Can international environmental cooperation be bought?

Cristina Fuentes-Albero\textsuperscript{a}, Santiago J. Rubio\textsuperscript{b,*}

\textsuperscript{a}University of Pennsylvania, Department of Economics, 3718 Locust Walk, Philadelphia, PA 19104, USA
\textsuperscript{b}University of Valencia, Department of Economic Analysis, Edificio Departamental Oriental, Avda. de los Naranjos s/n, 46022 Valencia, Spain

\textbf{A R T I C L E   I N F O}

\textbf{Article history:}
Received 10 July 2007
Accepted 6 May 2009
Available online 15 May 2009

\textbf{Keywords:}
Self-enforcing international environmental agreements
Linear environmental damages
Emissions
Transfers

\textbf{A B S T R A C T}

In this paper a two-stage game of international environmental agreement formation with asymmetric countries is analytically solved. The equilibrium of the game makes it possible to determine the size and composition of a stable agreement. Two cases are studied. In the first case, countries differ only in abatement costs, while in the second case, they differ in environmental damages. In both cases, two different institutional settings, one without transfers and another with transfers, are considered. The results establish that the asymmetry assumption has no important effects on the scope of cooperation in comparison with the symmetric case if transfers are not used or abatement costs represent the only difference among countries. However, when the only difference is in environmental damages, the level of cooperation that can be bought through a self-financed transfer scheme increases with the degree of asymmetry.

© 2009 Elsevier B.V. All rights reserved.

\section{1. Introduction}

Climate change, the depletion of the ozone layer, and the loss of biological diversity are some of the most important environmental problems facing contemporary societies. One of the main characteristics of this kind of problems is their international dimension because of the common property of global environmental resources and the transboundary effects of many polluting activities. Hence, managing environmental issues requires transnational cooperation. However, the lack of a supranational authority with enough coercive power over sovereign nations determines that international environmental cooperation must be reached by voluntary agreements. Therefore, International Environmental Agreements (IEAs) should be designed in such a way that they will be not only profitable, but also self-enforcing, i.e. there must be incentives for countries, while acting in their own self-interest, to join or to remain part of an agreement.

One of the earliest definitions of a self-enforcing agreement used in the literature on IEAs was the stability concept proposed by D’Aspremont et al. (1983) in their analysis of cartel formation. According to this definition an IEA will be stable if no signatory country has incentives to leave the agreement and if no non-signatory country has incentives to join the IEA, taking the membership decisions of all other countries as given.\textsuperscript{1} Models based on this concept include the seminal papers by Carraro and Siniscalco (1991, 1993) and Barrett (1994) where it is assumed that countries are identical. Carraro and Siniscalco (1991) have shown that if signatory countries act in Cournot fashion with respect to non-signatories, then a stable IEA consists of three countries when marginal environmental damage is constant (i.e., when countries’ best-reply functions are orthogonal), and of two countries when marginal damage increases with emissions (i.e., when the best-reply functions have a negative slope), in both cases regardless of the number of countries affected.\textsuperscript{2} In the paper published by Barrettin (1994), it is shown that when the best-reply functions have negative slope and signatory countries act in Stackelberg fashion, a stable IEA may achieve a high degree of cooperation, but only when the gains of cooperation are

\textsuperscript{1} There is also a stream of literature that adopted cooperative games as an analytical framework to deal with this issue. Among the different contributions to this literature we would like to stand out the papers written by Chander and Tulkens (1995, 1997). These authors use the gamma-core concept to argue that the grand coalition can be stable using appropriately defined transfers. Other papers that belong to this literature are Helm (2001), Eyckmans and Tulkens (2003), Germain et al. (2003), Forgó et al. (2005) and Flam (2006). More recently Chander (2007) has elaborated a reinterpretation of the gamma-core in terms of an infinitely repeated game.

\textsuperscript{2} These results hinge on a particular specification of the net benefit function from emissions. See Carraro and Siniscalco (1991, pp. 20–21). Other specifications of the net function have been studied by Finus (2001). His results show that although the level of participation is sensitive to the specification of the net benefit function and also to the parameter values, the number of signatories for the three types of net benefit functions studied by the author is not greater than three. Consult chapter 13 in Finus’ book.
small. Nevertheless, Carraro and Siniscalco (1993) have also shown that if the countries belonging to a stable agreement commit to cooperate, the agreement can be enlarged using ex-post transfers. This means that after the coalition has formed, transfers are used to bring new signatories into the agreement. Botteon and Carraro (1997) have extended this analysis for the asymmetric case, finding that ex-post transfers may succeed in expanding a stable coalition even without commitment. They obtain this result using an empirical model with five countries (regions) and constant marginal damages. Later on they adapted the empirical model to investigate the scope of their result with increasing marginal damages. See Botteon and Carraro (2001). More recently, Carraro et al. (2006) have used the integrated assessment simulation model of climate change elaborated by Eyckmans and Tulkens (2003) to show how optimal transfers may induce almost all countries into signing a self-enforcing IEA. An optimal transfer rule, according to the definition given by Eyckmans and Finus (2004), gives every signatory at least his free-rider payoff and allocates the remaining surplus in a proportional way. They also find that optimal transfers exhaust all possibilities for expansion by (ex-post) internal transfers.

In this paper, we exploit the definition of what is a potentially internally stable coalition, given by Eyckmans and Finus (2004), to show that the stability analysis can be developed without the necessity of specifying previously any kind of transfer rule. To know whether an agreement can be (internally) stabilized through transfers, it is enough to check whether the countries that gain with the agreement have enough resources to buy the cooperation of the countries that lose through according to the agreement. This will occur when the aggregate payoff of signatories is greater than the aggregate payoff they would receive if they quit the agreement and act as free-riders. Moreover, this condition can be checked without the necessity of specifying a transfer rule, just checking whether the surplus of the agreement over the free-rider payoffs of its members is positive. Therefore, an agreement can be stable with transfers if the previous condition is satisfied, but there are not enough resources to buy the cooperation of any non-signatory which will make the agreement potentially externally stable as well. Obviously, if these conditions are satisfied the transfer schemes that make possible to stabilize the agreement belong to the class of optimal or almost ideal transfer schemes defined by Eyckmans and Finus (2004). It is clear that in order to implement the agreement it will be necessary to specify an optimal transfer rule, but as we show in this paper, this is not necessary to know which can be the scope of cooperation when heterogeneous countries use transfers. Besides, to analyze this issue, we also investigate the relationship between the degree of asymmetry, participation and the gains to an IEA.

For a model with two types of countries and an arbitrary number of countries of each type, our findings establish that the asymmetry among countries has no relevant effects on the scope of environmental cooperation in comparison with the symmetric case if transfers are not allowed. With transfers the effects depend on the kind of asymmetry considered. If the countries differ in abatement costs alone, the result is that only limited cooperation can be bought. However, when countries differ in terms of environmental damages, the level of cooperation increases with the heterogeneity in marginal environmental damages, which implies that an agreement with a high degree of participation may be self-enforcing if the degree of asymmetry among the countries is sufficiently large. What we find in our analysis is that although an increase in asymmetry rises the incentives to deviate for the signatories with lower damages also rises the incentives to stay for the signatories with greater damages yielding, once the degree of asymmetry is above a certain threshold, a positive surplus of the agreement over the incentives to deviate. We also find that this threshold increases with the number of signatories with lower damages which explains the positive relationship between the degree of asymmetry and the level of participation. The same kind of result is obtained with three types of countries. Finally, we would like to highlight that the level of cooperation in our model only depends on the degree of asymmetry. Being more precise, we find that the level of cooperation only depends on the relative value of the marginal damages so that if a change in the absolute values is considered that does not alter their relative value, the change has no effect on participation. Thus, we find that it is not possible to establish any systematic relationship between the gains to full cooperation and the participation in the agreement since an increase in the relative value of the marginal damages can be accompanied with different variations of the gains to full cooperation. In fact, an increase in the degree of asymmetry can increase, decrease or keep constant the gains to cooperation depending on which are the changes in the absolute values of the marginal damages associated with the increase in their relative value.

Although a lot of papers have been published on the stability of IEAs, only a few have addressed this issue within a theoretical framework. These papers are now briefly reviewed. A first paper to quote is Hoel (1992). In this paper a numerical simulation is used to show that without transfers the number of signatories in a self-enforcing IEA is low. The result is obtained for a model where the countries only differ in marginal environmental damages that, on the other hand, are assumed constant. Using a generalization of this model, Petrakis and Xepapadeas (1996) extend the results of Carraro and Siniscalco (1993) to the case in which countries are completely asymmetric. However, they maintain the stable coalition commitment assumption in their analysis. Hoel and Schneider (1997), also in a model with linear environmental damages where all the countries are identical except for some non-environmental costs that depend on the number of signatories and represent the effects of social norms, point out that the prospect of receiving a transfer tends to reduce the incentive a country might have to join the agreement. On comparing their results with those obtained in this paper, it seems clear that this depends critically on their specification of non-environmental costs. Another contribution worth mentioning is Barrett (1997). In this paper, a seven-nation model is solved numerically where four nations are low benefit, low cost (type 1) and the remaining three are high benefit, high cost (type 2). With ex-ante transfers determined by Shapley values and signatories acting as a leader, the author finds that a self-enforcing IEA always exists although it never consists of more than three signatories.

More recently, Barrett (2001) has investigated the possibilities of buying cooperation in a linear model with ex-ante transfers and an arbitrary number of countries of two types which can only choose between two actions: to pollute or to abate. The countries have identical abatement costs, but differ in the benefit stemming from total abatement. In this framework, Barrett concludes that international cooperation can be bought and that transfers become the vehicle for increasing participation. In this paper, we extend this result for a non-linear model with two types of countries that differ not only in environmental damages but also in abatement costs. Although we assume that environmental damages are linear, the non-linearity of abatement costs is enough to change the nature of the game and the kind of equilibrium that characterizes the solution in comparison with Barrett’s (2001) model. Linearity of the net benefit function induces a bang-bang solution: if any signatory were to quit from the equilibrium agreement, all the remaining signatories would play pollute, like the non-signatories. However, in the non-linear model used in this paper this is not the case. Although the signatories increase their emissions when one signatory quits the agreement, they do not choose the same level of emissions as non-signatories. As the two models are different, both yield different predictions for the symmetric case and...
also for the asymmetric case without transfers. For instance, in Barrett’s (2001) paper, cooperation without transfers depends on the degree of asymmetry, see his examples in Section 4. However, in our model the maximum level of cooperation without transfers consists of three countries of the same type regardless of the degree of asymmetry. Moreover, we highlight in this paper that the stability of an agreement can be checked without the previous specification of a transfer scheme as occurs in Barrett’s (2001) paper.

Finally, we would like to mention the interesting papers written by Weikard (2009) and McGinty (2007), where special cases in the class of optimal transfers are used to analyze the stability of IEAs among asymmetric countries.3 Weikard (2009) suggests also for a model with quadratic abatement costs and completely asymmetric linear benefits, that large agreements may well be stable and shows that the grand coalition can be stable if the degree of asymmetry is large enough. Note that in order to ascertain whether or not the grand coalition is stable, it is enough to check whether it is internally stable. In this paper, we complete Weikard’s (2009) analysis by showing that between the grand coalition and the fully non-cooperative equilibrium there can be partial coalitions that can be also self-enforcing, depending on the differences in environmental damages between the countries. Moreover, we also show that the degree of asymmetry in environmental damages, not in abatement costs, is determinant when it comes to explaining the level of cooperation. We also investigate the potential scope of cooperation under asymmetry without transfers. In McGinty’s (2007) paper a model with twenty countries is numerically solved. His numerical simulations show that symmetric models may vastly underestimate the degree of abatement achievable by a stable agreement with transfers, but on the other hand they confirm the result obtained by Barrett (1994) for the symmetric model which establishes that there is a trade-off between the gains to an IEA and the number of signatories. However, we do not find any systematic relationship between the gains to full cooperation and the number of signatories of a stable IEA. As we have just pointed out, our findings establish that an increase in the gains to full cooperation can increase (decrease) participation provided that it comes with an increase (decrease) in the degree of heterogeneity. Nevertheless, on comparing our results with those obtained by McGinty (2007), one should take into account that besides assuming a quadratic benefit function, McGinty (2007) also assumes that the agreement is the Stackelberg leader in the abatement game. Our conjecture is that it is this last assumption that could explain the differences in results, as it occurs for the symmetric model.4 A first numerical example we have developed for the model with quadratic environmental damages gives support to this conjecture. The numerical example shows that an increase in the gains to full cooperation, accompanied with an increase in the degree of asymmetry, increases the level of cooperation.5

Next, we would like to devote a few lines to the implications of our results for international environmental policy. The first thing that seems pretty obvious is that our results are more optimistic than those obtained for the symmetric case, as we find that asymmetry can induce more cooperation if accompanied by transfers provided that the degree of asymmetry is sufficiently large. In other words, cooperation can be bought. However, only one class of transfer schemes can eliminate the incentives of some countries to act as free-riders. This introduces a new element into the debate about the properties that a transfer scheme should have. It should be an optimal transfer scheme. A first contribution in this line is the paper by Carraro et al. (2006) mentioned above. In this paper, they compare simple transfer schemes such as the Shapley Value, the Nash Bargaining solution or the Chander and Tulkens’ transfer scheme with optimal transfer schemes. The result is that the participation in a self-enforcing IEA with simple transfer schemes is lower than that which could be obtained using an optimal transfer scheme. This result suggests the possibility of a conflict between equity and stability, an issue that would deserve more attention in the future. In particular, the issue is of great significance when transfers are implemented through a system of tradable pollution permits, as the allocation of permits that maximizes cooperation may not satisfy other desirable properties.6

To conclude, we would like to clarify one issue. Following the approach adopted by Carraro and Siniscalco (1993), Barrett (1994), Chander and Tulkens (1997) and by many scholars afterwards, we focus on the case where only one IEA is formed and the questions remaining are the size and composition of the self-enforcing agreement. We are aware that this approach eliminates the possibility of different countries forming different agreements, i.e., the possibility of equilibrium with more than one agreement. Nevertheless, we believe that this approach may be reasonable for global environmental problems, such as climate change, for which two global environmental agreements have been launched by the United Nations (the UNFCCC and the Kyoto Protocol). Formally, it could be interpreted as an institutional constraint.

The structure of the paper is as follows. In Section 2, the model is set up and the definition of a stable IEA is presented. In Section 3, the Nash equilibrium of the emission game is solved for a given number of signatories. In Sections 4 and 5, the stability of IEAs with and without transfers is analyzed considering, in Section 4, that the only difference among countries appears in abatement costs and, in Section 5, in environmental damages. Section 6 summarizes our conclusions.

2. Self-enforcing international environmental agreements

2.1. The model

Consider N countries that pollute a common environment and negotiate the emission control of a specific pollutant. We define $x_i \geq 0$ as the level of emissions generated by a country $i$, and $X = \sum_{i=1}^{N} x_i$ as total emissions generated by all $N$ countries. Each country derives a gross benefit from its emissions so that the reduction of the emissions for controlling pollution implies some abatement costs, denoted by $(c_i/2)(d_i - x_i)^2$. The value of parameter $d_i > 0$, that stands for the business-as-usual emission level, depends on national technology, economic structure and the level of development. Parameter $c_i > 0$ represents the marginal cost of national abatement and depends on national technology and measures the intensity of the use of the pollutant for the production of goods and services. Each country also suffers environmental damages which depend on total emissions according to the following expression $m_i X$, where $m_i > 0$ is the marginal environmental damage which depends on the country’s environmental (natural)

---

3 Besides the definition of the optimal transfer scheme and the potentially internally stable coalitions, Eickmans and Finus (2004) show that any transfer scheme from the class of optimal transfers can make the coalition with the highest global welfare stable. However, they do not analyze the potential scope of the cooperation under this class of transfer schemes.

4 Remember that the differences in the results on participation obtained by Carraro and Siniscalco (1991) and Barrett (1994) are only justified by the fact that Barrett assumes that signatories enjoy a first-mover advantage in the abatement game, whereas Carraro and Siniscalco (1991) compute a Nash equilibrium.

5 The solution to the model with quadratic environmental damages and the numerical example are available as supplementary material at the EJOR website.

6 In McGinty’s (2007) paper, pollution permits are allocated in order to implement an optimal transfer but the focus of the paper is only on the degree of cooperation that can be reached through this particular transfer scheme not on the comparison with other transfers schemes.
endowment. Then, for each country, the cost function is: \(^7\)
\[ C_i = \frac{1}{4}C_i(\delta_i - x_i)^2 + mX \]
where \( i = 1, \ldots, n \) and \( x_i \leq \delta_i \).

2.2. A self-enforcing IEA

We model the formation of an IEA as a two-stage game. We will describe each game briefly, in reverse order, as we compute the subgame-perfect equilibria of this two-stage game by backward induction.

2.2.1. The emission game

Thus, the first part of agreement Nash equilibrium signatories as given. Thus, emissions are provided by the ing the emissions of all other countries as given in order to mini-

tation transfers, \(^8\)coalition structure. If \( C \) and \( C_i \) endowment. Then, for each country, the cost function is:

\[ C_i = \frac{1}{4}C_i(\delta_i - x_i)^2 + mX \]

where \( i = 1, \ldots, n \) and \( x_i \leq \delta_i \).

The first inequality, which is also known as the internal stability condition, simply means that any signatory country is at least as well-off staying in the IEA as withdrawing from it, assuming that all other countries do not change their membership decisions. The second inequality, which is also known as the external stability condition, similarly requires any non-signatory to be at least as well-off remaining a non-signatory as joining the IEA, assuming once again, that all other countries do not change their membership decisions.

Next, we rewrite this definition to take into account the possibility of transfers among signatories.

Definition 2. An agreement \( K \) is potentially self-enforcing if there is at least one self-financed transfer vector such that \( C_{i}(K) + T_i(K) < C_{i}(K \setminus \{ i \}) \forall i \in K \) and there is no self-financed transfer vector such that \( C_{i}(K) \geq C_{i}(K \cup \{ i \}) + T_i(K \cup \{ i \}) \forall i \not\in K \).

In other words, an agreement consisting of \( n \) countries is stable if the countries that gain from the agreement can buy the cooperation of the countries that are interested in signing the agreement only if they are adequately compensated, and when it is not possible to do the same for an agreement consisting of \( n + 1 \) countries. We assume a strict inequality in the first part of the definition, because it seems reasonable that, when transfers are included in international negotiations, the countries that benefit from the agreement will not be interested in buying cooperation if this means that all their gains must be transferred to the other countries and that the countries that lose from the agreement will not be interested in selling cooperation if they do not obtain, at least, a marginal gain. In the second part of the definition, we use a weak inequality because if an agreement consisting of \( n + 1 \) countries is not stable, the agreement does not form and transfers do not apply. As a result, it does not seem necessary to be as strict as in the other case.\(^8\)Thus, the first part of Definition 2 would operate as the internal stability condition with transfers and the second part as the external stability condition.

Given this definition it is straightforward to show the following:

Lemma 1. An agreement \( K \) is potentially self-enforcing through transfers if and only if \( \sum_{i \in K} C_i(K) - C_i(K \setminus \{ i \}) < 0 \) and \( \sum_{i \in K \cup \{ j \}} C_i(K \cup \{ j \}) - C_i(K) \geq 0 \forall j \not\in K \).

Proof. Suppose that for an agreement consisting of \( n \) countries \( \sum_{i \in K} C_i(K) - C_i(K \setminus \{ i \}) < 0 \) but that the internal stability conditions of Definition 1 are not satisfied for all signatories. In this case there must be a set of signatories \( W \subset K \) of size \( n \) for which \( C_i(K) - C_i(K \setminus \{ i \}) < 0 \) and a set of signatories \( L \subset K \setminus W \) of size \( n - \bar{n} \) for which \( C_i(K) - C_i(K \setminus \{ i \}) > 0 \) such that \( \sum_{i \in W} C_i(K) - C_i(K \setminus \{ i \}) > 0 \) holds. Then there must be at least one vector of self-financed transfers through which the \( \bar{n} \) signatories can buy the cooperation of the rest of signatories \( n - \bar{n} \). For these transfers \( C_i(K) + T_i(K) < C_i(K \setminus \{ i \}) \forall i \in K \) and the agreement with transfers is internally stable. Now we check that if an agreement consisting of \( n \) signatories is internally stable then condition \( \sum_{i \in W} C_i(K) - C_i(K \setminus \{ i \}) > 0 \) must be satisfied. The internal stability condition of Definition 2 can be written as \( T_i(K) < C_i(K \setminus \{ i \}) - C_i(K) \) for all signatories so that adding terms we get \( \sum_{i \in K} T_i(K) = 0 < \sum_{i \in K} C_i(K \setminus \{ i \}) - C_i(K) \) which implies that \( \sum_{i \in K \cup \{ j \}} C_i(K \cup \{ j \}) - C_i(K) < 0 \). Next, suppose that for any agreement consisting of \( n + 1 \) countries \( \sum_{i \in K \cup \{ j \}} C_i(K \cup \{ j \}) - C_i(K) \geq 0 \). If this is the case although there could

\(^7\) This is the cost function used by Botroon and Carraro (1997).

\(^8\) As we pointed out in the introduction we do not need to be more precise about how the countries decide the transfers at this stage of the game because we want to highlight in this paper, that the conditions for stability can be checked without the necessity of specifying in the first-stage of the game any kind of transfer rule.
be \( n \) countries with \( n \in \{0, n + 1\} \) for which \( C_i(K \cup \{j\}) - C_i(K) < 0 \), 
\[
\left\{ \sum_{i=1}^{n} W_i C_i(K \cup \{j\}) - C_i(K) \right\} \geq \sum_{i=1}^{n} C_i(K \cup \{j\}) - C_i(K) > 0 \text{ does not hold where now } W = 0 \text{ or } W \subseteq K \cup \{j\} \text{ and } i \in K \cup \{j\} . \] 
Then there is no self-financed transfer vector such that for all signatories \( C_i(K \cup \{j\}) + T_i^j(K \cup \{j\}) \leq C_i(K) \) and any agreement consisting of \( n + 1 \) signatories is internally unstable. Thus, by definition, the agreement consisting of \( n \) signatories is externally stable. Finally, we check if an agreement consisting of \( n \) signatories is externally stable. If this is the case, condition \( \sum_{i=1}^{n} C_i(K \cup \{j\}) - C_i(K) \geq 0 \) must be satisfied \( \forall j \neq K \). The external stability condition establishes that for all possible self-financed transfer vectors \( C_i(K) - C_i(K \cup \{j\}) \) holds \( \forall i \neq K \). This can be written as \( \frac{C_i(K) - C_i(K \cup \{j\})}{C_i(K)} \leq T_i^j(K \cup \{j\}) \) but according to Definition 2, country \( j \) stands for a non-signatory of the agreement \( K \) that we have represented by \( j \) in this Lemma, so that for adding with respect to the signatories of the agreement \( K \cup \{i\} \) we have adapted the notation and written \( \sum_{i=1}^{n} C_i(K \cup \{j\}) - C_i(K) \leq \sum_{i=1}^{n} T_i^j(K \cup \{j\}) = 0 \). Then \( \sum_{i=1}^{n} \left( C_i(K \cup \{i\}) - C_i(K) \right) \geq 0 \forall i \neq K \). Thus, we can conclude that if conditions \( \sum_{i=1}^{n} \left( C_i(K) - C_i(K \cup \{j\}) \right) < 0 \) and \( \sum_{i=1}^{n} \left( C_i(K \cup \{i\}) - C_i(K) \right) \geq 0 \forall i \neq K \) are satisfied, the internal and external stability conditions hold and the agreement is self-enforcing. \[\square\]

Each national difference, \( C_i(K) - C_i(K \setminus \{i\}) \), establishes the maximum payment that one country is willing to pay for cooperation if the difference is negative, or the minimum payment that has to be received to sell cooperation if the difference is positive. If the agreement is self-enforcing, then the sum of the maximum payments that some countries are willing to pay must be larger than the sum of the national payments that the other countries demand for selling cooperation. Hence in this case, there must be at least one transfer scheme that makes the agreement stable. In fact, there will be a class of transfer schemes that makes the agreement self-enforcing. This class was called the Almost Ideal or Optimal Transfer Schemes by Eyckmans and Finus (2004) and Carraro et al. (2006).

### 3. The Nash equilibrium of the emission game

In this section, we develop a model with two types of countries: type 1 and type 2 so that \( N = N_1 + N_2 \). Suppose that, as the outcome of the first-stage game, an agreement \( K = \{n_1, n_2\} \) of size \( n = n_1 + n_2 \) is formed. Then a representative non-signatory solves

\[
\min_{\chi_i} C_i = \frac{1}{2} \delta_i - \chi_i^2 + m_i X \quad i = 1, 2. 
\]

where \( X = \sum_{i=1}^{n} \left( n_i X_i + (N_i - n_i) X_i \right) \). The F.O.C. yields \( \chi_i = \delta_i - (m_i/c_i) \), \( i = 1, 2 \).

On the other hand, signatories are assumed to coordinate in order to minimize their aggregate costs taking the emissions of non-signatories as given.

\[
\min C(n_1, n_2) = \sum_{i=1}^{n} n_i C_i = \sum_{i=1}^{n} \left( \frac{1}{2} \delta_i - \chi_i^2 + m_i X \right). 
\]

The F.O.C. yield

\[
\chi_i = \delta_i - \frac{\sum_{i=1}^{n} n_i m_i}{c_i} \quad i = 1, 2. 
\]

Notice that signatory emissions decrease with the size of the agreement.

Aggregate emissions are

\[
X(n_1, n_2) = \sum_{i=1}^{n} \left( \delta_i - \frac{\sum_{i=1}^{n} n_i m_i}{c_i} \right) + \sum_{i=1}^{n} (N_i - n_i) \left( \delta_i - \frac{m_i}{c_i} \right). 
\]

which after developing the additions yields the following expression

\[
X(n_1, n_2) = \sum_{i=1}^{n} N_i \left( \delta_i - \frac{m_i}{c_i} \right) - \sum_{i=1}^{n} \left( m_i (N_i - n_i) - m_i \right), 
\]

\[j = 1, 2, \quad i \neq j. \]

This expression decreases with respect to the number of signatories of both types.\[10\]

\[
\frac{\partial X(n_1, n_2)}{\partial m_i} = -N_i \left( m_i (2N_i - 1) + m_i n_i \right) + \frac{m_i}{c_i} X(n_1, n_2) = 0, 
\]

\[i, j = 1, 2, \quad i \neq j, \quad n_i > 1. \]

Finally, we obtain the cost functions of signatories and non-signatories.

\[
C_i(n_1, n_2) = \frac{1}{2c_i} \left( \sum_{i=1}^{n} m_i n_i \right)^2 + m_i X(n_1, n_2), \quad i = 1, 2 
\]

\[
C_i(n_1, n_2) = \frac{m_i^2}{2c_i} + m_i X(n_1, n_2), \quad i = 1, 2. 
\]

These expressions show that both non-signatories and signatory cost functions depend on the size and composition of the agreement. Moreover, it is easy to check from (4) that non-signatory costs decrease with the size of the agreement as aggregate emissions decrease with the number of signatories. This means that in this game there are positive spillovers as an increase in the number of signatories reduces non-signatory costs. Consequently, we can conclude that if the internal stability condition is satisfied for a country, the profitability condition is also satisfied.\[11\] Moreover, it is easy to check that for a given number of signatories, non-signatory costs are lower than for signatories of the same type

\[
C_i(n_1, n_2) - C_i(n_1, n_2) = \frac{1}{2c_i} \left( n_i^2 - 1 \right) m_i^2 + 2m_i n_i n_j m_j + n_j^2 m_j^2 > 0. 
\]

for \( n_i > 1 \). \( i, j = 1, 2, \quad i \neq j. \)

Notice that this difference increases with respect to the number of signatories of both types.

Finally, we obtain that the gains to cooperation are given by:

\[
\sum_{i=1}^{n} (N_i - n_i) \left( C_i(0, 0) - C_i(n_1, n_2) \right) + \sum_{i=1}^{n} n_i \left( C_i(0, 0) - C_i(n_1, n_2) \right) 
\]

\[
= \sum_{i=1}^{n} \left( m_i X(n_1, n_2) \right) + \left( 2N_i - 1 \right) m_i^2 + 2 \left( (N_i - n_i) m_i n_j + N_i (n_i - 1) m_i m_j + n_j (n_j - 1) m_j^2 \right) 
\]

\[j = 1, 2, \quad i \neq j. \]

In this expression \( C_i(0, 0) \) stands for the cost of a country of type \( i \) at the fully non-cooperative equilibrium, \( C_i(n_1, n_2) \) for the cost of a non-signatory of type \( i \) and \( C_i(n_1, n_2) \)

\[10\] It is also easy to show that this is the case when the first signatory of type \( i \) enters the agreement.

\[11\] Remember that the internal stability condition requires that for \( K = \{n_1, n_2\} \) \( C_i(K) \leq C_i(K \setminus \{i\}) \) or \( C_i(K) + T_i^j(K \setminus \{i\}) \leq C_i(K \setminus \{i\}) \) then as \( C_i(K \setminus \{i\}) \leq C_i(K) \) where \( C_i(K \setminus \{i\}) \) stands for the costs corresponding to the singleton coalition structure or the fully non-cooperative equilibrium, we obtain that \( C_i(K) \leq C_i(K \setminus \{i\}) \) or \( C_i(K) + T_i^j(K \setminus \{i\}) \) so that the profitability condition holds. Remember that the profitability condition requires that signatories costs must be lower or at least equal to the fully non-cooperative equilibrium costs.
for the cost of a signatory of type $i$ in both cases when the agreement consists of $n_1$ countries of type 1 and $n_2$ countries of type 2. It is immediately observed that for $n_1 > 1$ (5) increases with respect to marginal environmental damages. Moreover, calculating the first derivatives of this expression with respect to $n_1$ and $n_2$ it can be concluded that the gains from cooperation also increase with the size of the agreement. However, an increase in marginal abatement costs has a negative impact on the gains from cooperation.

4. The Nash equilibrium of the membership game with heterogeneity in abatement costs

In this section we assume that $m_1 = m_2 = m$, $\delta_1 = \delta_2 = \delta$, and $c_1 > c_2$. Bearing these assumptions in mind, the costs functions for signatories and non-signatories according to their type can be calculated by means of the expressions in Section 3. Using these cost functions, we can conclude that the internal stability conditions will be satisfied when the following expressions are negative or zero.

$$C_1'(n_1, n_2) - C_1'(n_1 - 1, n_2) = m^2 \left( \frac{n^2 - 2n - 2n_1 + 3}{2c_1} - \frac{n_2}{c_2} \right),$$

$$C_2'(n_1, n_2) - C_2'(n_1, n_2 - 1) = m^2 \left( \frac{n^2 - 2n - 2n_2 + 3}{2c_2} - \frac{n_1}{c_1} \right).$$

Multiplying and dividing by $c_1$, these expressions can be written as follows.

$$C_1'(n_1, n_2) - C_1'(n_1 - 1, n_2) = \frac{m^2}{2c_1} \left( n^2 - 2n - 2n_1 + 3 - 2n_2 c \right), \quad (6)$$

$$C_2'(n_1, n_2) - C_2'(n_1, n_2 - 1) = \frac{m^2}{2c_1} \left( c(n^2 - 2n - 2n_2 + 3) - 2n_1 \right), \quad (7)$$

where $c$ is the ratio $c_1/c_2$. The same occurs for the external stability conditions. Thus, as both $n^2 - 2n - 2n_1 + 3$ and $n^2 - 2n - 2n_2 + 3$ are positive when the agreement consists of countries of both types, we have that for a given agreement consisting of countries of both types, the sign of these expressions only depends on the ratio $c$. This means that changes in the absolute values of parameters $c_1$ and $c_2$ that do not change their relative value will not have any effect on the stability or unstability of an agreement so that it can be established that

**Lemma 2.** The stability for a given agreement without transfers consisting of countries of both types depends only on the value of ratio $c$.

Using these conditions, stability has been analyzed for two different settings: with and without transfers.

4.1. Stable IEA without transfers

As shown in Section 3, if internal stability conditions hold, profitability conditions also hold. Consequently, in this section we develop the stability analysis using only the internal and external stability conditions. The result of this analysis is summarized in the following proposition:

**Proposition 1.** The maximum level of cooperation that can be reached through a self-enforcing IEA without transfers consists of three countries of the same type regardless of the heterogeneity in abatement costs.

The logic behind this result is that the internal stability condition for type 1 countries is satisfied when the differences in marginal abatement costs are large enough, whereas the contrary is required in order to satisfy the internal stability condition for type 2 countries. The conclusion is that there are no values for $c$ that satisfy the internal stability condition for an agreement with both types of countries. This result coincides with that obtained by Carraro and Siniscalco (1991) for the symmetric model with orthogonal reaction functions.

4.2. Stable IEA with transfers

Next, we explore the scope of cooperation when countries can use transfers. Therefore, the aim of this section is to find the values for $n_1$ and $n_2$ that satisfy the conditions of Lemma 1.

Using Lemma 1 the following result is obtained:

**Proposition 2.** Only an IEA consisting of two signatories of different types is potentially self-enforcing with transfers regardless of the heterogeneity in abatement costs.

If the countries in the agreement are of the same type, the free rider incentive is the same for all signatories and cooperation cannot be expanded using transfers. In this case, Proposition 1 applies. If the countries in the agreement are different, type 1 countries can buy the cooperation of type 2 countries according to Proposition 1, provided that the costs of the agreement are lower than the aggregate costs the signatories would receive if they quit the agreement and act as free-riders. Using (6) and (7) this difference yields the following expression:

$$n_1 \left( C_1'(n_1, n_2) - C_1'(n_1 - 1, n_2) \right) + n_2 \left( C_2'(n_1, n_2) - C_2'(n_1, n_2 - 1) \right)$$

$$= \frac{m^2}{2c_1} \left( n^2 - 4n + 3 \right) \left( \frac{n_1}{c_1} + \frac{n_2}{c_2} \right) \quad (8)$$

Thus, we find that when countries only differ in abatement costs, the possibility of buying cooperation depends only on the size of the agreement: $n$. In other words, although $c$ determines, according to Lemma 2, the sign of the stability conditions without transfers, it does not affect the sign of condition (8). Moreover, if the sign is negative, the greater $c_1$ or the greater $c_2$, the less the resources available to buy the cooperation. Obviously, this is related to the fact that the gains to an agreement decrease with marginal abatement costs as established in Section 3.

5. The Nash equilibrium of the membership game with heterogeneity in environmental damages

In this section we assume that $\delta_1 = \delta_2 = \delta$, $c_1 = c_2 = c$, and $m_1 > m_2$. Taking these assumptions into account, the cost functions for signatories and non-signatories according to their type can be calculated by means of the expressions in Section 3. Using these cost functions, we can conclude that the internal stability conditions will hold when the following expressions are negative or zero.

$$C_1'(n_1, n_2) - C_1'(n_1 - 1, n_2) = \frac{m^2}{2c_1} \left( n^2 - 4n_1 - 2n_2 + 3 \right) m_1^2$$

$$+ 2(n_1 - 1)n_2 m_1 m_2 + n_1^2 m_2^2,$$

$$C_2'(n_1, n_2) - C_2'(n_1, n_2 - 1) = \frac{m^2}{2c_1} \left( n^2 - 4n_1 - 2n_2 + 3 \right) m_2^2$$

$$+ 2(n_2 - 1)n_1 m_1 m_2 + n_2^2 m_1^2.$$

Dividing and multiplying by $m_2^2$, these expressions can be written as follows.

$$C_1'(n_1, n_2) - C_1'(n_1 - 1, n_2) = \frac{m^2}{2c_1} \left( n^2 - 4n_1 - 2n_2 + 3 \right) m_1^2$$

$$+ 2(n_1 - 1)n_2 m_1 m_2 + n_1^2 m_2^2,$$  \quad (9)

$$C_2'(n_1, n_2) - C_2'(n_1, n_2 - 1) = \frac{m^2}{2c_1} \left( n^2 - 4n_1 - 2n_2 + 3 \right) m_2^2$$

$$+ 2(n_2 - 1)n_1 m_1 m_2 + n_2^2 m_1^2,$$  \quad (10)
where $m$ is the ratio $m_1/m_2$. The same occurs for the external stability conditions. Thus the same kind of result that that obtained in Section 4, see Lemma 2, applies also when the countries differ in environmental damages: given an agreement, the sign of the stability conditions only depend on the relative value of marginal environmental damages. Next, using these conditions, stability has been analyzed in two different settings as in the previous section.

5.1. Stable IEA without transfers

Applying the stability conditions of Definition 1 we obtain the following result:

**Proposition 3.** The maximum level of cooperation that can be reached through a self-enforcing IEA without transfers consists of three countries of the same type regardless of the heterogeneity in environmental damages. An agreement consisting of one type 1 country and one type 2 country can also be self-enforcing if the differences in marginal environmental damages are not very large.

If the agreement consists of countries of the same type, there are once again no differences with respect to the symmetric case. However, an agreement consisting of one country of each type can be stable if the difference in environmental damages in relative terms is no greater than 40%. The reason for this result is that if the differences in environmental damages are not very large, cooperation between two countries of different types can be profitable for both, as the reduction in emissions that both have to bear will not be very large either. In this case, the profitability of the agreement implies its internal stability, as the alternative to being part of a bilateral agreement is bearing the costs corresponding to the fully non-cooperative equilibrium which in this case will be greater for both types of countries. However, with a second type 2 country in the agreement, the alternative for the type 2 countries to being part of the agreement is to be a free rider of an agreement with two signatories which with positive externalities yields lower costs making the agreement internally unstable. Thus, an agreement consisting of one country of each type is externally stable and given that it is also internally stable, we can conclude that it is self-enforcing. Hoel (1992) also obtains that only two countries cooperate in equilibrium in a model where all the countries have different marginal environmental damages but identical benefit functions. However, in his paper signatories agree upon the emission level which is the median value of their most preferred emission levels.

5.2. Stable IEA with transfers

We have just concluded that our candidates for being stable agreements without transfers are coalitions of three or two countries. In this section, we analyze if larger coalitions can be stable when transfers between countries are taken into account.

The results of our analysis can be summarized as follows:

**Proposition 4.** The level of cooperation that can be bought with transfers increases with the heterogeneity in environmental damages provided that cooperation is only bought by one or two countries of type 1.

As in the case studied in the previous section, the possibility of type 1 countries buying the cooperation of type 2 countries depends on the sign of the following difference:

$$n_1 \left( C'_1(n_1, n_2) - C'_1(n_1 - 1, n_2) \right) + n_2 \left( C'_2(n_1, n_2) - C'_2(n_1, n_2 - 1) \right)$$

$$- \frac{m_2^2}{2c} n_1 (n_1(n_1 - 2) - 2n + 3) m^2 + 2n_1(n - n_1)(n - 2)m$$

$$+ (n - n_1)(n^2 - 4n + 3 - n_1(n - 2)),$$

where $C'_1$ and $C'_2$ stand for the first derivative of the cooperation levels and $m$ is the ratio $m_1/m_2$.

Comparing this expression with that obtained in the previous case, see expression (8) in Section 4, one realizes that now the sign of this difference depends on the size of the agreement as in the previous case but also on the composition of the agreement and on the degree of asymmetry between both types of countries. Observe that the size and composition of the agreement determine the coefficients of a polynomial of second degree for $m$. Moreover, it is easy to check that the coefficient of $m^2$ is negative and that the coefficient of $m$ is positive when in the agreement there is only one or two countries of type 1. In that case, there will exist a lower bound for $m$ above which (11) is negative and the cooperation can be bought through transfers. Thus, for any agreement with a maximum of two signatories of type 1, it is possible to find an interval of values for $m$ for which the agreement is stable. Finally, this expression also defines a positive relationship between the level of participation in the agreement and the lower bound for $m$ so that the higher the heterogeneity in environmental damages, the higher the level of cooperation that can be bought using transfers.

**Table 1** shows the relationship between the size and composition of the agreements and the values of $m$ that allow to stabilize the agreement through transfers. We used expression (11) to obtain these values. The first column represents the size of the agreements and the second and fourth their composition where the first figure stands for the number of type 1 signatories and the second for the number of type 2 signatories. In the third column, the interval of values for $m$ that supports the corresponding level of cooperation when the agreement consists of only one type 1 signatory and $n - 1$ type 2 signatories is represented, and the fifth displays the interval of values for $m$ that supports the corresponding level of cooperation when the agreement consists of two type 1 signatories and $n - 2$ type 2 signatories. For each interval, the first figure is a lower bound that must be strictly satisfied in order for the type 1 country to have enough resources to buy the cooperation of type 2 countries. For instance, one type 1 country will only have enough resources to buy the cooperation of two type 2 countries if $m$ is higher than 1, and will only be able to buy the cooperation of three type 2 countries if $m$ is higher than 4.24, and so on.

The same can be said when there are two type 1 countries in the agreement. This means that, given a size and a composition of the agreement, this will be internally and externally stable only if $m$ belongs to the corresponding interval of Table 1. For instance, if $m$ is equal to 5, one type 1 country will have enough resources to buy the cooperation of type 2 and type 3 countries and also the cooperation of three. Then, the internal stability condition with transfers of Definition 2 will be satisfied for $n = 3$ and also for $n = 4$, but this will not be the case for the external stability condition. The agreement with two type 1 countries will only be stable if $m > 4.24$.

### Table 1

<table>
<thead>
<tr>
<th>n</th>
<th>Agreements</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>(1,2)</td>
<td>(1.424)</td>
</tr>
<tr>
<td>4</td>
<td>(1,3)</td>
<td>(4.246, 74)</td>
</tr>
<tr>
<td>5</td>
<td>(1,4)</td>
<td>(6.749, 20)</td>
</tr>
<tr>
<td>6</td>
<td>(1,5)</td>
<td>(9.201, 63)</td>
</tr>
<tr>
<td>7</td>
<td>(1,6)</td>
<td>(11.634, 06)</td>
</tr>
<tr>
<td>8</td>
<td>(1,7)</td>
<td>(14.06, 1649)</td>
</tr>
<tr>
<td>9</td>
<td>(1,8)</td>
<td>(16.49, 1981)</td>
</tr>
<tr>
<td>10</td>
<td>(1,9)</td>
<td>(18.91, 2133)</td>
</tr>
</tbody>
</table>
three signatories is not externally stable because the type 1 country can buy the cooperation of another type 2 country. Thus, the second part of Definition 2 does not hold. Consequently, an agreement consisting of one type 1 country and two type 2 countries is only externally and internally stable when \( m \) is greater than 1, but equal or lower than 4.24. The same argument applies for the other levels of participation and for an agreement that includes two type 1 countries. Table 1 shows that cooperation increases in both cases with the heterogeneity in environmental damages, as established in Proposition 4. However, the grand coalition cannot be self-enforcing unless there are only one or two type 1 countries.

Finally, we would like to point out that the reason why only one or two type 1 countries can belong to a self-enforcing agreement is very similar to the reason that explains the second part of Proposition 3: an agreement consisting of two countries is always internally stable if it is profitable.

Next we study the relationship between the degree of asymmetry, participation and the gains to cooperation. First, using (5) we calculate the gains to full cooperation obtaining the following expression

\[
2 \sum_{i=1}^{N} (C_i(0,0) - C_i(N_1, N_2)) - \frac{1}{2m} \left( N_1 m^2 + N_2 m^2 + (N-2)(N_1 m + N_2 m) \right),
\]

that multiplying and dividing by \( m^2 \) can be written as

\[
TC(0,0) - TC(N_1, N_2) = \frac{m^2}{2c} \left( N_1 m^2 + N_2 m^2 + (N-2)(N_1 m + N_2 m) \right),
\]

where \( TC(0,0) = \sum_{i=1}^{N} C_i(0,0) \) and \( TC(N_1, N_2) = \sum_{i=1}^{N} C_i(N_1, N_2) \). This expression says us that given \( m \) the gains to cooperation depend on the absolute value of parameter \( m \). Thus, a change in \( m \) can increase, decrease or keep constant the gains to cooperation depending on the change in the absolute value of parameter \( m_2 \) we consider. Now, using also (5) we calculate the gains to a given agreement \((n_1, n_2)\) that are given by

\[
\sum_{i=1}^{N} (N_i - n_i) \left( C_i(0,0) - C_i(n_1, n_2) \right) + \sum_{i=1}^{N} n_i \left( C_i(0,0) - C_i(n_1, n_2) \right)
\]

\[
= \frac{1}{2c} \left( n_1 m^2 + n_2 m^2 - n(n_1 m + n_2 m) \right) + 2(n-1)(N_1 m + N_2 m) (n_1 m + n_2 m),
\]

that multiplying and dividing by \( m^2 \) can be written as

\[
TC(0,0) - TC(n_1, n_2) = \frac{m^2}{2c} \left( N_1 m^2 + n_2 m^2 - n(n_1 m + n_2 m) \right) + 2(n-1)(N_1 m + N_2 m) (n_1 m + n_2 m),
\]

where \( TC(n_1, n_2) = \sum_{i=1}^{N} (N_i - n_i) C_i(n_1, n_2) + \sum_{i=1}^{N} n_i C_i(n_1, n_2) \), so that the relative gains to cooperation for a given agreement depends only on the heterogeneity of countries:

\[
\frac{TC(0,0) - TC(n_1, n_2)}{TC(0,0) - TC(N_1, N_2)} = \frac{n_1 m^2 + n_2 m^2 - n(n_1 m + n_2 m)}{N_1 m^2 + N_2 m^2 + (N-2)(N_1 m + N_2 m)^2}.
\]

Thus a change of \( m_2 \) keeping constant \( m \) will affect the gains to full cooperation but, according to Proposition 4, will not have any effect on the level of cooperation but besides, according to (14), will not have any effect on the relative gains of cooperation either so that we can conclude that

**Corollary 1.** Any variation in the absolute gains to full cooperation that it is not associated with a change in heterogeneity neither alter the level of cooperation nor the relative gains achieved by the agreement.

In fact, a reduction in the gains to full cooperation could yield an increase in the level of cooperation provided that this reduction is accompanied with an increase in heterogeneity. Summarizing, our analysis establishes that heterogeneity is the determinant of the level of cooperation independently of which are the gains to full cooperation associated with the degree of heterogeneity of countries. Thus, an increase in heterogeneity will have the same effect on the level of cooperation both if the variation in heterogeneity comes with an increase in the gains to full cooperation and if it has no effect on such gains.

Finally, we investigate which are the effects of a change in the absolute value of parameter \( m_2 \) that does not change the relative value \( m \). From (9)–(11) is immediate that when \( m_2 \) raises for a constant \( m \), the incentives to join or exit from the agreement are going to increase as well as the resources available to buy the cooperation. Thus, the greater the mean of the environmental damages for a given \( m \), the greater the necessary minimum transfers to stabilize the agreement. Nevertheless, type 1 countries will have more resources to buy the cooperation as well. Moreover, according to (12), when \( m_2 \) raises, the gains to full cooperation will also raise yielding an increase in the absolute gains for each level of cooperation. Notice that the relative gains, according to Corollary 1, will not change so that an increase in the gains to full cooperation will imply an increase in the absolute gains achieved by the agreement. Next we present two numerical examples that illustrate these results.

In the first example we have considered changes in heterogeneity that do not modify the gains to full cooperation. This example is shown in Table 2. In order to build the table, we have selected some values for \( m \) belonging to the different intervals that appear in Table 1 and for each one of these values, that are in the first column of Table 2, we have calculated using (12) the corresponding value of \( m_2 \) that yields the gains to full cooperation of the symmetric case. For these calculations we have given the following values to the parameters: \( N_1 = 3 \), \( N_2 = 9 \) and \( c = 0.5 \), and \( m_1 = m_2 = 1 \) for the symmetric case. Thus, for all the combinations \((m, m_2)\) in the table, the gains to full cooperation are 1452, the gains of the symmetric case. Then we have calculated the gains from cooperation assuming that only one type 1 country buys the cooperation of type 2 countries. The absolute gains to the agreement (AG), calculated according to (13), appear in column four and the relative gains (RG), calculated according to (14) in column five. Moreover, in columns six and seven, we show the figures corresponding to (9) and (10). Finally, the sum of the internal stability conditions for the signatories given by (11) appears in the last column.

The table shows that both the participation and the gains from cooperation increase with the heterogeneity of countries. Moreover, it also shows that the maximum payment that a type 1 country is willing to pay for cooperation increases with respect to \( m \) and that the surplus that can be distributed among the different signatories also increases although at a different rate because the neces-

### Table 2

<table>
<thead>
<tr>
<th>( m )</th>
<th>( m_2 )</th>
<th>( \text{Agreements} )</th>
<th>( \text{AG} )</th>
<th>( % )</th>
<th>( \text{IIE}_1 )</th>
<th>( \text{IIE}_2 )</th>
<th>( \text{Surplus} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.63</td>
<td>0.71</td>
<td>1 (2)</td>
<td>129.66</td>
<td>8.93</td>
<td>-11.92</td>
<td>4.62</td>
<td>-2.68</td>
</tr>
<tr>
<td>5.50</td>
<td>0.47</td>
<td>1 (3)</td>
<td>229.71</td>
<td>15.82</td>
<td>-37.94</td>
<td>11.05</td>
<td>-4.78</td>
</tr>
<tr>
<td>7.97</td>
<td>0.36</td>
<td>1 (4)</td>
<td>329.46</td>
<td>22.69</td>
<td>-64.79</td>
<td>14.79</td>
<td>-5.63</td>
</tr>
<tr>
<td>10.41</td>
<td>0.30</td>
<td>1 (5)</td>
<td>429.65</td>
<td>29.59</td>
<td>-93.06</td>
<td>17.37</td>
<td>-6.19</td>
</tr>
<tr>
<td>12.84</td>
<td>0.25</td>
<td>1 (6)</td>
<td>530.12</td>
<td>36.51</td>
<td>-122.16</td>
<td>19.26</td>
<td>-6.59</td>
</tr>
<tr>
<td>15.27</td>
<td>0.22</td>
<td>1 (7)</td>
<td>630.89</td>
<td>43.45</td>
<td>-151.83</td>
<td>20.70</td>
<td>-6.92</td>
</tr>
<tr>
<td>17.70</td>
<td>0.19</td>
<td>1 (8)</td>
<td>731.66</td>
<td>50.39</td>
<td>-181.86</td>
<td>21.83</td>
<td>-7.20</td>
</tr>
<tr>
<td>20.12</td>
<td>0.17</td>
<td>1 (9)</td>
<td>832.72</td>
<td>57.35</td>
<td>-212.08</td>
<td>22.74</td>
<td>-7.38</td>
</tr>
</tbody>
</table>
Table 3
Changes in heterogeneity that modifies the gains to full cooperation. $N_{1} = 3$, $N_{2} = 9$, $c = 0.5$, $m_{1} = 1$.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Agreements</th>
<th>AG</th>
<th>RG (%)</th>
<th>IEC$_{1}$</th>
<th>IEC$_{2}$</th>
<th>Surplus</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.63</td>
<td>(1,2)</td>
<td>257.41</td>
<td>8.93</td>
<td>-23.67</td>
<td>9.18</td>
<td>-5.31</td>
</tr>
<tr>
<td>5.50</td>
<td>(1,3)</td>
<td>1044.75</td>
<td>15.82</td>
<td>-172.50</td>
<td>50.25</td>
<td>-21.75</td>
</tr>
<tr>
<td>7.97</td>
<td>(1,4)</td>
<td>2502.49</td>
<td>22.69</td>
<td>-492.17</td>
<td>112.34</td>
<td>-42.81</td>
</tr>
<tr>
<td>10.41</td>
<td>(1,5)</td>
<td>4888.14</td>
<td>28.59</td>
<td>-1058.70</td>
<td>197.65</td>
<td>-70.45</td>
</tr>
<tr>
<td>12.84</td>
<td>(1,6)</td>
<td>8429.49</td>
<td>36.51</td>
<td>-1942.40</td>
<td>306.27</td>
<td>-104.78</td>
</tr>
<tr>
<td>15.27</td>
<td>(1,7)</td>
<td>13361.18</td>
<td>43.45</td>
<td>-3215.40</td>
<td>438.41</td>
<td>-146.53</td>
</tr>
<tr>
<td>17.70</td>
<td>(1,8)</td>
<td>19912.40</td>
<td>50.39</td>
<td>-4948.60</td>
<td>594.09</td>
<td>-195.88</td>
</tr>
<tr>
<td>20.12</td>
<td>(1,9)</td>
<td>28289.81</td>
<td>57.35</td>
<td>-7205.70</td>
<td>772.73</td>
<td>-251.13</td>
</tr>
</tbody>
</table>

sary minimum transfers to stabilize the agreement also increase. See the last two columns in Table 2. Thus, the greater the heterogeneity, the greater the resources available to eliminate free-riding incentives. For instance, for $n = 10$, the agreement consists of nine type 2 countries ($N_{2} = 9$) and one type 1 country ($N_{1} = 1$). For this case, the signatory of type 1 supports an increase in costs equal to 212.08 if it quits the agreement, whereas each signatory of type 2 would obtain a reduction in costs equal to 22.74 if it quits. Thus, the signatory of type 1 can eliminate this incentive paying each signatory of type 2 this amount and its costs would still be lower than the costs that it would incur by quitting the agreement for an amount equal to 7.38.

In the second example we have considered variations in heterogeneity that raise the gains to full cooperation. See Table 3. For this example $m_{0} = 1$, so that according to (12) the greater $m$ the greater the gains to full cooperation.

However, comparing the two tables, we see that for a given degree of heterogeneity the increase in the gains to cooperation has no effect on participation. The unique difference we observe in the comparison is in the absolute gains achieved for the agreement and also in the surplus, as we have pointed above. Now for $n = 10$, the signatory of type 1 supports an increase in costs equal to 7205.70 if it quits the agreement, whereas each signatory of type 2 would obtain a reduction in costs equal to 772.73 if it quits. Thus, the signatory of type 1 can eliminate this incentive paying each signatory of type 2 this amount and its costs would still be lower than the costs that it would incur by quitting the agreement for an amount equal to 251.13.

6. Conclusions

This paper analyzes the stability of IEAs under the assumption of asymmetry when countries can use transfers. To model the formation of an IEA we propose a two-stage game that has been analytically solved. The timing of the game is as follows: in the first stage each country decides whether or not to join a unique IEA based on the assumption that signatories will not only choose their emissions jointly in the second stage, but can also implement a transfer vector. In this setting, an agreement will be self-enforcing if no signatory country has incentives to leave the agreement and if no non-signatory has incentives to join the agreement, taking the membership decisions of all other countries and the transfers that they receive or pay as given.

Our findings allow us to conclude that heterogeneity between countries has no relevant effects on the scope of environmental cooperation in comparison with the homogeneous case if transfers are not allowed. With transfers, effects depend on the kind of asymmetry. If abatement costs are different, only limited cooperation can be bought through transfers. On the contrary, if the countries differ in terms of environmental damages, the level of cooperation increases with the degree of asymmetry. This result confirms in a more general setting, that obtained by Barrett (2001), which establishes that a strong asymmetry can lead to a high degree of international cooperation if transfers are used. Moreover, our result formally establishes a positive relationship between the scope of cooperation and the degree of asymmetry independently of the effect that a change in the heterogeneity of the countries has on the gains to full cooperation.

The following step of this research could be to investigate whether our conclusions can be extended to the case of increasing marginal environmental damages. A first numerical example we have developed suggests that Proposition 4 could also hold for the case of quadratic environmental damages. Nevertheless, a more complete analysis is needed to confirm this result.

Acknowledgements

An initial version of this paper was presented at the XIV Annual Conference of the European Association of Environmental and Resource Economists, Bremen (Germany), June 23–26, 2005, the II Congress of the Spanish–Portuguese Association of Environmental and Natural Resources Economics, Lisbon (Portugal), June 2–3, 2006 and the VI Meeting on Game Theory and Practice Dedicated to Development, Natural Resources and the Environment, Zaragoza (Spain), July 10–12, 2006. We are grateful to J.C. Pereau and three anonymous referees for very useful comments. Usual caveats apply. Financial support from Fundación BBVA and Ministerio de Educación y Ciencia under grants SEJ2004-05704 and SEJ2007-67400 are gratefully acknowledged. Grant SEJ2004-05704 has received support from FEDER funds.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2009.05.006.

References