F \geq F_d$, conditions (i) and (ii) are satisfied and thus $\mathcal{M}(0)$ is the UFSS.

Comparing with Figure 2, nations with farsight are more likely to cooperate and more likely to form larger IEAs. While no abatement by any nation (zero sized IEA) is always a Nash equilibrium, it is *not* in the UFSS for certain values of F . In fact, if we ignore the effects of the integer constraint, whenever there is a Nash equilibrium with a positively sized IEA (e.g., when $F \in [F_1(N), F_c)$ and $F \in [F_2(N), F_b)$), this IEA also belongs to the UFSS. Further in these cases, the zero sized IEA, a Nash equilibrium, is *not* in the UFSS. For other values of F , while the zero sized IEA is the only NE, it is not the only element in the UFSS.

When some of the Nash equilibria do belong to the UFSS, they only represent the first or the smaller elements in the UFSS. The existence of larger IEAs in the UFSS means that farsighted nations may possibly cooperate and form a larger IEA. The concept of UFSS is silent on the procedure in which the nations negotiate and form a particular IEA. Imposing more structure on the negotiation process may lead to finer predictions. For example, Ray and Vohra (2001) assumes a bargaining procedure with an exogenous order of movements, and we can show that under their procedure, the largest element in the UFSS will be proposed by the first mover and will be accepted by all other nations.

The definition of the UFSS, and the game in which it arises, does not involve actions at different points in time; the game is not written in extensive form. However, the UFSS does have the “flavor” of subgame perfection, as Xue (1998) noted. It is as if nations performed a thought experiment to predict the consequences of their actions. If we pursue the analogy of subgame perfection a bit further, the UFSS implies that IEAs can unravel, but they can not be built up. A nation can get the ball rolling by defecting from a stable IEA, but it can

only get the ball rolling downhill. For example, suppose that we think of beginning at a particular “subgame”, a status quo that falls just short of a stable IEA; that is, the size of the IEA is $M^{j+1} - 1$, possibly as the result of a signatory’s defection from a stable IEA. Since $M^{j+1} \ll M^{j+1} - 1$ (because no non-signatory wants to join the IEA), the IEA cannot be built up to the next stable element. However, beginning with this “subgame”, signatories do want to leave, causing the IEA to unravel to the next smallest stable element.

4.2 Coalitional farsighted stable set

Under Assumption 2, a group of nations may act together to deviate from the status quo. The unilateral indirect dominance relation defined in Definition 1 can be extended to coalitional indirect dominance: the sequence of outcomes is generated by coalitional rather than unilateral deviations, and the preference relation at each step must hold for all members of the coalitions within the step. The coalitional farsighted stable set (CFSS) can be defined similar to Definition 2 except that the dominance relation is coalitional rather than unilateral.

Since unilateral deviations are still allowed in determining the CFSS, if IEA M' unilaterally indirectly dominates IEA M , M' also coalitionally indirectly dominates M' . Thus, IEAs not in the UFSS are not in the CFSS either: the CFSS is a subset of the UFSS. The next proposition derives the CFSS from the UFSS.

Proposition 3 *For any F , the CFSS is a singleton which is the largest element of the associated UFSS.*

If coalitional deviations are possible, non-signatories to a smaller IEA may want to join the IEA as a group, and thus enjoy the higher benefit of the larger IEA. In our model, allowing for coalitional deviations raises the incentives of groups of non-signatories to join the IEA, but does not affect the incentives of groups of signatories to withdraw from the IEA. The possibility of binding agreements in the negotiation stage can only be welfare improving.

5 Discussion

We discuss the effects of the model parameters.

6 Conclusion

A Nash equilibrium in the participation game: proofs

In this appendix, we formally state and prove the results described in Section 3.1. Table 2 collects definitions of notation.

$h(x)$	The smallest integer no less than x
F_2	The locus above which signatories in a high cost state prefer to abate in a type 2 NE
F_1	The locus below which signatories in a low cost state prefer to prefer to pay the fine in a type 0 NE
$\tilde{F}$	The locus on which signatories have expected payoff of 0 in a type 1 NE
M_b	The value of M above which signatories have a higher payoff in a type 2 IEA than they would as non-signatories in a type 1 IEA.
F_0	The locus above which signatories' expected payment is higher in a type 2 than a type 1 NE

Table 2: Definitions and notation

Figure 2 embodies three parametric assumptions.

Assumption 3 *Point ‘a’ lies above the curve F_2 .*

Assumption 4 *The horizontal distance between curves F_2 and F_1 is greater than 2 at point ‘e’.*

Assumption 5 $M_b - \bar{\theta} > 2$ (equivalently, $\bar{\theta} > \frac{p+1}{1-p}$).

These conditions are appropriate for a model that describes the problem of forming an IEA to control GHGs. (i) A sufficient condition for Assumption 1 is that transactions costs are positive but small. (ii) Even if nations were certain that abatement costs are low, an agreement would have to contain at least several members in order for them to benefit from abatement.⁹ In addition, there is a non-negligible difference between high and low abatement costs. These conditions are sufficient for Assumption 2. (iii) Finally, the probability of high abatement costs is moderate or small (e.g., less than 0.5) so that Assumption 3 is satisfied.

Assumption 2 implies that if, at the participation stage, a signatory considers defecting from a candidate NE on F_2 below point e (i.e., for an IEA larger than $\bar{\theta}$), then the signatory knows that the resulting outcome at the abatement stage will be a type 1 NE rather than a type 0 NE. Assumption 3 insures that for any candidate NE on $\tilde{F}$ consisting of fewer than $\bar{\theta}$ members, no non-signatory wants to defect by joining the IEA. (If it were to join the IEA, the membership

⁹That is, θ_L must be “moderately large”, e.g. $\theta_L \geq 4$. Given our normalization, $\theta_L = 4$ means that in the low cost state, at least four nations would have to abate in order for their joint welfare to be higher than if they did not abate. As is clear from the lemmas below, this sufficient condition is very strong; the horizontal distance between the two graphs, at point b , can be greater than 2 even if this condition does not hold.

would still be lower than M_b , the critical value below which a nation prefers being a non-signatory to an IEA that results in a type 1 NE in the abatement stage, rather than a signatory to an IEA that results in a type 2 equilibrium.)

We first show that $M_b > M_N$. By inspection of Figure 2, this inequality implies that point b lies above the graph of F_0 .

Lemma 2 $M_b > M_N$.

Proof. Since $F_0 - F_1$ is a decreasing function of M for $M > 1$, and $F_0 - F_1 = 0$ at M_N , the lemma is true iff $F_0 - F_1 < 0$ at M_b . We have

$$F_0 - F_1 = \frac{-M - M\theta_H + M\theta_L + \phi\theta_H + M^2 - M\phi\theta_L}{(\phi - 1)(M - \phi)}.$$

This expression is negative iff the numerator is positive. Evaluating the numerator at $M = M_b$ and simplifying yields

$$\frac{1}{p^2} (\theta_L - 1) [p + \theta_L + p\theta_H - p\theta_L - p\phi\theta_L - p^2\phi\theta_H + p^2\phi\theta_L - 1],$$

so we require the term in square brackets to be positive. This term equals

$$\begin{aligned} (1 - \phi p) \bar{\theta} - (1 - p) &= (1 - p + p - \phi p) \bar{\theta} - (1 - p) \\ &= (1 - p) (\bar{\theta} - 1) + p(1 - \phi) \bar{\theta} > 0. \end{aligned}$$

■

Assumption (3) states that the intersection of $\tilde{F}$ and F_0 , denoted point a , lies above the curve F_2 . This assumption is equivalent to

$$\frac{\phi - p}{1 - p} > \frac{\theta_L}{\theta_H}. \quad (14)$$

This inequality holds if the transaction cost is small (ϕ close to 1) and/or if there is a substantial cost difference in the two states.

Our characterization of the NE of the participation game requires that the horizontal distance between F_2 and F_1 be greater than 2 for relevant values of F . We provide the necessary and sufficient condition for Assumption (4), and then show that this condition implies that the two graphs are “far enough apart”.

We begin with the following

Lemma 3 *The horizontal distance between F_2 and F_1 is a decreasing function of F .*

Proof. Using the definitions in section 2.2, we have $F_2 = M \frac{-1+\theta_H}{-M+\phi}$, so the inverse of this function is $M = -F_2 \frac{\phi}{-F_2-1+\theta_H}$. Also, $F_1 = -(-1 + \theta_L) \frac{M}{-M+\phi}$, so the inverse of this function

is $M = -F_1 \frac{\phi}{-F_1 - 1 + \theta_L}$. The horizontal distance between the two functions is

$$D \equiv F\phi \frac{-\theta_L + \theta_H}{(-F - 1 + \theta_H)(-F - 1 + \theta_L)}.$$

The derivative of D is

$$\frac{dD}{dF} = -\phi(-\theta_L + \theta_H) \frac{F^2 - 1 + \theta_L + \theta_H - \theta_H\theta_L}{(F + 1 - \theta_H)^2 (F + 1 - \theta_L)^2}$$

The sign of this derivative equals the negative of the sign of $F^2 - 1 + \theta_L + \theta_H - \theta_H\theta_L$, so we need to show that this expression is positive. Use the fact that $F > \theta_H - 1$, which implies that it is sufficient to show that $(\theta_H - 1)^2 + \theta_L - 1 > \theta_H(\theta_L - 1)$, or $(\theta_H - 1)^2 > (\theta_H - 1)(\theta_L - 1)$, or $(\theta_H - 1) > (\theta_L - 1)$. This inequality is true because $\theta_H > \theta_L$. ■

To state the next result we use the following definitions:

$$\begin{aligned} \alpha &\equiv (\theta_L - \theta_H)^3 < 0 \\ \beta &\equiv -2(\theta_L - 1.5)(\theta_L - \theta_H)^2 \\ \gamma &\equiv (3\theta_L - \theta_L^2)\theta_H + (\theta_L^3 - 3\theta_L^2 + 2\theta_L - 2). \end{aligned} \tag{15}$$

The next lemma shows that Assumption (4) is equivalent to the following inequality:

$$R \equiv \alpha p^2 + \beta p + \gamma < 0. \tag{16}$$

Lemma 4 (i) $F_2^{-1}(F_e) - F_1^{-1}(F_e) - 2 > 0 \iff R < 0$. (ii) A sufficient condition for $R < 0$ (when $p \geq 0$) is that $\beta < 0$ and $\gamma < 0$. Both of these inequalities are satisfied if

$$\theta_L > 3 \text{ and } \theta_H > \theta_L + \frac{2(\theta_L - 1)}{\theta_L(\theta_L - 3)}. \tag{17}$$

Proof. (Part i) Using the definition of D we have

$$D - 2 = \frac{-(4F - 2\theta_L - 2\theta_H - 2F\theta_L - 2F\theta_H + 2\theta_L\theta_H + 2F^2 + F\theta_L\phi - F\theta_H\phi + 2)}{(F - \theta_L + 1)(F - \theta_H + 1)}.$$

The denominator of this expression is positive, so $D > 2$ iff the numerator is positive. In view of Lemma (3), this inequality requires that the numerator is positive evaluated at $F = -M_e \frac{-1 + \theta_H}{-M_e + \phi}$ (where $M_e = \bar{\theta}$). Evaluating the numerator at this point and simplifying gives the expression

$$-\phi \frac{K}{(-\bar{\theta} + \phi)^2}, \text{ with } K \equiv (2(\theta_L - 1) + \bar{\theta}(\theta_H - \theta_L))\phi + \bar{\theta}(\bar{\theta} - 2)(\theta_L - \theta_H).$$

Since $-\frac{\phi}{(-\bar{\theta} + \phi)^2} < 0$ we need to show that $K < 0$. Note that K is an increasing function of ϕ . Therefore we need to establish that K evaluated at $\phi = 1$ is negative. Denote this value as L :

$$L = (2(\theta_L - 1) + \bar{\theta}(\theta_H - \theta_L))\phi + \bar{\theta}(\bar{\theta} - 2)(\theta_L - \theta_H).$$

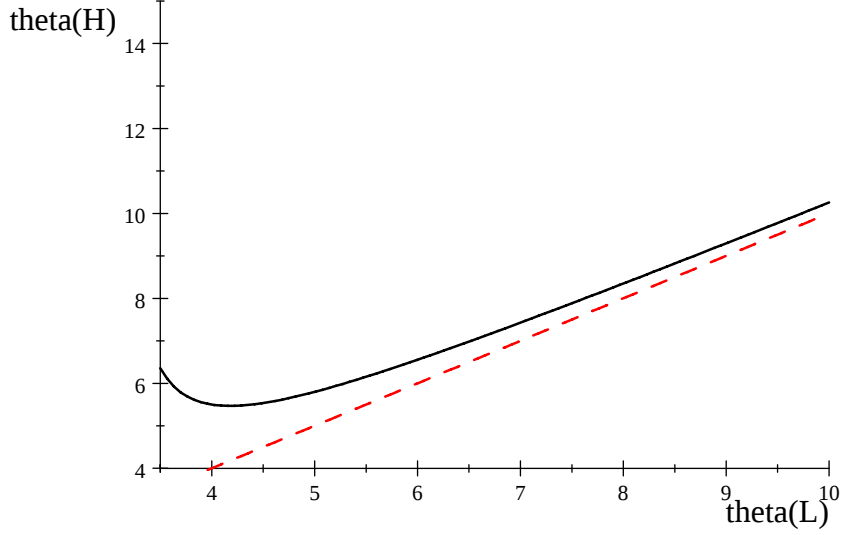


Figure 4: The graph of $\theta_L + \frac{2(\theta_L-1)}{\theta_L(\theta_L-3)}$

We write this expression as a function of p using $\bar{\theta} = p\theta_H + (1-p)\theta_L$, using the definitions in equation (15), to obtain the expression R given in equation (16). To establish part (ii) note that R is concave in p , and at $p = 0$, R is decreasing if $\theta_L > 1.5$ and $R = \gamma$. Therefore, a sufficient condition for $R < 0$ for $p \geq 0$ is $\theta_L > 1.5$ and $\gamma < 0$. Suppose that $\theta_L > 3$. Then $\gamma < 0$ if and only if

$$\theta_H > \theta_L + \frac{2(\theta_L - 1)}{\theta_L(\theta_L - 3)}.$$

■

Figure (4) shows the graph of $\theta_L + \frac{2(\theta_L-1)}{\theta_L(\theta_L-3)}$ and the (dashed) line where $\theta_H = \theta_L$. For $\theta_L > 4$, $R < 0$ for all p and for nearly all $\theta_H > \theta_L$. Given our normalization, $\theta_L = 4$ means that in the low cost state, at least four nations would have to abate in order for their joint welfare to be higher than if they did not abate.

The following corollary gives a necessary and sufficient condition for the horizontal distance between F_2 and F_1 to be greater than 2 for values of F less than F_e .

Corollary 1 $F_2^{-1}(F) - F_1^{-1}(F) - 2 > 0$ for $F \in (\theta_H - 1, F_e)$ for all $0 < \phi < 1$ if and only if $R < 0$.

Proof. The proof is immediate from the previous two lemmas. ■

We now formally state and prove the characterization of the NE to the participation game.¹⁰

¹⁰If nations could costlessly disband the IEA, there would be additional NE to the participation game. For example, for $F \geq F_2(\bar{\theta})$, $M = \bar{\theta}$ is a NE. For F between F_c and the horizontal coordinate of the intersection between $\tilde{F}$ and F_2 , $\tilde{F}^{-1}(F)$ is a NE, leading to a type 1 equilibrium in the abatement stage.

Proposition 4 We adopt Assumptions (3), (4), and (5). The IEA uses an escape clause with fine F that is taken as given at the participation stage.

(i) For $N \geq F_2^{-1}(F_b)$, and for $F \in [F_2(N), F_b]$ there exists a NE to the participation game consisting of $h(F_2^{-1}(F))$ members. The resulting abatement-stage NE is type 2.

(ii) The smallest IEA leading to a type 2 abatement-stage NE consists of M_b members, induced by fine F_b .

(iii) For $N \geq F_1^{-1}(F_c)$, and for $F \in [F_1^{-1}(N), F_c]$ there exists a NE to the participation game consisting of $h(F_1^{-1}(F))$ members. This NE induces a type 1 equilibrium in the abatement stage.

(iv) Define F_k to satisfy $h(\tilde{F}^{-1}(F_k)) - F_1^{-1}(F_k) = 1$; define $F_q = \sup \left\{ F \mid h(\tilde{F}^{-1}(F)) \leq F_2^{-1}(F) \right\}$; finally define $F_g = \min \{F_k, F_q\}$. It must be case that $F_c < F_g \leq F_d$. For $F_c \leq F \leq F_g$ there is a NE with $h(\tilde{F}^{-1}(F))$ members. This NE induces a type 1 equilibrium in the abatement stage. For $F > F_g$ there is no NE to the participation game that induces a type 1 equilibrium in the abatement game.

(v) If F can be chosen at stage 0, it is feasible to induce a type 2 equilibrium at the abatement stage iff $N \geq M_b$. If it is feasible to induce a type 2 equilibrium, it is optimal to do so.

Proof. (Proposition 2) We begin by explaining the meaning of $M_b = \frac{\bar{\theta}-1+p}{p}$ and $F_2(M_b)$, the coordinates of point b in Figure 2. At these values, a signatory in a type 2 equilibrium has the same payoff that it would obtain if it left the IEA and became a non-signatory in a type 1 IEA:

$$\begin{aligned} \pi_{s,2}(M) \begin{cases} < \\ = \\ > \end{cases} \pi_{n,1}(M-1) \\ \iff \\ M \begin{cases} > \\ = \\ < \end{cases} M_b \end{aligned} \tag{18}$$

For points on the curve F_2 below point b , a signatory in a type 2 equilibrium would not want to leave the IEA if that defection induced a type 1 equilibrium; for points on F_2 above b , a signatory would want to defect if the result was a type 1 equilibrium.

Since $M_b > \bar{\theta}$, $F_2(M_b) < F_e$; therefore, by Corollary (1) the horizontal distance between F_2 and F_1 is greater than 2:

$$F_2^{-1}(F) - F_1^{-1}(F) - 2 > 0 \text{ for } F \in (\theta_H - 1, F_2(M_b)]. \tag{19}$$

We now prove the claims in Proposition 2.

(i) For $F \in [F_2^{-1}(N), F_b]$ consider the candidate equilibrium consisting of $h(F_2^{-1}(F))$ members. If a member of the IEA defects, the resulting equilibrium to the abatement game is type 1, in view of Corollary 1. The defector's payoff is lower, in view of equation (18). No non-signatory wants to defect from the candidate by joining the IEA, because $\pi_{s,2}(M+1) < \pi_{n,2}(M)$. Therefore, the candidate is a NE.

(ii) For $F > F_b$ we need to consider two possibilities. Consider first $F \in (F_b, F_e)$. Clearly $M > h(F_2^{-1}(F))$ is not an equilibrium: a member would want to defect by leaving the IEA, since the resulting IEA would still induce a type 2 equilibrium in the abatement game. Similarly, $M = h(F_2^{-1}(F))$ is not an equilibrium: by equation (18), a member would want to leave the IEA, inducing a type 1 equilibrium in the abatement game. Next, consider $F \geq F_e$. Over this range of F , it is easy to see that the only candidate equilibrium is $h(\bar{\theta})$. However, this cannot be an equilibrium, since defection by a member would induce either a type 1 or a type 2 equilibrium in the abatement game (depending on the magnitude of F). By equation (18), a signatory who leaves the IEA (becoming a non-signatory) has a higher payoff than at the candidate equilibrium.

(iii) For $F \in [F_1^{-1}(N), F_c]$, consider a candidate NE at $M = h(F_1^{-1})$. Signatories' payoffs are positive at $h(F_1(M))$ because this point is to the right of $\tilde{F}$ (except at the endpoint F_c where the payoff is 0). If any signatory were to defect by leaving the IEA, the resulting NE in the abatement game is type 0, where a non-signatory obtains a 0 payoff. Therefore, no signatory wants to defect.

We now need to show that non-signatories do not want to defect from the candidate equilibrium by joining the IEA. Since $h(F_1^{-1}(F)) + 1 < F_1^{-1}(F) + 2 < F_2^{-1}(F)$ by inequality (19)) the defection induces a type 1 equilibrium in the abatement stage. The defector's payoff is lower at the new point than at the candidate, because $\pi_{s,1}(M+1; F) < \pi_{n,1}(M)$.

(iv) First note that if $F > F_q$ the only candidate NE to the participation game that could result in a type 1 equilibrium in the abatement game, is $M = h(\tilde{F}^{-1}(F))$, since smaller values would result in negative payoffs for signatories, and larger values would not be immune from defection by signatories. However for $F > F_q$ the candidate $M = h(\tilde{F}^{-1}(F))$ results in a type 2 equilibrium in the abatement stage. Therefore, NE to the participation game that lead to type 1 equilibria must have fines $F \leq F_q$.

Next consider candidates $M = h(\tilde{F}^{-1}(F))$ for $F > F_k$. A signatory would want to defect from this candidate, since the resulting abatement stage equilibrium would still be type 1. Therefore, NE to the participation game that lead to type 1 equilibria must have fines $F \leq F_k$.

Thus, for $F \leq F_g$ the candidate $M = h(\tilde{F}^{-1}(F))$ is consistent, in that it leads to a type 1 equilibrium, and it is immune from defection by signatories. Therefore we need only show

that this candidate is immune from defection by a non-signatory. By virtue of Assumption (5), if a non-signatory defects by joining the IEA, the resulting abatement stage equilibrium is still type 1. Thus, the defecting non-signatory has a lower payoff. Therefore, the candidate is a NE to the participation game.

(v) When all countries are in the IEA, aggregate welfare equals the joint welfare of IEA members. Define M_N to satisfy $F_0(M_N) = F_1(M_N)$. On F_0 , signatories' payoffs are the same in a type 1 or a type 2 equilibrium consisting of all nations. Therefore, at M_N :

$$\pi_{s,2}(M_N) = \pi_{s,1}(M_N; F_1^{-1}(M_N)).$$

Recall that $\pi_{s,2}(M)$ is independent of F . For $M < M_N$ the point $(M; F_1^{-1}(M))$ lies below the line F_0 , so at that point $\pi_{s,2}(M) < \pi_{s,1}(M; F_1^{-1}(M))$. Therefore, for $M < M_N$ aggregate welfare is higher in a type 1 equilibrium consisting of all nations, than in a type 2 equilibrium consisting of all nations. The argument is reversed when $M > M_N$.

By part (ii) above, the smallest IEA that results in a type 2 NE in the abatement stage consists of M_b members. By Lemma (2), this value is greater than M_N , the value above which it is optimal to induce a type 2 equilibrium. ■

B Farsighted stable sets: model details

In this section, we provide the model details for Section 4.

B.1 Circular decisions under farsightedness

Consider a Nash equilibrium IEA of the participation game identified in Section 2, and suppose the IEA's size is M . Being a NE implies that $\pi_s(M; F) \geq \pi_n(M-1, F)$. Suppose a signatory, say nation i , withdraws from the IEA, and consider the reaction of a non-signatory, say nation j . Since $\pi_s(M; F) \geq \pi_n(M-1, F)$, nation j has incentive to join the IEA, and if it joins, i 's payoff becomes $\pi_n(M; F) > \pi_s(M; F)$. Anticipating j 's reaction, i thus has incentive to withdraw: it can become a free-rider since j will join in its place. But this argument can go on forever, since every signatory has incentive to withdraw in order to become a free-rider.

B.2 Proof of the propositions

Proposition 2. (Proof.) To demonstrate “only if” we merely show that if any one of the conditions are not satisfied, then the set $\mathcal{M}(M_0)$ is either not internally or not externally stable. This claim is straightforward, and we do not provide the details. The proof demonstrates the “if” part of the proposition, following the outline given in the paragraph below the statement

of the Proposition. Step 1 uses conditions (i) and (ii) to establish that $\mathcal{M}(M_0)$ is an internally stable set, and Step 2 uses condition (iii) to confirm its external stability.

Step 1 (Internal Stability): Payoffs are monotonic in M . In addition, in order to move from an IEA of size M^j to an IEA of size M^{j+s} with $s > 1$ it is necessary to “move through” an IEA of size M^{j+1} . Therefore, the fact that M^{j+1} does not indirectly dominate M^j implies that larger IEAs also do not indirectly dominate M^j . Similarly, the fact that M^j does not indirectly dominate M^{j+1} implies that smaller IEAs also do not dominate M^{j+1} . These facts allow us to demonstrate internal stability by showing that neither of the IEAs M^j nor M^{j+1} indirectly dominate each other.

No element of $\mathcal{M}(M_0)$ can indirectly dominate a larger element of the set. For example to move from M^{j+1} to M^j , one signatory has to begin the process by leaving the IEA. The “first deviator’s” payoff is no higher (except for knife-edge cases, strictly lower) when it becomes a non-signatory at M^j instead of remaining a signatory at M^{j+1} . (For the knife-edge case, recall our assumption that in the case of a tie, a nation prefers to abate.)

To complete the argument for internal stability, we need only show that no element of $\mathcal{M}(M_0)$ can indirectly dominate a smaller element of the set. We do this by showing that M^{j+1} does not indirectly dominate M^j .

First consider the case where M^j and M^{j+1} are both positive and both of the same type. Recall from Lemma (1(ii)) that in order for $M^j \ll M^{j+1}$, it must be true that $\pi_n(m) \leq \pi_s(M^{j+1})$ for all $m = M^j, M^j + 1, \dots, M^{j+1} - 1$. Thus, for this step, all we need to establish is that this inequality does *not* hold for some m . We establish this inequality for the case of $m = M^{j+1} - 1$.

Recall that $\pi_{s,i}(1) < \pi_{n,i}(0)$, $i = 0, 1, 2$; a nation never wants to be the sole member of an IEA. Equations (7) - (11) imply that $\partial \pi_s(M, i) / \partial M = \partial \pi_n(M, i) / \partial M$. The two conditions above imply that, *if IEAs of sizes M and $M - 1$ are of the same type,*

$$\pi_{s,i}(M) < \pi_{n,i}(M - 1), \quad \text{or} \quad M \ll M - 1, \quad M = 1, \dots, N. \quad (20)$$

If M^j and M^{j+1} are of the same type, $M^{j+1} - 1$ and M^j are of the same type as well, implying that $M^{j+1} \ll M^{j+1} - 1$.

Next consider the case where M^j is a type 1 IEA and M^{j+1} is a type 2 IEA. (We know that there can be no type 1 IEAs larger than the smallest type 2 IEA because the curve F_2 lies above F_1 .) By condition (ii), $M^{j+1} < M_b$. By the definition of M_b , a nation prefers to be a non-signatory to an IEA of size $M^{j+1} - 1$ rather than a signatory to an IEA of size M^{j+1} , so again $M^{j+1} \ll M^{j+1} - 1$.

Finally, consider the case where $M_0 = 0$, so that $\pi_n(M_0) = 0$. If it were the case that the IEA with $M^2 - 1$ were a type 0 equilibrium, then a nonsignatory would want to join that IEA,

because joining increases its payoff from 0 to a non-negative level. In that case, M^1 indirectly dominates M^2 , violating internal stability. However, if the IEAs with $M^2 - 1$ and M^2 are both the same type, $M^2 \ll M^2 - 1$. The “last signatory” does not want to join, so M^2 does not indirectly dominate M^1 . Also, if M^2 is a type 2 IEA and $M^j - 1$ is type 1, then by virtue of condition (ii) we again have $M^2 \ll M^2 - 1$.

Step 2: (External stability) We need to show that each element in $\mathcal{N}$ in the complement of $\mathcal{M}(M_0)$ (i.e. $\mathcal{N}/\mathcal{M}(M_0)$) is indirectly dominated by some element of $\mathcal{M}(M_0)$. The set $\mathcal{N}/\mathcal{M}(M_0)$ is the union of two sets, IEAs that are smaller than, or larger than M_0 . Denote these as $A = \{M \mid M < M_0, M \in \mathcal{N}/\mathcal{M}(M_0)\}$ and $B = \{M \mid M > M_0, M \in \mathcal{N}/\mathcal{M}(M_0)\}$.

Consider set A . When $M_0 = 0$, $A = \emptyset$, so for this subset we need only consider $M_0 > 0$. In this case, condition (iii) states that IEAs smaller than M_0 are indirectly dominated by the IEA of size M_0 .

Now consider set B . We show that M^j indirectly dominates IEAs with sizes between M^j and M^{j+1} for $j + 1 \leq k$ (Recall that k is the index of the largest element of $\mathcal{M}(M_0)$.) That is, $M \ll M^j$ for all $M = M^j + 1, \dots, M^{j+1} - 1$. In addition, for $j = k$, $M \ll M^j$ for all $M = M^j + 1, \dots, N$. We provide details only for the case of $j < k$; the proof is similar when $j = k$.

From (13), we know $\pi_s(m^{j+1}) = \pi_n(M^j)$. We need to consider two cases: where M^j and M^{j+1} are the same type of IEA, and where they are different types. Suppose first that IEAs of sizes M^j and M^{j+1} are of the same type i , i.e., $\pi_{s,i}(m^{j+1}) = \pi_{n,i}(M^j)$. Since $\pi_{s,i}(\cdot)$ is strictly increasing in M , the equation means that $\pi_{s,i}(M^{j+1} - 1) < \pi_{n,i}(M^j)$, which in turn implies that $\pi_{s,i}(M) < \pi_{n,i}(M^j)$, for all $M = M^j + 1, \dots, M^{j+1} - 1$. Since all these IEAs are of the same type i , we know $\pi_s(M) < \pi_n(M^j)$ and thus $M \ll M^j$ for all $M = M^j + 1, \dots, M^{j+1} - 1$.

Suppose instead that IEAs of sizes M^j and M^{j+1} are of different types. Here there are two possibilities. Either (i) $M^j = 0$ and M^{j+1} is a type 1 or type 2 IEA, or (ii) M^{j+1} is a type 2 and M^j is a type 1 IEA. (As explained in the text, there can be no positive elements of the stable set that are type 0.)

First consider the possibility $M^j = 0$, which occurs when $j = 1$ and $M_0 = 0$. In this case, $\pi_s(M) < 0$ for $1 \leq M < M^2$, so $M \ll M^j = 0$ for all $M = 1, 2, \dots, M^{j+1} - 1$. Next consider the case where M^{j+1} is a type 2 and M^j is a type 1 IEA. Let $M' \leq M^{j+1}$ be such that the IEA of size $M' - 1$ is of type 1 but that of M' is of type 2. For IEAs between M' and M^{j+1} , Because $\pi_{s,2}(M^j - 1) < \pi_n(M^j)$, we know $\pi_{s,2}(M) < \pi_n(M^j)$ for all $M \in [M', M^{j+1} - 1]$. That is, $M \ll M^j$ for all $M \in [M', M^{j+1} - 1]$. For IEAs between M^j and $M' - 1$ we know from (13) that $\pi_{s,1}(M' - 1) < \pi_s(M^j)$; if this inequality did not hold, $M' - 1$ instead of M^{j+1} would have been the next element in $\mathcal{M}(M_0)$ after M^j . Again, since $\pi_{s,1}(\cdot)$ is increasing, we know $\pi_{s,1}(M) < \pi_s(M^j)$ for all $M \in [M^j + 1, M' - 1]$. Therefore, $M \ll M^j$ for all

$$M \in [M^j + 1, M' - 1]. \blacksquare$$

Confirmation of Example 1

(i) If $F < F_1(N)$, the IEA is of type 0 and its signatory earns negative payoffs. Signatories to an IEA of any positive size have incentive to withdraw, i.e., $M \ll 0$ for all $M > 0$. In this case, the UFSS has only one element $M = 0$.

(ii) Let $M' = h(F_1^{-1}(F))$. Recall that the IEA of size M' is a Nash equilibrium, i.e., $\pi_{s,1}(M') \geq \pi_{n,1}(M' - 1)$, implying that $\pi_{s,1}(M') \geq \pi_{n,1}(M)$ for all $M \leq M' - 1$ or $M \ll M'$. Since IEAs of size $M \geq M'$ are of the same type 1, Corollary ?? implies that M' is not indirectly dominated by any $M > M'$. Thus, M' is the smallest farsighted stable set, or $M_0 = M'$.

(iii) Suppose $F \in [F_2(N), F_b)$. The proof is the same as the case of (ii) since the IEA of size $M' = h(F_2^{-1}(F))$ is also a Nash equilibrium, and IEAs of sizes $M \geq M'$ are of the same type.

(iv) Suppose $F \in [F_b, F_d)$. To show that $M_0 = 0$, we only need to show that M_0 is not indirectly dominated by any $M > 0$. But we know from Figure 3 (and from Assumption 5) that there are two possible switches of IEA types: from type 0 to type 1 along curve F_1 , which is to the left of curve $\tilde{F}$, and from type 1 to type 2 along curve F_2 , which is to the left of M_b . Thus, the conditions in Proposition 2(ii) are satisfied, and $M_0 = 0$ is not indirectly dominated by $M > 0$ from Corollary 2

(v) Suppose $F \geq F_d$. The proof is similar to case (iv): $M_0 = 0$ since the conditions in Proposition 2(ii) are satisfied for all three possible switches of IEA types.

Proposition 3. (Proof) Consider first the set $\mathcal{M}_i(M_0)$ defined in Definition 3, with the largest element being M_i^k . From (13), we know

$$\pi_{s,i}(M_i^k) \geq \pi_{n,i}(M_i^{k-1}) > \pi_{n,i}(M_i^{k-2}) > \dots > \pi_{n,i}(M_i^1). \quad (21)$$

Consequently, at any IEA M_i^j , $j < k$, a group of $M_i^k - M_i^j$ non-signatories have incentive to join the IEA together, and earn $\pi_{s,i}(M_i^k)$ instead of $\pi_{n,i}(M_i^j)$. That is, M_i^k coalitionally indirectly dominates all other elements in the set $\mathcal{M}_i(M_0)$.

As shown in Proposition ??, the UFSS contains $\mathcal{M}_i(M_0)$ and possibly other IEAs in the UFSS with sizes smaller than M_0 . But from the proposition, we know $\pi_s(M_0) \geq \pi_n(M)$ for $M \in UFSS$, $M < M_0$. Then from (21), we know $\pi_{s,i}(M_i^k) > \pi_n(M)$, and thus M_i^k coalitionally indirectly dominates M . $\blacksquare$

References

- BARRETT, S. (1994): "Self-enforcing international environmental agreements," *Oxford Economic Papers*, 46, 878–894.
- (2003): *Environment and Statecraft*. Oxford University Press.
- BATABYAL, A. (2000): *The Economics of International Environmental Agreements*. Ashgate Press.
- BLOCH, F. (1997): "Noncooperative models of coalition formation in games with spillovers," in *Dew Directions in the Economic Theory of the Environment*, ed. by C. Carraro, and D. Siniscalco, pp. 311–352. Cambridge University Press.
- CARRARO, C., AND D. SINISCALCO (1993): "Strategies for the International Protection of the Environment," *Journal of Public Economics*, 52, 309–328.
- CHWE, M. S. (1994): "Farsighted Coalitional Stability," *Journal of Economic Theory*, 63, 299–325.
- DIAMANTOUDI, E., AND E. SARTZETAKIS (2002): "International Environmental Agreements - The Role of Foresight," University of Aarhus working paper 2002-10.
- DIXIT, A., AND M. OLSON (2000): "Does Voluntary Participation undermine the Coase Theorem," *Journal of Public Economics*, 76, 309 – 335.
- FINUS, M. (2001): *Game Theory and International Environmental Cooperation*. Edward Elgar.
- KOPP, R., R. MORGENSTERN, W. PIZER, AND F. GHERSI (2002): "Reducing Cost Uncertainty and Encouraging Ratification of the Kyoto Protocol," in *Global warming and the Asian Pacific*, ed. by C. Chang, R. Mendelsohn, and D. Shaw, pp. 231–46. Academia Studies in Asian Economies, Cheltenham, U.K.
- MARIOTTI, M. (1997): "A Model of Agreements in Strategic Form Games," *Journal of Economic Theory*, 74, 196–217.
- PIZER, W. (2002): "Combining Price and quantity controls to mitigate global climate change," *Journal of Public Economics*, 85, 409 – 34.
- RAY, D., AND R. VOHRA (2001): "Coalitional Power and Public Goods," *Journal of Political Economy*, 109(6), 1355 – 1382.

- VICTOR, D. (2003): "International agreements and the struggle to tame carbon," in *Global Climate Change*, ed. by J. M. Griffin, pp. 204–240. Edward Elgar, Cheltenham, UK.
- XUE, L. (1998): "Coalitional Stability under Perfect Foresight," *Economic Theory*, 11, 603–627.