Can Justice and Fairness enlarge the Size of International Environmental Agreements?

Christine Grüning[1] and Wolfgang Peters

April 2007

[1]Corresponding author: Christine Grüning, PO Box 1786, 15207 Frankfurt (Oder), Germany, cgruening@euv-frankfurt-o.de. Helpful comments by Michael Finus, Alexander Haupt, Silke Gottschalk, Michael Grüning and the participants of the IIPF and EAERE conference as well as workshops in Berlin, Bonn, and Rostock are gratefully acknowledged. We also want to thank the German Research Foundation (DFG) for support through the SPP 1142 program on "Institutional Design of Federal Systems".

Abstract

The literature on International Environmental Agreements (IEA) predicts a rather low number of signatories to an IEA. This is in sharp contrast to empirical evidence. As experimental economics provides some evidence for more complex human behavior, extending the theory of IEA to a broader class of preferences is clearly promising. The present paper shows that where countries' preferences incorporate justice and fairness there will be a strong incentive for them to choose similar abatement policies within and outside an IEA. Consequently, free-riding at the expense of the signatory states diminishes and participation in an IEA becomes a more successful strategy, so that the size of stable IEAs increases.

JEL classification: C7, D62, D63, H77

Keywords: International Environmental Agreements, coalition formation, justice and fairness.

1 Motivation

Where supranational institutions are absent voluntary cooperation between countries is designed to restrict harmful impacts of global environmental problems like the ozone layer, global warming or greenhouse gas emissions. Although not all countries cooperate, we can nevertheless observe several International Environmental Agreements (IEAs) where the number of signatories is decidedly large, e.g. the Montreal or the Kyoto Protocol.[1] However, empirical evidence is in sharp contrast to theoretical prediction. It is a well-known result in the theory on IEAs that the number of cooperating countries that are actively engaged in an IEA is likely to be very small.[2] To reconcile empirical evidence and economic theory the standard literature must be extended so that the huge participation and compliance in International Climate Change Agreements can be explained.[3]

Standard IEA models assume that each country is concerned only with its own welfare, defined as environmental benefits minus abatement costs. Nevertheless, experimental economics does provide some evidence for human behavior that is more complex than pure selfishness.[4] As IEAs have to be mutually beneficial and reasonably fair for all participants, enlarging the theory of IEAs to a broader class

[1] Evaluating the effectiveness of an IEA depends not only on the coalition size. Murdoch & Sandler (1997) maintain that there are agreements with a huge number of signatories and a rather lax abatement level similar to the non-cooperative Nash outcome.

[2] For details see Barrett (1992, 1994), Hoel (1992), Carraro & Siniscalco (1993) or Finus (2001).

[3] See Barrett (2002), Barrett & Stavins (2003) or Buchholz & Peters (2005) for recent approaches to reinforce the incentives to engage in IEAs. A strong recommendation for IEAs is given by Stern (2006).

[4] Cf. Rabin (1993), Fehr & Schmidt (1999), Bolton & Ockenfels (2000), Falk, Fehr & Fischbacher (2003), or Alesina & Angeletos (2005) and the literature cited in these papers. For a more general view on interdependent preferences and reciprocity, see Sobel (2005). Introducing justice and fairness bears the risk of modelling the extended preferences arbitrarily, as a broad range of observations can be used to explain any type of behavior, cf. Postlewaite (1998).

of preferences is clearly promising. As even governments are not just concerned about welfare alone recent papers on IEAs extend governmental decisions to issues of justice and fairness.[5] An empirical investigation of Lange et al. (2007) focuses on the relevance of equity considerations during the process of negotiations about international climate agreements. In their study almost all the experts involved in the Kyoto negotiation procedure stated that fairness play an important role.[6]

Justice put some pressure on governments to accept similar responsibilities. Thus, countries apply relatively conform measures or strategies which can be implemented in the theoretical analysis either by new instruments or by extending governments' objectives. The former strategy was used by Hoel (1992) or Finus & Rundshagen (1998) who focus on uniform emission reductions or uniform quotas. Here an institutional restriction obliges countries to behave alike but there is still no endogenous motivation for such an institutional rule. Hence, extending the class of preferences aims at an endogenous explanation of conform behavior.

Hoel & Schneider (1997), followed by Jeppesen & Andersen (1998), first applied a broader class of preferences to IEAs. They assume that countries are not only concerned with their own welfare. Countries' well-being is also related to the behavior of the other countries so that becoming a member of an IEA is an end in itself. As countries in this setting prefer to join an agreement, governments exhibits some

[5] According to Albin (2003) fairness has multiple facets (e.g. altruism or reciprocity) and is not unambiguously defined. Both concepts put some pressure on governments to behave conform and thus enforces behavioral or social norms. Following Lindbeck (1997) norms have an impact on rational behavior. As shown in Wooders et al. (2007) self-interested behavior, conformity and social norms need not be inconsistent. This is in line with Elster (1989, p. 102) who noticed that individual "actions typically are influenced both by rationality and by norms". For more details about social norms and private provision of public goods see Rege (2004).

[6] Lange et al. (2007) asked in their questionnaire for both, experts' own view on equity as well as the perception of equity views in different countries or country groups.

kind of conform attitude.

In a recent paper Lange & Vogt (2003) integrate altruism as another motivation for cooperation by applying Bolton & Ockenfels' (2000) ERC preferences.[7] While ERC preferences focus on an international welfare which is not directly observable, fairness is often a matter of verifiable measures. Lange et al. (2007) support empirical evidence that countries first focus on abatements[8] they observe so that free-riding at the expense of others conflicts with fairness.

We integrate two types of offsetting behavior that have an impact on an IEA: either only member states bear the abatement measures or all countries, both inside and outside the IEA, actively do so (independent of the coalition size). As we will show in what follows, complete or partial free-riding (doing nothing or applying a moderate policy only) plays a crucial role.[9] The larger the number of IEA members, the more the signatories internalize the global externality. So the incentive to provide additional measures outside the coalition diminishes and, consequently, any non-signatory state becomes a complete free-rider. Therefore, integrating both types of free-riding provides the possibility to study the offsetting behavior outside correctly.

Countries with fairness-oriented preferences compare their own measures or abatement costs with that of all other countries. If costs are relatively heterogenous among these countries, this will be seen as rather 'unfair'. Hence, governments try to avoid cost dispersions, which in our context is reflected by the variance in abatement costs. This concept is in line with Fehr & Schmidt (1999, F&S). Contrasting to ERC pref-

[7] ERC stands for equity, reciprocity and competition.

[8] See Victor & Coben (2005) for a different view of countries' conform behavior. They suggest that quantity strategies are favored by the diplomatic community over price instruments. While equal treatment can easily be granted with the former instrument, the latter often results in heterogenous quantity effects among countries.

[9] Lange & Vogt (2003) do not distinguish between these offsetting strategies, so that their analysis ignores complete free-riding out the IEA.

erences, which prefer a behavior close to the average, F&S assume that countries dislike economic differences between countries. If *first world* countries were to finance a significant part of the abatement strategies of the *third world,* this would be favored by the *second world* in the case of F&S, while, for ERC preferences, the *second world* countries are indifferent as such a measure does not change their average position significantly. Hence, F&S is slightly more in favor of a moderate redistribution, or a just and fair division of abatement duties.[10] All countries will do their share when they believe that all, or sufficiently many of them, will do theirs and they dislike own and other countries' deviations from such a conform behavior.[11]

The economic intuition for coalition formation distinguishes between three underlying motives: the traditional ones, i.e. countries' *individual gain from free-riding*, and the *collective efficiency gain* which measures the internalization of the environmental externality, and in addition the impact of justice and fairness, which favors similar behavior with respect to abatements (*conform behavior*). The last aim can be best met when each country has a similar participation strategy. This results in either the grand coalition or complete failure of the negotiation process. As the traditional effects work in opposite directions and fairness destabilizes medium size coalitions, it is the interplay of all three effects that determines the equilibrium of the entire game.

The paper is organized as follows: In section two we present the economic framework. Subsequently, in section 3 we analyze the policy game on abatements, which is followed by the formation of an IEA through a coalition of countries. Finally, we present some concluding remarks.

[10] Moreover, Engelmann & Strobel (2004) demonstrated that F&S preferences perform better than ERC preferences in explaining the observations from experiments.

[11] This point of view directly corresponds to Rawls' (1971, p. 236) theory of justice.

2 Economic Framework

In what follows we study in a complete information world a standard coalition formation game like that introduced by Barrett (1992, 1994) or Carraro & Siniscalco (1993). The aim is to explain international cooperation for $N \geq 4$ identical countries in case of an IEA.[12] In the first stage, countries can choose whether or not to join an IEA. This decision process results in S signatory states with the remaining $(N - S)$ countries behaving non-cooperatively. Subsequently, in stage two, the signatories and the outsiders of the IEA decide simultaneously on their abatement measures.[13]

The choices at both stages are determined through rational behavior of all countries. Instead of following the traditional approach, which focuses on the benefit of global abatement strategies minus private costs of the environmental policy, we additionally rely on preferences which directly integrate fairness considerations. As governments can easily observe countries' policies, they can see whether other countries are more engaged in environmental concerns. Then, their own deviations from a homogeneous strategy as well as foreign ones can be seen as a welfare loss. Hence, justice and fairness focuses on the differences in observable abatement strategies on cost dispersion. As a consequence, country j's payoff consists of the benefit minus

[12] As we tackle global environmental problems, the number of countries involved is sufficiently large and $N \geq 4$ is a rather weak assumption. If there are less than four countries there is no incentive to stay outside an IEA. However, we focus on agreements where the participation is endogenous and should therefore not be predetermined.

[13] Contrary to Barrett (1992, 1994) we simplify the strategic interaction between the insider and outsider by neglecting Stackelberg behavior by the coalition. Thus, we focus on simultaneously acting countries like in Carraro & Siniscalco (1993). However, as we deal with strategic substitutes, the coalition's weak position produces a more engaged IEA. Consequently, the abatement activities of a member and an outsider are more polarized than under the Stackelberg assumption. We will discuss this alternative approach in more detail in subsection 4.3.

costs[14] – represented by a quasi-linear logarithmic function – minus a term which measures heterogeneity by means of the variance in all abatement strategies

$$P_j = \ln\left(\sum\nolimits_i a_i\right) - a_j - \theta \cdot \sigma\left(a_1, ..., a_N\right), \tag{1}$$

where $(a_i, a_j) \geq 0$ correspond to the abatement levels of country i and j, while $\sigma(a_1, ..., a_N)$ measures the variance in the environmental policy of all countries'.[15] A country's payoff is strictly concave in its own strategy and continuous in that of the opponents. Moreover, in order to analyze the impact of justice and fairness, we introduce a parameter, $\theta \geq 0$, that represents the preference intensity for the welfare loss due to cost dispersion. Thus, in the case of $\theta = 0$, the payoff function of country j coincides with that in the traditional approaches which focus on pure selfishness. Increasing θ corresponds with a stronger concern for 'just or fair' cost shares.

Following the literature on IEA, we have a two-stage game, where the countries decide at the first stage whether to sign an IEA, given the decision of all other countries. Such a decision has to be based on what countries do after the signatories of the IEA and the outsiders are determined. For this reason, countries need to anticipate the level of abatements of the countries both inside and outside the coalition $(a_1, ..., a_N)$, which will be established at the second stage. Furthermore, IEAs are voluntary alliances of at least two countries ($S \geq 2$). All signatory states S behave cooperatively among themselves, whereas $(N - S)$ singletons behave non-cooperatively towards both the coalition and each other. The voluntary nature of an IEA implies that a country joins a coalition only if this is a profitable strategy

[14] For simplicity, we assume that costs are linear in abatements.

[15] Following Alesina & Angeletos (2005) the variance is a good measure for fairness and justice. The variance in the abatement strategies is defined as $\sum_i \frac{1}{N}\left(a_i - \overline{a}\right)^2$, where $\overline{a}$ is the global average of all countries' environmental policies. According to Rege (2004) the global average $\overline{a}$ can be seen as a norm for conform behavior.

for the potential signatory. Needless to say, equilibrium participation in an IEA requires both internal and external stability, i.e. no insider and no outsider has an incentive to deviate from the chosen participation strategy.[16]

Our principal objective is to analyze whether the number of IEA members is positively correlated with issues of justice and fairness in countries' preferences. However, before studying stability, we solve the game by backward induction.

3 Policy Game: The Second Stage

The countries simultaneously determine their abatement strategies at the second stage of the game. In the presence of a positive global environmental spillover, the voluntary cooperation of some countries improves the situation of the remaining singletons and creates an incentive to free-ride. If countries have identical preferences, signatories are – in equilibrium – more engaged than outsiders ($a_s > a_o$). As singletons can decide whether, and how, to engage in environmental concerns, we distinguish between complete ($a_o = 0$) and partial ($a_o > 0$) free-riders. Existence of a Nash equilibrium in the abatement game is guaranteed as the payoff function is strictly concave in the own strategy and continuous in the opponents' strategies. Additionally, the equilibrium is unique as we can apply a more general proof for the coalitional equilibria of Finus, v. Mouche & Rundshagen (2005) to our framework.[17] Consequently, in accordance with identical preferences and uniqueness, all signatories and outsiders are symmetric.

The cooperatively acting members of an IEA maximize their joint payoff, given

[16] This definition corresponds with that of cartel stability presented in the oligopoly literature by d'Aspremont & Gabszwicz (1986).

[17] An exception from uniqueness is given for $\theta = 0$. In that case, only aggregate abatement of the non-signatory states is unique, but we have a continuum of equilibria because of quasi-linearity. To simplify the analysis, in case of $\theta = 0$ we stick only to the symmetric solution.

the abatement levels of the non-cooperative countries a_o. Due to symmetry, the coalition maximizes the payoff of a representative IEA member, P_s. The coalition at least is engaged in abatements $a_s > 0$, which typically exceed the global average $\overline{a}$. Therefore, we have the following first-order condition, where the marginal benefit is balanced against marginal costs and the impact on cost dispersion

$$S\left[\frac{1}{Sa_s + (N-S)\,a_o} - \frac{2\theta\,(a_s - \overline{a})}{N}\right] - 1 = 0. \tag{2}$$

Singletons, in contrast, behave non-cooperatively towards the coalition and the other outsiders. To determine their best responses, they maximize their own payoff P_j given the abatement strategies of all countries $i \neq j$. In equilibrium, symmetry implies that the abatement strategies of each outsider a_o will be the same. We obtain the first-order condition for the non-signatories

$$\frac{1}{Sa_s + (N-S)\,a_o} - \frac{2\theta\,(a_o - \overline{a})}{N} - 1 \leq 0. \tag{3}$$

While marginal costs both inside and outside the IEA are identical, marginal benefits between signatories and outsiders deviate by the factor S. Thus, due to the internalization of the environmental externality through the formation of an IEA, the abatement activity of a non-signatory falls short of that of a signatory state.

In equilibrium the abatement activities[18] of the countries inside and outside the

[18] If we assume an upper bound for emissions exceeding one, equilibrium abatements fall short of the maximum emission.

IEA are

$$a_s^* = \begin{cases} \frac{1}{N-S+1} + \frac{1-S}{2\theta} + \frac{N(S-1)}{2\theta S} & \theta > \widetilde{\theta} \\ & \text{for} \\ \frac{\sqrt{N^4+8\theta N^2 S(N-S)}-N^2}{4\theta S(N-S)} & \theta \leq \widetilde{\theta} \end{cases} \tag{4}$$

and

$$a_o^* = \begin{cases} \frac{1}{N-S+1} + \frac{1-S}{2\theta} & \theta > \widetilde{\theta} \\ & \text{for} \\ 0 & \theta \leq \widetilde{\theta}. \end{cases} \tag{5}$$

The threshold level $\widetilde{\theta} = 0.5\,(S-1)\,(N-S+1)$ separates partial from complete free-riding of the outsiders when at least some signatories form an IEA.[19] While for sufficiently strong fairness considerations, $\theta > \widetilde{\theta}$, the non-signatories are partial free-riders, they behave as complete free-riders for rather weak preferences, $\theta \leq \widetilde{\theta}$. Whether a country outside the IEA becomes a complete or partial free-rider depends on countries' preferences and the number of signatories.[20]

In equilibrium, for a given coalition size S, the aggregate abatement activities A corresponds with the sum of abatements of the signatories and outsiders

$$A\,(S) = Sa_s^* + (N-S)\,a_o^*. \tag{6}$$

[19]Note, even for countries with identical preferences, in equilibrium we end up with different participation strategies and thus asymmetric abatement levels (eqs. (4) and (5)). Thus, although all countries are homogenous *ex ante*, they become heterogenous *ex post*.

[20]This threshold level increases with the total number of countries, $\partial\widetilde{\theta}/\partial N > 0$ respectively. The more countries that are faced with the environmental problem, the stronger is the incentive to free-ride. Thus, partial free-riding of the outsider countries requires a relatively strong θ. If the majority of the countries behave non-cooperatively (cooperatively), an increasing number of coalition members results in an increasing (decreasing) threshold level, $\partial\widetilde{\theta}/\partial S = \frac{N+1}{2} - S$.

Subsequently, these equilibrium values are used to analyze the participation strategies at the first stage.

If preferences are purely selfish, an outsider is a complete free-rider. This result changes if justice and fairness enters the scene. In our framework, the countries achieve more homogeneity, measured by the variance in economic behavior, through the choice of similar abatement strategies.

Proposition 1 *Abatements inside and outside the coalition.*
i) For signatories, stronger fairness preferences result in smaller abatement activities. If θ exceeds the threshold level $\widetilde{\theta}$ even an outsider becomes active. The stronger θ, the more abatements an outsider carries out. In the limit (for $\theta \to \infty$) there is no difference between an insider and an outsider.
ii) The aggregate does not significantly change in θ. For $\theta < \widetilde{\theta}$, fairness has a negative impact on global abatements, while $A(S)$ remains constant for all θ exceeding the threshold $\widetilde{\theta}$.[21]

Stronger fairness attitudes result in more homogeneity as countries inside and outside an IEA adjust their abatement levels to each other. In case of $\theta < \widetilde{\theta}$, outsiders are complete free-riders and more homogeneity requires a reduction in coalition's abatements. The price for less unequal cost sharing among countries is a loss in environmental quality as the global abatement measures $A(S)$ are reduced.

As long as both signatories and outsiders adopt an active measure, i.e. for $\theta > \widetilde{\theta}$, the aggregate abatement $A(S)$ does not change. For stronger θ, countries' abatements become similar which results in a redistribution of cost shares from the members of an IEA to outsider countries. While the non-signatory states reinforce their environmental policy a_o, measures of the IEA members a_s are reduced, cf. figure 1. This behavior is driven by the wish not to deviate too much from the

[21] The proof for i) and ii) follows immediately from eqs. (4), (5) and (6).

abatement measures the other countries realize. Thus, although countries choose different participation strategies they behave rather conform with respect to the abatement policy itself. Consequently, justice and fairness reduce the incentive to leave a coalition.

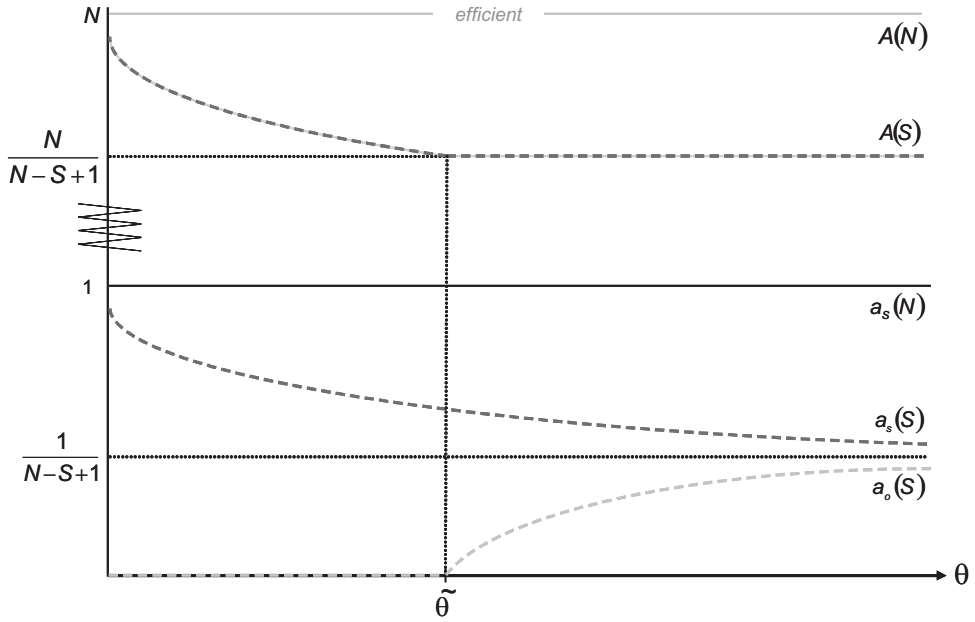

Figure 1: Abatements for coalition size S and N

Summarizing, for each given coalition size S, justice and fairness enforce similar abatement strategies even at the expense of a reduction in overall measures. Conform behavior in environmental policy results at the second stage of the entire game. However, what are the consequences for the participation strategies? Does the driving force for similar policy measures stabilize larger coalitions as governments feel a stronger incentive to join an agreement? As we are interested in the impact of our extended preferences on the size of an IEA, these questions will be analyzed in what follows.

4 Signing an IEA: The First Stage

Introducing a measure for countries' fairness preferences means that similar behavior becomes decisive. Intuitively speaking, either nearly all, or almost none, of the countries are expected to sign an IEA. Whether this conjecture holds true will be analyzed in what follows.

Although the stability of an agreement depends on the payoffs resulting from the policy game at the second stage, it is worthwhile taking a closer look at the variance in the environmental policy, as this is the origin of our new insights into coalition formation.

4.1 Cost Dispersion

It is the interplay of S and θ which becomes decisive for cost dispersion measured by the variance in abatements. Both parameters have an impact on the extent of global abatements and the distribution of the cost shares as they determine countries' (partial or complete) free-riding behavior.

Inserting all information about the abatements (eqs. 4 and 5), the variance is given by

$$\sigma(S,\theta) = \begin{cases} \frac{(N-S)(S-1)^2}{4\theta^2 S} & \theta > \widetilde{\theta} \\ & \text{for} \\ \frac{S(N-S)[a_s^*(S,\theta)]^2}{N^2} & \theta \leq \widetilde{\theta}, \end{cases} \tag{7}$$

where the threshold level $\widetilde{\theta}$ is defined as before. Obviously, the more the countries dislike an unfair cost dispersion, the smaller is the variance in their abatement activities. The following lemma summarizes the impact of S and θ on the variance.

Lemma 2 *The variance $\sigma(S,\theta)$ is decreasing in θ and single-peaked in S.*[22]

[22] A technical proof for the lemma is presented in the appendix.

Intuitively speaking, starting from a small IEA, an increase in the coalition size yields more heterogeneity in countries' abatements as the variance accounts for all pair-wise differences in countries' policy measures, i.e. the number of *insider meets outsider* increases until both groups are of nearly equal size. For $S \approx 0.5N$, we have two groups of nearly equal size, such that, with a further increase of S, the variance declines until it vanishes for $S = N$.

For justice and fairness alone, a coalition consists ideally either of all countries or none.[23] Subsequently, we show how this effect changes the results on coalition formation previously analyzed in the traditional literature.

4.2 Stability Analysis

The stability analysis for the coalition formation depends on the equilibrium payoffs inside and outside the coalition that result from the policy game at stage 2, given the coalition size S. As all countries are assumed to be identical, the players' payoffs depend only on the coalition size and the type (signatory $P_s(S)$ or outsider $P_o(S)$). The stability analysis is based on the status quo relative to the prevailing alternatives. External stability requires that no outsider has an incentive to join the coalition, i.e. $P_s(S+1) - P_o(S) \leq 0$, while internal stability is fulfilled if no insider wants to leave the coalition, i.e. $P_s(S) - P_o(S-1) \geq 0$. Note, an extreme coalition formation with either $S = 1$ (complete failure of an IEA) or $S = N$ (grand coalition) only requires one of the above relations. For $S = N$ it is sufficient to check internal stability and, for $S = 1$, only external stability is relevant.

For the equilibrium size S^* external and internal stability must be fulfilled simultaneously. These two payoff differences are implicit functions of θ and consider

[23] Hoel & Schneider (1997) extend governments preferences so that becoming an IEA member is an end itself and encourages participation. Justice and fairness, however, destabilize medium size coalitions.

both types of offsetting behavior. However, for partial free-riding we can solve these relations for θ analytically (which is described by a function δ),[24] while for complete free-riding only a numerical solution exists (henceforth called Δ). Given δ and Δ, we can state the following equilibrium conditions

$$\begin{aligned} &\delta(S^*-1) < \theta < \delta(S^*) \qquad \text{for partial free-riding, and} \\ &\Delta(S^*-1) < \theta < \Delta(S^*) \qquad \text{for complete free-riding, respectively.} \end{aligned} \tag{8}$$

Each relation on the left side guarantees that no signatory wants to withdraw from the IEA, and the relations on the right prevent an outsider from joining the contract. Obviously, except for corner solutions $S = 1$ or $S = N$, an interior equilibrium can only be stable where δ, respectively Δ, are increasing in S.

Therefore, both complete and partial free-riding seem to be important. Still, for a relatively small number of countries ($N < 12$) that are concerned with an international problem (like the water quality in the Baltic Sea area), complete free-riding is not relevant for the stability analysis. Complete free-riding seems to be a phenomenon which becomes active if the environmental externality affects more and more countries. The externality increases in N and thus provides an incentive for outsiders to behave as complete free-riders.[25]

Lemma 3 *The stability of a coalition for $N < 12$ is exclusively determined through $\delta(S)$, while both stability conditions $\delta(S)$ and $\Delta(S)$ become relevant for $N \geq 12$.*

Proof: $\delta(S)$ proves for stability if $\delta(S) > \widetilde{\theta}(S)$, while $\Delta(S)$ determines stabil-

[24] The δ-function is defined as follows:

$$\delta(S) = \frac{3S^2(N-1) + S(N-1) - 2S^3 - N}{4S(S+1)\left[\ln\left(1+\frac{1}{N-S}\right) - \frac{1}{(N-S)(N-S+1)}\right]} > 0.$$

Both the numerator and the denominator are strictly positive and finite for all $1 \leq S \leq N-1$. Furthermore, for all S, the term in square brackets in the denominator does not exceed one.

[25] This result shows that, contrary to Lange & Vogt (2003), complete free-riding cannot be left out of the analysis without loss of generality.

ity if the opposite holds. i) Inserting all integer numbers for $N \in [4, 12)$ shows that the relation $\delta(S) > \widetilde{\theta}(S)$ is satisfied for all S. ii) The following relation $\delta(N-1) > \widetilde{\theta}(N-1)$ shows that partial free-riding is always relevant for $N-1$, while $\delta(N-2) < \widetilde{\theta}(N-2)$ hints at the relevancy of complete free-riding. Both relations hold true for $N \geq 12$. Q.E.D.

Note that, for $N \geq 12$, the free-riding behavior of the potential outsider in the case of larger coalitions (grand $S = N$ or all-but-one $S = N-1$) switches from partial to complete free-riding. Starting from the grand coalition, a single outsider reduces its abatements but remains active, while, with two outsiders, both become complete free-riders. As we will show in what follows, this effect can enforce the stability of an all-but-one coalition. However, before presenting specific results, we have to focus on the general analysis.

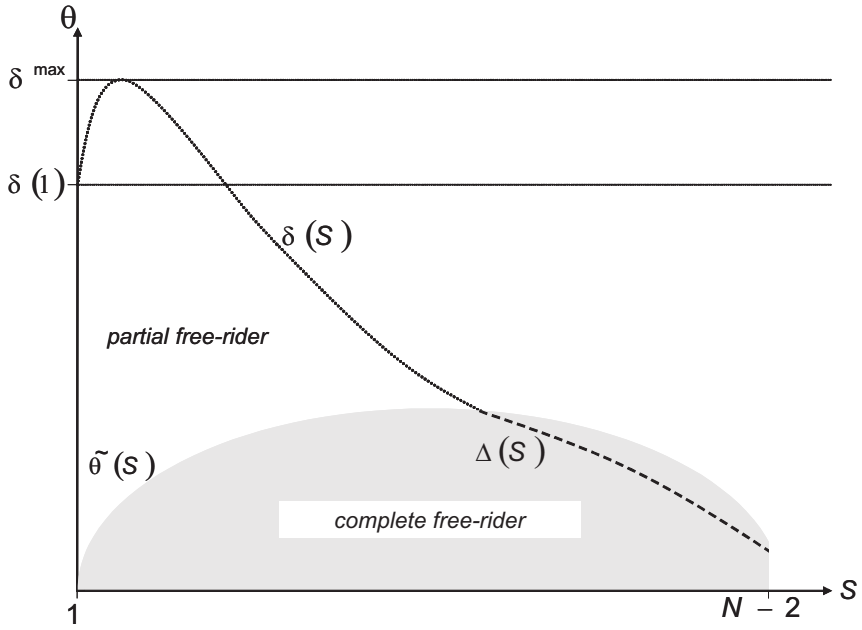

Figure 2: Stability analysis for $N = 100$

Although the number of countries N is decisive for the type of offsetting behavior, the main results with respect to coalition formation are qualitatively the same.

Therefore, we present our findings by using a numerical example with $N = 100$.[26] The analysis contains two parts.

First, figure 2 presents the δ-Δ-function for coalition sizes up to $(N-2)$. It shows the relevant parts of the functions $\delta(S)$ as a dotted line and $\Delta(S)$ as a dashed line that are separated from each other through $\widetilde{\theta}$. While the δ part shows a maximum, the Δ function is monotonously decreasing up to $(N-2)$. Within this range, offsetting behavior changes from partial to complete free-riding.

Second, following Lemma 3, offsetting changes go back to partial free-riding when almost all countries join the IEA, i.e. from $(N-2)$ to $(N-1)$. Here, we have to distinguish between two cases: According to our simulations, for $N < 28$ the δ-Δ-function increases in the relevant range, while it declines for $N \geq 28$.

As stable interior equilibria are located where the combined δ-Δ-function is increasing, we distinguish different areas for θ that are separated by $\delta(1)$, $\delta^{\max}$, $\Delta(N-2)$, and $\delta(N-1)$. Summarizing, we end up with four types of equilibria: two corner solutions ($S^* = 1$ and $S^* = N$) and two interior equilibria (a small coalition and an all-but-one coalition $S^* = N-1$). In order to test for an all-but-one coalition, we have to check whether $\delta(N-1)$ exceeds $\Delta(N-2)$. In either case, $\delta(1)$ and $\delta^{\max}$ are the relevant thresholds for θ that distinguish the areas for a failure of an IEA and a small coalition.

[26] We have checked the shape of the δ-Δ-function for all integer numbers N up to 100. Even for $N > 100$, there is no hint that qualitative results may change.

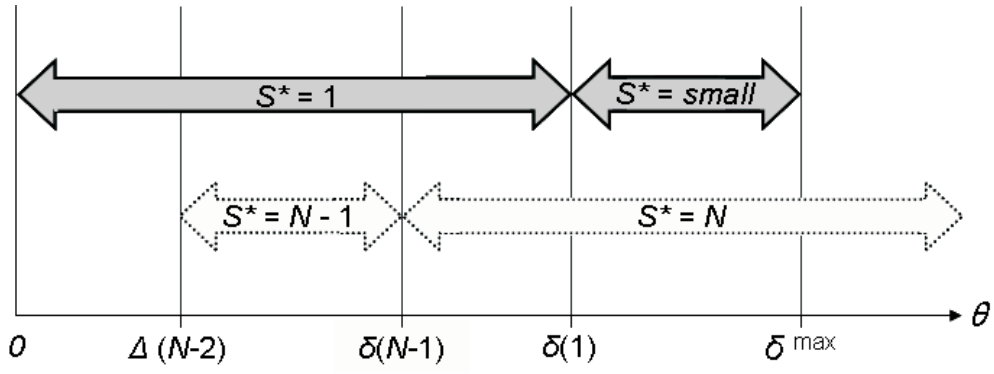

Figure 3a: Equilibria with all-but-one

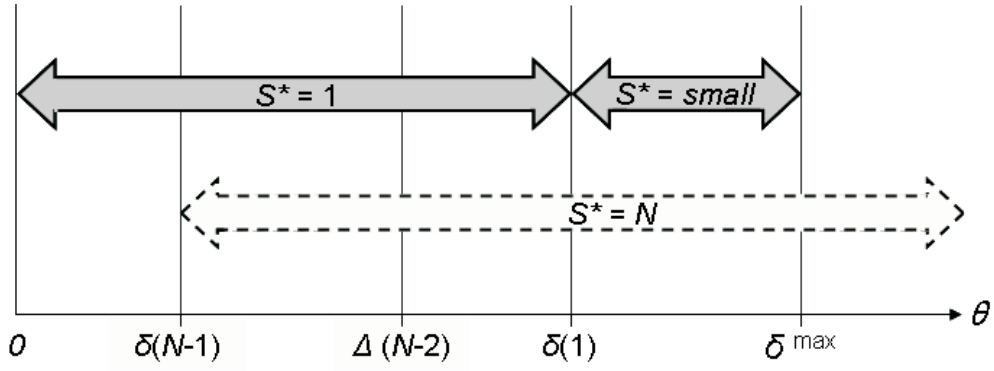

Figure 3b: Equilibria without all-but-one

As shown in figure 3, there are two different cases. If – as in figure 3a – $\Delta(N-2)$ falls short of $\delta(N-1)$, there is some range for all-but-one equilibria as well as for the grand coalition. If the opposite relation holds (figure 3b), there is only scope for the grand coalition. A unique equilibrium of the entire game only exists for rather low and very high θ. In the case of a medium size θ, coalitions always coexist with almost all countries and almost none.

Proposition 4 *Stable equilibria of an IEA.*

i) The corner solution $S^* = 1$ *is stable for* $\theta \in [0, \delta(1)]$.

ii) The stability of the grand coalition $S^* = N$ *requires* $\theta > \delta(N-1)$.

iii) For $\theta \in (\delta(1), \delta^{\max})$ *we obtain a rather small coalition* $S^* > 1$. *The coalition size for* $N \leq 100$ *does not exceed 6.*

iv) The all-but-one coalition $S^* = N-1$ *becomes stable if* $\delta(N-1)$ *exceeds* $\Delta(N-2)$ *and* $\theta \in [\Delta(N-2), \delta(N-1)]$. *The interval is non-empty for* $N < 28$. *For* $N \geq 28$ *this type of stable IEA does not exist.*

v) While uniqueness of the entire equilibrium requires $\theta > \delta^{\max}$ *for the grand coalition and* $\theta < \min\{\Delta(N-2), \delta(N-1)\}$ *in the case of a complete failure of an IEA, interior equilibria never occur alone. For an intermediate* θ, *the entire game has multiple stable equilibria.*

Proof: As the relation $\min\{\Delta(N-2), \delta(N-1)\} < \delta(1) < \delta^{\max}$ holds true in any case, we can distinguish the four areas for θ given in (i) up to (iv) . The first relation follows from inserting all information in the δ function, the second from the property of a maximum. The coalition size up to 6 in (iii) and the non-empty interval for $N < 28$ in(iv) can only be checked numerically, which we did for $N \leq 100$. Uniqueness (v) results as a consequence of (i) to (iv). Q.E.D.

As proposition 4 shows, a stable coalition can never be medium size. All kinds of equilibria are rather like those in the *battle of the sexes.* Adopting your partner's participation strategy is better than any other behavior.

According to the standard literature almost all countries defect. This behavior stems from the pure selfishness of all governments that is the driving force for free-riding at the expense of others. If preferences are extended, justice and fairness behavior stabilizes not only small, but even larger coalitions.

In general, there are three effects which determine stability: *individual free-riding*, *collective internalization,* and *conform behavior*, where the first two are the traditional Barrett effects. The individual gain from free-riding is rather selfish and depends on the difference between the insider's and the outsider's abatement. The

gain from internalization can be measured by the change in global abatements when a single country leaves or enters the coalition. Finally, justice and fairness favor similar policy measures, irrespective of the chosen participation strategy. According to the single-peaked variance in S, conform behavior destabilizes medium size coalitions as they provide an incentive either to leave or to join a coalition. Free-riding favors leaving a coalition, the efficiency argument goes in the opposite direction, while fairness prefers homogenous behavior. Consequently, it is the interplay of these three effects that determines the equilibrium coalition size.

Justice and fairness favor coalitions where either almost all or almost none sign an IEA. While for small coalitions conform behavior and free-riding are complementary in providing an incentive for staying outside an IEA, when the coalition size becomes large they favor opposite participation strategies. As in the traditional literature, it is the internalization gain which stabilizes IEAs, while the free-riding effect hinders larger coalitions. Thus, stronger fairness preferences are needed to overcome the instability of the grand coalition. In the case of the all-but-one coalition, the free-riding gain works to destabilize the grand coalition, but it is not strong enough to countervail the ones from internalization and conform behavior.[27]

4.3 IEA as Stackelberg Leader

While our previous result is based on the assumption that the members of an IEA do not intend, or are not able, to play a leading role vis-à-vis the non-signatory states, a complementary framework focuses on an IEA as a Stackelberg leader against the outsiders. As we have strategic substitutes and positive externalities, the members of an IEA can reinforce outsiders' abatements through a reduction in their own ac-

[27] In the case of complete free-riding a non-signatory's choice is restricted to $a_o = 0$. Hence, the gain from free-riding becomes smaller than for interior abatement strategies. As free-riding behavior changes from $(N-2)$ to $(N-1)$ this stabilizes the all-but-one coalition.

tivities. This directly favors the member states at the expense of the outsiders and thus provides an incentive to join an IEA. In the case of justice and fairness, Stackelberg leadership stabilizes larger coalitions even more as the economic behavior of insiders and outsiders turn out to be rather homogeneous. Like in our previous analysis, we can identify Stackelberg behavior, which also reduces the variation in the abatement measures, as a driving force for the formation of larger coalitions.

5 Conclusion

In the standard literature on International Environmental Agreements (IEA) empirical and theoretical predictions are inconsistent if we focus on the number of signing countries. While theory only proves the existence of small coalitions, there is empirical evidence for larger agreements such as the Kyoto or Montreal Protocols. By extending countries' preferences to incorporate issues of fairness and justice, governments try to avoid welfare losses due to cost dispersion, measured by the variance in countries' abatement policies. Such preferences provide an incentive for countries to behave in the same way, either almost all, or almost none, of the countries form an IEA. In both cases, the participation decisions are similar, which stabilizes both, larger and smaller coalitions but destabilizes medium-size coalitions.

Furthermore, the US staying outside an international agreement for free-riding reasons can be seen as an example of a stable all-but-one coalition. However, the applicability of our model should not be stressed too much. In reality countries are not identical in income, technology and preferences and the real world problems of forming an IEA are due to differences in these parameters. Hence, improving the empirical robustness of a theoretical model requires that at least some of these heterogeneities are integrated into the analysis.

6 Appendix: Single-peakedness of the variance

Obviously, $\sigma(S,\theta)$ is decreasing in θ as can be checked by a close look at eqs. (4) and (7). In the following, we prove the single-peakedness of σ in S in two steps: first, we determine the maximum $\sigma(S,\theta)$ in S for both partial and complete free-riding. Second, for a given θ outsiders can change their behavior from partial to complete free-riding, or the other way round, as the threshold $\widetilde{\theta}$ depends on S. Therefore, we distinguish between these two scenarios. As shown in figure 4, for low θ an increase in the coalition size S results in complete free-riding. For moderate θ, the free-riding behavior changes twice: from partial to complete free-riding and back. If θ exceeds a certain threshold, there is only partial offsetting.

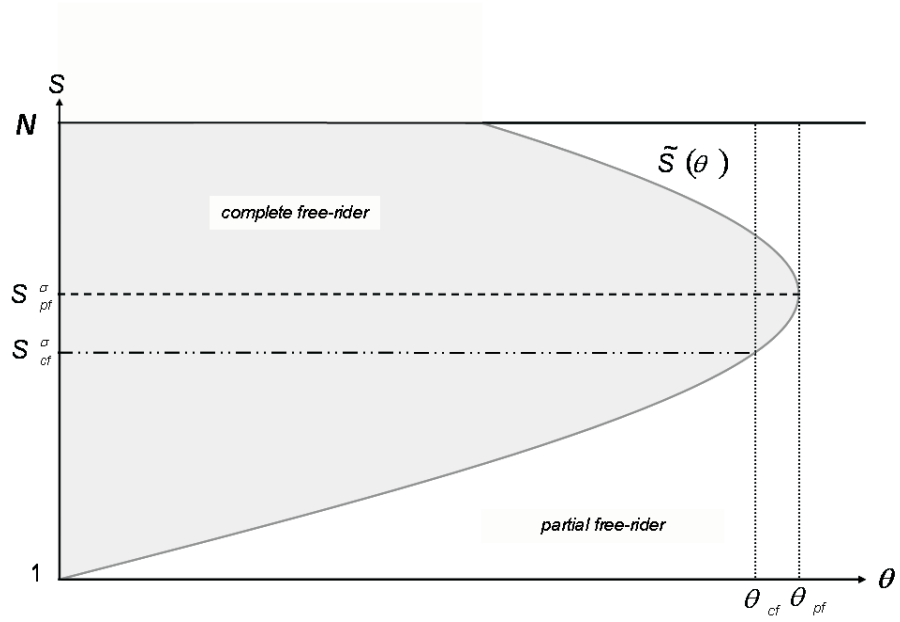

Figure 4: Free-riding and variance in abatements

We look at the two parts of σ (with partial and complete free-riding) separately. According to (7), in the case of partial free-riding the first-order condition

$$(N-S)\left(S^2-1\right)-S\left(S-1\right)^2=0, \tag{9}$$

characterizes a unique solution for the maximum variance at

$$S_{pf}^{\sigma} = \frac{N}{4} + \frac{N}{4}\sqrt{1 + \frac{8}{N}}. \tag{10}$$

If we look at the analogous condition in the case of complete free-riding, we obtain

$$(N - 2S)\, a_s^* + (N - S)2S\frac{\partial a_s^*}{\partial S} = 0. \tag{11}$$

After inserting (4) and the derivative $\partial a_s^*/\partial S$, we can rearrange terms so that the solution to our first-order condition becomes obvious:

$$[N - 2S]\, \frac{a_s^*}{\sqrt{N^4 + 8\theta N^2 S\,(N - S)}} = 0. \tag{12}$$

As the fraction in (12) is strictly positive, the term in brackets is decisive for the unique maximum of the variance, which requires

$$S_{cf}^{\sigma} = 0.5N. \tag{13}$$

Obviously, the maximum variance in the case of complete free-riding S_{cf}^{σ} falls short of that for partial free-riding S_{pf}^{σ}. We can distinguish three cases, which differ with respect to the extent of θ :

- According to figure 4, S_{cf}^{σ} maximizes the variance for small θ, i.e. $\theta \in [0\ ;\ \theta_{cf}]$, where $\theta_{cf} = 0.125N^2 - 0.5$.

- For θ exceeding $\theta_{pf} = 0.125N^2$ partial free-riding dominates, and the variance increases until S_{pf}^{σ} and diminishes thereafter.

- Between these two areas, for $\theta \in (\theta_{cf}\ ;\ \theta_{pf})$ the maximum variance will be obtained for $S_{cf}^{\sigma} < S^{\sigma} < S_{pf}^{\sigma}$, where S^{σ} lies on the lower part of the $\widetilde{S}(\theta)$ curve.

Consequently, $\sigma\,(S)$ is single-peaked in S regardless of the different scenarios in θ. Q.E.D.

References

Albin, C. (2003) Negotiating international cooperation: global public goods and fairness, *Review of International Studies*, Vol. 29 pp. 365-385.

Alesina, A. and G.M. Angeletos (2005) Fairness and redistribution, *American Economic Review*, Vol. 95 pp. 960-980.

Barrett, S. (1992) "International environmental agreements as games" in *Conflicts and Cooperation in Managing Environmental Resources* by Barrett, S., Pethig, R. (Ed.), Berlin: Springer.

Barrett, S. (1994) Self-Enforcing International Environmental Agreements, *Oxford Economic Papers* 46, 878-894.

Barrett, S. (2002). Consensus Treaties, *Journal of Institutional and Theoretical Economics*, Vol. 158 pp. 529-554.

Barrett, S. and R. Stavins (2003) Increasing Participation and Compliance in International Climate Change Agreements, *International Environmental Agreements - Politics, Law and Economics*, Vol. 3 pp. 349-376.

Bolton, G.E. and A. Ockenfels (2000) ERC: A Theory of Equity, Reciprocity, and Competition, *The American Economic Review*, Vol. 90 pp. 166 - 193.

Buchholz, W. and W. Peters (2005) A Rawlsian Approach to International Cooperation, *Kyklos*, Vol. 58 pp. 25-44.

Carraro, C. and D. Siniscalco (1993) Strategies for the International Protection of the Environment, *Journal of Public Economics*, Vol. 52 pp. 309-328.

d'Aspremont, C.A. and Gabszwicz (1986) "On the stability of collusion" in *New developments in the analysis of market structure* by Stiglitz, J.E., and G.F. Mathewson, Eds. New York: Macmillan.

Elster, J. (1989) Social Norms and Economic Theory, *The Journal of Economic Perspectives*, Vol. 3 pp. 99-117.

Engelmann, D. and M. Strobel (2004) Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments, *The American Economic Review*, Vol. 94 pp. 857-869.

Falk, A., E. Fehr and U. Fischbacher (2003) Reasons for Conflict: Lessons from Bargaining Experiments, *Journal of Institutional and Theoretical Economics*, Vol. 159 pp. 171-187.

Fehr and Schmidt (1999) A Theory of Fairness, Competition and Cooperation, *Quarterly Journal of Economics*, Vol. 114 pp. 817-68.

Finus, M. (2001) Game Theory and International Environmental Cooperation, Cheltenham and Northampton: Edward Elgar.

Finus, M. and B. Rundshagen (1998) Towards a positive theory of coalition formation and endogenous instrumental choice in global pollution control, *Public Choice*, Vol. 96 pp. 145-186.

Finus, M., v. Mouche, P. and Rundshagen, B. (2005) Uniqueness of Coalitional Equilibria, FEEM discussion paper No. 23.05, March 2, 2005.

Hoel, M. (1992) International Environmental Conventions: The Case of Uniform Reductions of Emissions, *Environmental and Resource Economics*, Vol. 2 pp. 141-159.

Hoel, M. and K. Schneider (1997) Incentives to Participate in an International Environmental Agreement, *Environmental and Resource Economics*, Vol. 9 pp. 153-170.

Jeppesen, T. and P. Andersen (1998) "Commitment and Fairness in Environmental Games" in *Game Theory and the Environment* by Hanley, N. and H. Folmer (Eds.), Cheltenham Northampton, Massachusetts: Edward Elgar.

Lange, A. and C. Vogt (2003) Cooperation in international environmental negotiations due to a preference for equity, *Journal of Public Economics*, Vol. 87 pp. 2049 - 2067.

Lange, A., C. Vogt and A. Ziegler (2007) On the importance of equity in international climate policy: an empirical analysis, *Energy Economics*, forthcoming.

Lindbeck, A. (1997) Incentives and Social Norms in Household Behavior, *The American Economic Review*, Vol. 87 pp. 370-377.

Murdoch, J. C. and T. Sandler (1997) The voluntary provision of a pure public good: The case of reduced CFC emissions and the Montreal Protocol, *Journal of Public Economics*, Vol. 63 pp. 331-349.

Postlewaite, A. (1998) The social basis of interdependent preferences, *European Economic Review*, Vol. 42 pp. 779-800.

Rabin, M. (1993) Incorporating Fairness into Game Theory and Economics, *The American Economic Review*, Vol. 85 pp. 1281-1302.

Rege, M. (2004) Social Norms and Private Provision of Public Goods, *Journal of Public Economic Theory*, Vol. 6 pp. 65-77.

Rawls, J. (1971) *A Theory of Justice*, Revised Ed. Oxford University Press, Oxford (1999).

Sobel, J. (2005) Interdependent Preferences and Reciprocity, *Journal of Economic Literature*, Vol. 43 pp. 392-436.

Stern, N. (2006) The Stern Review of the Economics of Climate Change, HM Treasury, London UK.

Victor, D.G. and L.A. Coben (2005) A Herd Mentality in the Design of International Environmental Agreements?, *Global Environmental Politics*, Vol. 5 pp. 24-57.

Wooders, M., Cartwright, E. and R. Selten (2007), Behavioral conformity in games with many players, *Games and Economic Behavior*, Vol. 57 pp. 347-360.