Reciprocal climate negotiators*

Karine Nyborg[†]

Abstract

Do reciprocal preferences matter for participation in international environmental agreements? Combining the two canonical models of Barrett (2003) and Rabin (1993), I find that 1) the empty coalition is always stable; 2) the grand coalition is stable if reciprocity preferences are sufficiently strong; 3) if a minority of countries are not reciprocal, a majority coalition can still be stable, if reciprocity is sufficiently strong and widespread; and 4) there is also a stable minority coalition, if costs are moderate and the number of countries is not too small. The latter coalition is weakly larger than the maximum stable coalition with standard preferences, but barely improves on welfare, and is characterized by negative sentiments. The analysis illustrates that although models of reciprocity tend to be complex, the context of a specific applied game may reduce this complexity.

JEL codes: F53, H87, Q54.

Keywords: International Environmental Agreements, Reciprocity, Coalitions.

**Acknowledgements:* I am grateful to Geir Asheim, Michael Hoel, Jon Hovi, Simon Dietz, Alessandro Tavoni, Natalia Montinari, Aart de Zeeuw, an anonymous reviewer, and seminar/conference participants at the Paris Environmental and Energy Economics Seminar, the Grantham Institute, 5th World Congress of Environmental and Resource Economists, University of Lund, the Graduate Institute in Geneva, and the Annual Meeting of the Norwegian Economists' Association for helpful discussion and comments. Thanks to the Research Council of Norway for funding through the NORKLIMA programme, and to the Beijer Institute of Ecological Economics for hosting me during an early stage of this work. The author is part of CREE (Oslo Centre for Research on Environmentally friendly Energy) and ESOP (Centre for the Study of Equality, Social Organization, and Performance).

[†]*Address:* Department of Economics, University of Oslo, P.O.Box 1095 Blindern, N-0317 Oslo, Norway. Phone: +47 22857283, e-mail: karine.nyborg@econ.uio.no.

Reciprocal climate negotiators

Abstract

Do reciprocal preferences matter for participation in international environmental agreements? Combining the two canonical models of Barrett (2003) and Rabin (1993), I find that 1) the empty coalition is always stable; 2) the grand coalition is stable if reciprocity preferences are sufficiently strong; 3) if a minority of countries are not reciprocal, a majority coalition can still be stable, if reciprocity is sufficiently strong and widespread; and 4) there is also a stable minority coalition, if costs are moderate and the number of countries is not too small. The latter coalition is weakly larger than the maximum stable coalition with standard preferences, but barely improves on welfare, and is characterized by negative sentiments. The analysis illustrates that although models of reciprocity tend to be complex, the context of a specific applied game may reduce this complexity.

JEL codes: F53, H87, Q54.

Keywords: International Environmental Agreements, Reciprocity, Coalitions.

1 Introduction

Reciprocal preferences – the desire to repay mean intentions by mean actions and kind intentions by kind actions – seem prevalent among individuals (Camerer 2003, Falk and Fischbacher 2003, 2006, Falk et al. 2008, Sobel 2005). But could such preferences matter for international climate negotiations?

In the present paper, I study possible impacts for participation in international environmental agreements *if* countries act according to reciprocal preferences. I do so using a simple coalition participation model from Barrett (2003), while specifying preferences along the lines of Rabin (1993).¹

The assumption that countries could have reciprocal preferences might, at first thought, seem far-fetched: even if individuals may be reciprocal, it does not necessarily follow that groups, countries or governments behave accordingly.² Nevertheless, if reciprocity is widespread in the general population, it may be prevalent among negotiators, government officials and heads of state as well; and if the median voter in a democracy has reciprocal preferences, a government aiming for reelection may want to act as if it does so too.

The view that trust and positive sentiments are important for the success of international climate negotiations seems, indeed, to be quite wide-spread. For years, international climate negotiations were plagued by an atmosphere of mistrust, which was often mentioned as one reason for the talks' failure to deliver strong results.³ When the Paris agreement was finally reached in December 2015, many observers noted that the atmosphere of the talks had improved substantially, claiming that this was promising for the success of the continued negotiations.⁴ And in the wake of the 2016 US Presidential elections, commentators were worrying that a withdrawal or effort reduction by the US could trigger negative reciprocal reactions by other countries.⁵

If climate negotiators were faced with a coordination game, trust may obviously play a role by making negotiators coordinate on a cooperative (uncooperative) Nash equilibrium. Unfortunately, the game facing international climate negotiators seems to be of the Prisoners' Dilemma kind: although curb-

¹ Reciprocal preferences should not be confused with reciprocal *strategies*, like tit-for-tat. To distinguish the two, Sobel (2005) uses the term 'intrinsic reciprocity' for what I call reciprocal preferences or just reciprocity.

² For example, if reciprocity is driven by sentiments like anger and gratitude, individuals and governments may not be equally affected by such emotions. Further, experimental evidence confirms that groups, or individuals acting on behalf of a group, may act differently than individuals (see Charness and Sutter 2012); I return to this literature in Section 4.

³ "Mr. President, this is the worst meeting I've been to since the eighth-grade student council" (Secretary of State Hillary Clinton to President Barack Obama at his arrival at the Copenhagen climate change summit in 2009, as quoted by Landler and Cooper (2010). "In such a poisonous atmosphere, no meaningful progress is possible." Zammit-Lucia (2013), commenting on the COP 19 in Warsaw.

⁴ "The 'Spirit of Paris' was still alive and well in Bonn, [sending] a very encouraging signal in preparation for negotiations to resume in November in Marrakech." (Outgoing UNFCCC Executive Secretary Christiana Figueres after UNFCCC talks in Bonn, May 2016, as quoted by Climate Policy Observer (2016).

⁵ "But a US reversal would likely also cause other major carbon emitting nations, like China or India, to at least slow their own efforts" (Krieg 2016).

ing global warming would benefit all countries, individual countries' incentives to abate their own greenhouse gas emissions are weak, regardless of others' efforts (Barrett 2003). Twenty-five years of theoretical studies of international environmental agreements have produced a whole array of largely pessimistic results, indicating that stable climate treaties are likely to have very few signatories and/or involve unambitious emission reduction goals (Barrett 1992, 1994, 2003, Carraro and Siniscalco 1993, Hoel 1992; for a survey, see Marrouch and Chaudhuri 2016).⁶

The Paris agreement is basically a set of voluntary pledges to reduce national emissions, lacking strong enforcement mechanisms (Barrett 2015). Consequently, viewing this as a break-through achievement seems to involve an implicit assumption that there does exist a cooperative Nash equilibrium, even if the game is a Prisoners' Dilemma.⁷ One rationale for such assertions could be the idea that governments have reciprocal preferences.

In his seminal paper, Rabin (1993) showed that if players' reciprocity preferences are sufficiently strong, a one-shot, two-player Prisoners' Dilemma game in material payoffs becomes a coordination game in utilities: if each expects the other to cooperate, each may be willing to cooperate herself, even if this does not pay off in material terms. Mutual defection, however, is also an equilibrium outcome in Rabin's model.

On a similar note, it seems plausible that governments' willingness to do their part to curb global warming could be increasing in others' willingness to do the same – and that if so, this may expand the set of stable international climate coalitions, compared to the standard preferences case. The aim of the present paper is to explore under what conditions, if any, this is indeed the case.

Although experimental and behavioral economists have provided strong indications of the prevalence of reciprocity among individuals (e.g., Fehr and Gächter 2000a, 2000b, Falk et al. 2003, Sobel 2005, Croson et al. 2006, Croson 2007, Falk et al. 2008), formal models of reciprocity are rarely seen in the environmental economics literature. One reason is that although various theoretical approaches have been developed (Rabin 1993, Levine 1998, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006, Cox et al. 2007), reciprocity models tend to be analytically complex. Standard game theory defines preferences over outcomes only. If players care intrinsically about others' intentions, one may need to define preferences over beliefs – a problem Rabin (1993) and others have resolved by applying psychological game theory (Geanakoplos et al., 1989). With a large number of players, the set of potentially relevant first- and higher-order beliefs to handle in a psychological game becomes very large. Moreover, research on psychological games is still in its infancy (Dufwenberg 2008). It turns out, however, that the framework used below leads to a

⁶The literature based on repeated games includes some more optimistic results as well, see Froyen and Hovi (2008), Heitzig et al. (2011), Kratzsch et al. (2012), Hovi et al. (2015).

⁷"This generation has taken vital steps to ensure that our children and grandchildren will see that we did our duty in securing the future of our planet. [...] What is so special about this deal is that it puts the onus on every country to play its part" (UK Prime Minister Cameron, 2015).

relatively simple and tractable analysis.

I am not the first to study the impact of social preferences for participation in international environmental agreements. Hoel and Schneider (1997) show that with a non-environmental cost of breaking agreements, larger coalitions become feasible. Van der Pol et al. (2012) find that 'community altruism', where signatories care about other signatories but not about non-signatories, increases treaty participation. Lange and Vogt (2003) show that inequality aversion as proposed by Bolton and Ockenfels (2000) can increase coalition size; if the abatement choice is discrete, they find that even the fully cooperative outcome can be feasible. Lange (2006) allows heterogeneity between countries, and finds that inequality aversion with respect to abatement targets across industrialized countries makes larger coalitions feasible.⁸

The reciprocal player is not unconditionally kind or caring. Unlike the altruist, a reciprocal player may wish to damage others, even at a cost. Unlike the inequality averse, the reciprocal player may not care about others' material payoffs at all. Rather, reciprocity is concerned with concepts like anger, reward and retaliation; and while it can promote socially beneficial cooperation, it can also at times be very destructive (Rabin 1993, Sobel 2005).

To my knowledge, the only previous paper studying international environmental agreements with reciprocal preferences is Hadjiyiannis et al. (2012). While I study a coalition participation game, Hadjiyiannis et al. (2012) study a compliance game; while I study an N -player game with discrete abatement choices, theirs is a game with only two players but continuous abatement. They find that reciprocity can facilitate cooperation, but only if the abatement level which is viewed as 'fair' by the other player is low; if countries' fairness view is more demanding, Hadjiyiannis et al. (2012) find that reciprocity is detrimental to cooperation.

For the sake of tractability, the analysis below is simplified in many respects. I stay close to the assumptions of the original, canonical models of Barrett (2003) and Rabin (1993), which may facilitate understanding and transparency. My hope, however, is that the highly stylized framework developed here can be used as a starting point for future analyses incorporating more plausible assumptions, such as heterogeneous countries, explicitly sequential notions of reciprocity, and other notions of "kindness" (or even disagreement on what is considered "kind").

Below, I show that with reciprocal preferences, three stable coalition sizes may be feasible: the empty coalition with zero participation; a minority coalition which is weakly larger than the maximal stable coalition size with standard preferences; and the grand coalition. For the latter to be stable, however, reciprocity preferences must be strong. The minority coalition relies on less strong assumptions on preferences, but barely improves on welfare, compared to the no compliance situation – just as in the standard preferences case. In addition, if some countries are not reciprocal, a majority coalition of reciprocal countries can be stable. This requires that the number of reciprocal countries is strictly more

⁸See also Grüning and Peters (2007), Kolstad (2013), Burger and Kolstad (2009), Hovi et al. (2015), Buchholz and Sandler (2016).

than a majority, and that these countries have strong reciprocal preferences.

2 The non-cooperative abatement game

I first establish the non-cooperative abatement game with reciprocal preferences, before turning to the issue of coalition formation in the next section.

Consider the simple one-shot global abatement game with $N \geq 2$ identical countries described by Barrett (2003, Ch.7). Each country i can choose either to abate ($q_i = 1$) or to pollute ($q_i = 0$). The material payoff for country i , π_i , consists of its environmental benefits from abatement (compared to some baseline) less its own abatements costs:

$$\pi_i = b(Q_{-i} + q_i) - cq_i \quad (1)$$

where $b > 0$ is the environmental benefits to the individual country due to one unit of abatement (by any country), $c > 0$ is a fixed per unit abatement cost,

and $Q_{-i} = \sum_{j=1}^N q_j - q_i = \sum_{j \neq i} q_j$ is the sum of abatement by countries other

than i . Moreover, $b < c$ and $bN > c$ (i.e. $N > c/b$). If countries' preferences coincide with their material payoffs as given by eq. (1), the above constitutes an N -player Prisoners' Dilemma game. Pollute ($q_i = 0$) is then a strictly dominant strategy with non-cooperative play; nevertheless, each country would have been better off if everyone had chosen Abate instead.⁹ The dilemma is illustrated in Figure 1: regardless of how many others abate, country i 's material payoff is always strictly higher if it pollutes itself.

INSERT FIGURE 1 ABOUT HERE

2.1 Defining reciprocity

Assume now that country i 's utility u_i depends on its material payoff π_i as well as reciprocity concerns R_i , where linear separability is assumed for simplicity, and $\alpha > 0$ is the weight placed on reciprocity concerns as compared to material payoff:

$$u_i = \pi_i + \alpha R_i \quad (2)$$

Below, "payoff" will refer to material payoff π_i , while "utility" or "preferences" refer to u_i .

Rabin (1993) assumed that reciprocity consists, essentially, of two parts: First, the negative (positive) emotion of being treated badly (nicely); second, the satisfaction of repaying by being mean (kind) in return.

⁹This must be the case since $\pi(1, Q_{-i}) = bQ_{-i} + b - c$ while $\pi(0, Q_{-i}) = bQ_{-i}$. Thus $\pi(1, Q_{-i}) - \pi(0, Q_{-i}) = b - c < 0$ (by assumption). Hence, for any Q_{-i} , $q_i = 1$ yields strictly lower material payoff for i than $q_i = 0$. If all play Abate, payoff for each country is $bN - c > 0$. If all play Pollute, payoff for each is 0.

Let f_{ij} denote i 's kindness towards j , and let $\tilde{f}_{ji}$ be i 's belief about j 's kindness towards i (for $i \neq j$). $f_{ij} < 0$ (> 0) means that i is being mean (kind). Extending Rabin's 2-player normal form game to allow for $N > 2$ players, I define the reciprocal part of utility as follows:

$$R_i = \frac{1}{N-1} \left[\sum_{j \neq i} \tilde{f}_{ji} + \sum_{j \neq i} f_{ij} \tilde{f}_{ji} \right] \quad (3)$$

where the sums are over all $j = 1, \dots, N$ for whom $j \neq i$. That is, each binary reciprocity relationship consists of the two parts discussed above (the pain of being treated badly, represented by $\tilde{f}_{ji}$, plus the pleasure of repayment, represented by $f_{ij} \tilde{f}_{ji}$).¹⁰ R_i is given by the average of each bilateral reciprocity relation. That is, I assume that a country cares about the average relationship between itself and each other country, while being unconcerned about the relations between others.

How to define "kindness" from i to j , however, is by no means obvious. Intuitively, whether an action is judged as "kind" might depend on a whole range of factors, such as the equality or fairness of the realized payoff distribution, i 's available choices, i 's sacrifice to secure j a reasonable outcome, the extent to which i disappoints j 's expectations, and whether the opponent appears to be an "altruistic type" (Levine 1998, Falk and Fischbacher 2006). Different individuals may also disagree on what is "kind" or appropriate, for example if social norms differ between societies or groups.

To keep the analysis simple and transparent, I have chosen to follow the intuition of Rabin's (1993) well-known reciprocity model: the more material payoff I am trying to secure to you, compared to what I *could* have secured to you, the kinder I am.

Let $\tilde{Q}_{-i}$ denote i 's belief about the number of others who will be abating. Let $\pi_j(q_i, \tilde{Q}_{-i}) = b(\tilde{Q}_{-i} + q_i) - cq_j$ denote the material payoff j will get as a function of i 's abatement choice, given i 's beliefs about others' strategies. Then, $\pi_j(q_i, \tilde{Q}_{-i})$ is the material payoff i is trying to secure to j , given i 's beliefs.

Let $\pi_{ij}^{\max}$ denote the maximum of $\pi_j(q_i, \tilde{Q}_{-i})$ with respect to q_i (the most i could secure to j for a given set of beliefs), and let $\pi_{ij}^{\min}$ denote the minimum of $\pi_j(q_i, \tilde{Q}_{-i})$ with respect to q_i (the least i could secure to j for given beliefs). Then, I define kindness from i to j as

$$f_{ij} = \frac{\pi_j(q_i, \tilde{Q}_{-i}) - \pi_{ij}^e}{\pi_{ij}^{\max} - \pi_{ij}^{\min}} \quad (4)$$

where

$$\pi_{ij}^e = \frac{1}{2}(\pi_{ij}^{\max} + \pi_{ij}^{\min}). \quad (5)$$

¹⁰In many games, only the second part is behaviorally relevant: even if you are pained by someone else's (believed) bad intentions, you may be left to take those intentions as given. In the present analysis, both parts matter: each may be behaviorally relevant when coalitions behave cooperatively.

Note that although I have suppressed this in the notation, $\pi_{ij}^{\max}$, $\pi_{ij}^{\min}$, and π_{ij}^e are all functions of the beliefs about others' strategies.¹¹¹²

With this specification, kindness depends on the payoff i tries to secure to j , compared to a fair or "equitable" payoff π_{ij}^e . The "equitable" payoff is the average of the least and most i could have secured to j (given i 's beliefs). Finally, still following Rabin, this is normalized by $(\pi_{ij}^{\max} - \pi_{ij}^{\min})$, the range of payoffs i could have secured to j . The latter can be interpreted to mean that kindness is measured relatively to i 's power vis-a-vis j .

If I choose the strategy that gives you the highest possible material payoff, given my beliefs about your and others' strategies, I am being maximally kind. If I choose the strategy that gives you the least possible material payoff, given my beliefs about yours and others' strategies, I am being minimally kind. With this specification, thus, what matters is what I try to secure to you compared with the options I think I have, not how much I sacrifice to do so. While another specification of "kindness" might obviously influence the results below, exploring this would require a separate analysis and is not pursued further here.

Like in Rabin (1993), a limitation of the current analysis is that the sequence of moves is not explicitly accounted for in the kindness measure. While a sequential reciprocity model would have been preferable, this would add substantial complexity. The sequential reciprocity models of Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) rely on psychological game theory, and while Dufwenberg and Kirchsteiger (2004) use a simplified version of Rabin's (1993) kindness measure, their simplifications are not inconsequential in the current context.¹³ Thus, in the interest of tractability and transparency, I will be keeping closely to Rabin's original framework. This means that "repaying" good or bad intentions will not refer to players' responses to others' observed past actions; rather, it refers to reactions to beliefs about others' strategies and/or intentions.

2.2 Reciprocity in the non-cooperative abatement game

In the non-cooperative abatement game presented above, there is only one way for i to influence j 's payoff: i 's choice to pollute or abate.

Taking others' behavior as given, i can secure no more to j than $\pi_{ij}^{\max} = b(Q_{-i} + 1) - cq_j$. Similarly, i can secure no less to j than $\pi_{ij}^{\min} = b(Q_{-i}) - cq_j$. Defining the equitable payoff π_{ij}^e as the average between these two, according to (5), yields

¹¹One may postulate that if $\pi_{ij}^{\max} = \pi_{ij}^{\min}$, then $f_{ij} = 0$. However, in the current analysis, $\pi_{ij}^{\max} > \pi_{ij}^{\min}$ will always hold.

¹²Rabin (1993) distinguishes between the *minimum Pareto efficient* payoff a player could have secured to another, and the *minimum* payoff a player could have secured to another. I am disregarding this distinction here.

¹³See Dufwenberg and Kirchsteiger (2004, Section 5). In particular, they omit the first sum in eq. (3) as well as Rabin's normalization of kindness in eq. (4), both of which would matter for conclusions in the games studied below.

$$\pi_{ij}^e = b(Q_{-i} + \frac{1}{2}) - cq_j. \quad (6)$$

Inserting this in (4), country i 's kindness towards j simplifies to

$$f_{ij} = q_i - \frac{1}{2}. \quad (7)$$

Since environmental quality is a pure public good, i is always equally kind or mean to everyone else. Moreover, i 's kindness does not depend on others' strategies.¹⁴

Thus, i 's belief about j 's kindness, $\tilde{f}_{ji}$, can quite naturally be assumed to depend, similarly, on j 's strategy only:

$$\tilde{f}_{ji} = q_j - \frac{1}{2}. \quad (8)$$

Inserting from eqs. (7) and (8), using $Q_{-i} = \sum_{j \neq i} q_j$ and that $f_{ij} = f_{ik}$ for all $j, k \neq i$, R_i can now be written as a function of own and others' strategies as follows:

$$R_i = (\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}) \quad (9)$$

This expression says that reciprocity concerns depend on the average kindness of others – judged by their abatement choices – and one's own abatement choice.

Inserting this into the utility function (2) defines reciprocal utility as a function of one's own and others' strategies:

$$u_i = u(q_i, Q_{-i}) = b(Q_{-i} + q_i) - cq_i + \alpha(\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}). \quad (10)$$

Segal and Sobel (2007) demonstrated that some psychological games can be reformulated by assuming that players have preferences over strategies, rather than beliefs, and they developed solution concepts for these games. Below, I will be using Segal and Sobel's (2007) definition of Nash equilibria in such games.

2.3 Nash equilibria

Let us first turn to abatement decisions in the case where all countries act non-cooperatively. Given others' strategies, $q_i = 1$ (abate) is (weakly) preferred to $q_i = 0$ (pollute) if $u(1, Q_{-i}) - u(0, Q_{-i}) \geq 0$, or

$$\frac{Q_{-i}}{(N-1)} \geq \frac{c-b}{\alpha} + \frac{1}{2}, \quad (11)$$

implying that the share of others who abate must be at least $(c-b)/\alpha + \frac{1}{2}$. This corresponds to strictly more than a majority (since $c > b$ and $\alpha > 0$).

¹⁴This is due to the assumed linearity of the environmental benefits (and costs being independent of others' efforts).

Define now $\hat{Q}_{-i}$ as the number of others abating that would make i exactly indifferent between abating and polluting:

$$\hat{Q}_{-i} = \left(\frac{c-b}{\alpha} + \frac{1}{2}\right)(N-1) \quad (12)$$

Whenever the number of other countries that abate exceeds $\hat{Q}_{-i}$, reciprocal concerns are sufficiently strong to outweigh the material incentive to free-ride. Note, however, that whenever $Q_{-i} < 1/2$, reciprocity *reinforces* the incentive to pollute as compared to the model with standard preferences.

Observe that $\hat{Q}_{-i}$ is strictly decreasing in α : the stronger the reciprocity preferences, the lower the number of abating others needed to make abatement individually preferable. Nevertheless, $\hat{Q}_{-i}$ is always strictly more than half of the others. If $\alpha < 2(c-b)$, there exists no $\hat{Q}_{-i}$ such that $\hat{Q}_{-i} \leq N-1$, and pollution is individually preferred regardless of others' abatement.

Following Segal and Sobel (2007), a Nash equilibrium is a strategy profile for which every agent i 's strategy maximizes his utility, given that i 's expectations about how his opponents will play the game are considered fixed. Let Q be the total number of countries that abate. The following proposition then demonstrates that although the symmetric one-shot climate game is a Prisoners' Dilemma in material payoffs, it can be a coordination game in utilities.

Proposition 1 *In the non-cooperative abatement game with identical, reciprocal countries, i) $Q = 0$ is a Nash equilibrium. ii) If $\alpha > 2(c-b)$, $Q = N$ is a Nash equilibrium. iii) If $\alpha > 2(c-b)$, the following situation is a Nash equilibrium: every country i uses a mixed strategy such that $q_i = 1$ with probability p and $q_i = 0$ with probability $1-p$, where $p = (c-b)/\alpha + 1/2$. In this situation, every country i is indifferent between abate and pollute. There is no pure strategy Nash equilibrium in which countries use different strategies.*

Proof. See the Appendix. ■

These results are illustrated in Figure 2. $\hat{Q}_{-i}$ represents a tipping point: if at least $\hat{Q}_{-i}$ others abate, the reciprocal benefits from abatement are sufficiently large to make it individually rational for every remaining country to abate too. Thus, if reciprocity is strong enough, abatement by every country is a Nash equilibrium. Conversely, if fewer than $\hat{Q}_{-i}$ others abate, the reciprocal benefits from abatement are too small to make abatement individually rational. Note also that in the no abatement equilibrium, utility is strictly negative: there are no costs and no environmental benefits, but countries are unhappy about others' unkindness.

The last sentence of Proposition 1 may be somewhat surprising. If everyone has the same preferences and still uses different pure strategies in Nash equilibrium, it must be because each is indifferent between the two pure strategies. In the present game, the latter situation is in fact not possible, because of the way utility depends on what *others* do (see the Appendix for details). Nevertheless, the Nash equilibrium described in Proposition 1 iii) is unstable in the sense that a small deviation from this situation would make one of the pure strategies strictly preferable for every country.

INSERT FIGURE 2 ABOUT HERE

Using eq. (10), it is easy to establish that the Nash equilibrium $Q = N$ is Pareto superior to $Q = 0$: If $Q = 0$, the utility of each country is

$$u_i = u(0, 0) = -\frac{1}{4}\alpha < 0 \quad (13)$$

while if $Q = N$, we have

$$u_i = u(1, N-1) = bN - c + \frac{3}{4}\alpha > 0. \quad (14)$$

2.4 Nash equilibria when some countries are not reciprocal

If only some countries are reciprocal, while the others care only about material self-interest π_i , $Q = N$ cannot be a Nash equilibrium. However, if reciprocity is strong enough and widespread enough, a high abatement Nash equilibrium, in which a majority of countries abate, still exists.

Proposition 2 *Assume that preferences are given by*

$$u_i = \pi_i + \alpha_i R_i$$

where $\alpha_i \in \{0, \alpha\}$. Let $A \leq N$ be the number of countries with $\alpha_i = \alpha$, while $N - A$ is the number of countries with $\alpha_i = 0$. Then, in the non-cooperative abatement game, i) $Q = 0$ is a Nash equilibrium. ii) If $\alpha \geq \frac{2(c-b)(N-1)}{(2A-N-1)}$, then the situation $Q = A$, where all countries with $\alpha_i = 0$ play Pollute and all countries with $\alpha_i = \alpha$ play Abate, is a Nash equilibrium. iii) If $\alpha \geq \frac{2(c-b)(N-1)}{(2A-N-1)}$, then the following situation is a Nash equilibrium: all countries with $\alpha_i = 0$ play the pure strategy Pollute. Every country i with $\alpha_i = \alpha$ plays a mixed strategy such that $q_i = 1$ with probability r and $q_i = 0$ with probability $1 - r$, where

$$r = \left(\frac{c-b}{\alpha} + \frac{1}{2} \right) \frac{(N-1)}{(A-1)}, \quad (15)$$

and each country i with $\alpha_i = \alpha$ is indifferent between Abate and Pollute. There is no pure strategy Nash equilibrium in which countries with $\alpha_i = \alpha$ use different strategies.

Proof. See the Appendix. ■

Note that $A > (N+1)/2$ is a necessary condition for Proposition 2 ii) to hold.¹⁵ Hence, $Q = A$ can never be Nash equilibrium unless a strict majority

¹⁵This follows since $\alpha \geq \frac{2(c-b)(N-1)}{(2A-N-1)}$, which cannot hold if the denominator on the right hand side is negative.

of countries are reciprocal. As demonstrated above, the tipping point $\hat{Q}_{-i}$ is always strictly larger than a majority; hence, if fifty percent of countries or less are reciprocal, no tipping point can be reached. In addition, each reciprocal country must have an even stronger preference for reciprocity than required for the full abatement equilibrium in Proposition 1.¹⁶

3 Coalition participation with reciprocity

Let me now turn to the treaty participation game extensively studied in the literature on international environmental agreements. Consider a three-stage game as follows (Barrett 2003, Ch. 7):

Stage 1: Every country i chooses whether or not to be part of the coalition;

Stage 2: Signatories decide their strategies collectively, aiming to maximize the coalition's total payoff;

Stage 3: Non-signatories choose their strategies non-cooperatively.

Assume that only one coalition forms, and that non-signatories do not form coalitions.

3.1 The standard preferences case

Consider first the standard preferences case where $u_i = \pi_i$ (see e.g. Barrett 2003, Wagner 2001). The game is solved by backward induction. In Stage 3, pollute is a strictly dominant strategy for non-cooperative players, so every non-signatory will pollute. Given this, the joint payoff of a coalition S of k countries is $\sum_{s \in S} \pi_s = k(bk - c)$ if they all abate, and 0 if they all pollute. Hence, in Stage 2, the coalition will prefer its members to abate if $k \geq c/b$. Given this, countries decide in Stage 1 whether to join S .

A coalition of size k is said to be stable if it satisfies the requirements of internal as well as external stability (d'Aspremont et al. 1983; see Wagner 2001, Marrouch and Chaudhuri 2015). Internal stability requires that when $k - 1$ others are members, and you are a member, it is better for you to stay than to leave. External stability requires that if k others are members, but you are not, it is better for you to stay outside.

Following Wagner (2001), let

$$\Pi_l(k) = b(kq_s + (N - k)q_n - cq_l) \text{ for } l = s, n \quad (16)$$

denote the material payoffs to signatories ($l = s$) and non-signatories ($l = n$), respectively, as a function of the number of signatories k . Then internal stability requires $\Pi_s(k) \geq \Pi_n(k - 1)$, while external stability requires $\Pi_n(k) \geq \Pi_s(k + 1)$.

If no other country takes part in the coalition, there is no coalition for country i to join, so $k = 0$, the empty "coalition", is stable. However, with standard preferences, there is another possibility as well. Let k^0 be the smallest integer such that $k^0 \geq c/b$. A coalition of size k^0 is stable (Barrett 2003,

¹⁶ $2(c - b)(N - 1)/(2A - N - 1) \geq 2(c - b)$, which holds with equality only if $A = N$.

Ch. 7.6): a country expecting $k^0 - 1$ others to join will join too, because its participation is required to make the other signatories abate (which they will do only if $c/b \leq k < c/b + 1$); for the same reason, a signatory of a coalition of size k^0 will stay. If the coalition is larger than k^0 , the individual signatory will prefer to leave, since it is not pivotal for the coalition's abatement.¹⁷

The implication is, unfortunately, that coalition formation can improve the sum of countries' payoffs only very slightly compared to the non-cooperative outcome of no abatement. This is illustrated in Figure 3.

INSERT FIGURE 3 ABOUT HERE

If $k^0 = c/b$, the coalition provides no net benefits at all to its members compared to the no abatement case, since their environmental benefits exactly outweigh their abatement costs. There is still a net benefit to non-members, who free-ride on the coalition's efforts. Signatories cannot gain from leaving because if one of them does, the coalition collapses (does not abate); hence, the relevant alternative for a signatory country is that no-one abates. If $k^0 > c/b$, the existence of the coalition secures a strictly positive gain to coalition members as well.

For example, assume that $N = 100$, $b = 2$ and $c = 3$. Then, $c/b = 3/2$, hence $k^0 = 2$. With a coalition of 2 countries, each signatory gets a payoff of 1, while each non-signatory gets a payoff of 4. Had all 100 countries abated, each of them would instead have received a payoff of 197.

3.2 Defining reciprocity in the three-stage game

Assume now that every country i has reciprocal preferences as given by eq. (2) above, where $\alpha > 0$, and where R_i is given by eqs. (3) - (5). Suppose also that a coalition S , if formed, collectively maximizes the sum of its members' utilities

$$\sum_{s \in S} u_s = \sum_{s \in S} (\pi_s + \alpha R_s) \quad (17)$$

with respect to q_s for every signatory $s \in S$.

A strategy for country i now consists of a plan, for any given beliefs about others' strategies and/or intentions, of whether to join the coalition in Stage 1 and, if a non-signatory, whether to abate or pollute in Stage 3. If i 's strategy implies joining, i 's abatement is determined by the coalition's policy in Stage 2.

Recall that "kindness" from i to j depends on the material payoff j actually gets, as well as the payoff i *could have* secured to j . In the coalition formation game, a country may influence others' payoffs via two channels: first, through its own abatement choice (which is implied by the participation choice for signatories), and second, through its influence on *others'* abatement, since a country's participation may affect the entire coalition's abatement choice.

¹⁷If k is larger than $c/b + 1$, signatory s faces the same freeriding incentive as in the non-cooperative game: the coalition abates regardless of whether s stays, and s 's abatement cost c is not outweighed by the corresponding benefits to s , b .

Since countries' possible influences on each other are now different than in the non-cooperative game, the reciprocity part of utility may need to be specified differently too. To check this, we need to go back to the basic relationships (2) – (5).

If, given i 's beliefs about others' strategies and/or intentions, i 's participation is decisive for the coalition's choice of whether or not to abate, country i is pivotal for the coalition's abatement. A country that is not pivotal in this sense has exactly the same options to influence others as in the non-cooperative game; consequently, kindness can be expressed according to eq. (7) as before in such cases. If i is pivotal, on the other hand, its power is considerably larger. It turns out, however, that even in such cases, the kindness function (4) can be simplified into precisely the same expression as before: $f_{ij} = q_i - \frac{1}{2}$. While a more detailed demonstration of this claim has been relegated to the Appendix, the intuitive reason is the following: Like Rabin (1993), I have used a normalization of the kindness measure making kindness relative to a player's power. In the coalition formation game, a pivotal player's choice has a larger impact on others than a non-pivotal player, but the impact of this on kindness is cancelled out by this normalization.¹⁸ Thus, the utility function (10) derived above still applies in all cases.

3.3 Stable coalitions with reciprocity

Let us now turn to the issue of stable coalitions with reciprocal players.¹⁹

For the analysis below, it is helpful to note that (10) can be rewritten as

$$u_i = [b + \frac{\alpha(2q_i + 1)}{2(N-1)}]Q_{-i} - \frac{\alpha + 2(c-b)}{2}q_i - \frac{\alpha}{4}.$$

As before, let k be the number of signatories. Let q_n be the abatement of a representative non-signatory, let q_s be the coalition's abatement policy (the abatement of a representative signatory), and let $U_s(q_n, q_s, k)$ denote the utility of a signatory country:

$$U_s(q_n, q_s, k) = [b + \frac{\alpha(2q_s + 1)}{2(N-1)}]Q_{-s} - \frac{\alpha + 2(c-b)}{2}q_s - \frac{\alpha}{4} \quad (18)$$

where $Q_{-s} = Q_n + (k-1)q_s$ and $Q_n = (N-k)q_n$. Similarly, the utility of a non-signatory country h can be written

$$U_h(q_h, q_n, q_s, k) = [b + \frac{\alpha(2q_h + 1)}{2(N-1)}]Q_{-h} - \frac{\alpha + 2(c-b)}{2}q_h - \frac{\alpha}{4} \quad (19)$$

where $Q_{-h} = kq_s + (N-k-1)q_n$ for $k = 2, \dots, N-1$, and $Q_{-h} \equiv 0$ for $k = N$. Let $U_n(q_n, q_s, k) = U_h(q_n, q_n, q_s, k)$ denote the utility of a representative non-signatory country.

¹⁸With a different specification of kindness, results might of course change; exploring this would require a separate analysis.

¹⁹I am grateful to an anonymous reviewer for highly helpful suggestions for improving the analysis of this section.

In the following, I look for coalitions which are externally and internally stable in the following sense:

Definition 1. A coalition of size k is internally stable if $U_s(q_n, q_s, k) \geq U_n(q_n, q_s, k-1)$.

Definition 2. A coalition of size k is externally stable if $U_n(q_n, q_s, k) \geq U_s(q_n, q_s, k+1)$.

The empty "coalition", in which no country is a signatory and all non-signatories countries pollute, is stable: if no-one joins, there are no signatories and everyone plays non-cooperatively in Stage 3, and we can apply Proposition 1, part i.

Proposition 3 *The no-cooperation situation $k = 0$, in which no country is a signatory and all countries pollute, is stable.*

Proof. See the Appendix. ■

If the preference for reciprocity is sufficiently strong, the grand coalition is also stable. This may not be surprising, given that this is a possible Nash equilibrium even in the non-cooperative game.

Proposition 4 *Assume $N > 2$, $\alpha > 2(c-b)$ and $c \leq b(N-1)$. Then, the grand coalition (the coalition abates, and $k = N$) is stable.*

Proof. See the Appendix. ■

Consequently, extremely cooperative as well as extremely uncooperative outcomes can be stable. If sufficiently many others take part in an abating coalition, a sufficiently strongly reciprocal player prefers to do so too. If others do not cooperate, a reciprocal player does not want to cooperate either.

Moreover, as established by the proposition below, there may exist a third stable coalition size k^1 . This resembles the small, stable coalition size k^0 from the payoff-maximizing countries case, but k^1 is weakly larger than k^0 . Nevertheless, k^1 is always a minority coalition – a small, but not necessarily minimal "coalition of the willing".

Proposition 5 *Assume that $N > 13$ and that $c \leq b(N+2)/3$. Let k^1 be defined as the smallest integer such that $k^1 \geq \underline{k}$, where $\underline{k} = \frac{2c(N-1)+\alpha(N+2)}{2b(N-1)+3\alpha}$. Then, there exists an externally and internally stable coalition size of k^1 countries, such that $\frac{N-1}{2} > k^1 \geq k^0 \geq c/b$, where the coalition abates while non-signatories pollute.*

Proof. See the Appendix. ■

That is, if the number of countries is not too small, and costs are not too high compared to environmental benefits, we know that a stable minority coalition of size k^1 exists, regardless of the strength of reciprocity concerns.²⁰ The number k^1 itself is weakly increasing in α (with an upper boundary at $\frac{N+2}{3} + 1$); if

²⁰The condition $N > 13$ in Proposition 5 is sufficient but not necessary. If it did not hold, a coalition of size k^1 may exist, but would not necessarily be stable. Stability requires that k^1 is weakly larger than $\underline{k}$, smaller than $\underline{k} + 1$, and smaller than $\hat{Q}_{-i}$ (see Figure 4). For sufficiently large N this is always feasible.

α becomes sufficiently small, k^1 coincides with k^0 . Since k^1 coincides with k^0 when α is sufficiently small, Proposition 5 can be interpreted as saying that the standard model is robust to small changes in preferences.

INSERT FIGURE 4 ABOUT HERE

While the details are given in the proof in the Appendix, the main intuition is illustrated in Figure 4. Note first that $\underline{k}$ corresponds to c/b in the standard preferences case in the following sense: $\underline{k}$ is the lowest k for which a country is indifferent between *everyone*, including itself, polluting, and being a signatory in an abating coalition. k^1 is the lowest integer weakly above $\underline{k}$. It is easily seen that when $k = \underline{k}$, non-signatories are better off than signatories. However, like in the standard preferences case, a signatory considering to leave cannot take others' abatement as given: if it leaves, the coalition size becomes too small for the coalition to prefer abatement.

Since k^1 is a minority coalition, others' behavior is, on average, mean. When $k = k^1$, no-one really wants to abate: not only is it materially unprofitable from a non-cooperative point of view, but each country would also like to punish others for their unkind behavior. So how can k^1 be *larger* than k^0 and still be stable?

To understand this, consider the situation of a signatory when $k = k^1$. There is, after all, a small group of $k^1 - 1$ others who are behaving nicely to you. They are too few to make you want to be nice yourself. Still, they represent a small island of kindness in a mean world. If you stop being nice to them, they will stop being nice to you. The island of kindness will disappear; there will be only meanness left in the world.

3.4 Coalition participation if some countries are not reciprocal

If only some countries are reciprocal, the grand coalition cannot be stable. Stable, abating coalitions of strictly positive size may still exist, but the situation is now more complex, because it matters *who* is a signatory.

If the conditions for Proposition 2 hold, i.e. if A and α are sufficiently large, we know that in the non-cooperative game, there is a Nash equilibrium in which every reciprocal country abates, while every payoff-maximizing country pollutes. Under the same assumptions, there is a corresponding stable majority coalition $k = A$ in the three-stage game.

Proposition 6 *Assume that preferences are given by*

$$u_i = \pi_i + \alpha_i R_i$$

where $\alpha_i \in \{0, \alpha\}$. Let $A \leq N$ be the number of countries with $\alpha_i = \alpha$, while $N - A$ is the number of countries with $\alpha_i = 0$. Then, if $\alpha \geq \frac{2(c-b)(N-1)}{(2A-N-1)}$,

there is stable coalition of size A such that all countries with $\alpha_i = \alpha$ are signatories; all other countries are non-signatories; and the coalition abates, while non-signatories pollute.

Proof. See the Appendix. ■

If reciprocity preferences are too weak and/or the number of reciprocal countries is too small, no such majority coalition can be stable.

Stable minority coalitions may exist as well. In particular, if the assumptions for Proposition 5 hold, if $A \geq k^1$ and all signatories are reciprocal, then the arguments given in the proof of Proposition 5 imply that there is a stable minority coalition size $k^1 \geq k^0$ such that the coalition abates while non-signatories pollute.

4 Discussion

The results that a grand or majority coalition may be stable are, at the outset, encouraging: in a hypothetical world corresponding to the model assumptions, a 'Paris agreement', in which countries simply pledged to abate voluntarily with no external enforcement, could conceivably be successful.

However, this conclusion should be interpreted with a healthy dose of skepticism. First, the model is obviously highly stylized. Introducing more realistic assumptions, like continuous abatement and heterogeneous country size, would require separate analyses. In particular, assumptions implying that a country's payoff from own abatement would depend on others' abatement (e.g., due to non-linear environmental damages) may considerably complicate the analysis, since the role of beliefs would then become more prominent. Generally, any change in the structure and determinants of countries' realized and potential payoffs would affect the endogenously derived specification of the utility function (eq. 10 in the current model).

Second, even within the context of the simple model above, strictly more than half of the countries must be reciprocal in order for a grand or majority coalition to be stable, and reciprocity must be quite strong (Proposition 4).

With this in mind, a glance at the experimental and behavioral economics literature is not exclusively encouraging. While groups sometimes manage to cooperate in Prisoners' Dilemma-like situations, both in the lab and in the field, it is also true that cooperation often fails (Ostrom 1990, Camerer 2003, Zelmer 2003). In repeated public good games with exogenous group formation, cooperation rarely prevails over time (Ledyard 1995).²¹ Several recent experimental

²¹In lab experiments, a substantial share of subjects are conditional cooperators, whose willingness to contribute to the public good is increasing in the contributions they expect from others (Fischbacher et al. 2001, Fischbacher and Gächter 2006, 2010, Croson et al. 2006). Since such behavior is consistent with reciprocal preferences, it is worth noting that the share of conditional cooperators in experimental studies often, but not always, exceed fifty percent, varying somewhat between countries: Martinsson et al. (2013) found the share of conditional cooperators to be 63% in Colombia and 50% in Vietnam; numbers for other countries include 50% in Switzerland (Fischbacher et al. 2001); Denmark 70 % (Thöni et al.

studies lend support to the rather bleak predictions from the theoretical literature on international environmental agreements (Barrett and Dannenberg 2012, 2016a, b).

Third, even if reciprocal preferences may seem relatively widespread among individuals, this does not necessarily imply that countries – or individuals negotiating on countries’ behalf – act according to such preferences.

While some evidence on actual negotiators’ preferences is available (Lange et al. 2010, Dannenberg et al. 2010, Hafner-Burton et al., 2014), none of this is, to my knowledge, concerned with reciprocity. Economic experiments confirm, however, that groups, and individuals acting as group representatives, often behave differently from single individuals (Charness and Sutter 2012). Compared to individuals, groups tend to be more competitive and better at maximizing profits (Kocher and Sutter 2007), more cognitively sophisticated and successful in coordinating on efficient equilibria (Charness and Sutter 2012), and less spiteful (Balafoutas et al., 2014). Groups have been found to be less generous (Hamman et al. 2010) and less trusting (Song 2008, Kugler et al. 2007), although results on the latter are somewhat mixed (Kocher and Sutter 2007). For linear public good games, whose incentive structure is similar to the non-cooperative game discussed above, groups have been found to contribute equally or even *more* than individuals (Hauge and Røgeberg 2015, Huber et al. 2014, Auerswald et al. 2013).

Fourth, even if reciprocal preferences were sufficiently strong and widespread, this may not be enough: if the grand or a majority coalition is stable, so is the situation with no or very small (k^1) participation. Negative expectations may be self-fulfilling: if countries expect others to pollute, cooperation may not arise in the first place, confirming the low expectations. Hence, countries may be stuck in a Pareto inferior state, even if the grand coalition would have been stable.

The latter argument can be questioned, however.²² Although the non-cooperative game discussed above is a Prisoners’ Dilemma in material payoffs, it is a coordination game in utilities. When parties in coordination games can communicate, they might simply agree to coordinate on a Pareto efficient Nash equilibrium. Similarly, in the three-stage game, communicating parties might just agree to discard any Pareto inferior solutions, establishing immediately the grand or majority coalition.

During international climate negotiations, there is obviously extensive communication. In other words: if no grand or majority coalition was observed in a world similar to the one described in the model, wouldn’t that be a strong indication that a stable grand or majority coalition was not, in fact, feasible?

To address such considerations formally, one would need a different model; the essentially static model above is silent on issues of equilibrium selection. One may envisage, for example, a more comprehensive framework in which not

2012); Russia 56% (Herrmann and Thöni 2009); USA 81%, Switzerland 44%, and Japan 42% (Kocher et al. 2008).

²²I am grateful to a reviewer for bringing up this discussion.

only abatement and participation choices, but also the choice to contribute efforts in attempts to change players' coordination, were modelled explicitly. One important issue would then be the possibility of path dependency: if realization of a Pareto inferior solution had already triggered negative sentiments, would these sentiments reduce reciprocal players' willingness to engage in coordination improvement initiatives?²³ ²⁴ Without a formal model, however, these arguments remain speculative.

Finally, note that my results have much in common with those of Lange and Vogt (2003), who assume that countries have inequality aversion. In simple, symmetric games, inequality aversion and reciprocal preferences often yield similar predictions; after all, equal sharing is often viewed as most fair or "kind" in such games. Inequality aversion is, in fact, often used as a simplified representation of reciprocity, due to the complexity of reciprocity models (Fehr and Schmidt 1999, Blanco et al. 2011) – even if reciprocal behavior cannot in general be explained by aversion to inequality alone (Falk et al. 2003, 2008, Blanco et al. 2011). In more complex contexts, however, there is little reason to expect predictions from the two types of models to coincide. As mentioned above, it is my hope that the present analysis can be used as a starting point for further analyses employing more realistic assumptions, such as heterogeneous countries. In such models, inequality aversion and reciprocity are less likely to produce similar conclusions.

While Lange and Vogt (2003) find that with discrete abatement and sufficiently strong inequity aversion, the grand coalition is stable, they also find that countries behave just like in the standard preferences case when the abatement choice is continuous. Studying the implication of continuous abatement in the present model would thus be of interest. This would require a separate analysis, however; among other things, the kindness measure would need to be revised.

5 Conclusions

In the debate on international climate policies, good will, fairness, and intentions to do "one's part" are recurrent themes. Thus it seems relevant to investigate whether reciprocal preferences may affect participation in international environmental agreements. The present analysis represents a first attempt to explore this issue. While the model is highly stylized, it is my hope that it can serve as a helpful starting point for those wanting to explore the topic further, introducing more realistic and complex assumptions.

If only a minority of other countries abate, a reciprocal country is *less* willing to abate than a country with standard preferences. Hence, I find that with

²³Casual observation from fighting with one's spouse may seem to support to this idea: in the presence of self-reinforcing negative sentiments, being allowed to communicate is not always sufficient to make parties reach out to each other and agree to coordinate on a Pareto superior equilibrium.

²⁴Note also that if preferences were private information (a complication not considered in my model), and if countries have a strategic interest in misrepresenting them, countries might not even know (or trust) whether a Pareto superior equilibrium (or stable coalition) existed.

reciprocal preferences, the situation in which everyone pollutes is stable in the three-stage climate coalition formation game.

If reciprocal preferences are sufficiently strong and widespread, however, the grand coalition is stable too. Moreover, even if a minority of countries are not reciprocal, a majority coalition can be stable – but this requires that the reciprocal majority is sufficiently large, and that these countries have relatively strong reciprocity preferences.

In addition, if costs are not too large and the number of countries is not too small, there exists a stable minority coalition. This coalition size is weakly larger than the maximal coalition in the standard preferences case. This situation, however, is characterized by mutually negative sentiments, and barely improves on welfare compared to no abatement. A signatory stays in this coalition because if it leaves, other signatories will prefer polluting. Compared to the case if the signatory stays, this will reduce environmental benefits - and increase the pain of being faced with a mean world.

References

- [1] Auerswald, H., C. Schmidt, M. Thum, G. Torsvik (2013): Teams Punish Less, CESifo Working Paper Series 4406, CESifo Group, Munich.
- [2] Balafoutas, L., R. Kerschbamer, Martin Kocher, M Sutter (2014): Revealed Distributional Preferences: Individuals vs. Teams, *Journal of Economic Behavior and Organization* 108, 319-330.
- [3] Barrett, S. (1992): "International Environmental Agreements as Games", in R. Pethig (Ed.): *Conflict and Cooperation in Managing Environmental Resources*, Berlin: Springer, 11-37.
- [4] Barrett, S. (1994): Self-Enforcing International Environmental Agreements, *Oxford Economic Papers* 46, 878-894.
- [5] Barrett, S. (2003), *Environment and Statecraft: The Strategy of Environmental Treaty-Making*, Oxford: Oxford University Press.
- [6] Barrett, S. (2015): Global Consensus on Climate Change Is a Good Start, *Yale Global Online* 15. December 2015 (<http://yaleglobal.yale.edu/content/global-consensus-climate-change-good-start>, accessed 30.05.16).
- [7] Barrett, S., and A. Dannenberg (2012): Climate negotiations under scientific uncertainty, *Proceedings of the National Academy of Sciences of the United States of America* 109 (43), 17372-17376.
- [8] Barrett, S., and A. Dannenberg (2016a): Tipping versus Cooperating to Supply a Public Good, *Journal of the European Economics Association*, forthcoming.

- [9] Barrett, S., and A. Dannenberg (2016b): An Experimental Investigation into 'Pledge and Review' in Climate Negotiations, *Climatic Change* (first online, DOI: 10.1007/s10584-016-1711-4).
- [10] Blanco, M., D. Engelmann, H.T. Normann (2011): A within-subject analysis of other-regarding preferences, *Games and Economic Behavior* 72 (2), 321–338.
- [11] Bolton, G.E, and A. Ockenfels (2000): ERC – A theory of equity, reciprocity, and competition, *American Economic Review* 90(1), 166-93.
- [12] Buchholz, W., and T. Sandler (2016): Successful Leadership in Global Public Good Provision: Incorporating Behavioural Approaches, forthcoming (online Jan. 30, 2016), *Environmental and Resource Economics*.
- [13] Burger, N.E., and C.D. Kolstad (2009): Voluntary Public Goods Provision, Coalition Formation, and Uncertainty, NBER Working Papers 15543, National Bureau of Economic Research.
- [14] Camerer, C. (2003): *Behavioral Game Theory. Experiments in strategic interaction*, Princeton University Press/Russell Sage Foundation.
- [15] Cameron, D. (2015): Historic global deal on climate change. Online post on LinkedIn (<https://www.linkedin.com/pulse/historic-global-deal-climate-change-david-cameron>), 12.12.15, accessed 30.05.16.
- [16] Carraro, C., and D. Siniscalco (1993): Strategies for the International Protection of the Environment, *Journal of Public Economics* 52, 309-328.
- [17] Charness, G. and M. Sutter (2012): Groups Make Better Self-Interested Decisions, *Journal of Economic Perspectives* 26, 157-176.
- [18] Climate Policy Observer (2016): UNFCCC talks resume in Bonn under the “Paris Spirit”. 28th May 2016 (<http://climateobserver.org/9986-2/>, accessed 01.06.16)
- [19] Cox, J.C., Friedman, D., and S. Gjerstad (2007): A tractable model of reciprocity and fairness, *Games and Economic Behavior* 59(1), 17-45.
- [20] Croson, R., 2007. Theories of commitment, altruism and reciprocity: evidence from linear public goods games. *Economic Inquiry* 45, 199–216.
- [21] Croson, R., Fatas, E., Neugebauer, T., 2006. Reciprocity, matching and conditional cooperation in two public goods games. *Economics Letters* 87, 95–101.
- [22] d’Aspremont, C.A., J. Jacquemin, J. Gabszewis, J.A. Weymark (1983): On the stability of collusive price leadership, *Canadian Journal of Economics* 16, 17-25.

- [23] Dannenberg, A., B. Sturm, and C. Vogt (2010), Do Equity Preferences Matter for Climate Negotiators? An Experimental Investigation, *Environmental and Resource Economics* 47, 91-109.
- [24] Dufwenberg, M. (2008): "Psychological games". In S.N. Durlauf and L.E. Blume (Eds.): *The New Palgrave Dictionary of Economics Online*, Second Edition, Palgrave Macmillan, 28 January 2014, doi:10.1057/9780230226203.1358.
- [25] Dufwenberg, M., and G. Kirchsteiger (2004): A Theory of Sequential Reciprocity, *Games and Economic Behavior* 47(2), 268-98.
- [26] Falk, A., E. Fehr, U. Fischbacher (2003): On the Nature of Fair Behavior, *Economic Inquiry* 41(1), 20-26.
- [27] Falk, A., E. Fehr, U. Fischbacher (2008): Testing theories of fairness – Intentions matter, *Games and Economic Behavior* 62 (1), 287–303.
- [28] Falk, A., and U. Fischbacher (2006): A Theory of Reciprocity, *Games and Economic Behavior* 54, 293-315.
- [29] Fehr, E. and U. Fischbacher (2002): Why Social Preferences Matter - the Impact of Non-Selfish Motives on Competition, Cooperation and Incentives, *Economic Journal* 112, C1-C33.
- [30] Fehr, E., and S. Gächter (2000a): Fairness and Retaliation: The Economics of Reciprocity, *Journal of Economic Perspectives* 14(3), 159-181.
- [31] Fehr, E., and S. Gächter (2000b): Cooperation and punishment in public goods experiments, *American Economic Review* 90 (4), pp. 980–994.
- [32] Fehr, E., and S. Gächter (2002): Altruistic Punishment in Humans, *Nature* 415, 137-140.
- [33] Fehr, E., and K. Schmidt (1999): A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114, 817-868.
- [34] Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71, 397–404.
- [35] Fischbacher, U., Gächter, S., 2006. Heterogeneous social preferences and the dynamics of free riding in public goods. *IZA Discussion Papers* 2011.
- [36] Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of freeriding in public goods. *American Economic Review* 100, 541–556.
- [37] Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71, 397–404.

- [38] Froyen, C.B., and J. Hovi (2008): A climate agreement with full participation, *Economics Letters* 99, 317–319.
- [39] Geanakoplos, J., D. Pearce, E. Stacchetti (1989): Psychological games and sequential rationality, *Games and Economic Behavior* 1, 60-79.
- [40] Grüning, C., and W. Peters (2007): Can Justice and Fairness Enlarge the Size of International Environmental Agreements? European University Viadrina (http://www.wiwi.europa.uni.de/de/lehrstuhl/fine/fiwi/team/gruening/GrueningPeters_04_07.pdf).
- [41] Hadjiyiannis, C., D. Iris, C. Tabakis (2012): International Environmental Cooperation under Fairness and Reciprocity, the B.E. Journal of Economic Analysis & Policy (Topics), 12(1), Article 33.
- [42] Hafner-Burton, E., B.L. LeVeck, D.G. Victor, and J.H. Fowler (2014): Decision Maker Preferences for International Legal Cooperation, *International Organization* 68, 845-876.
- [43] Hauge, K. E. and O. Røgeberg (2015): Representing Others in a Public Good Game, *Games* 6, 381-393.
- [44] Heitzig, J., K. Lessmann and Y. Zou (2011): Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games, *PNAS* 108 (38), 15739–15744.
- [45] Herrmann, B., and C. Thöni (2009): Measuring conditional cooperation: A replication study in Russia. *Experimental Economics*, 12, 87–92.
- [46] Hoel, M.O. (1992): International Environment Conventions: The Case of Uniform Reductions of Emissions. *Environmental and Resource Economics* 2, 141-159.
- [47] Hoel, M.O., and K. Schneider (1997): Incentives to Participate in an International Environmental Agreement, *Environmental and Resource Economics* 9(2), 153–170.
- [48] Hovi, J., H. Ward, F. Grundig (2015): Hope or Despair? Formal Models of Climate Cooperation, *Environmental and Resource Economics* 62 (4), 665–688.
- [49] Huber, S., J. Model, S. Städter (2013): Team Behavior in Public Goods Games with Ostracism, Mimeo, University of Regensburg.
- [50] Kocher, M.G., T. Cherry, S. Kroll, R.J. Netzer, M. Sutter (2008): Conditional cooperation on three continents, *Economics Letters* 101, 175–178.
- [51] Kocher, M., and M. Sutter (2007): Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments, *Empirica* 34(1), 63-88.

- [52] Kolstad, C.K. (2013): International Environmental Agreements with Other-Regarding Preferences, unpublished paper, Stanford University.
- [53] Kratzsch, U., G. Sieg and U. Stegemann (2012): An international agreement with full participation to tackle the stock of greenhouse gases, *Economics Letters* 115 (3), 473–476.
- [54] Krieg, Gregory (2016): Climate change worries escalate as Trump elevates top deniers, *CNN Politics* 16.11.16 (<http://edition.cnn.com/2016/11/16/politics/climate-change-donald-trump/>, accessed 18.11.16).
- [55] Kugler, T., G. Bornstein, M.G. Kocher, M. Sutter (2007): Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy, *Journal of Economic Psychology* 28(6), 646–657.
- [56] Landler, M., and H. Cooper (2010): After a bitter campaign, forging an alliance. *New York Times* March 18, 2010, accessed 10.04.14 at http://www.nytimes.com/2010/03/19/us/politics/19policy.html?pagewanted=all&_r=0.
- [57] Lange, A. (2006): The Impact of Equity-Preferences on the Stability of International Environmental Agreements, *Environmental and Resource Economics* 34, 247–267.
- [58] Lange, A., A. Löschel, C. Vogt and A. Ziegler (2010): On the Self-interested Use of Equity in International Climate Negotiations, *European Economic Review* 54, 2010, 359–375.
- [59] Lange, A., and C. Vogt (2003): Cooperation in International Environmental Negotiations due to a Preference for Equity, *Journal of Public Economics* 87, 2049–2067.
- [60] Ledyard, J.O. (1995): Public Goods: a Survey of Experimental Research. In: Kagel, J.H., Roth, A.E. (Eds.): *The Handbook of Experimental Economics*. Princeton University Press, Princeton, New Jersey, pp. 111–194.
- [61] Levine, D.K. (1998). Modeling Altruism and Spitefulness in Experiments, *Review of Economic Dynamics* 1, 593–622.
- [62] Marrouch, W., and A.R. Chaudhuri (2016): International Environmental Agreements: Doomed to Fail or Destined to Succeed? A Review of the Literature, *International Review of Environmental and Resource Economics* 9 (3–4), 245–319.
- [63] Martinsson, P., N. Pham-Khanh, C. Villegas-Palacio (2013): Conditional cooperation and disclosure in developing countries, *Journal of Economic Psychology* 34, 148–155.
- [64] Ostrom, E. (1990): *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge: Cambridge University Press.

- [65] Rabin, M. (1993): Incorporating Fairness into Game Theory and Economics, *American Economic Review* 83, 1281-1302.
- [66] Segal, U., and J. Sobel (2007): Tit for tat: Foundations of preferences for reciprocity in strategic settings, *Journal of Economic Theory* 136, 197-216.
- [67] Sobel, J. (2005): Interdependent Preferences and Reciprocity, *Journal of Economic Literature* 43, 392-436.
- [68] Song, F. (2008): Trust and reciprocity behavior and behavioral forecasts: Individuals versus group-representatives, *Games and Economic Behavior* 62, 675–696.
- [69] Thöni, C., J. -R. Tyran, and E. Wengström (2012): Microfoundations of social capital, *Journal of Public Economics* 96(7-8), 635-643.
- [70] Van der Pol, T., H.-P. Weikard, E. van Ierland (2012): Can altruism stabilise international climate agreements? *Ecological Economics* 81, 112-120.
- [71] Wagner, U.J. (2001): The Design of Stable International Environmental Agreements: Economic Theory and Political Economy, *Journal of Economic Surveys* 15 (3), 377–411.
- [72] Zammit-Lucia, J. (2013): COP19: the UN’s climate talks proved to be just another cop out. *Guardian* 02.12.13, <http://www.theguardian.com/sustainable-business/cop19-un-climate-talks-another-cop-out> (accessed 27.01.15). (<http://www.theguardian.com/sustainable-business/cop19-un-climate-talks-another-cop-out>).
- [73] Zelmer, J. (2003): Linear public good games: a meta-analysis, *Experimental Economics* 6, 299–310.

6 Appendix: Proofs

Proof of Proposition 1:

Proof. i) For $Q = 0$ to be a Nash equilibrium, it must be the case that $u_i(0, 0) \geq u_i(1, 0)$ for every i . Since countries are identical, it is sufficient to demonstrate that this holds for one i . Using eq. (10), $u_i(0, 0) \geq u_i(1, 0)$ is equivalent to

$$\alpha \geq 2(b - c)$$

which will always hold with $\alpha > 0$, because $b - c < 0$.

ii) For $Q = N$ to be a Nash equilibrium, it must be the case that $u_i(1, N - 1) \geq u_i(0, N - 1)$ for every i . Using (10), and that $Q = N$ implies $Q_{-i} = N - 1$, this gives

$$\begin{aligned} bN - c + \frac{3}{4}\alpha &\geq b(N - 1) + \frac{1}{4}\alpha \\ \alpha &\geq 2(c - b). \end{aligned}$$

iii) Consider first the possibility of a Nash equilibrium in pure strategies in which a share p of countries, where $0 < p < 1$, plays the pure strategy Abate, a share $1-p$ plays the pure strategy Pollute, and where all i are indifferent between Abate and Pollute. This would require, first, that $\hat{Q}_{-i}$ is an integer, otherwise $Q_{-i} = \hat{Q}_{-i}$ is not possible (and if $Q_{-i} \leq \hat{Q}_{-i}$, i is not indifferent between the pure strategies). Assume that $\hat{Q}_{-i}$ is an integer. However, if countries play different pure strategies, it cannot be the case that Q_{-i} is identical for all i . For a given Q , if $q_j = 1$ and $q_h = 0$, we must necessarily have $Q_{-j} = Q - 1$ and $Q_{-h} = Q$, hence $Q_{-j} < Q_{-h}$. Thus, the only possibility for all i to be indifferent is if they all play a mixed strategy.

Consider next the possibility that a share p play Abate, strictly preferring Abate, while a share $1-p$ play Pollute, strictly preferring Pollute. Define $\hat{Q}_{-i}$ such that $u_i(1, \hat{Q}_{-i}) = u_i(0, \hat{Q}_{-i})$. Then $Q_{-i} > \hat{Q}_{-i}$ is required for Abate to be strictly preferred by i , while $Q_{-i} < \hat{Q}_{-i}$ is required for Pollute to be strictly preferred. Hence we would need that for any j who Abates, $Q_{-j} > \hat{Q}_{-i}$, while for any h who Pollutes, $Q_{-h} < \hat{Q}_{-i}$. This implies $Q_{-j} > Q_{-h}$. But since, as demonstrated above, $Q_{-j} < Q_{-h}$, this cannot hold.

From eq. (12), we know that when $Q = \hat{Q}_{-i}$, the *share* of others playing Abate is $\frac{1}{2} + \frac{c-b}{\alpha}$. Consider now the possibility that every country i plays a mixed strategy such that $q_i = 1$ with probability $p = \frac{1}{2} + \frac{c-b}{\alpha}$ (and $q_i = 0$ with probability $1-p = \frac{1}{2} - \frac{c-b}{\alpha}$). Then, the expected number of others playing $q_i = 1$ equals $p(N-1) = \hat{Q}_{-i}$ for every i . In this situation, i is indifferent between Abate and Pollute. By the assumptions $c > b$ and $\alpha \geq 2(c-b)$, we know that $\frac{1}{2} < p < 1$. Hence, for every i , given that every other country plays Abate with probability $p = \frac{1}{2} + \frac{c-b}{\alpha}$, using the same strategy is a best response for i . The expected number of abating countries in this equilibrium is given by $N(\frac{1}{2} + \frac{c-b}{\alpha})$. ■

Proof of Proposition 2:

Proof. i) For $Q = 0$ to be a Nash equilibrium, it must be the case that $u_i(0, 0) \geq u_i(1, 0)$ for every i . For reciprocal countries with $\alpha_i = \alpha$, the proof is exactly as in Proposition 1, part i). For countries with $\alpha_i = 0$, this holds because the game is a Prisoners' dilemma and abate is strictly dominated by pollute, see footnote 1.

ii) For $Q = A$ to be a Nash equilibrium, it must be the case that $u_i(0, A) \leq u_i(1, A)$ for A players and $u_i(0, A) \geq u_i(1, A)$ for the remaining $N - A$ players. The latter follows because abate is a strictly dominated strategy for all $N - A$ players who have $\alpha_i = 0$. What remains to be shown is that $u_i(0, A) \leq u_i(1, A)$ for the A players who have $\alpha_i = \alpha$. When $Q = A$, $Q_{-i} = A - 1$ for any player for whom $q_i = 1$. By eq. (11), abate is preferred by an i with $\alpha_i = \alpha$ when $Q_{-i} = A - 1$ if

$$A \geq \left(\frac{c-b}{\alpha} + \frac{1}{2}\right)(N-1) + 1$$

or equivalently,

$$\alpha \geq 2(c-b) \frac{N-1}{2A-N-1}. \quad (20)$$

Since $\alpha > 0$, this can only hold if $2A - N - 1 > 0$, i.e.

$$A > \frac{N+1}{2}.$$

iii) Assume that $\alpha \geq 2(c-b)(N-1)/(2A-N-1)$. Due to i) and ii) above, there must then exist a tipping point $\hat{Q}_{-i}^\alpha \leq A$ such that $u_i(1, \hat{Q}_{-i}^\alpha) = u_i(0, \hat{Q}_{-i}^\alpha)$ for an i with $\alpha_i = \alpha$. The case where $\hat{Q}_{-i}^\alpha = A$ is covered by part ii) above. Consider now the case where $\hat{Q}_{-i}^\alpha < A$. Below, let $0 < r < 1$.

Consider first the possibility of a Nash equilibrium in pure strategies as follows: The $N-A$ players with $\alpha_i = 0$ play the pure strategy Pollute, $(1-r)A$ players with $\alpha_i = \alpha$ play the pure strategy Pollute, and rA players with $\alpha_i = \alpha$ play the pure strategy Abate. For the $N-A$ players with $\alpha_i = 0$, $u_i(0, \hat{Q}_{-i}^\alpha) > u_i(1, \hat{Q}_{-i}^\alpha)$, since Pollute strictly dominates Abate for such players. However, we must also have $u_i(1, \hat{Q}_{-i}^\alpha) \geq u_i(0, \hat{Q}_{-i}^\alpha)$ for the rA players with $\alpha_i = \alpha$ who play Abate and $u_i(1, \hat{Q}_{-i}^\alpha) \leq u_i(0, \hat{Q}_{-i}^\alpha)$ for the $(1-r)A$ players with $\alpha_i = \alpha$ who play Pollute. This cannot be satisfied by indifference, i.e., $\hat{Q}_{-i}^\alpha = rA$ and $u_i(1, \hat{Q}_{-i}^\alpha) = u_i(0, \hat{Q}_{-i}^\alpha)$ for all players with $\alpha_i = \alpha$: that would require, by eq. (11), $\frac{Q_{-i}}{(N-1)} = \frac{c-b}{\alpha} + \frac{1}{2}$ for all i with $\alpha_i = \alpha$, which cannot hold in because for any two countries j and h such that $q_j = 1$ and $q_h = 0$, we have $Q_{-j} < Q_{-h}$. Hence $\frac{Q_{-i}}{(N-1)} = \frac{c-b}{\alpha} + \frac{1}{2}$ cannot hold for both. Thus, the only possibility for all i with $\alpha_i = \alpha$ to be indifferent while still acting differently is if they all play a mixed strategy.

Consider next the possibility that rA countries with $\alpha_i = \alpha$ play Abate, strictly preferring Abate, while the remaining countries play Pollute, strictly preferring Pollute. Assume that $\hat{Q}_{-i}^\alpha$ is not an integer, and that consequently $rA \neq \hat{Q}_{-i}^\alpha$. For a polluting country in this situation, $Q_{-i} = rA$. For an abating country, $Q_{-i} = rA - 1$. Again, the $N-A$ players with $\alpha_i = 0$ would always strictly prefer Pollute. A polluting country i with $\alpha_i = \alpha$, however, would strictly prefer Pollute only if $Q_{-i} < \hat{Q}_{-i}^\alpha$, i.e., if $rA < \hat{Q}_{-i}^\alpha$. An abating country i with $\alpha_i = \alpha$ would strictly prefer Abate only if $Q_{-i} > \hat{Q}_{-i}^\alpha$, i.e., if $rA - 1 > \hat{Q}_{-i}^\alpha$. Thus, for the considered situation to be a Nash equilibrium, we must have $rA < \hat{Q}_{-i}^\alpha < rA - 1$, which cannot be the case.

Consider now the following possible Nash equilibrium: the $N-A$ players with $\alpha_i = 0$ play the pure strategy Pollute, and every country i with $\alpha_i = \alpha$ plays a mixed strategy such that $q_i = 1$ with probability r (and $q_i = 0$ with probability $1-r$). Then, the expected number of others playing $q_i = 1$ equals $r(A-1)$ for every i . In this situation, every country i with $\alpha_i = \alpha$ is indifferent between Abate and Pollute if $r(A-1) = \hat{Q}_{-i}^\alpha$.

What remains is to establish the level of r . Countries with $\alpha_i = \alpha$ have exactly the same preferences as those used to calculate $\hat{Q}_{-i}^\alpha$, the number of others abating making i indifferent between abating and polluting, in eq. (12).

Hence $\hat{Q}_{-i}^\alpha = \hat{Q}_{-i} = (\frac{c-b}{\alpha} + \frac{1}{2})(N-1)$, and

$$r = \frac{\hat{Q}_{-i}}{(A-1)} = (\frac{c-b}{\alpha} + \frac{1}{2}) \frac{(N-1)}{(A-1)}. \quad (21)$$

■

Proof of the claim that eq. (7) formalizes kindness in the three-stage game

Proof. The claim to be demonstrated is that eq. (7), $f_{ij} = q_i - \frac{1}{2}$, applies even in the three-stage game.

In Stage 3, non-signatories play non-cooperatively. They thus have the same options to influence others as in the non-cooperative game. It follows that the kindness of a non-signatory i towards any other country j can be expressed as in eq. (7): $f_{ij} = q_i - \frac{1}{2}$.

Turn then to Stage 2. For a given abatement policy of the coalition, a signatory's influence on others' payoff goes solely through the country's own contribution to the coalition's abatement, chosen implicitly when deciding in Stage 1 whether to join. Assume that there is a coalition of size k . If the abatement policy of a coalition of size $k-1$ is identical to the abatement policy of a coalition of size k , no signatory is pivotal, and their kindness to any other country j is again given by $f_{ij} = q_i - \frac{1}{2}$ (where q_i in this case is determined by the coalition's policy). However, if the abatement policy of a coalition of size $k-1$ differs from the abatement policy of a coalition of size k , every signatory is pivotal in the sense that if it leaves, this will affect all other signatories' abatement.

Assume now that a coalition of k members is abating, that the loss of one member would have made the coalition pollute, and that non-signatories pollute. Every individual signatory i is then pivotal in the sense that i 's participation is decisive for the coalition's abatement policy, given others' strategies. In this case, if signatory i stays, every other signatory gets a payoff of $bk - c$, while every non-signatory gets a payoff of bk . If i leaves, the coalition does not abate and everyone gets a payoff of 0. The equitable payoff for other signatories are then, according to eq. 5,

$$\pi_{is}^e = \frac{1}{2}(bk - c) \quad (22)$$

and for non-signatories

$$\pi_{in}^e = \frac{1}{2}(bk). \quad (23)$$

Using this and eq. (4), i 's kindness to another signatory if staying is thus given by

$$f_{is} = \frac{(bk - c) - \frac{1}{2}(bk - c)}{(bk - c)} = \frac{1}{2} \quad (24)$$

and if leaving

$$f_{is} = \frac{0 - \frac{1}{2}(bk - c)}{(bk - c)} = -\frac{1}{2}.$$

Moreover, i 's kindness to a non-signatory if joining is given by

$$f_{in} = \frac{bk - \frac{1}{2}bk}{bk} = \frac{1}{2} \quad (25)$$

and if not joining,

$$f_{in} = \frac{0 - \frac{1}{2}(bk)}{(bk)} = -\frac{1}{2}.$$

A similar argument can be made for the non-signatory considering to join.

Consequently, even for pivotal countries, kindness can be expressed as $f_{ij} = q_i - \frac{1}{2}$, and the reciprocity function (eq. 9) and utility function (eq. 10) can be applied as before. ■

Proof of Proposition 3:

Proof. Assume $k = 0$. Then, in Stage 3, all countries are non-signatories and thus play non-cooperatively. We can then use the results from the non-cooperative game. By Proposition 1, part i, we know that $u_i(0, 0) \geq u_i(1, 0)$ and that $Q = 0$ is a Nash equilibrium in the non-cooperative game; thus if $k = Q_{-i} = 0$ in the participation game, each non-signatory pollutes in Stage 3.

If $k = 0$, there is no coalition to decide in Stage 2 whether to Abate or not. If one country still joined in Stage 1, this "coalition" would decide the strategy of only one country and thus correspond to a non-cooperative player, whose best response to others' Pollution would be to Pollute (see the proof of Proposition 1, part i). Given this, there is no incentive to join in Stage 1.

External stability requires that $U_n(q_n, q_s, k) \geq U_s(q_n, q_s, k+1)$, i.e., $U_n(0, 0, 0) \geq U_s(0, 0, 1)$, which was verified above. Internal stability is not an issue here, since no country is a signatory and a coalition of -1 countries is not feasible. ■

Proof of Proposition 4:

Proof. Note first that non-signatory h 's best reply $q_h = B_h(q_n, q_s, k)$ is given by

$$B_h(q_n, q_s, k) = \arg \max_{q_h} U_h(q_h, q_n, q_s, k) = \begin{cases} 1 & \text{if } Q_{-h} \geq (\frac{1}{2} + \frac{c-b}{\alpha})(N-1) \\ 0 & \text{otherwise.} \end{cases} \quad (26)$$

This can be seen as follows: for $q_h = 1$ to maximize $U_h(q_h, q_n, q_s, k)$, we must have $U_h(1, q_n, q_s, k) \geq U_h(0, q_n, q_s, k)$. Inserting from eq. (19), this implies

$$\begin{aligned} \frac{3\alpha}{2(N-1)}Q_{-h} - \frac{\alpha + 2(c-b)}{2} &\geq \frac{\alpha}{2(N-1)}Q_{-h} \\ Q_{-h} &\geq (\frac{1}{2} + \frac{c-b}{\alpha})(N-1). \end{aligned}$$

The coalition's best reply q_s is given by

$$B_s(q_n, k) = \arg \max_{q_s} U_s(q_n, q_s, k) = \begin{cases} 1 & \text{if } Q_n \geq \hat{Q}_n \\ 0 & \text{otherwise,} \end{cases} \quad (27)$$

where $\hat{Q}_n = [\frac{1}{2} + \frac{c-b}{\alpha} - (\frac{b}{\alpha} + \frac{3}{2(N-1)})(k-1)](N-1)$. This can be seen as follows: for $q_s = 1$ to maximize $U_s(q_n, q_s, k)$, we must have $U_s(q_n, 1, k) \geq U_s(q_n, 0, k)$. Inserting from eq. (18) and using that $Q_{-s} = Q_n + (k-1)q_s$, this implies

$$\begin{aligned} [b + \frac{3\alpha}{2(N-1)}](Q_n + (k-1)) - \frac{\alpha + 2(c-b)}{2} &\geq (b + \frac{\alpha}{2(N-1)})Q_n, \text{ or} \\ Q_n &\geq [\frac{1}{2} + \frac{c-b}{\alpha} - (\frac{b}{\alpha} + \frac{3}{2(N-1)})(k-1)](N-1). \end{aligned}$$

To show that the grand coalition is stable, we need to establish that abatement is the best response of a coalition of size N , and that no signatory would be better off by leaving.

First, if there is a coalition of size $k = N$, eq. (27) implies that $B_s(q_n, N) = 1$ iff

$$\begin{aligned} 0 &\geq \frac{1}{2} + \frac{c-b}{\alpha} - (\frac{b}{\alpha} + \frac{3}{2(N-1)})(N-1), \text{ or} \\ 0 &\geq \frac{c-bN}{\alpha} - 1 \end{aligned}$$

which holds because $c < bN$ and $\alpha > 0$.

A signatory who stays in this coalition gets utility

$$\begin{aligned} U_s(q_n, 1, N) &= [b + \frac{3\alpha}{2(N-1)}]Q_{-s} - \frac{\alpha + 2(c-b)}{2} - \frac{\alpha}{4} \\ &= bN - c + \frac{3\alpha}{4}. \end{aligned}$$

To calculate the utility a country can get if it leaves, we need to know the best response of a coalition of size $N-1$. From (27) we have that $B_s(0, N-1) = 1$ iff

$$\begin{aligned} 0 &\geq [\frac{1}{2} + \frac{c-b}{\alpha} - (\frac{b}{\alpha} + \frac{3}{2(N-1)})(N-2)](N-1) \\ \frac{3(N-2)\alpha}{2(N-1)} - \frac{\alpha}{2} &\geq c - b(N-1) \\ \alpha &\geq \frac{2(N-1)}{(2N-5)}(c - b(N-1)) \end{aligned}$$

which holds because by assumption, $\alpha > 0$, $c \leq b(N-1)$ and $N > 2$.

If $k = N-1$ and $q_n = 1$, $Q_n = 1$. Thus, (27) implies that $B_s(1, N-1) = 1$ iff

$$\begin{aligned} \frac{1}{(N-1)} &\geq \frac{-2N+5}{2(N-1)} + \frac{(c-b(N-1))}{\alpha} \\ 0 &\geq \frac{-2N+3}{2(N-1)} + \frac{(c-b(N-1))}{\alpha} \end{aligned}$$

which holds because $N \geq 2$ and, by the assumption of the Proposition, $c \leq b(N-1)$. Hence, abating is a dominant strategy for a coalition of size $N-1$.

External stability is not an issue in this case, since there are no non-signatories. If country h leaves and abates, it gets utility (eq. (19))

$$\begin{aligned} U_h(1, q_n, 1, N-1) &= [b + \frac{3\alpha}{2(N-1)}](N-1) - \frac{\alpha + 2(c-b)}{2} - \frac{\alpha}{4} \\ &= bN - c + \frac{3\alpha}{4} = U_s(q_n, 1, N). \end{aligned}$$

If country h leaves and pollutes, it gets utility

$$U_h(0, q_n, 1, N-1) = [b + \frac{\alpha}{2(N-1)}](N-1) - \frac{\alpha}{4}.$$

For the coalition of size $k = N$ to be internally stable, we must have $U_s(q_n, 1, N) \geq U_h(0, q_n, 1, N-1)$, or

$$\begin{aligned} bN - c + \frac{3\alpha}{4} &\geq [b + \frac{\alpha}{2(N-1)}](N-1) - \frac{\alpha}{4}, \text{ i.e.,} \\ \alpha &\geq 2(c-b) \end{aligned}$$

which holds by assumption. Hence, the grand coalition is stable. ■

Proof of Proposition 5:

Proof. Define $\underline{k}$ such that if there exists a coalition of size $\underline{k}$, and non-signatories pollute, the coalition is exactly indifferent between abating and polluting; that is, $U_s(0, 0, \underline{k}) = U_s(0, 1, \underline{k})$.

Due to eq. (27), $B_s(0, k) = 1$ iff $0 \geq \hat{Q}_n$, i.e., if

$$k \geq \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha}. \quad (28)$$

Hence, we must have that

$$\underline{k} = \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha}. \quad (29)$$

Assume now that there exists a coalition of size k^1 , where k^1 is defined as the smallest integer such that $k^1 \geq \underline{k}$. Assume further that $q_n = 0$ and $q_s = 1$. For such a coalition to be stable, both the coalition and the non-signatories must be playing their best responses; no coalition member must be better off by leaving (internal stability); and no non-signatory must prefer to join (external stability).

By eq. (27) and the definition of $\underline{k}$, $B_s(0, \underline{k}) = 1$. Further, if $k^1 > \underline{k}$, eq. (28) implies that $B_s(0, k^1) = 1$. Hence in the situation described in the Proposition, the coalition is playing its best response.

Next, I show that non-signatories are playing their best response too, i.e., $B_h(0, 1, k^1) = 0$. When $k = k^1$, $q_n = 0$, and $q_s = 1$, we have $Q_{-h} = k^1$. By

(26), h 's best reply $B_h(0, 1, k) = 1$ iff $\frac{k^1}{N-1} \geq \frac{1}{2} + \frac{c-b}{\alpha}$. Since $c > b$ and $\alpha > 0$, this requires that $\frac{k^1}{N-1} > \frac{1}{2}$. However, this cannot be the case: under the given assumptions, k^1 is a minority coalition. To see this, note that $\underline{k}$ can be characterized as follows. When $c \leq b(N+2)/3$, which holds by assumption, $\underline{k}$ is increasing in α :

$$\frac{\partial \underline{k}}{\partial \alpha} = \frac{(N+2)(2b(N-1) + 3\alpha) - 3((2c + \alpha)(N-1) + 3\alpha)}{(2b(N-1) + 3\alpha)^2}$$

i.e., $\frac{\partial \underline{k}}{\partial \alpha} \geq 0$ iff

$$\begin{aligned} (N+2)(2b(N-1) + 3\alpha) - 3(2c(N-1) + \alpha(N+2)) &\geq 0, \text{ or} \\ b(N+2)/3 &\geq c. \end{aligned} \quad (30)$$

Second, when α goes to infinity, $\underline{k}$ goes to $\frac{N+2}{3}$:

$$\lim_{\alpha \rightarrow \infty} \underline{k} = \lim_{\alpha \rightarrow \infty} \frac{2c(N-1)/\alpha + (N+2)}{2b(N-1)/\alpha + 3} = \frac{N+2}{3} \quad (31)$$

Thus, $\frac{N+2}{3}$ is an upper boundary for $\underline{k}$ under the given assumptions. Since k^1 is the smallest integer weakly larger than $\underline{k}$, then for any $\alpha > 0$ we must have $k^1 < \frac{N+2}{3} + 1$, or equivalently, $k^1 < \frac{N+5}{3}$. Combining this with the result from above that $B_h(0, 1, k) = 1$ iff $k^1 \geq (\frac{1}{2} + \frac{c-b}{\alpha})(N-1)$, we have that $B_h(0, 1, k) = 1$ iff

$$\frac{N+5}{3} > k^1 \geq \left(\frac{1}{2} + \frac{c-b}{\alpha}\right)(N-1) \quad (32)$$

$$\text{or} \quad (33)$$

$$\alpha(13-N) > 6(c-b)(N-1)$$

which can never hold for $N > 13$, since $c > b$ and $\alpha > 0$. Consequently, $B_h(0, 1, k) = 0$, and both the coalition and the non-signatories are playing their best response in the situation described in the Proposition.

Next, consider whether this situation is internally stable: no member wants to leave.

From eq. (26), we know that $B_h(0, 0, k) = 1$ iff $-\frac{1}{2} \geq \frac{c-b}{\alpha}$, which cannot hold given that $\alpha > 0$; i.e., for any k , $B_h(0, 0, k) = 0$. Hence, if $q_n = q_s = 0$, polluting is a dominant strategy for non-signatory h , and $B_h(0, 0, \underline{k} + 1) = B_h(0, 0, \underline{k}) = B_h(0, 0, \underline{k} - 1) = 0$.

By definition of k^1 , $B_s(0, k^1 - 1) = 0$. Thus, if there is a coalition of size $k = k^1$ and one member leaves, the remaining coalition of size $k^1 - 1$ will not abate. The best response to this by a non-signatory h is to pollute. Internal stability thus requires $U_s(0, 1, k^1) \geq U_n(0, 0, k^1 - 1)$.

From eqs. (19) and (18) it can be seen that for any k , $U_n(0, 0, k) = U_n(0, 0, k - 1) = U_s(0, 0, k) = U_s(0, 0, k - 1) = -\frac{\alpha}{4}$. Hence $U_s(0, 1, k^1) \geq$

$U_n(0, 0, k^1 - 1)$ is equivalent to $U_s(0, 1, k^1) \geq U_s(0, 0, k^1)$. But we know from eq. (28) that $B_s(0, k^1) = 1$, which implies $U_s(0, 1, k^1) \geq U_s(0, 0, k^1)$, iff

$$k^1 \geq \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha}. \quad (34)$$

which holds by the definition of k^1 . Hence a coalition of size k^1 is internally stable.

It remains to be confirmed that a coalition of size k^1 is externally stable, i.e., no non-signatory wants to join.

If a non-signatory joins, the best reply of the expanded coalition of $k^1 + 1$ would still be to abate, i.e., $B_s(0, k^1 + 1) = 1$, due to eq. (28) and $k^1 + 1 > \underline{k}$. The remaining non-signatories would still prefer not to abate, since we have from eq. (26) that $B_h(0, 1, k^1 + 1) = 1$ iff

$$\begin{aligned} k^1 + 1 &\geq \left(\frac{1}{2} + \frac{c-b}{\alpha}\right)(N-1) \\ k^1 &\geq \frac{\alpha(N-3) + 2(c-b)(N-1)}{2\alpha} \\ k^1 &\geq \frac{(N-3)}{2} + \frac{(c-b)(N-1)}{\alpha} \end{aligned}$$

which cannot hold when $\frac{N-1}{2} > k^1$ and $N > 13$.

Consequently, a coalition of size k^1 as described in the Proposition is externally stable if $U_n(0, 1, k^1) \geq U_s(0, 1, k+1)$. Using (19),

$$U_n(0, 1, k^1) = \left[b + \frac{\alpha}{2(N-1)}\right]k^1 - \frac{\alpha}{4}$$

while (using (18)),

$$U_s(0, 1, k+1) = \left[b + \frac{3\alpha}{2(N-1)}\right](k^1) - \frac{\alpha + 2(c-b)}{2} - \frac{\alpha}{4}$$

$U_n(0, 1, k^1) \geq U_s(0, 1, k+1)$ thus requires that

$$\begin{aligned} \frac{\alpha + 2(c-b)}{2\alpha}(N-1) &\geq k^1 \\ &\text{or} \\ \hat{Q}_{-i} &\geq k^1 \end{aligned}$$

which must be the case, since $\hat{Q}_{-i}$ (defined by eq. (12)) is a strict majority and k^1 is a minority.

Finally, recall that k^0 is the smallest integer weakly larger than c/b . If $\alpha = 0$, we would have $\underline{k} = c/b$ (see eq. (29)). Further, $\underline{k}$ is increasing in α . Since k^1 is the smallest integer such that $k^1 \geq \underline{k} > c/b$, it follows that $k^1 \geq k^0$. ■

Proof of Proposition 6

Proof. To show that the coalition is stable, we need to establish that abatement is the best response of a coalition of size A consisting of reciprocal countries;

that no reciprocal signatory would be better off by leaving; and that no non-reciprocal non-signatory would be better off by joining.

First, if there is a coalition of size $k = A$, eq. (28) and the definition of $\underline{k}$ imply that $B_s(0, A) = 1$ iff

$$A \geq \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha} = \underline{k}. \quad (35)$$

This holds because, from the assumption that $\alpha \geq \frac{2(c-b)(N-1)}{(2A-N-1)}$, it follows that $2A - N - 1 \geq 0$, i.e., $A \geq \frac{N+1}{2}$. That is, A is a strict majority, while $\underline{k}$ is a strict minority (see the proof of Proposition 4). It follows also that $B_s(0, A-1) = B_s(0, A+1) = 1$.

A non-signatory country h is thus not pivotal for the coalition's decision to abate. Since non-signatories have $\alpha_h = 0$, their dominant strategy is to pollute.

A reciprocal signatory who stays in this coalition gets utility (using eq. (10))

$$U_s(0, 1, A) = bA - c + \frac{3\alpha}{2} \left(\frac{A-1}{N-1} - \frac{1}{2} \right).$$

If it leaves and pollutes, it gets utility

$$U_n(0, 1, A-1) = b(A-1) + \frac{\alpha}{2} \left(\frac{A-1}{N-1} - \frac{1}{2} \right).$$

Hence $U_s(0, 1, A) \geq U_n(0, 1, A-1)$ if

$$\begin{aligned} bA - c + \frac{3\alpha}{2} \left(\frac{A-1}{N-1} - \frac{1}{2} \right) &\geq b(A-1) + \frac{\alpha}{2} \left(\frac{A-1}{N-1} - \frac{1}{2} \right) \\ &\text{or} \\ \alpha &\geq \frac{2(c-b)(N-1)}{2A-N-1} \end{aligned}$$

which holds by assumption.

Finally, since a non-signatory is not pivotal for the coalition's abatement policy, and non-signatories are not reciprocal, no non-signatory prefers to join. ■

Figure 1: Payoff of country i , given that Q_{-i} others abate.

Red solid line: payoff if i pollutes.

Green dashed line: payoff if i abates.

Figure 2: Utility of a reciprocal country i , given that Q_{-i} others abate.

Red solid line: utility if i pollutes.

Green dashed line: utility if i abates.

Figure 3: Payoff of country i (standard preferences), given that Q_{-i} others abate.

Red solid line: payoff if i pollutes.

Green dashed line: payoff if i abates.

c/b : the minimum k for which the coalition prefers to abate.

k^0 : the smallest integer weakly larger than c/b .

Figure 4: Utility of a reciprocal country i , given that Q_{-i} others abate.

Red solid line: Utility if i pollutes.

Green dashed line: Utility if i abates.

Orange dotted line: Utility if no one abates.

$\underline{k}$: the minimum k for which the coalition prefers to abate.

k^1 : the smallest integer weakly larger than $\underline{k}$.