A Proposal to Reform the Kyoto Protocol: the Role of Escape Clauses and Farsight*

Larry Karp
University of California, Berkeley

Jinhua Zhao
Iowa Sate University

February 23, 2007

Abstract

A reform to the Kyoto Protocol that allows signatories to pay a fine instead of meeting the target level of abatement would achieve three 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. Third, it would make it easier to force signatories to comply with their obligations. 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". We compare our results to a prominent model of International Environmental Agreements (IEAs) which finds that the equilibrium size of an IEA tends to be small when the benefits of cooperation are large. We show that with an escape clause and a properly chosen level of fine, a large IEA can be formed even when the benefits of cooperation are large.

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

JEL classification numbers C72, H4, Q54

*We benefitted from comments from Roger Guesnerie, and from seminar participants at Montreal, UC Berkeley, Texas A&M, the University of Central Florida, the University of Nebraska at Lincoln, the Heartland Environmental and Resource Economics Conference 2006, and the Public Economics Theory Conference 2006 .

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 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. We propose a reform to the Kyoto Protocol that allows signatories to avoid achieving the target level of abatement upon payment of a fine. This "escape clause" would achieve three 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. Third, it would simplify the problem of enforcing signatories to comply with the treaty obligations. We study the second goal, specifically the participation game under an escape clause using both a Nash equilibrium solution 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 for only five years; the duration of its successor is also likely to be fairly short. During this period, we will learn much about 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)) treat an IEA as the non-cooperative Nash equilibrium (NE) of 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 the second stage (the abatement game), the IEA chooses the signatories' abatement level. The critical assumption is that when nations decide whether to participate in the IEA they anticipate that in the next stage the IEA's abatement decision will maximize members' joint welfare. Hereafter, by "standard model" we mean the model in which the abatement decision maximizes members' joint welfare conditional on the number of IEA members.

An important result from the standard model is that the equilibrium size of the IEA tends to be small in circumstances where the potential benefit from cooperation is large. Holding fixed the benefit of abatement, the potential benefit of cooperation is large when abatement costs are small. Other things equal, increased efficiency and reduced abatement cost (weakly) *reduces* the equilibrium size of the IEA, potentially reducing the welfare. Section 4 discusses this somewhat counter-intuitive result.

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 enable Kyoto members to achieve a given level of abatement more cheaply. After initial resistance, Kyoto signatories accepted the use of tradeable permits, joint implementation, and the clean development mechanism.[1] Another possible policy uses tradeable permits with a price ceiling; a regulator has the power to increase the allocation of permits to defend the price ceiling (Kopp, Morgenstern, Pizer, and Ghersi (2002), Victor (2003)). The current carbon reduction agreement amongst northeastern U.S. states uses a similar arrangement. This policy caps abatement costs and therefore reduces expected costs. If applied to Kyoto, this kind of policy would have 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 inherent difficulty of inducing sovereign nations not to free-ride, rather than design flaws such as the failure to accommodate the possibility of unexpectedly high abatement costs. Design changes that reduce these costs might even be counterproductive. A corollary, discussed by Barrett (2003) (Chapter 15) and Stiglitz (2006), is that a successful IEA requires some kind of external punishment. One proposal involves reforming the World Trade Organization to permit the use of trade sanctions against countries that do not abide by a climate change IEA.

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 unrealistic assumption that the IEA makes the abatement decision conditioned on membership size, in order to maximize members' joint welfare. We modify the standard model by replacing the assumption that signatories commit (only) to maximizing members' joint welfare, with the assumption that they can sign a simple binding contract. In our setting, with uncertain abatement costs, the IEA agreement is a contract that contains an escape clause. This contract consists of two parameters, a prescribed level of abatement and a cost of exer-

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

cising the escape clause (a "fine") that exempts the signatory from the requirement to abate. Revenue from the escape clause payments are returned equally to all signatories, except for a transactions cost. Given these terms of the contract, in the first stage nations decide whether to join the IEA (the participation game), understanding that the contract will be binding on signatories. In the second stage (the abatement game) each signatory observes the outcome of the participation game (the number of signatories) and learns its abatement cost, which cannot be verified by courts. The signatory then decides whether to abate or to invoke the escape clause.

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 model's description of an IEA. Our proposal differs from Kyoto by including the escape clause. This modification has the three desirable effects mentioned above: it protects signatories from unexpectedly high abatement costs, and thus answers one of the U.S. objections; unlike other cost-reducing reforms, such as trade in permits, the escape clause can increase equilibrium membership;[2] and it makes it easier to enforce signatories' compliance. The first effect is obvious. Our paper focuses on the second effect, and briefly discusses the third effect in the conclusion.

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 simple 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.

The standard model, and our first set of results, uses the notion of subgame perfect Nash equilibrium. Modern games of coalition formation, including Chwe (1994), Mariotti (1997), Xue (1998) and Ray and Vohra (2001), are based on a more sophisticated interpretation of rationality, in which agents understand how their provisional decision to join or leave a coalition would affect other agents' participation decisions; nations are "farsighted". Diamantoudi and Sartzetakis (2002) adopt a similar notion of foresight to study IEAs, without establishing a precise relationship with the earlier theoretical literature. Eyckmans (2001) and de Zeeuw (2005) apply this notion to IEA models. Our description of farsighted nations builds directly on Chwe (1994). We show that farsighted nations may be able to form larger IEAs.

Section 2 sets out our model. Section 3 analyzes the one-shot NE, and Section 4 compares our model with the standard model. Section 5 studies the non-cooperative participation game

[2] 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 nations are farsighted. In all of these settings, the equilibrium is subgame perfect.

2 The Model

Each of N homogenous nations decides whether to join an IEA to reduce a global pollutant.[3] 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.[4]

In the case of 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), changes in the flow of GHGs will not lead to significant changes in the stock. The first order approximation of expected damages equals a constant plus $b(dS)$, where dS is the change in the stock, i.e., the flow. This approximation is "adequate" if dS is very small – as is the case in our setting.[5] Hereafter we choose units of the value of abatement so that $b = 1$.

When nations decide whether to join the IEA, they do not know their true abatement costs. 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 the probability that $\theta_i = \theta_H$ is $p \in (0, 1)$. The distributions of θ_i, $i = 1, \ldots, N$, are independent.

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, and p are common knowledge. In stage one, nations play a *participation game* in which they simultaneously 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 5). The outcome of this game is a partition of nations between signatories and non-signatories. Nations understand how their participation decision affects the equilibrium outcome in stage two. In stage two, each nation

[3] We assume that there is at most only *one* IEA for this particular pollutant.

[4] We ignore leakage, which decreases the benefit resulting from a nation's abatement. Leakage might cause the total benefit to be either concave or convex in the number of signatories who abate. Thus, our assumption of linearity is "neutral."

[5] A growing body of literature that uses a second order approximation of damages to compare the use of taxes and quantity restrictions, points out that the estimated expected damage function (for GHGs) is relatively flat even for substantial changes in stocks (Hoel and Karp 2001).

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, except for a fraction $0 < 1 - \phi < 1$ that is lost to transactions costs.[6] Each of the signatories receives a transfer of $\frac{M-m}{M}\phi F$.

2.1 Stage two: the abatement game

We assume that the required abatement level of the IEA exceeds the individually rational abatement even when $\theta = \theta_L$. This assumption is equivalent to

$$\theta_L > 1. \tag{1}$$

This inequality implies that in the abatement game, a non-signatory has a dominant strategy of not abating. Non-signatories never abate regardless of their cost realizations.

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 $M \in \mathcal{N} \equiv \{0, 1, \ldots, N\}$. Its strategy is a mapping from $\Theta \times \mathcal{N}$ to its action set, {do not abate, abate}. 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 the two actions breaks the tie by abating.

2.2 Types of NE

We identify the combinations of F and M that support each of the three types of NE. As we show below, each signatory's optimal decision depends on its own cost realization and on the *number* of other signatories, but not on the *actions* of other signatories. Therefore, conditional on M, each signatory has a dominant strategy.[7] The net benefit of abating given cost θ is $1 - \theta$

[6] We explored the possibility of sharing the fine revenue only amongst the abaters, rather than amongst all signatories. That version of the model is sensitive to the rule that determines how the fine revenue is allocated when no member abates.

[7] It is important that costs be non-verifiable or non-contractable, so that nations are unable to sign contracts that condition their abatement action on their cost realization. Since agents have dominant strategies in the abatement game, it does not matter if costs are private or public information.

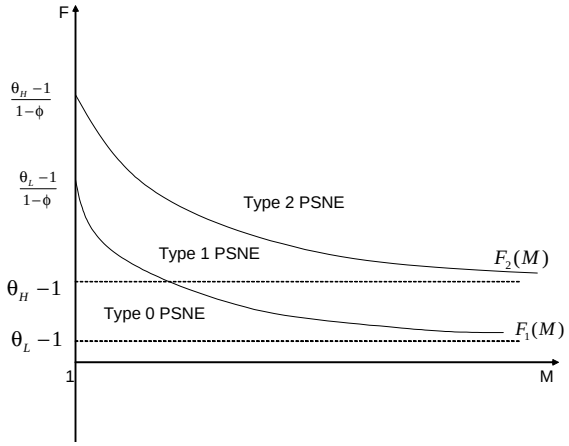

Figure 1: Three types of Nash Equilibria

(the additional benefit of a nation's own abatement minus the cost of abatement); the net benefit of not abating is $-F + \frac{\phi F}{M}$ (the fine minus the rebate). These two are equal at $F = M\frac{\theta_i - 1}{M - \phi}$. Signatory i will not abate given cost θ_L if and only if

$$F < F_1(M) \equiv \frac{\theta_L - 1}{1 - \phi/M}, \tag{2}$$

and it will abate given cost θ_H if and only if

$$F \geq F_2(M) \equiv \frac{\theta_H - 1}{1 - \phi/M}. \tag{3}$$

Since nations are symmetric and they have dominant strategies, 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 $F_2(M)$. The figure shows the regions of the different types of NE, conditional on M and F. For brevity, we will sometimes refer to a "type i IEA" to mean an IEA that results in a type i NE at the abatement stage, $i = 0, 1, 2$. 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 $(\phi = 0)$, the cost of exercising the fine and therefore the cost of joining the IEA 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 (before learning the cost realizations) of a signatory and a non-signatory in a type i IEA with M members and fine F. (Subscript s denotes "signatory" and n denotes "non-signatory.") Since no signatories abate in a type 0 NE, $\pi_{s,0} = -(1-\phi)F$ and $\pi_{n,0} = 0$.

Calculations reported in Appendix A show the expected payoffs in a type 1 NE:

$$\pi_{s,1}(M;F) = M(1-p) - \left((1-p)\theta_L + Fp(1-\phi)\right), \tag{4}$$

$$\pi_{n,1}(M) = M(1-p). \tag{5}$$

In a type 1 equilibrium, the expected fraction of signatories that abate is $(1-p)$, resulting in an expected abatement benefit of $M(1-p)$. This value equals the expected payoff of a non-signatory (who never abates). A signatory's expected abatement cost is $(1-p)\theta_L$ and its expected fine payment net of reimbursements is $Fp(1-\phi)$.

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

$$\pi_{s,2}(M) = M - \bar{\theta}, \tag{6}$$

$$\pi_{n,2}(M) = M \tag{7}$$

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

The "membership cost" of the IEA is the reduction in expected payoff for a non-signatory that decides to join the IEA, under the assumption that other nations' decisions remain fixed. Inequality (1) implies that the membership costs ($(1-\phi)F$ in a type 0 IEA, $(1-p)(\theta_L - 1) + Fp(1-\phi)$ in a type 1 IEA, and $\bar{\theta} - 1$ in a type 2 IEA) are always positive, i.e.

$$\pi_{n,i}(M) - \pi_{s,i}(M+1;F) > 0, \qquad i = 0, 1, 2. \tag{8}$$

Consequently, a nation wants to join an IEA *only if by joining, it changes the abatement equilibrium*, e.g. from a type 0 to a type 1 or from a type 1 to a type 2 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.

3 The participation game: Nash equilibrium

This section describes the subgame perfect NE to the participation game. We show that for an interval of fine F, a reduction in membership costs (smaller F) increases equilibrium membership size. We then discuss the welfare implications of different fines.

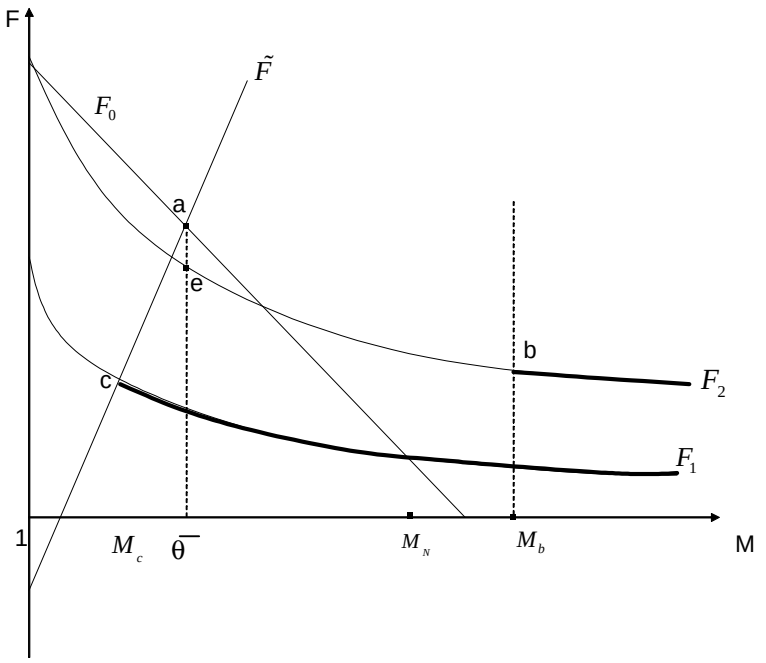

Figure 2: The NE to the Participation Game without Integer Constraints.

On $\tilde{F}$ type 1 IEA members have 0 payoff. On F_0 members have equal payoffs in type 1 and type 2 IEAs.

3.1 The subgame perfect NE

Here we consider the NE to the participation game when the existence of the escape clause and the size of the fine are taken as given. Figure 2 reproduces Figure 1 and includes the graph of $\tilde{F}(M) \equiv (M - \theta_L)\frac{p-1}{p(\phi-1)}$, the locus of points at which signatories in a type 1 IEA have zero payoff. Signatory payoffs are negative to the left of this line. The Figure includes a vertical line at $M = \bar{\theta}$; note that $\pi_{s,2}(\bar{\theta}) = 0$ so any type 2 equilibrium with fewer than $\bar{\theta}$ members gives signatories a negative payoff. It also includes a vertical line at $M_b \equiv \frac{\bar{\theta}-1+p}{p}$, derived from $\pi_{s,2}(M_b) = \pi_{n,1}(M_b - 1)$: a nation is indifferent between being a signatory to a type 2 IEA of size M_b and a non-signatory to a type 1 IEA with $M_b - 1$ members. If a type 2 IEA is smaller than M_b, a member has incentive to leave the IEA if its defection switches the IEA from type 2 to type 1.

This model contains five parameters: the number of nations N, the cost parameters θ_L and θ_H, the probability of a high cost p, and the transactions cost parameter ϕ. There are many combinations of parameter values that lead to different NE configurations. Since a complete taxonomy would be uninteresting, we discuss the NE set under assumptions that are reasonable for the control of GHG, represented by Figure 2. These assumptions ensure that the horizontal distance between the curves F_1 and F_2 is sufficiently great. Appendix A shows

that the following are sufficient conditions for the configuration that we describe: (a) there is a non-negligible difference between high and low costs, (b) the probability of high abatement costs is less than 0.5, and (c) θ_L is at least 4. Given our normalization, condition (c) means that an IEA needs at least four members in order for abatement in the low cost state to improve the members' welfare.

The "integer constraint" states that the number of members in any IEA, and the number of defectors from any proposed equilibrium, must be an integer. The integer constraint greatly complicates the analysis, without adding much economic insight. The formal analysis in Appendix A respects the integer constraint, but here (for simplicity) we ignore it. For example, if we state that a point x on the curve F_1 is a NE, we mean that when $F = F_x$ (the vertical coordinate of point x) then an IEA consisting of the smallest integer no less than $F_1^{-1}(F_x)$ is a NE. With this understanding, the heavily shaded curves in Figure 2 "indicate" the NE correspondence for positive F. As we show in Appendix A, respecting the integer constraint brings two modifications to the NE correspondence. First, the actual NE correspondence includes a step function on or to the right of these heavy lines. Second, the NE set could be slightly larger (corresponding to one more possible membership size for each of the two types).

Four facts help identify the equilibrium set. First, a membership size of 0 (not shown in Figure 2) is always a NE. Second, in view of inequalities (1) and (8), a member is deterred from leaving the IEA only if its defection would change the equilibrium type 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 eliminates all points not on curves F_1 and F_2 as candidates. Third, an outcome with negative expected payoffs for IEA members cannot be a NE to the participation game. This fact eliminates all points above c on F_1 (points to left of $\tilde{F}$) and all points above e on F_2 (where $M < \bar{\theta}$). Fourth, points between e and b on F_2 are not NE since a member wants to defect: defection switches the IEA from type 2 to type 1 and raises the defector's payoff, because $M < M_b$.

From the discussion preceding equation (8), a signatory's expected cost of joining an IEA (i.e., the membership cost) is weakly increasing in F. Figure 2 shows that smaller fines increase the equilibrium size of the IEA when $F < F_c$ (for type 1 IEAs) and when $F < F_b$ (for type 2 IEAs). In both of these cases, a reduction in the fine (weakly) decreases membership cost while (weakly) increasing equilibrium membership size. We regard this as a "natural" comparative statics result: lower membership costs should promote membership. As we note in Section 4, the standard model implies that lower membership costs reduce membership.

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 become flat lines. In this case, a member's equilibrium

action in the abatement stage is independent of the number of members. Then nations have no incentive to join the IEA in the participation stage, because their decision has no effect on other members' behavior; the equilibrium size of the IEA is 0.

3.2 Welfare implications

In this section we allow nations to choose the fine in period 0 in order to maximize the expected aggregate welfare. We denote the resulting outcome and the corresponding level of welfare as the *NE optimum*, because it is chosen with the understanding of how the participation and the abatement games unfold. Since agents are homogenous before the realization of individual costs, there would be no reason to disagree on the fine. We assume that $N > \theta_L$, so that it is socially optimal to have a low cost nation abate.

Even when nations can choose F in stage 0, there are two obstacles to achieving the first best outcome: the participation constraint that arises from non-cooperative behavior in the participation game, and the incentive compatibility constraint that occurs because costs are not verifiable in the abatement game.[8] We characterize the NE optimum and compare it to two benchmarks. In the *full information first best* benchmark, a social planner observes costs and instructs nations whether to abate. This social planner overcomes both the participation constraint and the incentive compatibility constraint. In the second benchmark, the *constrained information second best*, the social planner is able to require all nations to join the IEA (i.e., it overcomes the participation constraint), but abatement costs are private information and thus, as in our model, the incentive compatibility constraint has to be satisfied.

Define the function $F_0(M) \equiv \frac{\theta_H - M}{1-\phi}$ (graphed in Figure 2) as the locus of points at which signatories' payoffs are equal in type 1 and type 2 IEAs: $\pi_{s,1}(M; F_0(M)) = \pi_{s,2}(M)$. Below this line, signatories' payoffs are higher in a type 1 equilibrium. The point M_N satisfies $F_0(M_N) = F_1(M_N)$, as shown in the figure. The horizontal coordinate of point c, denoted M_c, satisfies $\tilde{F}(M_c) = F_1(M_c)$.

Since the nations are *ex ante* identical, the IEA consists of all nations or of zero nations in each of the three optima. Depending on the cost structure and the number of nations, there are three possible outcomes in the abatement stage: nations abate for all cost realizations, only low cost nations abate, and no nation abates. The first two cases correspond to an IEA of N nations while in the last case, an IEA does not exist. Figure 3 shows the abatement patterns for

[8] A slightly more complicated mechanism than the constant fine that we propose can solve the participation constraint. For example, if the fine is very small for $M < N$ and very large for $M = N$, nations know that in equilibrium there would be no abatement unless all nations join. In this case, a type 2 IEA with $M = N$ is a NE provided that $N \geq \bar{\theta}$. If both the fine and the reimbursement depend on M, it is possible to solve the participation constraint and to induce a type 1 NE.

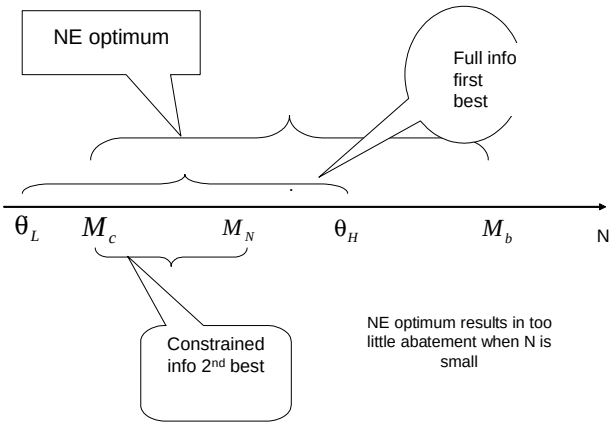

Figure 3: Abatement patterns depending on N

different ranges of N (identified by three brackets, one for each of the three different scenarios) When N falls within a bracket, only low cost nations abate. For all values of N above the bracket, nations abate for both high and low costs. For values of N below the bracket, nations do not abate for either high or low costs and thus an IEA is not formed. Lemma 4 in Appendix A establishes the ordering of the five points on the N axis.

The top bracket in Figure 3 means that in the NE optimum, for $N < M_c$, no nation abates; for $M_c \leq N < M_b$, only low cost nations abate; and for $M_b \leq N$, both high and low cost nations abate. From Figure 2, the smallest IEA that can be supported as a NE is M_c. For $N < M_c$ the unique NE contains 0 members, regardless of F. For $N > M_c$, it is feasible, and optimal, to induce all nations to participate in the IEA, so the only question is whether this is a type 1 or a type 2 IEA. The proof of the following Proposition shows that for $N > M_c$, a type 2 IEA has higher welfare than a type 1 IEA (both with N members) if and only if $N > M_N$; however, it is feasible to induce a type 2 IEA in the participation game if and only if $N \geq M_b$.

Proposition 2 *If F can be chosen at stage 0, it is feasible to induce a type 2 equilibrium at the abatement stage if and only if $N \geq M_b$. If it is feasible to induce a type 2 equilibrium, it is optimal to do so.*

The middle bracket in Figure 3 indicates that in first best full information scenario only low cost nations abate when $N < \theta_H$ and all nations abate for $N \geq \theta_H$. The optimality of this outcome is obvious.

The lowest bracket in Figure 3 means that in the constrained information second best outcome, no nation abates if $N < M_c$; for $M_c \leq N < M_N$, only low cost nations abate; and for

$M_N \leq N$, all nations abate. In the constrained information second best, the social planner can require nations to join the IEA. Thus, by choice of the fine the planner can induce either a type 1 or a type 2 IEA consisting of all nations. However, if $N < M_c$, a positive fine leads to a negative expected payoff for any type of IEA,[9] so it is (constrained) optimal to have no nation join the IEA (or equivalently, set the fine equal to 0). It is optimal to induce a type 2 equilibrium if and only if $N \geq M_N$, just as with the NE optimum. The social planner's ability to require nations to join the IEA eliminates the participation constraint. Thus, all nations abate in the constrained information second best outcome when $N \geq M_N$.

Comparison of these three outcomes shows the separate effects of the participation constraint and the incentive compatibility constraint. For example, suppose that we begin with the full information first best, and then add (only) the incentive compatibility constraint. If $N < M_c$, this change eliminates abatement, which is positive in the first best outcome (where only low cost nations abate). However, if $M_N \leq N < \theta_H$, the incentive compatibility constraint increases the amount of abatement, because it causes all nations to abate, not only those nations with lows costs (as in the first best outcome). Both of these results are due to the fact that a type 1 equilibrium requires that high cost nations incur a transactions cost (i.e., the fine from invoking the escape clause is not fully rebated to the signatories), leading to lower welfare than in the first best outcome. In contrast, there are no transactions costs in a type 2 equilibrium. The desire to avoid this transactions cost means that in some circumstances (where the first best optimum has only low cost nations abating) the information-constrained social planner induces no nations to abate, and in other circumstances it induces all nations to abate. However, if $\phi = 1$, in which case the fine is fully rebated to the signatories without any transaction cost, we can verify that $M_c = \theta_L$ and $M_N = \theta_H$ ($\tilde{F}$ and F_0 become vertical lines at θ_L and θ_H respectively). Thus, without transaction costs, the full information first best and constrained information second best coincide.

If we maintain the informational constraint and now add the participation constraint, the amount of abatement falls when $M_N < N \leq M_b$. Here, the addition of the participation constraint causes the outcome to change from a grand type 2 IEA to a grand type 1 IEA. The participation constraint makes the grand type 2 IEA infeasible for an intermediate range of N.

In summary, if $N < M_c$ the incentive compatibility constraint is so tight that no nation abates, regardless of the fine, in the NE and constrained information second best outcome. If $M_c \leq N < \theta_H$, the optimal fine induces the socially optimal amount of abatement in both cases, but it does so by incurring a transactions cost. If $\theta_H \leq N < M_b$ the NE optimum results

[9] If F and N lead to a type 0 IEA (i.e., (F, N) is below curve F_1), the expected payoff is negative for any $F > 0$. If (F, N) leads to a type 1 IEA, the payoff is negative because (F, N) is to the left of $\tilde{F}$. For a type 2 IEA, the payoff is negative because $N < \bar{\theta}$.

in too little abatement, relative to the first best, and there is a transactions cost. If $N \geq M_b$ the optimal fine achieves the first best in the NE optimum.

4 Comparison with the "standard model"

The standard model assumes that the IEA determines abatement levels *after* nations have made their participation decisions, and the abatement levels are chosen to maximize members' joint expected welfare at the abatement stage. To be consistent with our setting, we again assume that abatement is a binary choice, perhaps because of some technological constraint. We first consider the simplest case where the IEA (which cannot verify cost realizations) sets an abatement level (0 or 1) for every signatory, without an escape clause. That is, the IEA decides whether to abate, independent of cost realizations, in order to maximize the expected welfare of the signatories.[10] We then consider the situation in which the IEA is able to use an escape clause, and that (unlike in our model) it chooses the fine conditional on M, to maximize members' welfare. These two settings differ only in the policy menu available to the IEA, not in the timing of the decisions. Conditional on membership size, the IEA's payoff is (weakly) higher when it has the option to use an escape clause. However, in the standard model, by reducing the membership cost, the escape clause lowers the equilibrium membership size and reduces global welfare. We compare the outcome in the standard model under the two policy menus to the outcome in our setting, i.e., when the IEA can commit to the *ex ante* levels of abatement and the fine for invoking the escape clause.

4.1 The IEA chooses whether to abate conditional on membership

We define the function $h(y)$ as the smallest integer not less than y. Suppose 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

$$\Pi(M) = \max\left\{0, M - \bar{\theta}\right\}. \tag{9}$$

[10] With cost uncertainty, the joint welfare maximization assumption means that high cost members' payoff, $h\left(\bar{\theta}\right) - \theta_H$ is negative for many combinations of parameter values. Thus, with cost uncertainty at the participation stage, the assumption of joint welfare maximization at the abatement stage requires that signatories are able to commit to taking a future action that is not in their *ex post* interest. In our model, the signatories to an IEA are not able to solve the collective action problem. However, they are able to commit to following a simple contingent contract. In a type 2 equilibrium, or in a type 1 equilibrium with a sufficiently low fine, a high cost signatory's payoff is greater than $h\left(\bar{\theta}\right) - \theta_H$. Thus, with cost uncertainy, the commitment problem is more severe in the standard model than in our model.

In the abatement stage, the IEA chooses to abate if and only if $M \geq \bar{\theta}$. In this case, the unique NE to the participation game is $M = h\left(\bar{\theta}\right)$. To confirm this, note that if there are $h\left(\bar{\theta}\right)$ members, each signatory's expected payoff is non-negative; no signatory wants to defect, because the resulting IEA would choose not to abate, leaving the defector with a zero payoff. No non-signatory wants to join, since in view of inequality (1), $h\left(\bar{\theta}\right) > h\left(\bar{\theta}\right) + 1 - \bar{\theta}$: the non-signatory's payoff in the NE exceeds its payoff if it joins the IEA. Further, $M > h(\bar{\theta})$ cannot be an NE because a signatory has incentive to defect: since $M - 1 \geq h(\bar{\theta})$, the remaining IEA still chooses to abate, leaving the defector with a higher payoff.

In this model, the membership cost (the expected abatement cost minus the benefit of increased abatement) equals $\bar{\theta} - 1$. The level of membership, $h\left(\bar{\theta}\right)$, weakly increases with the membership cost. When $N \geq h\left(\bar{\theta}\right)$, global welfare (the sum of member's and non-members' welfare) in the NE is $\left(N - \bar{\theta}\right) h\left(\bar{\theta}\right)$. As $\bar{\theta}$ increases between integers welfare falls, but welfare has an upward jump as $\bar{\theta}$ passes through an integer value. If all nations were compelled to join the IEA, global welfare (the "potential gain" from cooperation) is $\left(N - \bar{\theta}\right) N$. Relative to this grand IEA, the fraction of potential welfare achieved in equilibrium is $\frac{h\left(\bar{\theta}\right)}{N}$, a non-decreasing function of expected costs. This example illustrates why the standard model leads to a pessimistic view of IEAs: they achieve a substantial fraction of potential gains from cooperation (i.e. $h\left(\bar{\theta}\right)$ is close to N) only when potential gains are small (i.e. $\bar{\theta}$ is close to N). IEAs are effective only when they are unimportant.

Consider now a setting where the IEA commits to the abatement decision *before* agents decide whether to join (as is the case for Kyoto). In this case, the equilibrium membership size is 0: there is nothing to offset nations' temptation to free-ride since even if a signatory defects, the remaining signatories still abate. (Kyoto solved this free-rider problem by stipulating that the agreement would not enter into force unless a minimum level of participation was achieved.) Thus, without the escape clause, moving the abatement decision from before the participation stage to after the participation stage increases the equilibrium IEA size from zero to $h\left(\bar{\theta}\right)$.

4.2 The IEA chooses the fine conditional on membership

Suppose now that the IEA is able to use an escape clause with a fine, while leaving the abatement decisions to the signatories. As in our model, decisions in the abatement stage are made non-cooperatively. In this variation of the standard model, the IEA chooses the fine F *after* the participation stage. If it is optimal for an IEA of size M to induce a type 1 equilibrium, it chooses the smallest fine that will achieve this, $F_1(M)$. If it is optimal to induce a type 2 equilibrium, the IEA sets $F \geq F_2(M)$. To induce a type 0 equilibrium, the IEA sets the fine at

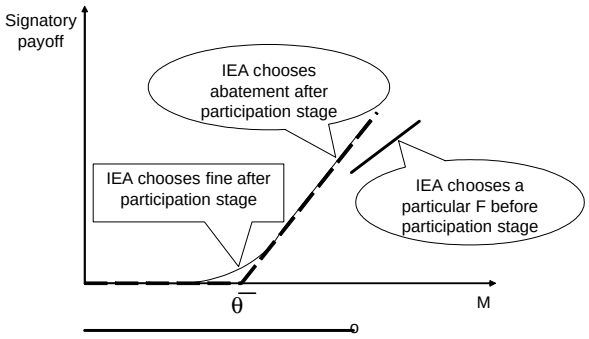

Figure 4: Signatory payoffs in different IEA games

zero. Thus, the expected payoff of a member is

$$\hat{\Pi}(M) = \max\left\{0, \pi_{s,1}(M; F_1(M)), \pi_{s,2}(M)\right\}, \tag{10}$$

where $\pi_{s,1}$ and $\pi_{s,2}$ are given in (4) and (6). It is straightforward to show that for a sufficiently small M it is optimal to set $F = 0$, and for sufficiently high M it is optimal to set $F \geq F_2(M)$. For a range of intermediate values of M it is optimal to induce a type 1 equilibrium at the abatement stage.

Equations (9) and (10) imply $\hat{\Pi}(M) \geq \Pi(M)$. 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. Let $\hat{M} = \sup\{M \mid \hat{\Pi}(M) = 0\}$. Following the same argument as in the previous case without the escape clause, we know that the equilibrium membership size is $h(\hat{M})$. From (10) and the fact that $\pi_{s,2}(\bar{\theta}) = 0$, we know $\hat{M} \leq \bar{\theta}$, i.e., $h(\hat{M}) \leq h(\bar{\theta})$: a reduction in membership costs (weakly) reduces equilibrium membership size and may reduce equilibrium global welfare. (An example is available on request.)

Figure 4 compares the payoffs under the two policy instruments, conditional on M. The thick-dashed lines represent the expected payoff of a signatory when the IEA chooses the abatement level, and the thin solid curve represents the payoff when the IEA chooses the fine.

Figure 4 also shows (the heavy solid line) the signatory's payoff in our model, where the IEA chooses the fine before the participation stage. In this example, the (possibly non-optimal) fine induces a type 1 NE. The equilibrium to this participation game occurs at the point of discontinuity of the solid line, a level greater than $\bar{\theta}$. Both member and non-member welfare is higher when the fine is chosen optimally before the participation decision, compared to when either the fine or the abatement level is chosen conditional on membership.

4.3 Summary of the comparison

Table 1 compares the equilibrium IEA size for games with different timing and different policy instruments. When the policy instrument is the abatement level (0 or 1), choosing this level after nations' participation decision (as in the standard model) increases the IEA membership and aggregate welfare. If the policy instrument is the escape clause with a fine, it is better to make the choice before the participation decision, as in our model.

	IEA Policy Instrument	
Decision Timing	Abatement Level	Level of Fine
Before participation	$M^* = 0$	M^* can be large
After participation	$M^* = h(\bar{\theta}) \geq 2$ (small)	$M^* = h(\hat{M}) \leq h(\bar{\theta})$ (even smaller)

Table 1: Equilibrium IEA size in four different games

If the IEA must make its choice before the participation decision, a fine leads to a higher payoff. If the IEA must make its choice after the participation decision, choosing abatement leads to a higher payoff. Of the four possibilities, both membership and aggregate welfare are highest when the IEA chooses a fine before the participation decision, the situation modeled in our paper.

Nations are willing to join the IEA only if they think that their membership will change the actions of *other* members. A well-designed IEA can promote membership by increasing a nation's leverage on other members.

5 Farsighted stability in the participation game

The Nash equilibria in the previous sections 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. Here we relax those assumptions.

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 candidate equilibrium 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' responses – 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 fewer than $\bar{\theta}$ members, and lying to the left of the curve $\tilde{F}$. In that outcome, each signatory's payoff is negative regardless of the type of the resulting IEA, 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 at $h(\bar{\theta})$. Thus, no signatory to the IEA of size $h\left(\bar{\theta}\right)$ wants to withdraw. We show that when nations are farsighted, this IEA is stable; in Section 3 we saw that it is not a NE.

The second implicit NE assumption, that each nation acts on its own, may or may not be reasonable depending on whether nations can credibly coordinate *in the negotiation process* before the IEA is signed. 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. 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 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.

Here we assume that the nations are farsighted. Depending on whether pre-IEA binding agreements are possible, we analyze the participation game under different assumptions:

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

Assumption 2 (Coalitional Farsight) *Coalitional deviations are allowed.*

We first obtain the stable set to the participation game under Assumption 1. We then show that a particular element of this set corresponds to the stable IEA under Assumption 2.

5.1 Unilateral farsighted stable set

The Nash equilibrium does not predict the kinds of IEAs that will form in the participation game when nations are farsighted. Chwe (1994)'s farsighted stable set (FSS) provides reason-

able prediction for this kind of behavior. 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 outcome b to a. We define a dominance relation between two outcomes that allows each nation to act 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, $z_0, \ldots, z_m$ with $z_0 = a$ and $z_m = b$ and nations indexed by $0, \ldots, m-1$ such that $z_j \rightarrow_j z_{j+1}$ and $z_j \prec_j b$ for all $j = 0, \ldots, m-1$.*

That is, starting with outcome a, m nations make sequential unilateral changes, generating a sequence of intermediate outcomes, $z_1 .... z_{m-1}$. Each nation j in the sequence prefers the final outcome b to z_j, the interim outcome that it faces. 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 the 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*: $\forall b \in Z \backslash V$, $\exists a \in V$ 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 this IEA is an element of 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 stable outcomes, and excludes all the outcomes that are indirectly dominated by some other stable 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 farsighted agents, and 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. (Details available on request.) 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.

We consider only outcomes in which there is a single IEA. Nations decide whether to join the IEA, rather than deciding which IEA to join. Section 5.3 discusses this restriction. Since the nations are *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, $Z = \{1, 2, ...N\} \equiv \mathcal{N}$, and each nation is either a signatory or a non-signatory. This observation simplifies the determination of the 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 $m = M, M-1, \ldots, M'+1$.
(ii) Suppose $M < M'$. Then $M \ll M'$ if and only if $\pi_n(m; F) \leq \pi_s(M'; F)$ for all $m = M, M+1, \ldots, M'-1$.

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 want to join the IEA, because otherwise circular decisions arise, resulting in large negative payoffs.

In searching for the UFSS of the participation game, we continue to use the NE outcomes in the abatement stage. In our setting, the unique NE of the abatement state consists of dominant strategies. One nation's abatement decision does not affect the relative payoffs of the two actions for other nations. Consequently, the NE to the abatement game coincides with the stable set in that game. We thus only need to find the UFSS for the participation game.

5.2 Finding the UFSS

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 $M^1 < N$, the set $\mathcal{M}(M^1)$ generated by M^1 is a finite and strictly increasing sequence of integers, indexed by M^j, $j = 1, 2, \ldots$, all less than or equal to N, with*

$$M^j = h(m^j), \quad \textit{where } m^j = \min\{m \in \mathcal{R} : \pi_s(m) \geq \pi_n(M^{j-1})\},\ j \geq 2. \tag{11}$$

The sequence $\mathcal{M}(M^1)$ depends on F, but we suppress that argument. Given an IEA of size M^1, the set $\mathcal{M}(M^1)$ is generated by a simple sequence of comparisons. Starting with $M^1 = M^1$, the next element M^2 is the smallest IEA such that a signatory's payoff in M^2 is no less than the non-signatory's payoff in M^1. Once we identify M^2, the next element M^3 is found through the same procedure. We repeat this process until the greatest possible element is reached.[11]

If we know the smallest element of the stable set V, we can simply set it equal to M^1 and use (11) to construct V. In Appendix A (Proposition 6) we formally prove that $\mathcal{M}(M^1)$ coincides with V if M^1 is appropriately chosen. Here we describe the intuition. Equation (11) and the monotonicity of $\pi_n(\cdot)$ and $\pi_s(\cdot)$ implies that, for any given M^j and M^{j+1}, $j = 1, 2, \ldots$, (i) a signatory's payoff at M^{j+1} is not less than a non-signatory's payoff at M^j, and (ii) a signatory's payoff at any IEA (with size) between M^j and M^{j+1} is lower than a non-signatory's payoff at M^j. Observation (i) means that M^{j+1} is not indirectly dominated by M^j, and (ii) means that all IEAs with sizes between M^j and M^{j+1} are indirectly dominated by M^j. Therefore, if we can establish that M^j is not indirectly dominated by M^{j+1}, and if M^1 is the smallest element of V, we know $\mathcal{M}(M^1)$ is both internally and externally stable and $\mathcal{M}(M^1) = V$.

In Appendix A (Proposition 6) we show how the smallest element of V, M^1, can be identified. Essentially, M^1 is chosen so that (i) for all $j = 1, 2, \ldots$, M^j is not indirectly dominated by M^{j+1}, and (ii) if $M^1 > 0$, IEAs smaller than M^1 are indirectly dominated by M^1 (for external stability). Below we use an example to illustrate the structure of a stable set and how the stable sets depend on the value of F. For each value of F, we verify stability of $\mathcal{M}(M^1)$ by showing that conditions (i) and (ii) are satisfied. In showing (i), we repeatedly use the following fact: if IEAs M^{j+1} and $M^{j+1} - 1$ are of the same type, then M^j is not indirectly dominated by M^{j+1}. In order for $M^j \ll M^{j+1}$, we must have $M \ll M^{j+1}$ *for all* $M^j \leq M < M^{j+1} - 1$ (from

[11] The fact that payoffs are discontinuous when M increases across the boundary F_2 (because the outcome jumps from a type 1 to a type 2 equilibrium) means that the equality $\pi_s(m) = \pi_n(M^{j-1})$ might have either two or no solutions. This fact requires that we use the $\min$ operator and the weak inequality in the middle line of (11).

Lemma 1). But from (8), this sequence of inequalities is violated at $M = M^{j+1} - 1$ if M^{j+1} and $M^{j+1} - 1$ are of the same type.

Figure 5 illustrates an example. As in Section 3.1, the stable set depends on the relative positions of the curves F_0, F_1, F_2, $\tilde{F}$ and M_b. Figure 5 uses the definitions and assumptions in Figure 2 (and Figure 6 in Appendix A); in addition, it assumes: (a) point c is above the line $F_2(N)$;[12] (b) the horizontal distance between $\tilde{F}$ and F_1 at $F = F_{b'}$ is not less than one; and (c) 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:

Summary 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 $\mathcal{M}(M^1)$, where $M^1 = h(F_1^{-1}(F))$. All IEAs in the UFSS are of type 1.
(iii) If $F \in [F_2(N), F_{b'})$ where $F_{b'} = F_2(h(M_b) - 1)$, the UFSS is the set $\mathcal{M}(M^1)$, where $M^1 = h(F_2^{-1}(F))$. All IEAs in the UFSS are of type 2.
(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 $\mathcal{M}(M^1)$ where $M^1 = 0$. The second element of $\mathcal{M}(M^1)$ is $M^2 = h(\tilde{F}^{-1}(F))$.
(v) If $F \geq F_d$, the UFSS is the $\mathcal{M}(M^1)$ where $M^1 = 0$. The second element of $\mathcal{M}(M^1)$ is $M^2 = h(\bar{\theta})$.

Figure 5 graphs the first one or two elements of the UFSS, which are represented by the bold lines (ignoring the integer constraint). When $F < F_1(N)$ (case (i)), $M = 0$ indirectly dominates all other IEAs because any IEA with a positive membership leads to negative expected payoffs for its signatories. Consider now case (ii) when $F \in [F_1(N), F_2(N))$. (The same reasoning applies to case (iii) when $F \in [F_2(N), F_{b'})$.) Since all IEAs in $\mathcal{M}(M^1)$ are of the same type (type one), M^{j+1} and $M^{j+1} - 1$ are of the same type and thus M^j is not indirectly dominated by M^{j+1}, $j = 1, \ldots$. We only need to verify that M^1, determined from curve $F_1(\cdot)$ given F, indirectly dominates all smaller IEAs. But from the previous section we know the IEA of size M^1 is a Nash equilibrium, so that $\pi_{s,1}(M^1) \geq \pi_{n,1}(M^1 - 1)$. The inequality implies that $\pi_{s,1}(M^1) \geq \pi_{n,1}(M)$ for all $M \leq M^1 - 1$ since $\pi_{n,1}(\cdot)$ is increasing in M.

When $F \geq F_{b'}$, $M^1 = 0$. To verify stability, we only need to show that M^j is not indirectly dominated by M^{j+1}, $j = 1, \ldots$. When $F \geq F_d$ (case (v)), applying the procedure in (11) to $M^1 = 0$ means that $M^2 = h(\bar{\theta})$. Since IEAs larger than $M^2 = h(\bar{\theta})$ are all of type 2, M^j is not indirectly dominated by M^{j+1} for $j = 2, \ldots$. To show that $M^1 = 0$ is not indirectly dominated

[12] For $\phi \approx 1$, 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.

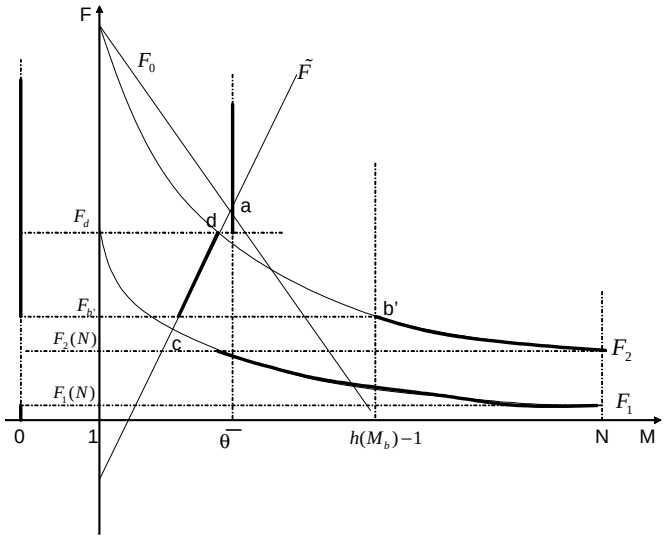

Figure 5: First Elements of the UFSS

by $M^2 = h(\bar{\theta})$, note that an IEA of size $M^2 - 1$ cannot be of type zero due to the assumption that the horizontal distance between $\bar{\theta}$ and F_1 at F_d is not less than one. If $M^2 - 1$ is of type 2, then $M^2 \ll M^2 - 1$. If $M^2 - 1$ is of type 1, the fact that $\bar{\theta} < M_b$ and the definition of M_b implies that $M^2 \ll M^2 - 1$. In both cases, M^1 is not indirectly dominated by M^2.

In case (iv) where $F \in [F_{b'}, F_d)$, applying (11) to $M^1 = 0$ implies that M^2 is on the line $\tilde{F}$ (or the integer form of it). The assumption that the horizontal distance between $\tilde{F}$ and F_1 is not less than one implies that $M^2 - 1$ is also of type 1, so that M^1 is not indirectly dominated by M^2. An additional complication arises when M^j is type 1 and M^{j+1} is type 2 for some $j \geq 2$. If $M^{j+1} - 1$ is also type 2, then $M^{j+1} \ll M^{j+1} - 1$ and thus M^j is not indirectly dominated by M^{j+1}. If $M^{j+1} - 1$ is type 1, the fact that the corresponding segment of F_2 lies to the left of $h(M_b) - 1$ in Figure 5 implies that $M^{j+1} < M_b$. Then from the definition of M_b, $M^{j+1} \ll M^{j+1} - 1$ and M^j is not indirectly dominated by M^{j+1}.

Summary 1 implies that $M^{j+1} - M^j > 1$.[13] 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.

[13]This claim follows directly from Definition 3 and (8) when successive elements of V result in the same type IEA. When successive elements result in different type IEAs as in cases (iv) and (v) of Summary 1, the claim follows from the discussion therein. For example, in case (v), if there is a j such that M^{j+1} is type 2 while M^j is type 1, it must follow that $M^{j+1} - M^j > 1$. The reason is that, from the discussion of case (v), $M^{j+1} < M_b$. Then if $M^{j+1} - M^j = 1$, the definition of M_b implies that M^j indirectly dominates M^{j+1}, violating internal stability.

5.3 The effects of unilateral farsight

Comparing Figures 2 and 5 shows that farsighted nations have greater ability to cooperate, leading to the possibility of larger IEAs for a given F. For example, Figure 5 shows that there is a range of F for which the IEA with no members (the "zero IEA") is not in V; the zero IEA is always a NE. For values of F such that Nash equilibria with positive membership exist (e.g., when $F \in [F_1(N), F_c)$ and $F \in [F_2(N), F_b)$), the NE with the largest membership also belongs to V. For sufficiently large values of F, the zero IEA is the only NE, but the are IEAs with positive membership in V. Although positive Nash equilibria are sometimes elements of V, there may be (and typically are) larger elements.

The concept of UFSS is silent on the negotiation procedure that leads to a particular IEA. Imposing more structure on the negotiation process may lead to finer predictions. Ray and Vohra (2001) study a negotiating protocol in a setting where there can be multiple IEAs (coalitions) and each coalition's action is conditioned on its own membership. This model shares an assumption of the "standard model": the action of the coalition/IEA is determined after nations decide whether to join. We considered imposing the Ray and Vohra's negotiating protocol onto our model, in which the fine and level of abatement are chosen before nations decide which (if any) coalition to join (so that payoffs are given in equations (4) – (7)). Each IEA collects fines only from its own members. The resulting equilibrium coalitional structure can simultaneously contain both type 1 and type 2 IEAs, and also have some nations outside any IEA.

Consider the example where the largest element of V is a type 2 IEA with $M = 10$, the smallest positive element is a type 1 IEA with $M = 3$, and $N = 14$ (so that one nation is outside any IEA). Suppose also $\theta_H > (1 - \phi) F$, so that costs to a type 1 IEA member are lower than costs to a type 2 IEA member. Since all nations have the same benefit of abatement, the payoffs to type 1 members are higher than to type 2 members, when both types of IEA coexist. (See equation (4) and (6).) In this example, under the Ray and Vohra protocol (with our payoffs) the first proposer would decide to remain outside any IEA. The second proposer would propose a coalition with a total of three members, and the third proposer would propose a coalition with the remaining 10 nations.

The coexistence of type 1 and type 2 IEAs (with the same fine and level of abatement) is implausible for Kyoto. For this reason, prior to Lemma 1 we limited the set of outcomes by assuming that each nation makes a binary choice (in or out of the IEA), rather than deciding which (if any) IEA to join. Had we not made that restriction, then (for the example above), the outcome (10,3,1) (two IEAs consisting of 3 and 10 members, with one nation outside the IEA) would have been stable.

The definition of the UFSS, and the participation 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 begin at a particular status quo that falls just short of a stable IEA, i.e. the size of the IEA is $M^{j+1} - 1$. Since $M^{j+1} \ll M^{j+1} - 1$, the IEA cannot be built up to the next stable element. However, beginning with this status quo, signatories do want to leave, causing the IEA to unravel to the next smallest stable element. "Rolling the ball uphill" requires coalitional deviation.

5.4 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 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 by modifying Definition 2 to replace unilateral with coalitional dominance relation.

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 in this setting.

6 Conclusion

In a non-cooperative setting (without side-payments), nations participate in IEAs in order to change the behavior of *other* nations. If an IEA chooses members' abatement levels before

the participation decision, a nation has little or no leverage over other nations' actions, and therefore has little incentive to join an IEA. If the IEA chooses members' abatement after the participation decision, a nation has some leverage, and therefore has an incentive to join. However, IEAs typically specify the membership requirements before, not after, nations decide whether to join. The assumption that the IEA choice occurs after the participation decision is therefore questionable. In addition, this assumption implies that a reform that lowers abatement costs (e.g. trade in permits) also lowers equilibrium membership, and may reduce global welfare. The policy implication is that reforms that make it possible to reach abatement targets more cheaply offer little hope for increasing participation in the IEA; instead, some kind of external threat, such as trade sanctions, would be needed in order to obtain cooperation. In our view, this policy implication is too pessimistic, and it is based on an implausible model of how IEAs are structured.

A re-designed Kyoto needs to accomplish (at least) three goals. It needs to address legitimate concerns arising from uncertainty about abatement costs: the cost of achieving a fixed target may be prohibitive. It also needs to attract more members. Finally, it needs a mechanism to help ensure signatory's compliance. The escape clause achieves all of these objectives, and is simple to implement.

The text emphasized the participation goal, but the compliance goal is also important. Kyoto has no explicit penalty against signatories that are non-compliant at the end of 2012. The only realistic penalty, under Kyoto, is to require non-compliant nations to enter the post-Kyoto agreement with a reduced level of emissions permits (to offset their Kyoto deficit). That remedy would discourage participation in the next round. Under our proposal, "non-compliance" has a different meaning. A signatory that neither abates nor pays the fine has defaulted on the promise to pay other signatories an explicit dollar amount. The international financial system has experience in dealing with defaults of this nature. The escape clause proposal not only encourages participation, but also discourages non-compliance.

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.

There are subgames at which high cost nations' (*ex post*) payoff is negative. This situation is due to of *ex ante* cost uncertainty and non-verifiability; it also arises (even more severely) in the standard model. Because sovereignty is an important constraint on nations' ability to make binding contracts, the magnitude of the signatory's payoff in the worst state of the world is relevant in assessing the plausibility of our proposal. However, nations are not able to costlessly abrogate – either unilaterally or in unison – agreements that they have signed. If disbanding

under unanimous agreement is an option that costs more than $(1-\phi)F$, the IEA would never disband. If unilateral withdrawal from the IEA costs more than the high cost signatory's loss in a type 1 IEA, unilateral defection does not occur.

Our conclusion that the use of an escape clause leads a large IEA relies on the fact that the net fine increases with the number of IEA members. The same result holds if, instead of paying a fine, signatories who do not abate are punished by those who do. For example, abating signatories might be allowed to withdraw a WTO-mandated trade concession from non-performing signatories. If all IEA signatories agree to this punishment, it would not violate the WTO (*unlike the suggestion to use trade sanctions to punish non-participants*). Since an increase in the number of IEA members increases the punishment, thereby increasing signatories' incentive to abate, this proposal also makes it possible to sustain a large IEA.

We assumed that nations have a 0-1 choice in abatement levels. In reality, a nation may decide to abate but not by enough to reach the IEA requirement. Our results still hold if a non-compliant signatory must pay the full fine regardless of the degree of non-compliance. In this case, if a signatory deviates, it undertakes zero abatement, as in our model. For future work, it is interesting to study IEA designs where the fine depends on the degree of deviation.

There are a number of other avenues for further research. We noted that the simplest "hybrid policy" (tradeable permits with a price ceiling) does not encourage participation. However, a modified version of that policy which auctions additional permits and then returns revenue to IEA members might be effective. We assumed that nations' costs are uncorrelated, although in fact they are likely to be positively correlated. We also assumed that nations are homogenous at the participation stage, whereas in fact there is considerable heterogeneity even amongst OECD members.

A Model details and proofs

A longer version of this appendix, containing more details, is available on request.

A.1 Derivation of expected payoffs, equations (4) and (5)

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

$$p_{m,M-1} \equiv \frac{(M-1)!}{m!\,(M-1-m)!}\,(1-p)^m\,p^{M-1-m}. \tag{12}$$

In a type 1 NE, a signatory with low costs chooses to abate, and has an expected payoff of

$$u_{a,1}(\theta_L) \equiv \sum_{m=0}^{M-1} p_{m,M-1}\left\{m+1-\theta_L+\frac{M-1-m}{M}\phi F\right\}. \tag{13}$$

The signatory with high cost chooses not to abate and has an expected payoff of

$$u_{na,1} \equiv \sum_{m=0}^{M-1} p_{m,M-1}\left\{m-F+\frac{M-m}{M}\phi F\right\}. \tag{14}$$

The expected payoff of the signatory is

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

which simplifies to (4) using the binomial formula in (12).

The expected payoff of a non-signatory is

$$\pi_{n,1}(M) = \sum_{m=0}^{M} p_{m,M}m.$$

Using the binomial formula, this equation simplifies to equation (5).

A.2 Statement and Proof of results for Nash equilibrium correspondence

The equilibrium number of members must be a nonnegative integer. Define $h(x)$ to be the smallest integer not less than x. Figure 6 describes the set of NE in the participation game respecting the integer constraint. Compared with Figure 2, the NE set is slightly larger: it includes the segment on $\tilde{F}$ between c and g, and point b is replaced by point b', the intersection between F_2 and the vertical line at $h(M_b)-1 < M_b$. The figure is based on the following two assumptions.

Assumption 3 *The horizontal distance between curves F_2 and F_1 is greater than 2 at point e.*

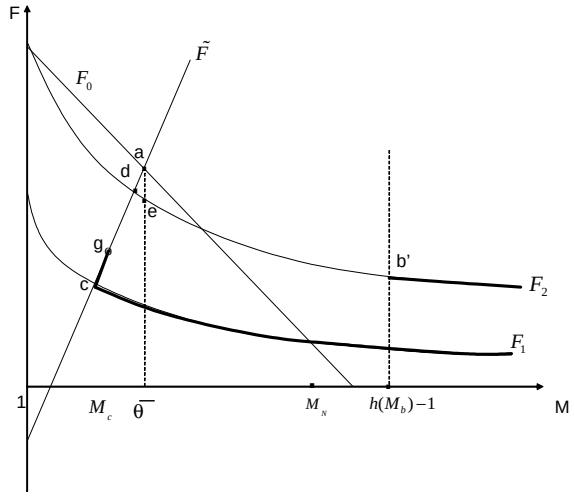

Figure 6: The Nash Equilibrium to the Participation Game

Assumption 4 $M_b - \bar{\theta} > 2$.

In additions to Assumptions 3 and 4, Figure 2 embodies two additional assumptions that have no effect on our analysis:: (i) point "a" lies above curve F_2, and (ii) point c is below point e (i.e., $F_c \leq F_e$).

The conditions in Assumption 3 and 4 are appropriate for a model that describes the problem of forming an IEA to control GHGs. Assumption 4 requires that the probability of high abatement costs is moderate or small: $p < (\bar{\theta} - 1)/(\bar{\theta} + 1)$. A sufficient condition is $p \leq 0.5$. Below we establish sufficient conditions for Assumption 3. They are (i) θ_L must be "moderately large:" $\theta_L > 3$. Given our normalization, $\theta_L > 3$ means that even if nations were certain that abatement costs are low, an agreement would have to contain at least four members in order for them to benefit from abatement,[14] and (ii) there is a non-negligible difference between high and low abatement costs. To show the conditions for Assumption 3, we first establish that

Lemma 2 *The horizontal distance between $F_2(M)$ and $F_1(M)$ in the $M - F$ plane is a decreasing function of F.*

Proof. (sketch) From equation (3), the inverse of $F_2(M)$ is $F_2^{-1}(F) = -F\frac{\phi}{-F-1+\theta_H}$, and from equation (2), the inverse of $F_1(M)$ is $F_1^{-1}(F) = -F\frac{\phi}{-F-1+\theta_L}$. Taking the derivative with respect to F of the distance $D(F) \equiv F_2^{-1}(F) - F_1^{-1}(F)$ establishes the lemma. ■

[14] 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.

To state the next result we use the following definitions:

$$\begin{gathered}\alpha \equiv (\theta_L - \theta_H)^3 < 0 \\ \beta \equiv -2(\theta_L - 1.5)(\theta_L - \theta_H)^2 \\ \gamma \equiv \left(3\theta_L - \theta_L^2\right)\theta_H + \left(\theta_L^3 - 3\theta_L^2 + 2\theta_L - 2\right) \\ R \equiv \alpha p^2 + \beta p + \gamma.\end{gathered} \tag{15}$$

Lemma 3 *(i)* $F_2^{-1}(F_e) - F_1^{-1}(F_e) - 2 > 0$ *if and only if* $R < 0$. *(ii) A sufficient condition for* $R < 0$ *is that* $\beta < 0$ *and* $\gamma < 0$. *Both of these inequalities are satisfied if*

$$\theta_L > 3 \quad \textit{and} \quad \theta_H > \theta_L + \frac{2(\theta_L - 1)}{\theta_L(\theta_L - 3)}. \tag{16}$$

Proof. (sketch) (i) Using D defined in the proof of Lemma 2 we evaluate $D - 2$ at $F = F_e$. We then show that this function is positive if and only if $R < 0$.

To establish part (ii), note that R is concave in p, and at $p = 0$, R is decreasing if $\theta_L > 1.5$. Therefore, a sufficient condition for $R < 0$ for $p \geq 0$ is $\theta_L > 1.5$ and $\gamma < 0$. Suppose $\theta_L > 3$. Then $\gamma < 0$ if and only if

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

■

Combining Lemmas 2 and 3, we immediately know

Proposition 4 *Conditions (16) are sufficient for Assumption 3, i.e.,* $F_2^{-1}(F) - F_1^{-1}(F) - 2 > 0$ *for* $F \leq F_e$ *and for all* $0 < \phi < 1$.

We now formally state and prove the characterization of the NE to the participation game in Figure 2. Part (iv) of the Proposition 5 uses three values of F, defined by

$$\begin{gathered}F_k: \quad h\left(\tilde{F}^{-1}(F_k)\right) - F_1^{-1}(F_k) = 1 \\ F_q = \sup\left\{F \mid h\left(\tilde{F}^{-1}(F)\right) \leq F_2^{-1}(F)\right\} \\ F_g = \min\{F_k, F_q\}\end{gathered}$$

For $F \leq F_k$, the horizontal distance between $\tilde{F}(\cdot)$ and $F(\cdot)$ does not exceed one. For $F \leq F_q$, an IEA of size $h(\tilde{F}^{-1}(F))$ is of type 1.

Proposition 5 *We adopt Assumptions 3 and 4.*
(i) Suppose $N > h(M_b) - 1$. *Then 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 of type 2.*
(ii) The smallest type 2 IEA consists of $h(M_b)$ *members, induced by fines* $F \in \left[F_2^{-1}(h(M_b)), F_{b'}\right)$.
(iii) Suppose $N \geq F_1^{-1}(F_c)$. *Then for* $F \in \left[F_1^{-1}(N), F_c\right]$ *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) 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 $h\left(\tilde{F}^{-1}(F)\right)$ 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.

We discuss some basic results and intuition before formally proving the Proposition. The intuition in explaining Figure 2 still applies to Figure 6 and the Proposition. To understand why the segment between c and g is NE, note that along the segment, defection by a signatory changes the IEA from type 2 to type 1 (due to the integer constraint), reducing the defector's payoff. To understand replacing point b by b', recall that M_b is defined by the equality $\pi_{s,2}(M_b) = \pi_{n,1}(M_b - 1)$. A simple (omitted) calculation establishes that

$$\pi_{s,2}(M) < \pi_{n,1}(M-1) \quad \text{if and only if} \quad M < M_b. \tag{17}$$

Further, points on curve F_2 below b' correspond to IEA with sizes of at least $h(M_b)$, while those at or above b' correspond to IEA with sizes of at most $h(M_b) - 1$. Thus, for points on 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.

By Assumption 4 $M_b - 1 > \bar{\theta} + 1$, which implies that $h(M_b) - 1 > \bar{\theta} + 1$. Since F_2 is a decreasing function, $F_b \equiv F_2(h(M_b) - 1) < F_2\left(\bar{\theta}\right) \equiv F_e$. Thus, by Proposition 4, the horizontal distance between F_2 and F_1 is greater than 2:

$$F_2^{-1}(F) - F_1^{-1}(F) > 2 \text{ for } F \in [F_b, F_e]. \tag{18}$$

Proof. (i) For $F \in [F_2(N), F_{b'})$, 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 Proposition 4. The defector's payoff is lower, in view of equation (17). 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) We only need to show that there is no type 2 NE when $F \geq F_{b'}$. Consider first $F \in [F_{b'}, F_e)$. Clearly $M > h\left(F_2^{-1}(F)\right)$ is not an equilibrium: a member would want to defect by leaving the IEA, since the resulting IEA would still be of type 2. Similarly, $M = h\left(F_2^{-1}(F)\right)$ is not an equilibrium: by equation (17), 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, the only candidate equilibrium is $h\left(\bar{\theta}\right)$. (Smaller IEAs have negative expected profits; if the IEA were larger, a member would want to

defect since the resulting IEA is still type 2.) However, $h\left(\bar{\theta}\right)$ 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). If the result is still a type 2 equilibrium, a signatory has incentive to defect because $\pi_{s,2}(M+1) < \pi_{n,2}(M)$. If the result is a type 1 equilibrium, by equation (17), a signatory who leaves the IEA (becoming a non-signatory) has a higher payoff than at the candidate equilibrium.

(iii) For $F \in \left[F_1^{-1}(N), F_c\right]$, consider a candidate NE at $M = h(F_1^{-1}(F))$. Signatories' payoffs are positive at 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 $h\left(F_1^{-1}(F)\right) + 1 < F_1^{-1}(F) + 2 < F_2^{-1}(F)$ by inequality (18) , 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) From the definitions of F_c, F_g and F_d, we know $F_c < F_g \leq F_d$. 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 $M = h\left(\tilde{F}^{-1}(F)\right)$, 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 $M = h\left(\tilde{F}^{-1}(F)\right)$ 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 $M = h\left(\tilde{F}^{-1}(F)\right)$ 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 $M = h\left(\tilde{F}^{-1}(F)\right)$ is immune from defection by signatories. To show that it is a type 1 NE, we need only show that this candidate is immune from defection by a non-signatory. If the non-signatory joins the IEA, then one of the following two possibilities occur. (a) The resulting IEA is still type 1. In this case the defector has a lower payoff, because a non-signatory's dominant strategy is not to abate. (b) The IEA becomes a type 2. In this case the defector has a lower payoff by virtue of Assumption 4. ■

A.3 Welfare in the Nash Equilibrium

Here we prove the welfare results of Section 3.2. We first show that $M_b > \theta_H > M_N$. By inspection of Figure 2, inequality $M_b > M_N$ implies that the point on curve F_2 corresponding to M_b lies above line F_0. (It is simple to establish that $M_N > M_c > \theta_L$. We omit these details.)

Lemma 4 $M_b > \theta_H > M_N$.

Proof. (sketch) Let

$$w(M) \equiv F_0(M) - F_1(M) = \frac{-M - M\theta_H + M\theta_L + \phi\theta_H + M^2 - M\phi\theta_L}{(\phi - 1)(M - \phi)}, \tag{19}$$

which implies that

$$w'(M) = -\frac{1}{1-\phi} + \frac{\phi(\theta_L - 1)}{(M-\phi)^2}. \tag{20}$$

We prove the Lemma in three steps.

Step 1: Direct calculation and some simplification establishes that under Assumption 3 $w'(M) < 0$ for $M \geq \bar{\theta}$.

Step 2: Using Step 1 and the fact that $w(M_N) = 0$, we show that $M_N < \theta_H$ by verifying that $w(\theta_H) < 0$.

Step 3: Since $M_b - \theta_H = (1-p)\frac{\theta_L - 1}{p} > 0$, it follows that $M_b > \theta_H$. ■

Proof of Proposition 2 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 $\left(M; F_1^{-1}(M)\right)$ 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 Proposition 5 part (ii), the smallest IEA that results in a type 2 NE in the abatement stage consists of $h(M_b)$ members. By Lemma (4), this value is greater than M_N, the value above which it is optimal to induce a type 2 equilibrium. ■

A.4 Detailed results on the stable sets

The following Proposition presents the necessary and sufficient conditions for $\mathcal{M}(M^1)$ to coincide with the UFSS V. It also provides a method of finding the smallest element of V, M^1.

Proposition 6 *$\mathcal{M}(M^1) = V$ if and only if M^1 is the smallest non-negative integer for which the following three conditions hold:*
(i) If $M^1 = 0$, then either a) M^2 and $M^2 - 1$ are the same type or b) M^2 is a type 2 IEA, $M^2 - 1$ is a type 1 IEA and $M^2 < M_b$.
(ii) Either all positive M^j in V are the same type IEA, or there is a switch from a type 1 to a type 2 IEA. In the latter case, M^*, defined as the smallest type 2 IEA in V, must satisfy at least one of the following two conditions: (a) $M^* < M_b$, or (b) $M^* - 1$ is a type 2 IEA.
(iii) If $M^1 > 0$, $\pi_s(M^1) \geq \pi_n(M)$ for all $M < M^1$.

We prove necessity by showing that if any of the three conditions fail, then $\mathcal{M}(M^1)$ violates either internal or external stability. To show sufficiency, we confirm that when all these conditions hold, $\mathcal{M}(M^1)$ is both internally and externally stable. For example, condition (ii) insures that no positive element of $\mathcal{M}(M^1)$ indirectly dominates a smaller positive element; conditions (i) and (iii) guarantee the M^2 does not indirectly dominate M^1. We use the definition of $\mathcal{M}(M^1)$ and the monotonicity of $\pi_{s,i}(M)$ and $\pi_{n,i}(M)$ to confirm external stability.

Proof. Necessity (i) Suppose, to the contrary of the proposition, that $M^1 = 0$, and M^2 and $M^2 - 1$ are of different types. Further, if M^2 is type 2 and $M^2 - 1$ is type 1, $M^2 \geq M_b$. If M^2 is type 1 but $M^2 - 1$ is type 0, $\pi_s(M^2) > \pi_n(M)$ for $M \leq M^2 - 1$, so $0 \ll M^2$. If $M^2 \geq M_b$ is type 2 and $M^2 - 1$ is type 1, the definition of M_b implies that $M^2 - 1 \ll M^2$, implying that $0 \ll M^2$. In both cases, a condition for internal stability of V is violated.

(ii) Recall the definition of M^*, the smallest type 2 IEA in V; define M^{**} as the largest type 1 IEA in V. If (ii) is not satisfied, then $M^* - 1$ is a type 1 IEA and $M^* \geq M_b$. In that case $\pi_s(M^*) > \pi_n(M^* - 1)$ by definition of M_b. Since $\pi_n(M)$ is an increasing function, $\pi_s(M^*) > \pi_n(M)$ for all $M < M^*$, so $\pi_s(M^*) > \pi_n(M^{**})$. These inequalities imply $M^{**} \ll M^*$, so V is not internally stable. (iii) If condition (iii) fails, then $M^1 > 0$, $\pi_s(M^1) < \pi_n(M^1 - 1)$. In this case M^1 does not indirectly dominate 0; consequently, $0 \in V$, contradicting $M^1 > 0$.

Sufficiency Step 1 uses conditions (i) and (ii) to establish that $\mathcal{M}(M^1)$ 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 (established below) 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} (because of the definition of the sequence $\mathcal{M}(M^1)$) 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.

By equation (11), no element of $\mathcal{M}(M^1)$ indirectly dominates 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 $\mathcal{M}(M^1)$ indirectly dominates a smaller element of the set. We do this by showing that M^{j+1} does not indirectly dominate M^j. The argument in the first paragraph of Step 1 then implies internal stability.

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* $m = M^j, M^j + 1, \ldots, M^{j+1} - 1$. 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 (4) - (7) 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*,

$$\pi_{s,i}(M) < \pi_{n,i}(M - 1), \quad \text{or} \quad M \ll M - 1, \quad M = 1, \ldots, N. \tag{21}$$

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 $M^{j+1} \ll M^{j+1} - 1$. That is, in the process of moving from an IEA of size M^j to an IEA of size M^{j+1}, the "last signatory" prefers not to join.

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.) If condition (iia) holds then $\pi_s\left(M^{j+1}\right) < \pi_n\left(M^{j+1} - 1\right)$ (in view of the definition of M_b); thus, M^{j+1} does not indirectly dominate M^j. If condition (iib) holds, then M^{j+1} and $M^{j+1} - 1$ are both type 2 IEAs, so $\pi_s\left(M^{j+1}\right) < \pi_n\left(M^{j+1} - 1\right)$ by inequality (8). Again, M^{j+1} does not indirectly dominate M^j.

Finally, consider the case where $M^1 = 0$, so that $\pi_n\left(M^1\right) = 0$. If condition (ia) holds, then $M^2 \ll M^2 - 1$. The "last signatory" does not want to join, so M^2 does not indirectly dominate $M^1 = 0$. If condition (ib) holds then the definition of M_b implies that $M^2 \ll M^2 - 1$.

Step 2: (External stability) We need to show that each element in $\mathcal{N}$ in the complement of $\mathcal{M}(M^1)$ (i.e. $\mathcal{N}/\mathcal{M}(M^1)$) is indirectly dominated by some element of $\mathcal{M}(M^1)$. The set $\mathcal{N}/\mathcal{M}(M^1)$ is the union of two sets, IEAs that are smaller than, or larger than M^1. Denote these as $A = \{M \mid M < M^1, M \in \mathcal{N}/\mathcal{M}(M^1)\}$ and $B = \{M \mid M > M^1, M \in \mathcal{N}/\mathcal{M}(M^1)\}$. We show that all elements of A are indirectly dominated by some element of V, and then show

the same for set B.

Consider set A. When $M^1 = 0$, $A = \varnothing$, so for this subset we need only consider $M^1 > 0$. In this case, condition (iii) states that IEAs smaller than M^1 are indirectly dominated by the IEA of size M^1.

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 $\mathcal{M}(M^1)$.) That is, $M \ll M^j$ for all $M = M^j + 1, \ldots, M^{j+1} - 1$. In addition, for $j = k$, $M \ll M^j$ for all $M = M^j + 1, \ldots, N$. We provide details only for the case of $j < k$; the proof is similar when $j = k$.

From (11), 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 were 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 $M = M^j + 1, \ldots, M^{j+1} - 1$. 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 $M = M^j + 1, \ldots, M^{j+1} - 1$.

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 and M^j is a type 1 IEA. (There can be no positive elements of the stable set that are type 0.)

First consider the possibility $M^1 = 0$. In this case, $\pi_s(M) < 0$ for $1 \leq M < M^2$, so $M \ll M^j = 0$ for all $M = 1, 2, \ldots, M^{j+1} - 1$. Next consider the case where M^{j+1} is a type 2 and 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. Consider 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 (11) 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 $\mathcal{M}(M^1)$ 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 $M \in [M^j + 1, M' - 1]$. ■

Proposition 6 enables us to find V by identifying its smallest element M^1, and then applying the recursive relation in equation (11). In particular, we *choose the smallest possible* M^1 in order to satisfy the three conditions. We start with $M^1 = 0$ and test whether conditions (i) and (ii) hold. If they hold, then the smallest element of V is zero. If they do not hold, then $M^1 \geq 1$ and we identify its value using conditions (ii) and (iii). To verify that a candidate $M^1 \geq 1$ is the smallest element of V we only need to verify that it is not indirectly dominated by smaller IEAs. The results in Summary 1 and Figure 5 are obtained following this procedure.

Proof of Proposition 3 Consider the set $V = \mathcal{M}(M^1)$ with the largest element being M^k. From (11), we know

$$\pi_s(M^k) \geq \pi_n(M^{k-1}) > \pi_n(M^{k-2}) > \ldots > \pi_n(M^1). \tag{22}$$

Consequently, at any IEA M^j, $j < k$, a group of $M^k - M^j$ non-signatories have an incentive to join the IEA together, and earn $\pi_s(M^k)$ instead of $\pi_n(M^j)$. That is, M^k coalitionally indirectly dominates all other elements in the set $\mathcal{M}(M^1)$. ■

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*. Ashagate 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.

DE ZEEUW, A. (2005): "Dynamic effects on the stability of international environmental agreements," Fondazione Eni Enrico Mattei, Nota de Lavoro 41.2005.

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.

EYCKMANS, J. (2001): "On the farsighted stability of the Kyoto Protocol," CLIMNEG Working Paper 40, CORE, Universite Catholique de Louvain.

FINUS, M. (2001): *Game Theory and International Environmental Cooperation*. Edward Elgar.

HOEL, M., AND L. KARP (2001): "Taxes and Quotas for a Stock Pollutant with Multiplicative Uncertainty," *Journal of Public Economics*, 82, 91–114.

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.

RAY, D., AND R. VOHRA (2001): "Coalitional Power and Public Goods," *Journal of Political Economy*, 109(6), 1355 – 1382.

STIGLITZ, J. E. (2006): "A New Agenda for Global Warming," *Economists' Voice*, pp. 1–4.

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.