Department of Economics
University of Southampton
Southampton SO17 1BJ
UK

Discussion Papers in Economics and Econometrics

INTERNATIONAL ENVIRONMENTAL AGREEMENTS WITH A STOCK POLLUTANT, UNCERTAINTY AND LEARNING

Alistair Ulph

No. 0207

This paper is available on our website
http://www/soton.ac.uk/~econweb/dp/dp02.html

INTERNATIONAL ENVIRONMENTAL AGREEMENTS WITH A STOCK POLLUTANT, UNCERTAINTY AND LEARNING

ALISTAIR ULPH
(University of Southampton)

Revised Version,
August 2002

Paper presented at conference "Risk and Uncertainty in Environmental and Resource Economics", Wageningen, The Netherlands, June 6-7, 2002.

ABSTRACT

I analyse how the possibility of future resolution of uncertainty about damage costs affects incentives to join a self-enforcing international environmental agreement. Unlike earlier literature, the model allows many countries, uses a solution concept which does not restrict the size of a stable IEA, and, since it is dynamic, analyses different membership rules: fixed (countries commit) or variable (countries decide each period whether to join). With fixed membership, learning yields at least as high membership and global welfare as no learning, unless both expected damage costs and uncertainty are high. With variable membership, learning leads to higher future membership but lower global welfare than no learning. Fixed membership leads to higher global welfare than variable membership if there are at least as many signatories who abate pollution in each period and state of the world as variable membership. This occurs for only 20% of parameter values with no learning and 2% with learning.

Key words: International Environmental Agreements, uncertainty, learning, fixed membership, variable membership, self-enforcing.

JEL Classification: F02, F18, Q4

1. INTRODUCTION

Trying to reach an international agreement on problems such as climate change is beset by the difficulties of dealing with a problem which is truly global, and hence ideally requires actions by all countries, by the time scale over which current reductions in greenhouse gas emissions might affect future climate, by the still considerable uncertainty about the possible benefits and costs of actions to reduce greenhouse gas emissions, both in aggregate and in terms of their distribution across countries, and the fact that we will get better information on these costs and benefits in the future. The recent IPCC (2001) reports give an excellent summary of the major uncertainties about the potential impacts of climate change and their associated degree of confidence, and the sequence of IPCC reports is an excellent illustration of how scientific information has improved over time.

The possibility of future learning leads to the important *timing* question: should we delay reductions in emissions until we get better information on impacts, or should we accelerate emissions reductions in case we learn bad news and have little time to respond to it? There is also the important question of how does the answer to this question depend on the significant irreversibilities involved, both in the accumulation of greenhouse gases and the accumulation of capital. There is now a significant literature on these questions both theoretical[1] and with empirical applications to climate change[2]. In general the theoretical arguments are ambiguous and the empirical literature suggests small effects, leading to the claim by Karp and Zhang (2001) that the issue of uncertainty and learning is a red herring for climate change policy and policy-makers should base their current emissions strategy on the best scientific information currently available[3].

However all this literature assumes there is a single global decision-maker. But a crucial feature of global environmental problems is the absence such a world authority, and hence the need to tackle such problems by means of international environmental agreements (IEAs), which, given the sovereignty of independent nation states, have to be designed to be self-enforcing. This leads to a second, and perhaps more important, timing question: how does the possibility that uncertainty will be resolved through future learning affect the incentives for countries to join international environmental agreements, both now and in the future. In this context it is worth noting that uncertainty about the impacts of climate change and the need for further research was one of the reasons the US gave for not ratifying the Kyoto Agreement.

Relative to the first timing question, there is much less literature which has addressed this second timing question. As Kolstad (2002) notes there are a number of informal

[1] See Fisher , Hanemann and Narain (2001) for an excellent survey.

[2] See, for example, Manne and Richels (1992), Peck and Teisberg (1993), Grubb, Chapuis and H-Duong (1995), Kolstad (1996), Nordhaus and Popp (1997), Ulph and Ulph (1997), Kelly and Kolstad (1999), Karp and Zhang (2000, 2001).

[3] The possibility of catastrophic risks such as the disintegration of the West Antartic or Greenland ice sheets or partial shutdown of the North Atlantic thermohaline circulation system does suggest the need for stronger current action to reduce emissions, see Fisher and Narain (2001) and Gjerde, Grepperud and Kverndokk (1999).

arguments suggesting that when there is uncertainty it might be easier to reach agreements before that uncertainty gets resolved; the informal argument is that if uncertainty is about winners and losers then it may be easier to get agreement before the identity of the winners and losers is known. On the other hand, as Helm (1998) notes, the broad distributional consequences of environmental problems are often known, and scientific uncertainty is then used as a 'figleaf' by distributional losers to delay reaching an agreement[4].

However although there is now a large formal literature on international environmental agreements (see Finus (2001) and Barrett (2002) for excellent overviews), there is very little formal analysis of the issues related to this paper. In some earlier work (Ulph and Ulph (1996), Ulph and Maddison (1997), Ulph (1998)) I used a two-country, two period model of a global stock pollutant where there was uncertainty about damage costs, and compared the non-cooperative and cooperative equilibria when either no information or full information about damage costs would become available at the end of period 1. Those papers showed that while the value of information was always positive in the cooperative equilibrium (essentially assuming a single decision-maker – for whom information must always be non-detrimental) it could be negative in the non-cooperative equilibrium if there was negative correlation in the information available to countries (i.e. one country learned it was going to suffer more from global warming the other suffer less than had been expected). That finding relates tangentially to the issues considered in this paper because the difference in the value of information between the cooperative and non-cooperative equilibria is exactly the same as the difference in gains from cooperation between learning and no learning. So the fact that information would have positive value in the cooperative equilibrium and negative in the non-cooperative equilibrium means that the gains to cooperation will be greater when it is known that learning will take place than when there is no learning. Of course in a two-country world the issue of coalition formation does not arise, but assuming the argument about positive and negative values of information generalised to many countries, that need not imply that more countries will join an international agreement when there is learning. To answer that requires an explicit model of coalition formation. In the well-known model due to Barrett (1994) the number of countries who join an agreement is *inversely* related to the gains from cooperation. Of course the Barrett model is static and ignores the issue of uncertainty and learning. The purpose of this paper is to extend his framework to allow for a stock pollutant, uncertainty and learning.

Kolstad (2002) also uses a two-country model and an indirect argument about how uncertainty and learning might affect the difficulty of reaching an international environmental agreement. His measure of difficulty is the size of the side payments that might be needed to ensure that no country is worse off in the cooperative equilibrium

[4] Helm also presents some formal modelling whose relevance to the argument I found it difficult to assess. Helm considers a unidirectional externality with a single exporter and single importer. In the one-shot Nash equilibrium the importer abates, the exporter does not. Because it is not possible for the importer to punish the exporter, replication by itself does not help. But if the exporter believes that there is a probability that at some stage in the future she will become an importer, then with a high enough probability replication will induce the exporter to abate. However I may have misunderstood the argument.

than in the non-cooperative equilibrium, with the notion that the larger the side payments necessary the less likely it is they would be agreed to. However, this is subject to a similar criticism as the one I made about my earlier work. Assuming that the result on the size of side payments generalises to more than two countries, without an explicit model of coalition formation it is not clear what this says about the number of countries that might join an international agreement. The role of side payments in helping to secure cooperation is contested. Carraro and Siniscalco (1993) argued that transfers from signatories to non-signatories would not help to increase the number of signatories. They assumed identical countries, but even with significantly asymmetric countries, Barrett (2001) shows that side payments, *by themselves,* do little to increase the number of signatories. What is required, if there are significant asymmetries between countries, is to change the rules of the game, and this, combined with side payments, can dramatically increase the number of signatories. Moreover these side payments would be self-enforcing. In this framework, the need for larger side payments could make it *more likely* that one will get a successful IEA.

Both the above cases point to the need for an explicit model of coalition formation with uncertainty and learning. To my knowledge the only paper to date with such a model is Na and Shin (1998). They consider a three-country model in which, *ex ante*, countries are identical. Uncertainty is only about the distribution of net benefits. They consider a model of coalition formation similar to Barrett (1994) in which countries first decide whether to join a coalition and then decide their emission level; in this second-stage game the interaction between signatories and non-signatories is a Nash equilibrium. They show that if countries have to decide whether or not to join before the uncertainty about the distribution of net benefits is resolved, then the grand coalition in which all three countries join is the unique stable IEA. But once the distribution of benefits is known, then the grand coalition is unstable, and either a two-country coalition will form or no coalition will form. So in their model, learning is bad for cooperation.

But their paper has a number of limitations. First it deals only with distributional uncertainty. It is straightforward to show that if, at the other extreme, uncertainty was solely about the extent of global net benefits, and countries were identical *ex post* as well as *ex ante*, then learning has no effect – the unique stable IEA is three countries. But this reflects the second limitation of their model. It is well known (Finus (2001)) that if the interaction between signatory and non-signatory countries takes the form of a Nash equilibrium, then, for the class of models used by Na and Shin, with symmetric countries the maximum number of countries who will join an IEA is 3. So it not very surprising in a three country world of *ex ante* symmetric countries to find that the grand coalition forms, and that this is unaffected if uncertainty resolution leaves countries *ex post* identical. Third, their model is a one-period model, whereas many important environmental problems involve stock pollutants. Moreover, learning is a dynamic process.

In this paper I will address the issue of how uncertainty and uncertainty and learning affects the incentives to join an international environmental agreement, using a model which overcomes the limitations of Na and Shin. The model extends a model by Rubio

and Ulph (2002a) to include uncertainty and learning, and the model by Rubio and Ulph had in turn extended the model of Barrett (1994) by allowing for a stock pollutant. I consider a model with N identical countries and a two-period model of stock pollution. Because, like Barrett and Rubio and Ulph, the way countries choose their emissions is modelled as a Stackelberg equilibrium, the stable IEA can involve between 2 and N countries depending on parameter values, so it overcomes the limitation of the Na and Shin model that there can be no more than 3 countries in a stable IEA. I focus on the case where uncertainty is about the extent of global damages, so countries will be identical *ex post* as well as *ex ante*, but unlike Na and Shin, I show that learning does affect the number of countries who sign up. Finally because the model is dynamic I can address issues which cannot be addressed in the static framework of Na and Shin, or indeed in much of the literature on IEAs. In a dynamic framework we could model IEAs as if countries have to decide at the outset whether or not to join an IEA and are committed to that decision for all future time periods and states of the world; I call this the fixed membership model. More consistent with the notion of national sovereignty, we could model countries as deciding each period and each state of the world whether to join an IEA; I call this the variable membership model. This allows analysis of how membership varies over time as information is refined. By comparing the fixed and variable membership models one can address the question - is it better to have countries commit at the outset to their membership decision or to allow them to decide each period?

I shall show that if membership is fixed, then, if expected damage costs are high and there is a relatively high degree of uncertainty about damage costs, learning leads to lower membership and lower global welfare than no learning; but otherwise learning leads to more members and higher welfare than if there is no prospect of learning. On the other hand if membership is variable, then, for the special case I consider, first period membership is unaffected by whether or not there is learning, but second period membership is (on average) higher with learning than no learning, but global welfare is lower. As to whether it is better to have fixed or variable membership, fixed membership results in higher expected global welfare than variable membership if it leads to at least as many signatories who abate pollution in each period and each state of the world. Otherwise, variable membership yields higher expected global welfare. The parameter values for which fixed membership is better than variable occur in 20% of cases with no learning but only 2% of cases with learning.

In section 2 I will set out the basic model with one period and no uncertainty so as to establish notation and key ideas. In section 3 I extend the model to two periods and introduce uncertainty and learning, and then analyse what stable IEA will form in four cases: No Learning with Fixed Membership; No Learning with Variable Membership, Learning with Fixed Membership and Learning with Variable Membership. In section 4 I use these results to compare IEA membership and global welfare when there is No Learning and Learning and when there is Fixed Membership or Variable Membership. Section 5 concludes and suggests obvious lines for further extensions.

2. THE BASIC MODEL.

To introduce the key concepts of the model, such as the notion of stable (self-enforcing) IEAs, I begin by setting out the basic version of the model with only one period and no uncertainty or learning. This reproduces the results in Barrett (2000) and Rubio and Ulph (2002a). There are N identical countries indexed by $i = 1,\ldots, N$ each of whom can emit q_i units of pollution. To keep things simple I assume that q_i can take one of only two values, which I normalise to be 0 or 1 and interpret as *abate* or *pollute*[5]. I denote by Q_i the total emissions of all countries other than i and by Q the total emissions of all N countries. I deal with a global pollution problem where it is the aggregate global level of emissions, Q, which determines damages in each county. I denote the net benefit function of country i by:

$$\pi(q_i, Q_i) \equiv bq_i - cQ = bq_i - c(q_i + Q_i) = b[q_i - \gamma(q_i + Q_i)]$$

where b is the (constant) benefit a country derives from a unit of emissions, c is the (constant) unit damage cost a country gets from each unit of global emissions, and γ is the *cost-benefit* ratio of the damage caused to a country by a unit of pollution divided by the benefit that country receives from emitting a unit of pollution. Since all that matters is the cost-benefit ratio, I henceforth normalise b to 1.

To make the problem interesting, I make the following assumption.

Assumption 1. $\frac{1}{N} < \gamma < 1$.

Then it is straightforward to derive:

Lemma 1 (Non-cooperative and Cooperative Equilibria) *Given Assumption 1: (i) if all countries act non-cooperatively then the dominant strategy for each country is to set* $q_i = 1$*, and each country will receive a net payoff 1-N*γ *<0; (ii) if all countries act cooperatively (jointly maximise aggregate net benefits) then each country will set* q_i *=0 and receive a net payoff of 0; (iii) the gain to cooperation (difference in payoff if all cooperate and if all do not cooperate) for each country is G ≡ N*γ*-1, which is increasing in N and* γ*.*

[5] We make this simplification because even in the one-period model with certainty, allowing for continuous emission levels, for example by having quadratic damage and abatement costs, makes the analysis of IEA stability quite complicated. In Barrett's seminal 1994 paper he resorted to numerical simulations to derive his results. Moreover he did not take account of the need to ensure that emissions must be non-negative. Taking proper account of such restrictions by allowing for corner solutions complicates the analysis further (see Rubio and Ulph (2002c)). Since my aim is to extend the basic Barrett model by having a stock pollutant and uncertainty and learning, I follow Rubio and Ulph (2002a) in taking the simplest version of the Barrett model which has discrete emission levels. I conjecture that the results are not sensitive to this simplification, but it will be important for future research to test that conjecture.

The first part of Lemma 1 ensures that, as long as $\gamma < 1$ (i.e. benefit to a country from emitting a unit of pollution exceeds the cost to it of the additional pollution) the non-cooperative global pollution game is a Prisoner's Dilemma. If $\gamma \geq 1$ then even countries acting non-cooperatively would want to abate pollution. The second part ensures that as long as $N\gamma > 1$ (the damage cost to all countries from an extra unit of pollution exceeds the benefit to a country from emitting that unit) then it is in the collective interest of all countries to abate pollution. If the inequality was not satisfied then even countries acting cooperatively would find that global damage costs were so low it was not worth abating pollution.

Lemma 1 covers the two extremes where all countries either cooperate or do not cooperate. I now analyse the intermediate case in which n countries, $2 \leq n \leq N$ join an International Environmental Agreement. This is modelled as a two-stage game. In the second-stage (the Emission Game), countries decide their emission levels, with each non-signatory country (denoted by f – fringe or free-rider) choosing its emission level to maximise its net-benefit, taking as given emissions of all other countries, while the signatory countries (denoted by s) choose their emissions to jointly maximise their aggregate net benefit, taking as given the emissions of non-signatory countries[6]. In the first-stage (the Membership Game) each country has to decide whether to sign or not to sign.

It is straightforward to show:

Lemma 2 (Emission Game - Strategies). *For any n signatories, the optimal strategy of the non signatories is to pollute no matter what the signatories do, while for signatory countries the optimal strategy is for each to abate if $n\gamma \geq 1$, and otherwise to pollute.*

The optimality of the strategy for non-signatories follows from Lemma 1 and the fact that polluting is a dominant strategy when countries act non-cooperatively. The optimality of the strategy for signatories follows from the fact that if one of its members abates, then the n signatory countries collectively save $n\gamma$ in avoided damage costs and forego 1 in benefits.

Define $n^* \equiv \Omega(1/\gamma)$, where $\Omega(x)$ denotes the smallest integer not less than x. It follows from Assumption 1 that $2 \leq n^* \leq N$. By Lemma 2, n^* is the smallest size of IEA (I will call it the threshold size) at which it pays the members of the IEA to abate. Then:

Lemma 3 (Emission Game – Payoffs). *For any number of signatory countries n, $2 \leq n \leq N$, the payoffs to signatories and non-signatories from the second-stage game are:*

[6] In the case of continuous emission levels, it is important to distinguish between the case where the signatory countries take as given the emission levels chosen by the non-signatories (the Cournot assumption) and the case where signatories take as given the *reaction functions* of non-signatory countries (the Stackelberg assumption) – see Finus (2000), Rubio and Ulph (2002c). In this simpler model of two emission levels, the dominant strategy of non-signatories is to pollute no matter what signatories do, so there is no difference between Cournot and Stackelberg.

$N \geq n \geq n^*$: $\pi^s(n) = -\gamma(N-n); \qquad \pi^f(n) = 1-\gamma(N-n)$

$n^* > n \geq 2$: $\pi^s(n) = \pi^f(n) = 1-\gamma N$

Finally turning to the first-stage game, I define a an IEA of size n as being *stable or self-enforcing* if it satisfies two stability conditions:

Internal Stability: $\pi^s(n) \geq \pi^f(n-1);$

External Stability: $\pi^f(n) \leq \pi^s(n+1)$

Internal stability is just the condition that no signatory country wishes to leave the IEA and become a non-signatory, while external stability is the condition that no non-signatory country wants to join the IEA and become a signatory. These conditions can also be interpreted as the conditions for a Nash equilibrium of the membership game in which each country takes as given the membership decisions of all other countries and decides whether to join or not join, knowing, by Lemma 3, what the payoffs to signatories and non-signatories will be. Note that in either interpretation we are considering only single unilateral decisions whether to join or not join the IEA, on the assumption that the membership decisions of all other countries remain the same.

Result 1 (Membership Game). *The unique stable IEA of the static model is of size n^*, and the payoffs to signatories and non-signatories are:*

$$\pi^s = -\gamma(N-n^*); \pi^f = 1-\gamma(N-n^*)$$

Proof:

Internal Stability: $\pi^s(n^*) = -\gamma(N-n^*) \geq \pi^f(n^*-1) = 1-\gamma N \Leftrightarrow \gamma n^* > 1;$ true by definition of n^*.

External Stability: $\pi^f(n^*) = 1-\gamma(N-n^*) \geq \pi^s(n^*+1) = -\gamma(N-n^*-1) \Leftrightarrow 1 > \gamma$ - true by Assumption 1. QED

Another interpretation of the external stability condition is that no IEA with more than n^* signatories is internally stable. If there was an IEA with strictly greater than n^* signatories then each signatory would have an incentive to defect because it calculates that the other signatories will remain as signatories, that, by Lemmas 2 and 3, they will continue to abate pollution, so all that happens if it defects is that aggregate pollution goes up by 1 unit, costing the defector γ, but gaining it 1. So no IEA of size greater than n^* is immune to defections. However, once the number of signatories reaches the critical threshold level, any further defection by a signatory will cause the remaining signatories to switch their emission policy to all polluting. So the cost of a further defection is now $n^*\gamma$, and by definition of n^*, $n^*\gamma \geq 1$, so defection no longer pays. As the proof shows, this is just the condition for internal stability of n^* to be satisfied. In summary, the incentives to free-ride ensure the size of a stable IEA cannot exceed the minimum threshold level at which it just pays signatories to abate; the threat that any further reduction in the size of the IEA below that threshold stops any further defections.

Corollary 1 *The size of the unique stable IEA is a decreasing function of γ, and as γ varies in the range given by Assumption 1, n^* varies between N and 2.*

From Lemma 1 and Corollary 1 the greater is the cost of pollution relative to the benefit of emitting pollution the greater are the gains from cooperation, but the smaller is the size of the stable IEA that can be formed. The intuition is straightforward. Free-riding means that the only stable IEA is the minimum size at which it just pays an IEA to abate pollution; the greater is the cost of pollution relative to the benefit of emission the smaller is the minimum size of IEA at which it pays signatories to abate rather than pollute.

There is another way to interpret Corollary 1. Define global welfare as the global average payoff to a country from a stable IEA:

$$\pi^g = (n^*\pi^s + (N - n^*)\pi^f)/N = -\gamma(N - n^*) + (N - n^*)/N .$$

Subtract the payoff each county would have got with no IEA: $1-\gamma N$ to obtain the gain per country from having an IEA of size n^*, denoted by: $\Delta = n^*(\gamma - \frac{1}{N})$. $n^*\gamma$ is the savings in damage costs each country will get from having an IEA of size n^*, and n^*/N is the average loss of output each country gets from having an IEA of size n^*. Since $\gamma > 1/N$, there are always strictly positive gains from a stable IEA of size n^*.

It might look as if Δ is increasing in n^*. But that is wrong, because of course n^* is endogenous and depends (inversely) on γ. A simple way to see what happens as γ and hence n^* varies is to note that, if we ignore the fact that membership has to be an integer, then $n^*\gamma \cong 1$. So, to a reasonable approximation, the savings in damage costs from having an IEA are *independent of the size of the IEA*. This just reflects the intuition that the stable IEA is the minimum size that makes abatement worthwhile for signatories, and this is inversely proportional to the costs of damage. On the other hand, as n^* increases, the average loss of output from the formation of an IEA increases. So the gains from having an IEA are decreasing in the size of the IEA, and increasing in γ.

3. STOCK POLLUTANT WITH UNCERTAINTY AND LEARNING

I now extend the model of the previous section to deal with a stock pollutant with uncertainty and learning. Following Rubio and Ulph (2002a), the simplest possible extension involves 2 periods of time, indexed $t = 1,2$ and the assumption that the damage costs in period t depend on the stock of pollution $Z_t = Z_{t-1} + Q_t$. Note that, for simplicity, I am ignoring any natural decay in the stock of the pollutant (I will also ignore discounting). I normalise by assuming that $Z_0 = 0$. I continue to assume constant unit benefit and damage costs, and that unit damage costs do not depend on the stock of pollution. In analysing how the dynamics of the stock pollutant affects membership of an IEA, following Rubio and Ulph (2002a,b) I consider two cases: *fixed membership (fm)*, where countries decide at the start of period 1 whether or not to join an IEA and are committed to being ether a signatory or a non-signatory for both periods, and *variable membership (vm)*, where countries decide each period whether or not to join an IEA.

To introduce uncertainty, I assume that unit damage costs (relative to benefits) are not known with certainty. For simplicity, I assume that with equal probability unit damage costs can be either high or low: $\gamma_h \equiv \gamma + \delta; \gamma_l = \gamma - \delta$, so that expected damage costs are γ, and δ is a measure of the dispersion of damage costs. Note that I continue to assume that all countries are homogeneous, so that damage costs are perfectly correlated across countries. To capture learning, again I assume the simplest form of learning (following (Ulph and Ulph (1996), Na and Shin (1998)) namely that if learning takes place then between period 1 and 2 all countries learn with certainty the true value of unit damage costs. If there is no learning then countries have to make their period 2 decisions before they know the true value of unit damage costs[7]. In analysing membership of an IEA I will consider the two cases of *learning (l)* and *no learning (nl)*. Finally I assume that all countries are risk neutral.

Putting the above together there are four cases I shall consider in the following order: no learning with fixed membership, no learning with variable membership, learning with fixed membership and learning with variable membership[8]. Before doing the analysis I make the following assumption about parameter values:

Assumption 2. $1/N < \gamma_l < \gamma < 0.5; \quad \gamma < \gamma_h < 1;$

i.e. $1/N < \gamma < 0.5; \quad 0 < \delta < (\gamma - 1/N) < 1 - \gamma$

Thus δ can take values between 0 and $\bar{\delta} \equiv [\gamma - 1/N]$; I write $\delta = \theta\bar{\delta}$ where $0 < \theta < 1$, and θ is a measure of the degree of uncertainty about damage costs which is independent of other parameters. The two key parameters in this model are γ and θ.

[7] Perhaps because it takes time to process the information about damage costs that occurred in period 1.
[8] The model with No Learning is identical to the model in Rubio and Ulph (2002a) with certainty.

3.1 No Learning, Fixed Membership.

Country *i*'s expected value function over the two periods are given by:

$$V_i(q_{i1}, q_{i2}, Q_{i1}, Q_{i2}) = q_{i1} - 2\gamma(q_{i1} + Q_{i2}) + q_{i2} - \gamma(q_{i2} + Q_{i2}) \quad (1a)$$

Summing over all *N* countries to get the expected global value function we have:

$$V^g \equiv \sum_i V_i(q_{i1}, q_{i2}, Q_{i1}, Q_{i2}) = q_{i1} + Q_{i2} - 2N\gamma(q_{i1} + Q_{i1}) + q_{i2} + Q_{i2} - N\gamma(q_{i2} + Q_{i2}) \quad (1b)$$

In both cases a unit of pollution emitted in period 1 has twice the expected damage cost of a unit of pollution emitted in period 2, because the damages last for 2 periods. It is straightforward to see that[9]:

Lemma 4. *Given Assumption 2: (i) if all countries act non-cooperatively then the dominant strategy for each country is to set* $q_{i1} = q_{i2} = 1$ *(i.e. pollute in both periods) and each country gets an expected payoff 2-3Nγ; (ii) if all countries cooperate then each will set* $q_{i1} = q_{i2} = 0$ *(i.e. abate in both periods) and each country gets an expected payoff 0; (iii) the expected gains to cooperation are 3Nγ - 2.*

As long as $2\gamma < 1$ then each country acting non-cooperatively will always want to pollute, even in period 1 when it is most costly, while as long as $N\gamma > 1$, then when all countries cooperate they will always want to abate pollution, even in period 2 when pollution is least costly.

Turning to the formation of an IEA, I again model this as a two stage game in which in stage 1 countries decide whether or not to join the IEA, to which they are committed for 2 periods, while in stage 2 they choose their emission levels over both periods. I start with the second-stage game.

Lemma 5 (Emission Game – Strategies). *For any number of signatories, n, the optimal strategy of the non-signatories is to pollute in both periods, no matter what the signatory countries do. The optimal strategy of the signatories is for each to abate in both periods if* $n\gamma \geq 1$*; to abate in period 1 and pollute in period 2 if* $2n\gamma \geq 1 > n\gamma$*; and to pollute in both periods if* $1 > 2n\gamma$*.*

Again, the optimal strategy for the non-signatories just follows from Lemma 4. The optimal strategy for the signatories reflects the fact that because pollution in period 2 is less costly than in period 1, the IEA needs a bigger number of signatories to justify abating pollution in both periods than to justify abating pollution just in period 1. I now define $\overline{n} \equiv \Omega(1/2\gamma) \leq n^* = \Omega(1/\gamma)$, where, by Assumption 2, $2 \leq \overline{n} \leq N/2; 3 \leq n^* \leq N$. n^* is the smallest threshold size at which pays IEA signatories to abate in both periods, and $\overline{n}$ is the smallest threshold size at which it pays to IEA signatories to abate in period 1. Then the expected payoffs are:

[9] All proofs are in the Appendix

Lemma 6 (Emission Game – Payoffs) *For any number of signatories, n, the expected payoffs to signatories and non-signatories are:*

$$N \geq n \geq n^* \qquad V^s(n) = -3\gamma N + 3\gamma n; \qquad V^f(n) = -3\gamma N + 2 + 3\gamma n$$
$$n^* > n \geq \bar{n} \qquad V^s(n) = -3\gamma N + 1 + 2\gamma n; \quad V^f(n) = -3\gamma N + 2 + 2\gamma n$$
$$\bar{n} > n \geq 2 \qquad V^s(n) = V^f(n) = -3\gamma N + 2$$

Turning to the first-stage membership game we have:

Result 2 (Membership Game) *For any value of γ there is a unique stable IEA of the fixed membership with no learning case, which can take one of two possible types as follows:*

$$0.4 \leq \gamma \leq 0.5: \quad n_{fmnl} = n^*; \quad V^s_{fmnl} = -3\gamma N + 3\gamma n^*; \qquad V^f_{fmnl} = -3\gamma N + 2 + 3\gamma n^*$$
$$1/N < \gamma < 0.4: \quad n_{fmnl} = \bar{n}; \quad V^s_{fmnl} = -3\gamma N + 1 + 2\gamma \bar{n}; \quad V^f_{fmnl} = -3\gamma N + 2 + 2\gamma \bar{n}$$

The intuition is that again free-riding will cause defection down to a minimum critical threshold at which IEA signatories will change their strategy. For relatively high values of γ the threat that the signatories will not abate in period 2 is sufficient to deter defections below n^*. However for lower values of γ it is not until one reaches the critical threshold, $\bar{n}$, below which the signatories would not abate in either period, that this is sufficient to deter any defections. As noted in Rubio and Ulph (2002a), this model explains why for a wide range of parameter values, an IEA may become ineffective in later stages of dealing with a stock pollutant. For future reference I will refer to the first type of stable IEA for this case as FMNL(i), in which there are n^* signatories who abate in both periods, and to the second type of stable IEA as FMNL(ii), in which there are $\bar{n}$ signatories who abate only in period 1.

3.2 No Learning – Variable Membership

In this case I continue to assume that there is no learning, but now allow that in each time period countries are free to join or leave the IEA. Of course countries in period 1 will need to understand how their decisions affect the stock of emissions and hence the membership and emission decisions and their associated payoffs in period 2. So I work backwards.

3.2.1 Period 2.

At the start of period 2 the world inherits a stock of pollution, Q_1, $0 \leq Q_1 \leq N$, from period 1. From (1), the period 2 expected payoff to any country *i* is :

$$\pi_{i2} = q_{i2} - \gamma(q_{i2} + Q_{i2}) - \gamma Q_1$$

The last term is a constant as far as period 2 decisions are concerned. So the model is isomorphic to the static model with certain cost-benefit parameter γ set out in Section 2, and Result 1 leads immediately to:

Result 3 (Period 2 –Membership) *The unique stable IEA in period 2 of the model with no learning and variable membership is of size n** $(=\Omega(1/\gamma))$ *and all signatories abate pollution in period 2. The expected period 2 payoffs to signatories and non-signatories are:*

$V_2^s(Q_1) = -\gamma(Q_1 + N - n^*); V_2^f(Q_1) = 1 + V_2^s(Q_1)$.

Note that membership in period 2, *n**, is independent of Q_1.

3.2.1. Period 1.

The first issue I need to address is what a country believes about how it's decisions in period 1 affect the payoff it will get in period 2. Obviously one link is through the effect of period 1 decisions on Q_1. But that is not enough to determine period 2 payoff, because as Result 3 notes, that payoff will also depend on whether a country is a signatory or not in period 2. Rubio and Ulph (2002a), discuss a number of ways a country in period 1 might form beliefs about whether it will be a signatory or not in period 2. I shall take their simplest approach – the Random Assignment Rule - of assuming that there is a random process of choosing signatories in period 2 in which each country has the same probability – *n*/N* – of being a signatory. A justification for this approach might be that, with homogeneous countries, all that the stability analysis can explain is how many countries will sign the IEA, it cannot explain which countries sign up. The implicit assumption that a country's chances of being a signatory in period 2 are independent of whether it was a signatory or not in period 1 reflects a view that countries are unable to make commitments about membership from one period to the next. As noted, Rubio and Ulph (2002a) explore other ways of modelling these beliefs, including one in which a country's chances of being a signatory in period 2 depend strongly on whether it was a signatory in period 1. I leave it for further research to explore the implications of using these other approaches in a model with uncertainty and learning.

The implication of the Random Assignment Rule is that each country has the same expected period 2 payoff:

$$\overline{V}_2(Q_1) \equiv \frac{n^*}{N}V^s(Q_1) + \frac{N-n^*}{N}V^f(Q_1) = \frac{N-n^*}{N} - \gamma(Q_1 + N - n^*)$$

Then the expected payoff to country *i* in period 1 is:

$$\pi_{i1}(q_{i1}, Q_{i1}) = q_{i1} - 2\gamma(q_{i1} + Q_{i1}) - (N - n^*)[\gamma - \frac{1}{N}]$$

The last term is a constant and so this is iso-morphic to the simple static model with cost-benefit parameter 2γ. Result 1, combined with Result 3, immediately yields:

Result 4 (Membership Game) *The unique stable IEA of the case of no learning with variable membership has memberships* n^1_{vmnl}, n^2_{vmnl} *in periods 1 and 2 where* $n^1_{vmnl} = \overline{n}, n^2_{vmnl} = n^*$ *and in each period all signatories abate pollution. The expected payoffs to signatories and non-signatories are:*

$V^s_{vmnl} = -2\gamma(N-\bar{n}) - (N-n^*)[\gamma - 1/N]; \quad V^f_{vmnl} = 1 - 2\gamma(N-\bar{n}) - (N-n^*)[\gamma - 1/N].$

3.3 Learning – Fixed Membership.

Between period 1 and period 2, countries learn whether damage costs are high or low, so countries are able to condition their emissions in period 2 on what they have learned. So now each country decides on 3 emission levels: q_{i1}, q_{i2h}, q_{i2l}. Thus the expected payoff to country *i* is now:

$V_i(q_{i1}, q_{i2h}, q_{i2l}, Q_{i1}, Q_{i2h}, Q_{i2l}) = q_{i1} - 2\gamma(q_{i1} + Q_{i1}) + 0.5\sum_{j=h,l}[q_{i2j} - \gamma_j(q_{i2j} + Q_{i2j})]$

It is then straightforward to see that:

Lemma 7 *Given Assumption 2 (i) if all countries act non-cooperatively then the dominant strategy for country is to set* $q_{i1} = q_{i2h} = q_{i2l} = 1$*, i.e. pollute in each period and each state of the world, with expected payoff 2-3γN; (ii) if all countries cooperate then each will set* $q_{i1} = q_{i2h} = q_{i2l} = 0$*. Expected payoffs and expected gains from learning are the same as with No Learning.*

$\gamma_h < 1$ ensures that countries acting non-cooperatively want to pollute even if damage costs turn out to be high, while $\gamma_l > 1/N$ ensures that when all countries cooperate they want to abate pollution even when damage costs are low.

Turning to the analysis of the IEA, again I start with the second-stage emission game, where now countries can fine-tune their emissions in period 2 to the information they have gained. It is straightforward to show:

Lemma 8 (Emission Game- Strategies) *For any number of signatories, n, the optimal strategy for the non-signatories is to pollute in both periods and both states of the world no matter what the signatories do. For signatories, if* $2n\gamma > n\gamma_h > n\gamma_l \geq 1$*, signatories will abate in both periods and in both states of the world; if* $2n\gamma > n\gamma_h \geq 1 > n\gamma_l$*, signatories will abate in period 1 and in period 2 if damage costs are high, but will pollute in period 2 if damage costs are low; if* $2n\gamma \geq 1 > n\gamma_h > n\gamma_l$*; finally if* $1 > 2n\gamma > n\gamma_h > n\gamma_l$ *signatories will pollute in both periods and both states of the world.*

I now define two new critical threshold IEA sizes: $\tilde{n} = \Omega(1/\gamma_l)$ and $\hat{n} = \Omega(1/\gamma_h)$, where, by Assumption 2, $\tilde{n} \geq n^* \geq \hat{n} \geq \bar{n}$. Lemma 8 says that membership needs to be relatively large ($n \geq \tilde{n}$) for signatories to find it pays to abate in all periods and states of the world, in particular in the low damage cost state in period 2. As membership declines, signatories start polluting in increasing order of damage costs: first ($\tilde{n} > n \geq \hat{n}$) only in low damage costs states in period 2; then ($\hat{n} > n \geq \bar{n}$) in both high and low damage cost states in period 2; and finally ($\bar{n} > n$) in both period 1 and period 2. This leads to:

Lemma 9 (Emission Game: Payoffs) *For any number of signatories, n, the expected payoffs to signatories and non-signatories are as follows:*

$n \geq \tilde{n}$ $\quad V^s = -3\gamma N + 3\gamma n \quad V^f = -3\gamma N + 3\gamma n + 2$

$\tilde{n} > n \geq \hat{n}$ $\quad V^s = -3\gamma N + n(2.5\gamma + 0.5\delta) + 0.5 \quad V^f = -3\gamma N + n(2.5\gamma + 0.5\delta) + 2$

$\hat{n} > n \geq \bar{n}$ $\quad V^s = -3\gamma N + 2\gamma n + 1 \quad V^f = -3\gamma N + 2\gamma n + 2$

$\bar{n} > n$ $\quad V^s = -3\gamma N + 2 \quad V^f = -3\gamma N + 2$

Note that in Lemma 9 it is assumed that $\tilde{n}$, $\hat{n}$ and $\bar{n}$ are distinct. But that may not be the case. For small values of δ, (low uncertainty) there will be little difference between γ_h, γ_l and hence it is possible to have $\tilde{n} = \hat{n} = n^* > \bar{n}$; on the other hand, as δ approaches γ, γ_h approaches 2γ, and it is possible to have $\bar{n} = \hat{n} < n^* < \tilde{n}$. Of course, for any n the precise rules set out in Lemma 8 determine which strategy the IEA will choose.

The fact that $\tilde{n}$, $\hat{n}$ and $\bar{n}$ may not all be distinct also complicates providing an analytical statement of the stability results, because the nature of the operator in taking smallest integers not less than a real number, makes it difficult to give precise statements on parameter values about the size of the differences between $\tilde{n}$, $\hat{n}$ and $\bar{n}$ and hence to relate these to parameter values needed to ensure stability of IEAs. The following result summarises what can be said analytically:

Result 5 *There are three possible stable IEAs with membership, and expected payoffs as follows:*

FML(i) $n_{fml} = \tilde{n}; \quad V^s_{fml} = -3\gamma N + 3\gamma\tilde{n}; \quad V^f_{fml} = 2 - 3\gamma N + 3\gamma\tilde{n};$

FML(ii) $n_{fml} = \hat{n}; \quad V^s_{fml} = -3\gamma N + 0.5[1 + \hat{n}(5\gamma + \delta)]; \quad V^f_{fml} = -3\gamma N + 0.5[4 + \hat{n}(5\gamma + \delta)];$

FML(iii) $n_{fml} = \bar{n}; \quad V^s_{fml} = -3\gamma N + 2\gamma\bar{n} + 1; \quad V^f_{fml} = -3\gamma N + 2\gamma\bar{n} + 2;$

(i) *There cannot be a stable IEA of size greater than $\tilde{n}$ or less than $\bar{n}$.*

(ii) *If $\tilde{n} - \hat{n} \geq 1; \hat{n} - \bar{n} \geq 2$ then the unique stable IEA will be $\bar{n}$.*

(iii) *If $\tilde{n} - \hat{n} \geq 1; \hat{n} - \bar{n} = 1$ then there are parameter values for which $\hat{n}$ is stable, and parameter values for which $\bar{n}$ is unstable.*

(iv) *If $\tilde{n} = \hat{n}$ then there are parameter values for which $\tilde{n}$ is stable and parameter values for which $\bar{n}$ is unstable.*

(v) *If $\hat{n} = \bar{n}$ then $\hat{n}$ is the unique stable IEA.*

Results 5(iii) and 5(iv) do not exclude the possibility that there may be either no stable IEA or multiple stable IEAs. Nor does Result 5 provide a clear indication of the parameter values for which the different stable IEAs arise. To make more progress I have conducted numerical simulations[10], which lead to the following:

Result 5′ *For the case of fixed membership with learning, for each set of parameter values for N, γ and θ (δ) there is a unique stable IEA which will be one of FML(i),*

[10] The numerical simulations used 6 values of N = 25, 50,…, 150, 1000 values of γ between 1/N and 1, and 1000 values of θ between 0 and 1.

FML(ii) or FML(iii). For any N the regions of γ, θ parameter space in which each is the unique stable IEA are as follows: Region A: FML(i); Region B: FML(i); Region C: FML(iii). Regions A, B and C are illustrated in Figure 1.

The first point to note is that Region B is not a connected region of parameter space. There is one connected area (which, for later purposes, I have split into two: B(i) and B(ii) depending on whether $\gamma \geq 0.4$ or $\gamma < 0.4$); and then, depending on *N*, there are a small number of areas like B(ii) and B(iii).

The intuition behind Result 5' is as follows. Note first that the stable IEA FML(iii) is the same as the stable IEA FMNL(ii). When θ =0, $\tilde{n} = \hat{n} = n^*$ and FML(i) is the same as FMNL(i). So, as Figure 1 shows, when θ = 0, the outcome of fixed membership with learning is the same as fixed membership with no learning, and FML(i) is the unique stable IEA if $0.5 \geq \gamma \geq 0.4$ while FML(iii) is the unique stable IEA if $1/N \leq \gamma < 0.4$. As shown in Figure 1, regions A and C, these remain the unique stable IEAs for small positive values of θ. From Result 5 (iv) FML(i) is likely to remain stable as long as $\tilde{n} = \hat{n}$, and this requires large values of γ and small values of δ and hence small values of θ, which get smaller as γ increases. For larger values of θ $\tilde{n}$ is no longer stable, and the stable IEA switches to FML(ii). From Results 5 (ii), (iii) and (v), FML(iii) is more likely to be stable than FML(ii) when there is a significant difference between $\hat{n}$ and $\overline{n}$, which again requires relatively low values of δ, and hence relatively low values of θ, but as γ gets smaller the values of θ for which FML(ii) can be stable get larger. But these last set of results depend quite sensitively on the distance between $\hat{n}$ and $\overline{n}$, and because these depend on the operator $\Omega(x)$ - taking the smallest integer not less than *x* - as γ and θ vary, the distance will not vary smoothly, and that is why disconnected sub-regions like B(iii) and B(iv) can arise.

The numerical simulations show that, to a first approximation, Region A accounts for about 2% of parameter space, Region B for about 40% and Region C for about 58%.

3.4 Learning – Variable Membership

3.4.1. Period 2

By the start of period 2, countries will know whether damage costs are high or low. Given a stock of pollution, Q_1, inherited from period 1, the payoff function to country *i* in period 2 in state *j=l,h* is: $\pi_{i2j} = q_{i2j} - \gamma_j(q_{i2j} + Q_{i2j}) - \gamma_j Q_1$. The last term is a constant for period 2 state *j*, so the model is again iso-morphic to the simple static model with certainty of section 2. Applying Result 1 yields:

Result 6 *The unique stable IEA in period 2, state j=l,h of the model with learning and variable membership has membership* $n_{2l} = \tilde{n} = \Omega(1/\gamma_l)$ *and* $n_{2h} = \hat{n} = \Omega(1/\gamma_h)$ *if damage costs are low and high respectively. Payoffs to signatories and non-signatories are:*

$$V_{2j}^{s}(Q_1) = -\gamma_j(Q_1 + N - n_{2j}); \quad V_{2j}^{f} = 1 + V_{2j}^{s}(Q_1), j=l,h$$

Note that if damage costs are high, there will be fewer signatories to an IEA in period 2 than if damage costs are low. This simply reflects the property of the static Barrett model that IEA membership is inversely related to the gains from cooperation. Note also that, as in Result 3, period 2 membership is independent of Q_1.

3.4.2 Period 1.

In period 1, countries have to form a view about whether they will be signatories or non-signatories in each state of the world in period 2. I continue to employ the Random Assignment Rule, and assume that in state $j = l,h$ each country has the same probability, n_{2j}/N, of being a signatory. Thus the expected period 2 payoff (taking expectation across likelihood of being a signatory or non-signatory) to each country in state $j=l,h$ is: $\overline{V}_{2j}(Q_1) = \frac{N - n_{2j}}{N} - \gamma_j(Q_1 + N - n_{2j})$. Taking expectations across the states of the world, each country has expected Period 2 payoff:

$$\overline{\overline{V}}_2(Q_1) = \frac{N - \overline{\overline{n}}_2}{N} - \gamma(Q_1 + N) + \phi \text{ where:}$$

$$\overline{\overline{n}}_2 = 0.5(n_{2l} + n_{2h}), \quad \phi = 0.5(\gamma_{2l} n_{2l} + \gamma_{2h} n_{2h})$$

$\overline{\overline{n}}_2$ is expected membership in period 2; ϕ is expected savings in damage costs from period 2 IEA membership, relative to having no IEA members in period 2 at all.

Then the period 1 expected payoff function for country i in period 1 is:

$$V_{i1}(q_{i1}, Q_{i1}) = q_{i1} - 2\gamma(q_{i1} + Q_{i1}) + \frac{N - \overline{\overline{n}}_2}{N} - \gamma N + \phi.$$

The last three terms are constants, and so this is again iso-morphic to the simple static model with certainty in Section 2, and so Result 1 again yields:

Result 7 *The unique stable IEA of the case with learning and variable membership has memberships:* $n_{vml}^{1} = \overline{n}$, $n_{vml}^{2l} = \tilde{n}$, $n_{vml}^{2h} = \hat{n}$ *in period 1 and period 2 states j=l,h respectively, and expected payoffs to signatories and non-signatories:*

$$V_{vml}^{s} = -2\gamma(N - \overline{n}) + \frac{N - \overline{\overline{n}}_2}{N} - \gamma N + \phi; \quad V_{vml}^{f} = 1 - 2\gamma(N - \overline{n}) + \frac{N - \overline{\overline{n}}_2}{N} - \gamma N + \phi.$$

Note that, for this simple model, with variable membership, period 1 membership is the same in period 1 with learning and no learning.

This completes the analysis of the stable IEAs for the four cases. In the next section I compare the outcomes in terms of membership and expected payoffs.

4. Comparison of Learning/No Learning, Fixed/Variable Membership

In this section I consider the questions: how does learning affect membership and expected payoffs and how does having membership fixed rather than variable affect membership and expected payoffs?

4.1 Comparison of Learning and No Learning.

4.1.1 Fixed Membership.

It will be useful to recap the properties of the different stable IEAs for Fixed Membership with No Learning and Learning, and the parameter values for which they arise, which I do with reference to Figure 1. With No Learning, there are two possible stable IEAs: FMNL(i): membership n^* with strategy (for signatories) (0,0,0); and FMNL(ii): membership $\bar{n}$ with strategy (0,1,1). FMNL(i) occurs when $0.4 \leq \gamma < 0.5$, i.e. area A and B(i) of Figure 1, FMNL(ii) otherwise. With Learning there are three possible stable IEAs: FML(i): membership $\tilde{n}$ with strategy (0,0,0); FML(ii): membership $\hat{n}$ with strategy (0,0,1); and FML(iii): membership $\bar{n}$ with strategy (0,1,1); they occur in Regions A, B and C of Figure 1 respectively.

The comparisons between No Learning and Learning are now given for each region:

Result 8 *The comparison between Learning and No Learning varies over the four regions of parameter space as follows:*

- *A. FML(i) involves the same strategy and hence same amount of pollution by each signatory as FMNL(i), but FML(i) has more signatories ($\tilde{n}$) than FMNL(i) (n^*); and hence global pollution is less with Learning than No Learning; signatories and non-signatories are better off with Learning than No Learning, and in terms of global welfare, Learning is better then No Learning.*
- *B(i) FML(ii) involves more pollution per signatory and has fewer signatories ($\hat{n}$) than FMNL(i) (n^*), and hence global pollution is higher with Learning than No Learning; signatories can be better off or worse off with Learning than with No learning, but non-signatories are unambiguously worse off with Learning than No Learning and global welfare is lower with Learning than No Learning. Learning worse than No Learning*
- *B(ii)- B(iv) FML(ii) involves less pollution per signatory and has at least as many signatories ($\hat{n}$) than FMNL(ii) ($\bar{n}$) and so global pollution is lower with Learning than No Learning; signatories can be better off or worse off with Learning than No Learning but non-signatories are unambiguously better off with Learning than No Learning and global welfare is higher with Learning than No learning*
- *C FML(iii) and FMNL(ii) are identical in all respects, so Learning and No learning are equivalent.*

Thus in Regions A and B(ii)-B(iv) Learning is better than No Learning, in Region B(i) No Learning is better than Learning and in Region C they are identical. In summary,

Learning produces higher global welfare when it leads to more signatories than No Learning, and who in addition may do more abatement (over time and states of the world) than with No learning; and *vice versa* when No Learning is preferred to Learning.

4.1.2 Variable Membership.

The analysis for this case is more straightforward. From Results 4, with No Learning membership in period 1 will be $\bar{n}$, in period 2 n^*. From Result 7, with Learning membership in period 1 is also $\bar{n}$, membership in period 2 is $\tilde{n}, \hat{n}$ $(\tilde{n} > n^* > \hat{n})$ if damage costs are low, high respectively, with expected membership in period 2 $\bar{\bar{n}}_2 = (\tilde{n} + \hat{n})/2$. In terms of expected payoffs, it is straightforward to show that:

$$V^s_{vml} - V^s_{vmnl} = V^f_{vml} - V^f_{vmnl} \equiv \Gamma_{vm} = (\phi - \gamma n^*) - (\bar{\bar{n}}_2 - n^*)/N$$
$$\text{where } \phi = 0.5*[\gamma_h \hat{n} + \gamma_l \tilde{n}].$$

I define Γ_{vm} as the expected gains from learning with variable membership. It can be divided into two parts. $\Gamma^1_{vm} \equiv \phi - \gamma n^*$ is the difference in the expected savings in damage costs from having members of the IEA in period 2 between Learning and No Learning. The second term, $\Gamma^2_{vm} \equiv -(\bar{\bar{n}}_2 - n^*)/N$, is the difference between the average loss of output in period 2 from having members of the IEA between Learning and No learning.

If it was possible to ignore the fact that membership has to be an integer, then $n^* = 1/\gamma; \tilde{n} = 1/\gamma_l; \hat{n} = 1/\gamma_h; \gamma n^* = 1; \phi = 1; \bar{\bar{n}}_2 = n^*(\frac{\gamma^2}{\gamma^2 - \delta^2}) > n^*; \Gamma^1_{vm} = 0; \Gamma_{vm} = \Gamma^2_{vm} < 0$
In words, expected period 2 membership with Learning will be larger than period 2 membership with No Learning, and this gap will increase as the degree of uncertainty increases; the expected savings in damage costs from having an IEA are the same with Learning and No Learning (being 1 in both cases); and expected loss of output from having an IEA is greater with Learning than with No Learning. So overall there will be lower welfare with Learning than No Learning. This is a straightforward application of the result in Section 2 that the bigger the IEA that forms the smaller are the gains from having an IEA; Learning results, on average, in a bigger IEA in period 2 (and the same size in period 1), so welfare falls with Learning. This difference increases as the degree of uncertainty increases.

Of course the approximation argument used in the last paragraph is not correct. What would be expected is that for small values of θ, the difference between $\bar{\bar{n}}_2$ and n^* will be small, and hence Γ^2_{vm} may be dominated by the approximation errors. As θ increases however, the gap in membership will widen and Γ^2_{vm} should become unambiguously negative. On the other hand, the approximation errors are likely to mean that Γ^1_{vm} is as likely to be positive as negative, and is likely to be close to zero no matter what value θ takes. This is borne out by simulation results. Taking N =100, for each value of θ =

0.1,…,0.9. I take 1000 values of γ, calculate $\Gamma^1_{vm}, \Gamma^2_{vm}$ and Γ_{vm}, and then calculate the average value of these indicators over the 1000 values of γ, and the proportion of these 1000 values for which $\Gamma^1_{vm}, \Gamma^2_{vm}$ and Γ_{vm} are positive. The results are shown in Table 1. As predicted, as θ increases, Γ^2_{vm} becomes unambiguously negative and increases in absolute size. On the other hand the approximation of Γ^1_{vm} to zero gets better as θ increases, with the average value tending to zero and the proportion of cases which are positive tending to 50%. On average overall gains from learning are negative, get more negative as θ increases, and have a decreasing % of cases of positive values. To summarise:

Result 9 *With variable membership, period 1 membership is the same whether there is Learning or No Learning. Expected period 2 membership is higher with Learning than with No Learning and expected payoffs (to both signatories and non-signatories) are lower with Learning than No Learning. Gains from learning are negative and decline as degree of uncertainty increases.*

It is worth noting that, when comparing Learning and No learning, there is an important difference between fixed and variable membership. With fixed membership, Learning produces higher global welfare than No Learning when membership is higher with Learning than with No Learning. But with variable membership, Learning generates lower global welfare than No Learning precisely because it leads to higher (expected) membership in period 2 than period 1. This is because the model with variable membership is essentially like a sequence of static models, and inherits the feature of the static model that IEAs with higher membership generate lower benefits, because membership adjusts so that savings in damage costs are (approximately) constant while bigger IEAs lead to lower output.

4.2 Comparison of Fixed and Variable Membership

In this subsection I am interested in the question whether it is better to fix membership (force countries to commit to being a signatory or non-signatory for both periods[11]) or to allow countries to decide each period (and each state of the world if learning takes place) whether to join or not. I consider first the case of No Learning and then consider Learning.

4.2.1 No Learning

Putting together Results 2 and 4, if $0.4 \leq \gamma \leq 0.5$, fixed membership will have n^* (=3) signatories who abate pollution in each period, while variable membership will have only $\bar{n}$ (=2) signatories who abate in period 1 and n^* who abate in period 2. So fixed membership will result in less global pollution than variable membership. If $1/N < \gamma < 0.4$, then fixed membership has $\bar{n}$ signatories in both periods, but they abate only in period 1, while variable membership again has $\bar{n}$ signatories who abate in period 1 but

[11] I ignore the issue of how such commitment could be enforced.

$n^* \geq \bar{n}$ signatories who abate in period 2. So variable membership results in less global pollution than fixed membership. A comparison in terms of expected payoffs is given in the following:

Result 10 *If there is No Learning, and* $0.4 \leq \gamma \leq 0.5$*, then with fixed membership, the number of signatories is higher in period 1 and no lower in period 2 than with variable membership; expected global welfare is higher with fixed membership than with variable membership; non-signatories will be better off with fixed membership than variable membership but signatories will only be better off with fixed membership than with variable membership if* $N < 3/(1-2\gamma)$*, for which a sufficient condition is* $N<15$*. If* $1/N < \gamma < 0.4$*, then with variable membership the number of signatories is higher in period 2 than with fixed membership and no lower in period 1; both signatories and non-signatories are better off with variable membership than with fixed, and expected global welfare is higher with variable membership than with fixed membership.*

4.2.2 Learning

Putting together Results 5' and 7, Table 2 summarises the number of signatories and whether they abate (0) or pollute (1) for each period and state of the world for the stable IEA with variable membership and learning and the three possible stable IEAs for fixed membership with learning: FML(i), FML(ii) and FML(iii).

Table 5 Number of Signatories and Their Emission

	Period 1	Period 2 (low)	Period 2 (high)
VML	$\bar{n}$ (0)	$\tilde{n}$ (0)	$\hat{n}$ (0)
FML(i)	$\tilde{n}$ (0)	$\tilde{n}$ (0)	$\tilde{n}$ (0)
FML(ii)	$\hat{n}$ (0)	$\hat{n}$ (1)	$\hat{n}$ (0)
FML(iii)	$\bar{n}$ (0)	$\bar{n}$ (1)	$\bar{n}$ (1)

Recalling that $\tilde{n} \geq \hat{n} \geq \bar{n}$, relative to VML, FML(i) will entail less global pollution in period 1 and period 2 (high) and the same in period 2 (low), so overall is less polluting. FML(iii) entails more global pollution in both states in period 2 and the same in period 1, so overall is more polluting. FML(ii) entails less pollution in period 1, more pollution in period 2 (low) and the same in period 2 (high), so it is ambiguous whether overall it is more or less polluting than VML. A comparison in terms of expected payoffs is given:

Result 11 *When there is Learning, then the comparison between fixed and variable membership is as follows:*

Region A: Fixed membership entails at least as many signatories as variable membership, and usually more in period 1 and period 2 high state; global pollution is lower with fixed membership; it is ambiguous whether signatories are better off with fixed or variable membership; non-signatories are unambiguously better off, and average global welfare is higher with fixed than variable membership.

Region B: Relative to variable membership, fixed membership usually entails more signatories and hence less pollution in period 1; fewer signatories, who all pollute, and hence significantly more pollution in period 2 (high); and the same number of signatories and pollution in period 2 (low); it is ambiguous whether payoffs are higher with variable than fixed membership, but signatories are more likely to be better off with variable than fixed membership than are non-signatories.

Region C: Variable membership entails at least as many signatories as fixed membership, and usually more in period 2; moreover signatories never pollute; so global pollution is lower with variable membership than with fixed. Signatories and non-signatories are unambiguously better off with variable than with fixed membership and so average global welfare is higher with variable than fixed.

To resolve the ambiguities in Result 11, I used similar simulations as the ones already reported and found that in Region A, signatories are always worse off with fixed than variable membership, while in Region B, welfare of signatories and global welfare are always higher with variable than fixed membership; non-signatories are also always better off with variable than fixed membership as long as $\theta > 0.3$, and for smaller θ they are better off for the majority of values of γ. Thus:

Result 11′ *In terms of global welfare, variable is preferred to fixed membership except in Region A.*

Taking together Results 10 and 11', and referring to Figure 1, in terms of global welfare, with No Learning, fixed membership is preferred to variable membership in Regions A and B(i), while with Learning fixed membership is preferred to variable membership only in Region A. The common feature of these results is, not very surprisingly, that fixed membership is preferred to variable membership as long as fixed membership has at least as many signatories in each period and each state of the world, and that signatories do not pollute. With No Learning this occurs in Regions A and B(i), i.e. about 20% of parameter space; with Learning this occurs only in Region A, about 2% of parameter space. So in a substantial majority of cases it will be better to have variable than fixed membership.

5. Conclusions

In this paper I have analysed the effects of uncertainty and learning on the incentives and timing for countries to join an International Environmental Agreement. The analysis extends previous analysis by (i) explicitly modeling the process of coalition formation; (ii) working with a model of coalition formation where, in principle, it is possible to get any number of countries up to the grand coalition to join an IEA depending on parameter values, independent of the number of countries involved; (iii) by working with a model which allows for the dynamics of stock pollution, and in which it is possible to allow for two types of IEAs – ones with fixed membership or with variable membership; (iv) working with a model where countries are identical *ex ante* and *ex post*, so uncertainty is not about the distribution of gains and losses across countries, but about how large damage costs are likely to be for all countries.

The conclusions are as follows. First I compared whether it was better to have Learning and No Learning. When there is fixed membership, if expected damage costs are high and the degree of uncertainty about damage costs is high, then the outcome with Learning is worse (in terms of global welfare) than the outcome with No Learning, but otherwise Learning is at least as good as No Learning. Whether Learning is better or worse than No Learning depends simply on whether it leads to more or less signatories (sometimes reinforced by whether the signatories do more abatement). However, when membership is variable, there is a strikingly different result. Period 1 membership does not depend on whether there is Learning or No Learning; Learning leads to more signatories in period 2 in the *low* damage cost state than in the *high* damage cost state; averaging across the two states, membership in period 2 is higher with Learning than with No Learning. But global welfare is lower with Learning than with No Learning, essentially because the variable membership model inherits a feature of the static model that the higher the membership, the smaller the benefits of the IEA. Second, I compare whether it is better to have fixed or variable membership. Fixed membership will be preferred to variable membership if it entails at least as many signatories who abate pollution in each period and each state of the world. With No Learning this occurs when there are high expected damage costs (about 20% of parameter range); with Learning it occurs when there is the same high expected damage costs, but only when there is a low degree of uncertainty about damage costs (about 2% of parameter space).

Of course the model is extremely simple in a number of important respects, and I mention five. First, the dynamics of stock pollution are very simple, so that while total damage costs depend on the stock of pollution, unit damage costs are independent of the stock of pollution. One consequence of this assumption is that the model of variable membership effectively decomposes into a sequence of static models, with a particular implication that IEA membership in period 2 is independent of how much pollution was generated in period 1, and, more importantly, that period 1 membership is independent of whether Learning takes place in period 2 or not. In a companion paper (Ulph (2002)) I allow for unit damage costs to depend on the stock of pollution in period 1, along the lines of Rubio and Ulph (2002a,b). This now makes membership in period 2 a decreasing function of pollution in period 1, and also means that membership in period 1 depends on

whether Learning takes place or not. However, while this makes the model richer, and somewhat complicates the analysis (e.g. with fixed membership there can be multiple stable IEAs) it does not change the broad features of the results outlined above. In particular membership in period 1 is not significantly affected by whether learning takes place in period 2.

Second the model is limited to 2 periods. With variable membership, membership rises over time when there is No Learning and rises on average when there is Learning. This just reproduces the result in Rubio and Ulph (2002a) without uncertainty. But this is an artefact of the two-period horizon, which effectively makes pollution in the second period less damaging than in the first (because it does not last as long). Rubio and Ulph (2002b) consider the infinite-horizon model without uncertainty and show that membership declines over time. An obvious line for future research is to extend the analysis with uncertainty and learning to an infinite-horizon model.

Third, in this paper I have focussed on the case where countries are identical *ex ante* and *ex post*, uncertainty is solely about the extent of global net benefits, and learning is a very simple process. As noted in the introduction, Na and Shin (1998) focussed only on the case where there was uncertainty about the distribution of known total global benefits. It would clearly be desirable to have a model which combined both features and where there are asymmetries between countries both *ex post* and *ex ante.* Moreover in a model of more than two periods it would be desirable to model much richer processes of learning, in which there could be active and passive forms of Bayesian learning (along the lines of Karp and Zhang (2000), Kelly and Kolstad (1999)). However, unlike those models which had a single regulator, it would be desirable to model which countries engage in active or passive learning, how are costs shared, and how does this affect incentives to join an agreement.

Fourth, in this model countries are restricted to only two actions (pollute or abate). I conjecture that this is not a major restriction on the model. Finally it would be desirable to consider other concepts of stable coalitions. Clearly, there is much to be done before we really understand how uncertainty and learning affect incentives to join international environmental agreements.

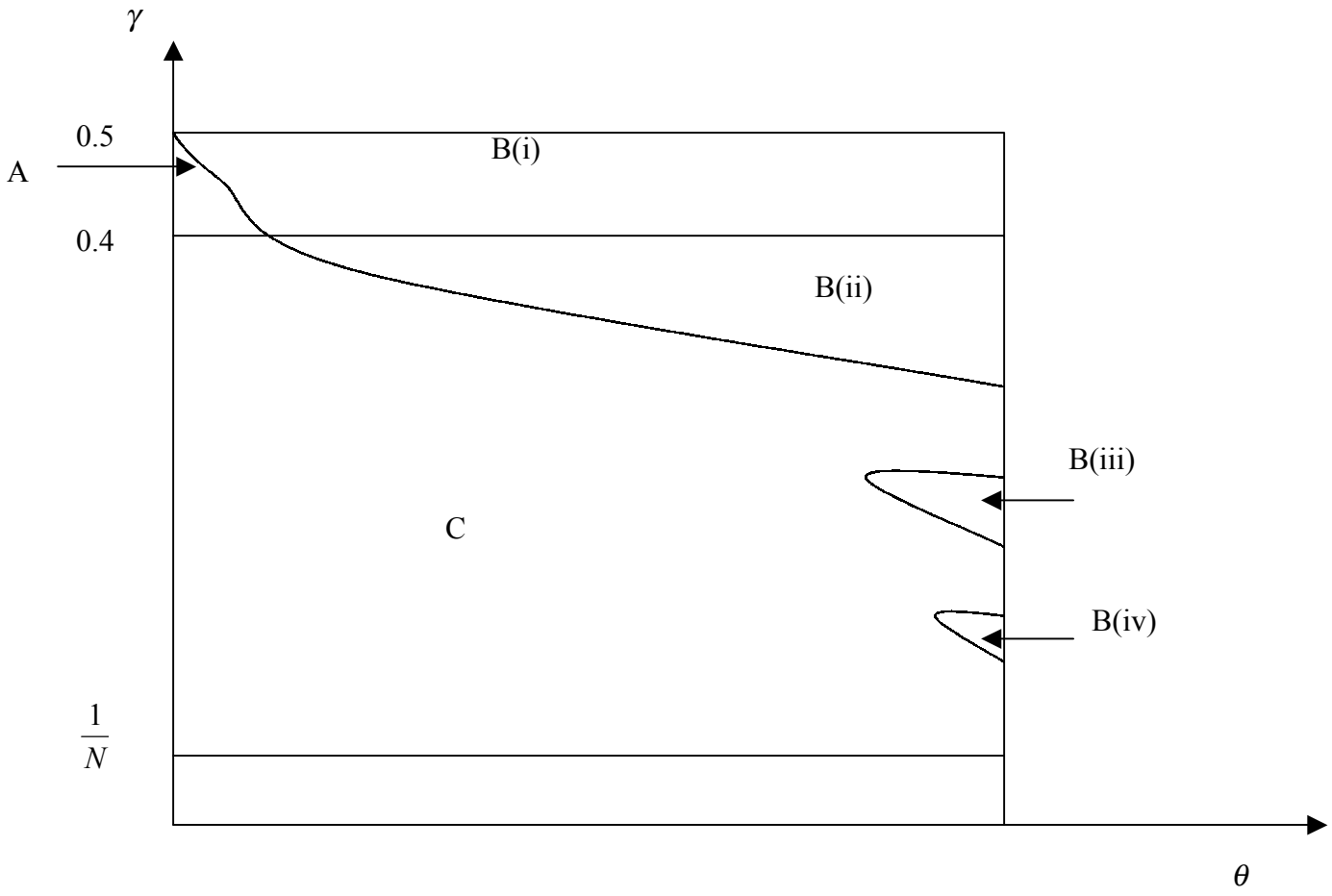


Figure 1: Fixed Membership Learning Regions of Parameter Space inWhich Different Stable IEAs Arise

Table 1: Gains from Learning- Variable Membership
($N = 100$)

	Γ^1_{vm}		Γ^2_{vm}		Gains from Learning (Γ_{vm})	
	Average	**% +ve**	**Average**	**% +ve**	**Average**	**% +ve**
θ						
0.1	**-0.0203**	**32.6**	**-0.0001**	**29.1**	**-0.0203**	**32.6**
0.2	**-0.0298**	**42.3**	**-0.0014**	**23.3**	**-0.0312**	**40.1**
0.3	**-0.0283**	**44.3**	**-0.0043**	**6.0**	**-0.0326**	**40.3**
0.4	**-0.0286**	**31.8**	**-0.0086**	**0.2**	**-0.0372**	**26.5**
0.5	**-0.0232**	**48.9**	**-0.0152**	**0.0**	**-0.0383**	**40.0**
0.6	**-0.0192**	**47.3**	**-0.0249**	**0.0**	**-0.0440**	**28.1**
0.7	**-0.0133**	**45.7**	**-0.0401**	**0.0**	**-0.0534**	**25.3**
0.8	**-0.0055**	**47.3**	**-0.0668**	**0.0**	**-0.0719**	**24.9**
0.9	**0.0040**	**55.6**	**-0.1256**	**0.0**	**-0.1216**	**14.4**

Appendix – Proofs

Lemma 4.

(i) From (1a) it is straightforward to see that, taking Q_{i1}, Q_{i2} as given, it will pay to set $q_{i1} = 1 \Leftrightarrow 1 > 2\gamma$, and $q_{i2} = 1 \Leftrightarrow 1 > \gamma$.

(ii) Similarly, from (1b), taking as given Q_{i1}, Q_{i2} it will pay to set $q_{i1} = 0 \Leftrightarrow 1 \leq 2N\gamma$ and $q_{i2} = 0 \Leftrightarrow 1 \leq N\gamma$.

Lemma 5. In period t = 1, 2, let q_{st} be the emissions of a particular signatory country, and Q_{st} the total emissions of all other signatory countries; the emissions by non-signatories are $N - n$ in each period. By analogy with (1b) the total expected payoff function for the n signatory countries is:

$$V^n(q_{s1}, q_{s2}, Q_{s1}, Q_{s2}) \equiv q_{s1} + Q_{s1} - 2n\gamma(q_{s1} + Q_{s1} + N - n) + q_{s2} + Q_{s2} - n\gamma(q_{s2} + Q_{s2} + N - n)$$

The IEA wants to determine q_{st} taking as given Q_{st} and the emissions of the non-signatories, (of course in equilibrium $Q_{st} = (n-1)q_{st}$) so as to maximise V^n. So the IEA will set $q_{s1} = 0 \Leftrightarrow 1 \leq 2n\gamma$ and $q_{s2} = 0 \Leftrightarrow 1 \leq n\gamma$.

Result 2. Note first that if $\frac{1}{3} \leq \gamma \leq \frac{1}{2}$, $n^* = 3$, $\bar{n}$ =2; while for all γ < 1/3, $n^* - \bar{n} > 1$.

I consider 2 cases.

(i) $\frac{1}{3} \leq \gamma \leq \frac{1}{2}$.

(a) Stability of n^*.

External Stability satisfied iff $2 + 3\gamma n^* \geq 3\gamma(n^* + 1)$, i.e. $\gamma \leq 2/3$; which is satisfied since $\gamma \leq 0.5$

Internal Stability satisfied iff $3\gamma n^* \geq 2 + 2\gamma(n^* - 1)$, i.e. iff $\gamma \geq 2/5$

(b) Stability of $\bar{n}$.

Internal stability satisfied iff $1 + 2\gamma\bar{n} \geq 2 \Leftrightarrow \bar{n} \geq 1/2\gamma$, true by definition of $\bar{n}$.

External stability satisfied iff $2 + 2\gamma\bar{n} \geq 3\gamma(\bar{n} + 1)$ iff $\gamma \leq 2/5$.

Thus n^* is the unique stable IEA if $\gamma \geq 2/5$, while $\bar{n}$ is the unique stable IEA if $\gamma < 2/5$.

(ii) $\gamma < 1/3$. External Stability of n^* and Internal Stability of $\bar{n}$ are the same as above. Internal Stability of n^* is $3\gamma n^* \geq 2 + 2\gamma(n^* - 1)$, i.e. $n^* \geq (2 - 2\gamma)/\gamma$. But n^* is smallest integer not less than $1/\gamma$. So condition

cannot be satisfied if $\frac{2-2\gamma}{\gamma}-\frac{1}{\gamma}>1$, i.e. $\gamma<1/3$. So n^* is not internally stable. External Stability of $\bar{n}$ is satisfied iff $2+2\gamma\bar{n}\geq 1+2\gamma(\bar{n}+1)$ iff $\gamma\leq 0.5$, which is satisfied. So for $\gamma<1/3$, only $\bar{n}$ is stable.

In summary, for $\gamma\geq 0.4$, n^* is the unique stable IEA. For $\gamma<0.4$, $\bar{n}$ is the unique stable IEA.

Lemma 6

(i) Given the separability of the payoff function it is clear that provided $1\geq 2\gamma>\gamma_h>\gamma_l$ then, taking as given Q_{i1},Q_{i2h},Q_{i2l} the benefit to country *i* of polluting in each period and each state is at least as great as the additional damage cost it incurs, so $q_{i1}=q_{i2h}=q_{i2l}=1$ is a dominant strategy.

(ii) Similarly, it is clear that provided $2\gamma N>\gamma_h N>\gamma_l N>1$, then when all countries cooperate the benefit of an extra unit of pollution is less than the marginal global damage cost in each period and each state, so that the optimal strategy when all countries cooperate is for each country to set $q_{i1}=q_{i2h}=q_{i2l}=0$.

(iii) Since neither of the above strategies involves different actions in different states in period 2, there is no difference in expected payoffs or gains from cooperation between learning and no learning.

Lemma 7

The proof is very similar to Lemma 5. The IEA wishes to determine q_{s1}, q_{s2l}, q_{s2h} to maximise:

$$V^n(q_{s1},q_{s2l},q_{s2h})=q_{s1}+Q_{s1}-2n\gamma(q_{s1}+Q_{s1}+N-n)$$
$$+0.5[q_{s2l}+Q_{s2l}-n\gamma_l(q_{s2l}+Q_{s2l}+N-n)]+0.5*[q_{s2h}+Q_{s2h}-n\gamma_h(q_{s2h}+Q_{s2h}+N-n)]$$

So it will set $q_{s1}=0\Leftrightarrow 1\leq 2n\gamma$; $q_{s2h}=0\Leftrightarrow 1\leq n\gamma_h$; $q_{s2l}=0\Leftrightarrow 1\leq n\gamma_l$.

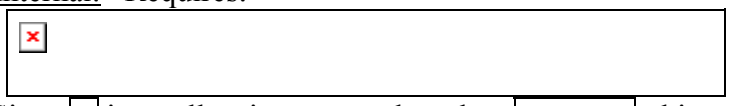
Result 5

(i) $\tilde{n}$: <u>External:</u> Satisfied since: $3\gamma\tilde{n}+2\geq 3\gamma(\tilde{n}+1)\Leftrightarrow\gamma\leq 2/3$
$\bar{n}$: <u>Internal:</u> Satisfied since: $2\gamma\bar{n}+1\geq 2\Leftrightarrow\bar{n}\geq 1/2\gamma$
So there can be no stable IEA greater than $\tilde{n}$ or less than $\bar{n}$.

To consider remaining stability conditions requires assumptions about relationship between $\tilde{n}$, $\hat{n}$ and $\bar{n}$.

(ii) I start be assuming that $\tilde{n}$, $\hat{n}$ and $\bar{n}$ are distinct and differ by at least 2. Since, $n^*\geq\hat{n}\geq\bar{n}$, and since from Result 2, a necessary condition for $n^*-\bar{n}>1$ was $\gamma<1/3$, that will be a necessary condition also for these inequalities to hold.

Internal: Requires:



Since [ ] is smallest integer not less than [ ], this condition will not be satisfied if: [ ]. So [ ] is not internally stable.

[ ]: External: Satisfied since:

[ ]

Internal: Requires:

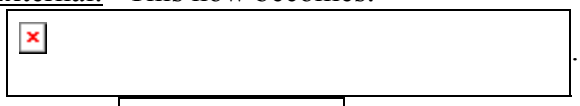
Since [ ] is smallest integer not less than [ ] this condition will not be satisfied if [ ], which holds since, the necessary condition [ ].

[ ]: External: Satisfied since: [ ].

So, if [ ] are distinct and differ by at least 2, then the only stable IEA is [ ].

(iii) I now assume that [ ] are distinct, but that either pair differ only by 1. This affects only the external stability conditions for [ ].

[ ]: External: This now becomes:



. This is satisfied since (4-6γ) > 1, and [ ]. So [ ] remains externally stable.

[ ]: External: This now becomes:

[ ]. This condition will not hold if [ ]. As [ ], it should be possible to find values of δ which satisfy this inequality while also ensuring that [ ].

While the internal stability condition for [ ] remains unchanged, the necessary condition ($\gamma < 1/3$) which was needed to ensure that [ ] differed by at least 2, and which ensured that the internal stability condition was not satisfied, no

longer holds. So there will be parameter values for which will be internally stable.

To summarise, if are distinct but differ by only 1 then there will be parameter values for which is internally stable and parameter values for which is externally instable.

(iv) Finally I consider the case where either (low values of δ) or (high values of δ).

In this case, the only critical threshold values of relevance are and . The internal stability condition for now becomes: , for which a sufficient condition is . As it will be possible to find values of δ which satisfy this inequality and allow . If , then the external stability condition for becomes: . Since , would not be externally stable.

So if there are parameter values for which is stable and parameter values for which is unstable.

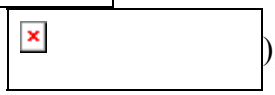
In this case the only critical threshold values of relevance are and , and they are will be distinct. So the internal stability condition for is as in (ii), and the external stability condition for is as in (ii) or (iii). So it remains to check the internal stability condition for which now becomes: where the last inequality is satisfied since .

So if , the unique stable IEA is .

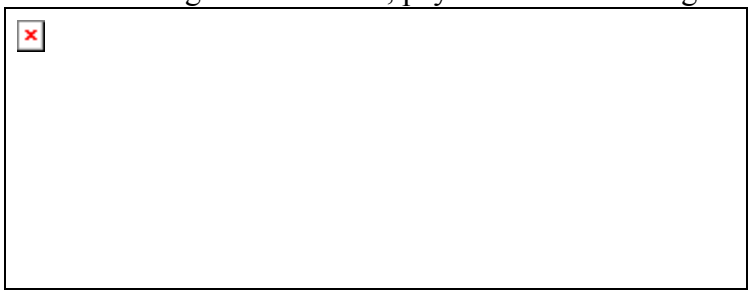
Result 8

A and D are obvious.

B. Ignoring a constant term $-3\gamma N$, payoffs to signatories, non-signatories and global (aggregate payoff over all countries divided by N) under learning and no learning are:

Last two inequalities hold unambiguously (recall ); first may not hold for some parameter values.

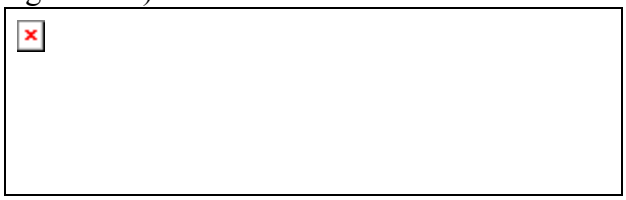
C. Payoffs for learning are same as B; payoffs for no learning are now:



Again, last two inequalities hold unambiguously; first may not hold for some parameter values.

Result 10

For the variable membership model, from Result 4 the expected payoffs to signatories, non-signatories and globally (average of signatories and non-signatories) are:



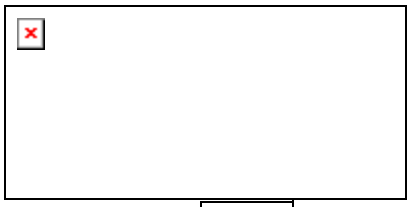
(1)

For the fixed membership model, the corresponding expected payoffs are:

$.4 \leq \gamma \leq .5$

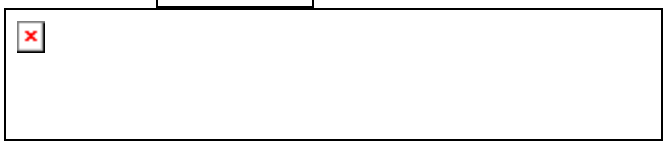
(2)

$\gamma < 0.4$



(3)

Recalling that [ ], it is immediately obvious, by comparing (1) and (3), that when $\gamma < 0.4$, [ ] so the variable membership model is unambiguously better than the fixed membership model. It is also obvious, by comparing (1) and (2), that when $0.4 \leq \gamma \leq 0.5$ [ ]. A little calculation shows that [ ] since



Finally recalling that, for this range of values of γ, n^* = 3, [ ] = 2, [ ], for which a sufficient condition is $N < 15$.

Result 11

Ignoring a constant term $-3\gamma N$ the expected payoffs to signatories, non-signatories and globally are summarised in the following table:

IEA type	Payoff to signatories	Payoff to non-signatories	Global average Payoff
VML	1+