Working Paper 20/2014

Reciprocal climate negotiators: Balancing anger against even more anger

Karine Nyborg

The CREE Centre acknowledges financial support from The Research Council of Norway, University of Oslo and user partners.
ISBN: 978-82-7988-190-2
ISSN: 1892-9680
http://www.cree.uio.no

Reciprocal climate negotiators: Balancing anger against even more anger[*]

Karine Nyborg[†]

Abstract

I explore possible impacts of reciprocal preferences on participation in international environmental agreements. Reciprocal countries condition their willingness to abate on others' abatement. No participation is always stable. A full or majority coalition can be stable, provided that reciprocity is sufficiently strong and widespread. In addition, a stable minority coalition can exist, even with weak reciprocity preferences. This latter coalition is weakly larger than the maximum stable coalition with standard preferences, but is characterized by mutually negative sentiments.

JEL codes: F53, H87, Q54.

Keywords: International Environmental Agreements, Reciprocity, Coalitions.

1 Introduction

> "Mr. President, this is the worst meeting I've been to since the eighth-grade student council." *Secretary of State Hilary Clinton to President Barack Obama at his arrival at the Copenhagen climate change summit in 2009 (as quoted by Landler and Cooper 2010).*

To date, international climate negotiations have yielded discouraging results. Economists ought perhaps not to be surprised by this, since the payoffs associated with countries' greenhouse gas emissions seem to be structured like a Prisoners' Dilemma game: While limiting global warming is crucial for all countries, individual countries' incentives to abate their own greenhouse gas emissions are weak. And since there exists no global institution that can enforce cooperation

[*] *Acknowledgements:* I am grateful to Aart de Zeeuw, Michael Hoel, Jon Hovi, and seminar participants at the Paris Environmental and Energy Economics Seminar and the Grantham Institute for helpful comments to previous drafts. 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 one 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.

by sovereign states, standard environmental policy instruments cannot readily solve the problem. In line with this, the theoretical literature on international environmental agreements has established a whole array of pessimistic results. One is that although stable climate treaties may be feasible, they are likely to have very few signatories, and/or to involve unambitious emission reduction goals (Barrett 1992, 1994, 2003, Carraro and Siniscalco 1993, Hoel 1992).[1]

Such pessimism has been questioned on the basis of results from behavioral and experimental economics (Grüning and Peters 2007, Burger and Kolstad 2009). A substantial body of research has established that both in the laboratory and in the field, human beings sometimes manage to cooperate in Prisoners' Dilemma-like situations, even in the absence of external enforcement (Ostrom 1990, Camerer 2003, Zelmer 2003). Nevertheless, the same literature confirms that cooperation frequently fails (in addition to the above-mentioned references, see e.g. Tavoni et al. 2011, Barrett and Dannenberg 2012).

Reciprocity is often suggested as a possible explanation to observed cooperation patterns (Fehr and Gächter 2000, Sobel 2005, Croson et al. 2006, Croson 2007). Reciprocity can be defined as a *preference* for repaying mean (kind) intentions by mean (kind) actions. This should not be confused with a reciprocal *strategy,* like tit-for-tat; to distinguish the two, Sobel (2005) uses the term 'intrinsic reciprocity' for what I call reciprocal preferences or just reciprocity. Suggested formal modelling approaches include Rabin (1993), Levine (1998), Dufwenberg and Kirschsteiger (2004), Falk and Fischbacher (2006) and Cox et al. (2007).

My aim here is to explore theoretically the impacts of reciprocal preferences on participation in international environmental agreements. Since reciprocity models tend to become analytically complex, I will do so by means of a very simple model – too simple, in fact, to capture several of the potentially most interesting aspects of reciprocity in the climate negotiations context. Still, I believe some useful insight can be derived even from the highly stylized analysis below.

The idea of studying social preferences within the framework of international environmental agreement models is not new. Hoel and Schneider (1997) show that if there is some non-environmental cost of breaking agreements, the size of equilibrium coalitions is enlarged. 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 inequity aversion can increase coalition size; if the abatement choice is discrete, they find that even the fully cooperative outcome may be feasible. Lange (2006) allows heterogeneity between countries, and finds that inequity aversion with respect to abatement targets across industrialized countries makes larger coalitions feasible.[2]

Nevertheless, it is not clear that altruism and/or inequity aversion provide

[1] Within the literature based on repeated games, however, there is also a few papers with more optimistic results: see Froyn and Hovi (2008), Kratzsch et al. (2012), Heitzig et al. (2011).

[2] See also Grüning and Peters (2007), Kolstad (2013).

reasonable explanations of the cooperation sometimes observed in field and laboratory studies. Models of altruism typically predict voluntary contributions to public goods to be *decreasing* in others' contributions, while empirical evidence suggests that this relationshop is *increasing* (Nyborg and Rege 2003, Croson 2007). In public good game experiments, hardly any subjects are unconditional altruists who keep contributing if others do not. A substantial share of subjects, however, are conditional cooperators who contribute more the more others contribute (Ledyard 1995, Fischbacher et al. 2001, Fischbacher and Gächter 2006, 2010, Croson et al. 2006). Martinsson et al. (2013) summarize findings from experiments across the world, indicating that conditional cooperators tend to constitute a majority, or close to a majority, of subjects: their list comprises Colombia with 63%, Vietnam 50% (Martinsson et al. 2013); Switzerland 50% (Fischbacher et al. 2001); Denmark 70 % (Thöni et al. 2012); Russia 56% (Herrmann and Thöni 2009); USA 81%, Switzerland 44%, and Japan 42% (Kocher et al. 2008). Conditional cooperation may be consistent with inequity aversion (Fehr and Schmidt 1999), but cannot be explained by simple altruism models.

Altruism as well as inequity aversion models define preferences over material outcomes only. A growing body of experimental evidence indicates, however, that people also care about others' intentions (Fehr and Gächter 2000, Camerer 2003). In ultimatum games, for example, responders tend to reject inequitable offers from proposers; however, if offering an equitable share was not an option for the proposer, responders are considerably more likely to accept (Falk et al. 2003). Such behavior can be explained neither by altruism nor inequity aversion.[3]

A reciprocal person is not generally kind. Rather, reciprocity is about anger and gratitude, retaliation and reward. Although reciprocity may help secure cooperation, it can also be very destructive. An altruistic or inequity averse person would never destroy valuable resources for the sake of revenge; a reciprocal person might do precisely that (Sobel 2005 gives the example of an employee deliberately ruining the firm's computer system on his last day of work to punish the firm for firing him).

Rabin (1993) demonstrated, within the context of a two-player normal form game, that reciprocity may transform a Prisoners' Dilemma game, as measured in material payoffs, into a coordination game in utilities. Mutual defection will still be an equilibrium, but so can mutual cooperation. Below, I am exploring whether a similar result holds for an N-player non-cooperative game with global public goods, and how reciprocal preferences would affect participation in international environmental agreements.

In spite of the reciprocity concept's popularity in behavioral economics, formal models have rarely been employed in the applied literature, presumably because of their complexity and the unfamiliar methods often required. If players care intrinsically about others' intentions, one may need to define preferences

[3] The same is true for the behavior displayed in Frans de Waal's capuchin monkey fairness experiment posted on https://www.youtube.com/watch?v=-KSryJXDpZo&feature=player_detailpage. While the monkey could possibly be acting strategically, its behavior can be explained neither by altruism nor by inequity aversion.

over beliefs. Standard game theory, however, defines preferences over outcomes only. For this reason, Rabin (1993) applies psychological game theory (Geanakoplos et al., 1989) in his 2-player analysis.[4] With N players, the set of potentially relevant beliefs easily become excessively complex; furthermore, as pointed out by Dufwenberg (2008), research on psychological games is still in its infancy.

To keep the analysis tractable, I am going to use an extremely simple model of participation in international environmental agreements from Barrett (2003). In the model, abatement choices are binary, abatement costs as well as environmental benefits are linear, and all countries are identical. Concerning the modeling of reciprocity, I follow Rabin (1993) closely, but modify his approach to allow for more than two players. It turns out that within the game I study, reciprocity can be formalized as a function of own and others' strategies only, permitting me to use an approach by Segal and Sobel (2007) rather than psychological game theory.

To my knowledge, Hadjiyiannis et al. (2012) is the only preexisting formal analysis of reciprocal preferences in international environmental agreements. While participation is the topic of my discussion, Hadjiyiannis et al. (2012) study compliance. Their analysis also differs from mine in that they employ a 2-player framework and assume continuous abatement choices. They find that reciprocity can facilitate cooperation, but only if the abatement level required to be viewed as 'fair' by the other player is low. Whenever countries' fairness view is more demanding, Hadjiyiannis et al. (2012) find that reciprocity is detrimental to cooperation.

Even if individuals were reciprocal, it would not follow automatically that *countries* behave reciprocally. A democratic government wanting to be reelected may well act according to reciprocal preferences if it believes that the median voter is reciprocal; moreover, if government leaders or negotiating officials hold reciprocal preferences, this could influence their negotiation behavior.[5] Nevertheless, below I simply explore the consequences for participation in international environmental agreements *if* countries act according to a particular specification of reciprocal preferences. I do not claim that they *do* have reciprocal preferences.

My findings can be summarized as follows. No coalition participation is always a stable situation. When few others are expected to abate, reciprocal countries are even less willing to abate than countries with standard preferences. In the case where all countries have strong preferences for reciprocity, however, the grand coalition of all countries is stable. If not all countries are reciprocal, a coalition comprised of a substantial number of countries can be stable; this requires, however, that the share of reciprocal countries is strictly more than half, and that these countries' preferences are sufficiently strongly reciprocal.

With standard preferences, the maximal stable coalition is so small that very

[4] The same holds for Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006).

[5] It is also conceivable that reciprocal preferences might represent behavior of self-interested countries acting strategically in a bigger game of international relations in general. This remains to be explored, however. For a formal model of issue linkage, see e.g. Conconi and Perroni (2002).

little of the potential benefits of cooperation are realized. In the analysis below, this coalition size will be called k^0. With reciprocal preferences, a coalition of size k^0 is not generally stable: at such a low participation level, countries are so angry with the others that they take pleasure in hurting them, adding to the material disincentives to contribute.[6] Nevertheless, if the total number of countries is large and the cost-benefit ratio is relatively small, there exists a stable minority coalition size $k^1 \geq k^0$ – regardless of the strength of the reciprocity preference. This situation should not be confused with the possible full (or majority) coalition mentioned above. With the grand coalition, each country's behavior can be considered kind; with sufficiently strong reciprocal preferences, the grand coalition is thus enforced by the preference to repay kindness by kindness. Coalition size k^1 is very different: it represents a 'barely worth it'-situation enforced by similar mechanisms as k^0 of the standard preferences case (and even coincides with k^0 when reciprocity preferences are weak enough).

With coalition size k^1, most countries' behavior is considered unkind, and countries are so angry and disappointed with each other that they would derive satisfaction from punishing each other. In spite of this, if a country i expects $k^1 - 1$ others to stay in the coalition, i will prefer to stay in the coalition too. It will do so because this is the only way it can keep a small island of kindness in a world of meanness; if it leaves, the world becomes universally mean.

I do not claim that my model is particularly realistic. Among its ridiculously simplified features, the assumption of identical countries may, in the current context, stand out as particularly outrageous. While my analysis was motivated by a wish to understand how reciprocity may influence climate negotiations, the current paper thus provides no final answer to this question; it might, however, suggest a framework for beginning to explore it.

2 Defining reciprocity

Assume 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:

$$u_i = \pi_i + \alpha R_i \tag{1}$$

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

In his 2-player normal form game, Rabin (1993) used the following specification for the reciprocal part of player i's preferences:

$$R_i = \tilde{f}_{ji}(1 + f_{ij}) \tag{2}$$

where f_{ij} denotes i's kindness towards j, and $\tilde{f}_{ji}$ is i's belief about j's kindness towards i (for $i \neq j$). $f_{ij} < 0$ (> 0) means that i is being mean (kind). That is,

[6] My use of words like 'angry' in this paper should not be taken too literally, as the formal model need not necessarily reflect such emotions; I still use words like this because I think they help convey the intuition of the mechanisms I have in mind.

reciprocity consists, essentially, of two parts: First, the negative (positive) emotion of being treated badly (nicely), represented by $\tilde{f}_{ji}$; second, the satisfaction of repaying by being mean (kind) in return, represented by $\tilde{f}_{ji}f_{ij}$.[7]

To understand the intuition, consider the following story: Paul pays Ann's bill at a restaurant. Ann thinks Paul does this to insult her, which makes her feel bad (the first part). However, Ann's pain is reduced if, when leaving the restaurant, she tells Paul that he's a snobbish fool, insulting him back (the second part). Note, as illustrated by this example, that misunderstandings and differering norms, fairness views and/or cultures can complicate the relationship between reciprocal players considerably, even though the model presented below is too simple to represent this potentially quite relevant aspect of reciprocity.

To allow for more than 2 players, I assume that reciprocity is binary and additive, in the sense that a country cares about the average relationship between itself and each other country, while being unconcerned about the relations between others. I will define the reciprocal part of utility in the N-player model as the average of each bilateral reciprocity relation:

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

where the sums are over all $j = 1, ..., N$ for whom $j \neq i$.

We now need to operationalize what it means to be "kind" or "mean". Rabin (1993) defines "kindness" based on the material payoff i is *trying to* secure to j, and the range of payoffs i *could have* secured to j – given i's beliefs.

Let σ_i be i's strategy, let σ_{-i} denote the strategies of everyone other than i, and let $\tilde{\sigma}_{-i}$ denote i's belief about the strategies of everyone else. Further, let $\pi_j(\sigma_i, \sigma_{-i})$ denote the material payoff j will get as a function of i's and others' strategies. Then, $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ is the material payoff i is trying to secure to j.

Let $\pi_{ij}^{\max}$ denote the maximum of $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ with respect to σ_i, and let $\pi_{ij}^{\min}$ denote the minimum of $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ (for given $\tilde{\sigma}_{-i}$). Then, following Rabin (1993), I define kindness from i to j as

$$f_{ij} = \frac{\pi_j(\sigma_i, \tilde{\sigma}_{-i}) - \pi_{ij}^e}{\pi_{ij}^{\max} - \pi_{ij}^{\min}} \tag{4}$$

where

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

If $\pi_{ij}^{\max} = \pi_{ij}^{\min}$, then $f_{ij} = 0$.[8] 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, $\tilde{\sigma}_{-i}$.

[7] Usually, 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, I still need to include both parts, since each can be behaviorally relevant when coalitions behave cooperatively.

[8] Rabin (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.

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 latter is given by the average of the least and most i could have secured to j (given i's beliefs). Finally, the whole thing is normalized by the range of payoffs i could have secured to j.

If I choose the strategy that gives you the highest possible material payoff, given my beliefs about yours 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. Note thus that with this specification, what matters is what I try to secure to you compared with the options I think I have, not how much I sacrifice myself.

3 Non-cooperative play with reciprocity

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, given by

$$\pi_i = b(Q_{-i} + q_i) - cq_i \tag{6}$$

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 $Q_{-i} = \sum_{j=1}^{N} q_j - q_i = \sum_{j \neq i} q_j$ denotes 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. (6), and this is common knowledge, the above constitutes a N-player Prisoners' Dilemma game. Pollute ($q_i = 0$) is then a strictly dominant strategy with non-cooperative play; nevertheless, had everyone chosen Abate instead, each country would have been better off.[9] The dilemma is illustrated in Figure 1, making clear that regardless of how many other countries abate, country i's material payoff is always strictly higher if it pollutes itself.

INSERT FIGURE 1 ABOUT HERE

3.1 Defining reciprocity in the non-cooperative game

In the non-cooperative game presented above, the only way i can influence j's payoff is through i's choice of pollute versus abate. Given the behavior of

[9] This 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.

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

$$\pi_{ij}^{e} = b(Q_{-i} + \frac{1}{2}) - cq_j. \tag{7}$$

Thus, eq. (4) simplifies to

$$f_{ij} = q_i - \frac{1}{2}. \tag{8}$$

Since environmental quality is a pure public good, i is always equally kind or mean to everyone else, and i's kindness does not depend on others' strategies.[10] Thus, i's belief about j's kindness, $\tilde{f}_{ji}$, can quite naturally be assumed to depend, similarly, on j's strategy only:

$$\tilde{f}_{ji} = q_j - \frac{1}{2}. \tag{9}$$

Inserting from eqs. (8) and (9), using $Q_{-i} = \sum_{j \neq i} q_j$ and that $f_{ij} = f_{ik}$ for all $j, k \neq i$, R_i can be written as a function of own and others' strategies as follows:

$$R_i = (\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}) \tag{10}$$

where the first factor reflects the average kindness of others.

Inserting this into the utility function (1) defines reciprocal utility as a function of own and others' strategies:

$$u_i = u(q_i, Q_{-i}) = b(Q_{-i} + q_i) - cq_i + \alpha(\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}). \tag{11}$$

3.2 Nash equilibria

Let us now turn to abatement decisions in the case where all countries act noncooperatively. 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

$$\frac{Q_{-i}}{(N-1)} \geq \frac{c-b}{\alpha} + \frac{1}{2}, \tag{12}$$

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 $\alpha > 0$).

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

$$\hat{Q}_{-i} = (\frac{c-b}{\alpha} + \frac{1}{2})(N-1) \tag{13}$$

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

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. Whenever $Q_{-i} < 1/2$, reciprocity *reinforces* the incentive to pollute as compared to the model with standard preferences.

Note 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 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 U_i, given that i's expectations about how his opponents will play the game are considered fixed.[11] Let Q be the total number of countries that abate. Then we have the following result:

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. ■

The above result is illustrated in Figure 2. $\hat{Q}_{-i}$ represents a tipping point in the model. 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. Conversely, if fewer than $\hat{Q}_{-i}$ others abate, the reciprocal benefits from abatement are too small to make abatement individually rational. Although the symmetric one-shot climate game is a Prisoners' Dilemma in material payoffs, it consequently becomes a coordination game in utilities.[12]

INSERT FIGURE 2 ABOUT HERE

Using eq.(11), 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

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

while if $Q = N$, we have

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

[11] Segal and Sobel (2007) demonstrated that many psychological games – including the one studied by Rabin (1993) – can be reformulated by assuming that players have preferences over *strategies* rather than beliefs, and developed solution concepts for such games.

[12] The last sentence of the Proposition may be somewhat surprising. If everyone has the same preferences and they still use two different pure strategies in Nash equilibrium, everyone must be indifferent between the two pure strategies. In the present game, this is not possible because utility depends on what *others* do. See the proof for details.

3.3 What if some countries do not have reciprocal preferences?

If only some countries are reciprocal, $Q = N$ cannot be a Nash equilibrium. A high abatement Nash equilibrium, in which a majority of countries abate, may still exist.

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, if

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

and

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

there are two pure strategy Nash equilibria in the non-cooperative abatement game, represented by $Q = 0$ and $Q = A$, respectively.

Proof. See the Appendix. ■

Note that Proposition 2 requires that a strict majority of countries are reciprocal. As demonstrated above, the tipping point $\hat{Q}_{-i}$ is always strictly larger than a majority; hence, if less than half are reciprocal, the tipping point cannot be reached. Furthermore, each reciprocal country must have an even stronger preference for reciprocity than what was required for the full abatement equilibrium in Proposition **1.**

4 Coalition participation with reciprocity

Let us 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.

4.1 The standard preferences case

Consider first the standard case where each country maximizes its own payoff π_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, see Wagner (2001). 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 $\Pi_s(k)$ denote the material payoff of a signatory country as a function of the number of signatories k. Similarly, let $\Pi_n(k)$ denote the material payoff of a non-signatory country 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, country i will not prefer to form an abating coalition on its own (if it is at all meaningful to speak of a coalition of one). Thus, $k = 0$, the coalition of zero members, 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 exactly size k^0 is stable (Barrett 2003, Ch. 7.6): a country i will join, and stay, because at this coalition size its participation is required to make the other signatories abate, which they will do only if $c/b \leq k < c/b + 1$.[13]

The implication is, unfortunately, that coalition formation can improve the sum of countries' payoffs only slightly compared to the non-cooperative outcome. This is illustrated in Figure 3.

INSERT FIGURE 3 ABOUT HERE

If $k^0 = c/b$, the coalition will provide no net benefits at all to its members compared to the no abatement case, since their environmental benefits exactly outweigh their abatement costs. There will still be a net benefit to non-members, who free-ride on the coalition's efforts. A member of a coalition of size k^0 knows, however, that if it leaves, the coalition collapses (does not abate); hence, the relevant alternative for a signatory country is the payoff it gets by polluting if *no-one* else abates, which is zero. If $k^0 > c/b$, coalition members will strictly prefer staying in the coalition to leaving.

[13] If 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 outweighted by the corresponding benefits to s, b.

4.2 Defining reciprocity in the three-stage game

Assume now that every country i has reciprocal preferences as given by eq. (1) 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

$$\sum_{s \in S} u_s = \sum_{s \in S} (\pi_s + \alpha R_s) \tag{14}$$

with respect to q_s for every signatory $s \in S$. Assume that the coalition always chooses the same abatement strategy q_s for every member s.

A strategy σ_i for country i now consists of a plan, for any given beliefs about others' strategies, 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.

In the three-stage coalition game, the impact i's actions has on j's payoff may depend on others' strategies. Thus, kindness, reciprocity and utility could also depend on others' strategies. The specification of reciprocity thus becomes more complicated, and must in principle be calculated conditionally on the set of expectations concerning everyone's strategies.

The kindness function (4) depends not just on what i tries to secure to j, but also on what i *could have* secured to j. In most situations in the three-stage game, i's power to change others' payoff is just as limited as it was in the non-cooperative game: then, i can effectively influence any other country j only through its abatement (pollution) choice, increasing (not increasing) j's payoff by b. The important exception to this rule occurs when i is pivotal for a coalition's willingness to abate. Then, i's power is considerably larger, since it can secure or cancel out abatement efforts not just from itself, but from many others as well.

It turns out, however, that when checking for the potential stability of coalitions below, the expression for reciprocity can be simplified in a manner rather similar to the expression used in the non-cooperative game. One important reason is that the normalization of kindness by $\pi_{ij}^{\max} - \pi_{ij}^{\min}$ adopted from Rabin (1993) essentially makes the kindness measure a relative one. In the present game, this implies that the kindness difference between joining and not joining for a pivotal country is just like the kindness difference between abating and not abating in the non-cooperative game.

Consider the case where i expects $k-1$ other countries to join the coalition, and i expects to be pivotal in the sense that given everyone else's strategy and beliefs, i's participation is required for the coalition to abate. In that case, if i joins, every signatory gets a payoff of $bk - c$, while every non-signatory gets a payoff of bk; if i does not join, everyone gets a payoff of 0. In the results presented below, these calculations are taken into account; for details, see the proofs in the Appendix.

4.3 Stable coalitions with reciprocity

Let $U_s(q_s, k)$ denote the utility of a signatory country as a function of the coalition's abatement policy for each of its members q_s and the number of signatories k. Similarly, let $U_n(q_n, k)$ denote the utility of a non-signatory country n as a function of its own abatement choice q_n and the number of signatories k.[14]

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

Definition 3 *A coalition of size k is internally stable if $U_s(q_s, k) \geq U_n(q_n, k-1)$, and expectations are correct in the sense that every $s \in S$ expects $k-1$ other countries to be signatories, while every $n \notin S$ expects k other countries to be signatories.*

Definition 4 *A coalition of size k is externally stable if $U_n(q_n, k) \geq U_s(q_s, k+1)$, and expectations are correct in the sense that every $s \in S$ expects $k-1$ other countries to be signatories, while every $n \notin S$ expects k other countries to be signatories.*

The result below establishes that the empty coalition is stable. Intuitively, if no-one joins, there are no signatories in Stage 3, which means that everyone plays non-cooperatively. Consequently, we can use the analysis from the non-cooperative case, including its specification of reciprocity. Proposition 1, part i, demonstrated that zero abatement is a Nash equilibrium in the non-cooperative case. Expecting this outcome, countries have incentives neither to join nor to abate.

Proposition 5 *The no-cooperation solution $k = 0$, 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 6 *If $\alpha > 2(c-b)$, the grand coalition (the coalition abates, and $k = N$) is stable.*

Proof. See the Appendix. ■

If non-signatories are expected to abate in Stage 3, it is straightforward to see that the coalition will prefer to Abate in Stage 2 too: By doing so, the coalition can achieve the Pareto superior outcome in which everyone abates. Since this corresponds to a Nash equilibrium in the non-cooperative case, countries expecting such behavior from others will prefer to join the coalition in Stage

[14] Note the distinction to the notation $u(q_i, Q_{-i})$ from the non-cooperative case: while $u(q_i, Q_{-i})$ gives i's utility as a function of i's own and others' behavior, $U_m(q_m, k)$ is a different function depending on whether $m = s$ or $m = n$, where m is i's coalition membership status; moreover, the second variable of $U_m(q_m, k)$ is the number of coalition members, which may or may not correspond to the number of others abating, Q_{-i}.

1.[15] Even if non-signatories are expected to pollute in Stage 3, a coalition with k signatories may prefer to abate in Stage 2. It will do so if this provides net benefits to its signatories, that is, if $U_s(1,k) \geq U_s(0,k)$. In Stage 1, a country joins if this gives higher utility than not joining; that is, if $k-1$ others are expected to join, i joins if $U_s(1,k) \geq U_n(0,k-1)$. Thus, even if non-signatories are expected to pollute, the grand coalition may form – provided that a large enough number of countries are expected to join.

The above demonstrated that with sufficiently strong reciprocity, extremely cooperative as well as extremely uncooperative outcomes may be feasible. The proposition below establishes that an intermediate case is also possible. In addition to the two stable coalition sizes $k=0$ and $k=N$, in which all countries act identically, there can be a third stable, but small coalition size, $k=k^1 \geq k^0$, in which the k^1 signatories abate, while the $N-k^1$ non-signatories pollute.

This situation resembles the small, stable coalition size k^0 from the payoff-maximizing countries case. If the number of countries is not too small, and the cost-benefit ratio is modest, the stable minority coalition of k^1 countries exists regardless of the level of α. If α becomes sufficiently small, k^1 coincides with k^0.

Proposition 7 *Assume that $N > 13$ and that $c/b \leq (N+2)/3$. Then, there exists an externally and internally stable coalition consisting of k^1 countries, such that $\frac{N-1}{2} > k^1 \geq k^0$, for which the coalition abates while non-signatories pollute, and where k^1 is defined as the smallest integer such that $k^1 \geq \underline{k}$, where $\underline{k} = \frac{2c(N-1)+\alpha(N+2)}{2b(N-1)+3\alpha}$.*

Proof. See Appendix B. ■

The two stable coalitions of zero and N countries correspond to the two pure strategy Nash equilibria in the non-cooperative game. The stable coalition size k^1 is new to the coalition participation game.

Although k^1 is weakly increasing in α, it always consists of a strict minority of countries. k^1 is a relatively small coalition size, consisting of a minority of the countries. In other words, k^1 is always strictly smaller than the tipping point $\hat{Q}_{-i}$.

The intuitive reason for this is as follows. If a large enough number of countries are expected to become signatories, a 'snowball' effect may arise.[16] If a sufficiently large majority coalition of $k < N$ arises, it is not externally stable: everyone would prefer to join, even the current $N-k$ non-signatories. If they did join, the coalition would develop into the grand coalition. Thus, once a coalition becomes sufficiently large, it will develop into the grand coalition.

The stable coalition size k^1 is of a different kind. It is the coalition one would end up with if initially, a relatively large number of countries are expected to

[15] More precisely, they will be indifferent between being signatories and non-signatories. However, if they stay outside, they will abate and thus behave, in terms of abatement as well as kindness, exactly as if they were part of the coalition.

[16] I am now speaking as if the model were dynamic; it is clearly not, so the following explanation must be taken only as a hint to understand the mechanism at hand.

join, but too few to make the 'snowball' effect start working towards the grand coalition. Countries would then find that the benefits of joining, including reciprocal benefits, do not justify its costs, until no more than k^1 expected signatories are left. Like k^0 in the standard preferences case, k^1 is sustained only because at this level of participation, the coalition is just on the verge of collapsing. If the coalition becomes slightly larger, the free-rider problem kicks in, making at least one country prefer to leave. Reciprocal preferences do not automatically eliminate the free-rider problem.

INSERT FIGURE 4 ABOUT HERE

This is illustrated in Figure 4. $\underline{k}$ is the lowest k for which a country is indifferent between being a polluting non-signatory, given that everyone else pollutes too, and being a signatory in an abating coalition of k members. 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, a signatory considering to leave cannot take others' abatement as given, because if it leaves, the coalition will collapse.

When $k = k^1$, no-one really wants to abate: not only is it materially unprofitable, but everyone would also like to punish the others for their polluting behavior. However, although both material free-rider incentives and reciprocity preferences would speak for pollution, there is, after all, a small group of $k^1 - 1$ others who are behaving nicely to each other (and to everyone else). They are too few to make you want to be nice yourself. Still, they do 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, and there will be only meanness left in the world.

It is not obvious that results would be the same with a different kindness measure. One may want to apply a less relative and more absolute measure, one may want to take i's own sacrifice into account when considering how kind i is, or importantly, one may want to explore the effects of different countries having different ideas of kindness. I leave this, however, for future research.

4.4 Coalition participation if some countries are not reciprocal

Finally, consider the case where only some countries are reciprocal. In this case, the grand coalition is not feasible, but there may still exist stable, abating coalitions of strictly positive size. In fact, although k^0 was generally not stable in the case with only reciprocal countries, it is now possible that k^0 as well as k^1 can be stable, in addition to $k = 0$ and, if reciprocity preferences are sufficiently strong and widespread, $k = A$.

Assume, like in Proposition 2, 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 $\alpha_i = \alpha$, and let $N - A$ be the number of countries with $\alpha_i = 0$.

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

If reciprocity preferences are too weak and/or the number of reciprocal countries is too small, no such majority coalition will be stable. Even if a stable coalition $k = A$ does exist, it will not necessarily be realized, since other, smaller coalition sizes are stable too. In particular, the no participation coalition is always stable (see the proof for Proposition 5).

If the assumptions for Proposition 7 hold, there will exist a stable minority coalition size $k^1 \geq k^0$ consisting of reciprocal countries. This holds whether the coalition of $k = A$ is stable or not.

However, when only some countries are reciprocal, a coalition size of k^0 can also be stable – provided that it consists of *non-reciprocal* countries. Reciprocal countries are not signatories to such a coalition, because they are too angry. Given that no reciprocal countries are expected to participate, however, a small coalition of payoff-maximizing countries can be stable, according to exactly the same reasoning as in the standard preferences case.

The above has implicitly assumed that preferences are common knowledge. If they are not, countries can have a strategic interest in misrepresenting their true preferences. This could make coordinating on a Pareto superior equilibrium substantially more difficult. Knowledge of the distribution of preferences is required to know which equilibria exist at all. If preferences are public information, a country may be hesitant to abate because it does not know which equilibrium others are trying to coordinate on. If preferences are private information, this potential cause of coordination failure will persist, and in addition one does not know which potential equilibria are there at all. If too few countries believe that a high abatement equilibrium exists, too few will try to coordinate on it. Similarly, if too few expect others to believe that a high abatement equilibrium exists, too few will try to coordinate on it. If countries are expected to strategically misrepresent their true preferences, allowing communication may not be sufficient to secure coordination.

5 Conclusions

In this paper, I have considered the role of reciprocal preferences in a simple climate treaty participation game.

If countries play non-cooperatively, the situation where all countries pollute is always a Nash equilibrium. In this situation, unwillingness to abate is even stronger than in the case with standard preferences. When no-one abates, countries are angry with each other to the extent that, in addition to the economic cost, abating would give them displeasure: they do not want to help others who are harming them.

With sufficiently strong reciprocity preferences, however, the situation where all countries abate is also a Nash equilibrium with non-cooperative play. When a sufficiently large number of countries abate, every other country will prefer to abate as well. If only some countries have reciprocal preferences, a Nash equilibrium in which a majority of countries abate can still exist; this requires, however, that reciprocity is sufficiently strong and widespread.

Nevertheless, regardless of how strong reciprocity preferences are, abatement can only be individually preferable if more than a majority of other countries abate. If only a minority of countries have reciprocal preferences, every country will thus end up polluting in the non-cooperative game.

In a three-stage game of coalition formation, the grand coalition consisting of all countries is stable, given that reciprocity preferences are sufficiently strong. Again, however, the situation with no cooperation is stable too.

In addition, there can exist a third stable coalition size k^1. This stable coalition consists of a minority of countries, but is weakly larger than k^0, the largest possible stable coalition in the standard preferences case. If the total number of countries is not too small, and the cost-benefit ratio is relatively small, a stable coalition size k^1 exists regardless of the strength of reciprocity preferences; as reciprocity preferences become weaker, k^1 coincides with k^0.

When $k = k^1$, no-one really wants to abate. Not only is it materially unprofitable; the average contribution of others is so low that reciprocal players would be willing to sacrifice own material benefits in order to punish others. However, being a member of a coalition of size k^1 can be thought of as a situation in which anger is exactly balanced against even more anger. There is, after all, a small group of $k^1 - 1$ others who are behaving nicely. 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 will disappear, and there will be only meanness left in the world. k^1 represents the point where each signatory is just on the limit of what it can take – and while signatories do contribute, they do so in anger.

The model presented here has been extremely simplified. Modifying it in a multitude of ways would obviously be required to make it relevant to actual international climate negotiations. Studying the implications of continuous abatement choices would be of interest; even more importantly, it is desirable to extend the model to allow more substantial heterogenity among countries. In the present model, there are no differences in incomes or emission histories, no competing fairness norms, no cultural differences, and essentially no room for misunderstandings. These matters are precisely the ones making reciprocity preferences potentially relevant to climate negotioations. With reciprocity, coordination on a better equilibrium may be prevented by mutual mistrust and anger. If a reciprocal country believes that others are behaving unfairly, this might, for example, impair not just its willingness to abate, but perhaps even its willingness to help overcome negotiation deadlocks. The model presented here, however, is too simple to study such hypoteses – which must thus be left as a topic for future research.

References

[1] Barrett, S. (1992): "International Environmental Agreements as Games", in R. Pethig (Ed.): Conflict and Cooperation in Managing Environmental Resources, Berlin: Springer, 11-37.

[2] Barrett, S. (1994): Self-Enforcing International Environmental Agreements, Oxford Economic Papers 46, 878-894.

[3] Barrett, S. (2003), Environment and Statecraft: The Strategy of Environmental Treaty-Making, Oxford: Oxford University Press.

[4] 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.

[5] 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.

[6] Camerer, C. (2003): Behavioral Game Theory. Experiments in strategic interaction, Princeton University Press/Russell Sage Foundation.

[7] Carraro, C., and D. Siniscalco (1993): Strategies for the International Protection of the Environment, Journal of Public Economics 52, 309-328.

[8] Conconi, P., and C. Perroni (2002): Issue linkage and issue tie-in in multilateral negotiations, Journal of International Economics 57, 423–447.

[9] Cox, J.C., Friedman, D., and S. Gjerstad (2007): A tractable model of reciprocity and fairness, Games and Economic Behavior 59(1), 17-45.

[10] Croson, R., 2007. Theories of commitment, altruism and reciprocity: evidence from linear public goods games. Economic Inquiry 45, 199–216.

[11] Croson, R., Fatas, E., Neugebauer, T., 2006. Reciprocity, matching and conditional cooperation in two public goods games. Economics Letters 87, 95–101.

[12] 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.

[13] Dufwenberg, M., and G. Kirchsteiger (2004): A Theory of Sequential Reciprocity, Games and Economic Behavior 47(2), 268-98.

[14] Falk, A., E. Fehr, U. Fischbacher (2003): On the Nature of Fair Behavior, Economic Inquiry 41(1), 20-26.

[15] Falk, A., and U. Fischbacher (2006): A Theory of Reciprocity, Games and Economic Behavior 54, 293-315.

[16] 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.

[17] Fehr, E., and S. Gächter (2000): Fairness and Retaliation: The Economics of Reciprocity, Journal of Economic Perspectives 14(3), 159-181.

[18] Fehr, E., and S. Gächter (2002): Altruistic Punishment in Humans, Nature 415, 137-140.

[19] Fehr, E., and K. Schmidt (1999): A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics 114, 817-868.

[20] Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71, 397–404.

[21] Fischbacher, U., Gächter, S., 2006. Heterogeneous social preferences and the dynamics of free riding in public goods. IZA Discussion Papers 2011.

[22] Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of freeriding in public goods. American Economic Review 100, 541–556.

[23] Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397–404.

[24] Froyn, C.B., and J. Hovi (2008): A climate agreement with full participation, Economics Letters 99, 317–319.

[25] 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).

[26] 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.

[27] 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.

[28] Herrmann, B., and C. Thöni (2009): Measuring conditional cooperation: A replication study in Russia. Experimental Economics, 12, 87–92.

[29] Hoel, M.O. (1992): International Environment Conventions: The Case of Uniform Reductions of Emissions. Environmental and Resource Economics 2, 141-159.

[30] Hoel, M.O., and K. Schneider (1997): Incentives to Participate in an International Environmental Agreement, Environmental and Resource Economics 9(2), 153–170.

[31] 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.

[32] Kocher, M.G., T. Cherry, S. Kroll, R.J. Netzer, M. Sutter (2008): Conditional cooperation on three continents, Economics Letters 101, 175–178.

[33] Kolstad, C.K. (2013): International Environmental Agreements with Other-Regarding Preferences, unpublished paper, Stanford University.

[34] 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.

[35] Lange, A. (2006): The Impact of Equity-Preferences on the Stability of International Environmental Agreements, Environmental and Resource Economics 34, 247-267.

[36] Lange, A., and C. Vogt (2003): Cooperation in International Environmental Negotiations due to a Preference for Equity, Journal of Public Economics 87, 2049-2067.

[37] 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.

[38] Levine, D.K. (1998). Modeling Altruism and Spitefulness in Experiments, Review of Economic Dynamics 1, 593-622.

[39] Martinsson, P., N. Pham-Khanh, C. Villegas-Palacio (2013): Conditional cooperation and disclosure in developing countries, Journal of Economic Psychology 34, 148–155.

[40] Nyborg, K. and M. Rege (2003): Does Public Policy Crowd Out Private Contributions to Public Goods? Public Choice 115 (3), 397-418.

[41] Ostrom, E. (1990): Governing the Commons: The Evolution of Institutions for Collective Action, Cambridge: Cambridge University Press.

[42] Rabin, M: Incorporating Fairness into Game Theory and Economics, American Economic Review 83, 1281-1302.

[43] Segal, U., and J. Sobel (2007): Tit for tat: Foundations of preferences for reciprocity in strategic settings, Journal of Economic Theory 136, 197-216.

[44] Sobel, J. (2005): Interdependent Preferences and Reciprocity, Journal of Economic Literature 43, 392-436.

[45] Tavoni, A., A. Dannenberg, G. Kallis, A. Löschel (2011): Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proceedings of the National Academy of Sciences of the United States of America, 108 (29), 11825-11829.

[46] Thöni, C., J. -R. Tyran, and E. Wengström (2012): Microfoundations of social capital, Journal of Public Economics 96(7-8), 635-643.

[47] van der Pol, T., H.-P. Weikard, E. van Ierland (2012): Can altruism stabilise international climate agreements? *Ecological Economics* 81, 112-120.

[48] 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. (11), $u_i(0,0) \geq u_i(1,0)$ 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 (**??**), 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 $Q_{-i} \lesseqgtr \hat{Q}_{-i}$, 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. (13), 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 $p = \frac{1}{2} - \frac{c-b}{\alpha}$). 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. 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.

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 $Q = A$, $Q_{-i} = A-1$. By eq. (12), abate is preferred by i when $Q_{-i} = A-1$ if

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

or equivalently,

$$\alpha \geq 2(c-b)\frac{N-1}{2A-N-1}. \tag{15}$$

This is feasible given that $2A - N - 1 > 0$, i.e.

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

As long ■

Proof of Proposition 5:

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, so that a "coalition" consisting of 1 country came into existence, such a 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 $U_n(0,0) \geq U_s(1,1)$, 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 6:

Proof. Assume $k = N$. Then in Stage 3, there are by assumption no non-signatories.

In Stage 2, the coalition of N countries prefers to Abate if $U_s(1,N) \geq U_s(0,N)$. Consider first the case where a coalition of size $k = N - 1$ would prefer to abate. No individual signatory is then pivotal in the sense that its participation is decisive for the coalition's policy, and every signatory's kindness can be expressed as in the non-cooperative case, by eq. (11). Thus, the coalition will abate if

$$bN - c \geq -\alpha$$

which always holds since, by assumption, $bN - c > 0$ and $\alpha > 0$.

In Stage 1, country i will then join if, given the expectation that everyone else joins, it can do no better than joining. Proposition 1, part ii) demonstrates that if $N - 1$ others abate and countries play non-cooperatively, then country i can do no better than abate too, given that $\alpha > 2(c - b)$ (which is assumed in the current Proposition). Hence, with the expectation that $N - 1$ others join and the coalition abates, country i can do no better than abating too, which is equivalent to joining in Stage 1.

What if a coalition of size $N - 1$ is not expected to abate in Stage 2? Every individual signatory i would then be pivotal in the sense that given everyone else's strategy and beliefs, its participation is decisive for the coalition's policy. In that case, if i joins, everyone else gets a payoff of $bN - c$, while if it does not join, everyone else gets a payoff of 0. The equitable payoff is then, due to eq. (5),

$$\pi_{ij}^{e} = \frac{1}{2}(bN - c) \tag{16}$$

and according to eq. (4), i's kindness if joining is given by

$$f_{sj} = \frac{(bN - c) - \frac{1}{2}(bN - c)}{(bN - c)} = \frac{1}{2} \tag{17}$$

and if not joining

$$f_{nj} = \frac{0 - \frac{1}{2}(bN - c)}{(bN - c)} = -\frac{1}{2}$$

which means that even a pivotal country's kindness is given by eq. (8), i.e. $f_{ij} = q_i - \frac{1}{2}$, and utility can be expressed by eq. (11). Hence the grand coalition, if it exists, will abate in Stage 2. The rest of the analysis above thus goes through as before.

Note that country i will be indifferent between being a signatory and being a non-signatory that abates. Thus, any situation in which a share x of the N countries are signatories to an abating coalition and a share $1 - x$ are non-signatories who abate is also stable. However, $x < 1$ would not affect the coalition's decision to abate in Stage 2 (due to Proposition 1, part ii), hence any situation in which $x < 1$ is equivalent to the case where $x = 1$ both in terms of outcomes and utilities.

The above establishes internal stability. External stability is not an issue here, since no country is a non-signatory and a coalition of $N + 1$ countries is not feasible. ■

Proof of Proposition 7:

Proof. In Stage 3, non-signatories play non-cooperatively and thus have the same influence on 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. (8): $f_{ij} = q_i - \frac{1}{2}$.

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. Hence, for a non-pivotal signatory i, kindness to any other country j is also given by $f_{ij} = q_i - \frac{1}{2}$ (where q_i is determined by the coalition's policy).

Consider now the case where a coalition of k members is abating, and where, given everyone's strategies and beliefs, the loss of one member would have made the coalition pollute. Every individual signatory i is then pivotal in the sense that given everyone else's strategy and beliefs, i's participation is decisive for the coalition's policy in Stage 2. Assume further that non-signatories are expected to pollute in Stage 3. In this case, if i joins, every other signatory gets a payoff of $bk - c$, while every non-signatory gets a payoff of bk. If i does not join, everyone else gets a payoff of 0. The equitable payoff for other signatories would then, according to eq. 5, be

$$\pi_{is}^{e} = \frac{1}{2}(bk - c) \tag{18}$$

and for non-signatories

$$\pi_{in}^{e} = \frac{1}{2}(bk) \tag{19}$$

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

$$f_{ss} = \frac{(bk - c) - \frac{1}{2}(bk - c)}{(bk - c)} = \frac{1}{2} \tag{20}$$

and if not joining

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

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

$$f_{sn} = \frac{bk - \frac{1}{2}bk}{bk} = \frac{1}{2} \tag{21}$$

and if not joining,

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

Consequently, even for a pivotal signatory to an abating coalition, kindness can be expressed as $f_{ij} = q_i - \frac{1}{2}$. As a result, the reciprocity function (eq. 10) and utility function (eq. 11) can be applied as before.

In the situation described in the Proposition, non-signatories pollute in Stage 3. For a signatory, we will thus have $Q_{-i} = k - 1$ if the coalition abates and $Q_{-i} = 0$ if the coalition pollutes. A coalition of $k < N$ members will abate in Stage 2 if $U_s(1, k) \geq U_s(0, k)$.

Using eq. (11), this implies

$$\begin{aligned} bk - c + \frac{3}{2}\alpha(\frac{k-1}{N-1} - \frac{1}{2}) &\geq -\frac{1}{4}\alpha \\ k &\geq \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha} \end{aligned} \tag{22}$$

Define $\underline{k}$ as the coalition size making the coalition exactly indifferent between polluting and abating in Stage 2, i.e. $U_s(0, \underline{k}) = U_s(1, \underline{k})$, or

$$\underline{k} = \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha} \tag{23}$$

The coalition will abate in Stage 2 if $k \geq \underline{k}$. k^1 is defined as the smallest integer such that $k^1 \geq \underline{k}$. Thus, in Stage 2, a coalition of k^1 countries will abate, but a coalition of $k^1 - 1$ will not.

In Stage 1, a country will join if, given the expectation that $k - 1$ others join, it can do no better than joining; that is, $U_s(1, k) \geq U_n(0, k - 1)$. Consider a country that expects $k^1 - 1$ others to join. Since the coalition will abate when $k = k^1$, the utility of each signatory if it joins is

$$U_s(1, k^1) = bk^1 - c + \frac{3}{2}\alpha(\frac{k^1 - 1}{N-1} - \frac{1}{2}). \tag{24}$$

If the country does not join, the coalition will consist of $k^1 - 1$ signatories and will not abate, and each non-signatory's utility is

$$U_n(0, k^1 - 1) = -\frac{1}{4}\alpha \tag{25}$$

The country will thus join if $U_s(1, k^1) \geq U_n(0, k^1 - 1)$, i.e.

$$bk^1 - c + \frac{3}{2}\alpha(\frac{k^1 - 1}{N-1} - \frac{1}{2}) \geq -\frac{1}{4}\alpha \tag{26}$$

which is exactly the same problem as considered in eq. (22). Thus, the above inequality holds if $k^1 \geq k$, which holds by definition. That is, if i expects $k^1 - 1$ others to join, i can do no better than joining. Hence, a coalition of k^1 members is internally stable.

External stability requires that for $k = k^1$, no non-signatories want to join. The coalition abates regardless of whether $k = k^1$ or $k = k^1 + 1$. A country that expects k^1 others to join will join if $U_s(1, k^1 + 1) \geq U_n(0, k^1)$. Using eq. (11), this would imply

$$\begin{aligned} b(k^1+1) - c + \frac{3}{2}\alpha(\frac{k^1}{N-1} - \frac{1}{2}) &\geq bk^1 + \frac{1}{2}\alpha(\frac{k^1}{N-1} - \frac{1}{2}) \qquad (27) \\ \alpha(\frac{k^1}{N-1} - \frac{1}{2}) &\geq c - b \end{aligned}$$

Since $c > b$ and $\alpha > 0$, the above can only hold if $\frac{k^1}{N-1} \geq \frac{1}{2}$, or $k^1 \geq \frac{N-1}{2}$. However, this cannot be the case, given the assumptions of the Proposition.

To see this, note that $\underline{k}$ can be characterized as follows. First, if $b(N+2) \geq 3c$ (or $c/b \leq (N+2)/3$), $\underline{k}$ is increasing in α:

$$\begin{aligned} \frac{\partial \underline{k}}{\partial \alpha} &= \frac{(N+2)(2b(N-1)+3\alpha) - 3(2c(N-1)+\alpha(N+2))}{(2b(N-1)+3\alpha)^2} \qquad (28) \\ & \frac{(N+2)(2b(N-1)+3\alpha) - 3((2c+\alpha)(N-1)+3\alpha)}{(2b(N-1)+3\alpha)^2} \end{aligned}$$

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

$$\begin{aligned} (N+2)(2b(N-1)+3\alpha) - 3(2c(N-1)+\alpha(N+2)) &> 0 \qquad (29) \\ b(N+2) &> 3c. \end{aligned}$$

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

$$\lim_{\alpha\to\infty} \underline{k} = \lim_{\alpha\to\infty} \frac{2c(N-1)/\alpha + (N+2)}{2b(N-1)/\alpha + 3} = \frac{N+2}{3} \qquad (30)$$

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 $\underline{k}$, the upper boundary for k^1 is $\frac{N+2}{3} + 1 = (N+5)/3$. The question is whether we can have $k^1 \geq \frac{N-1}{2}$. This is only possible if N is relatively small:

$$\begin{aligned} \frac{N+5}{3} &\geq \frac{N-1}{2} \\ 13 &\geq N \end{aligned}$$

Consequently, under the given assumptions, $k^1 < \frac{N-1}{2}$, which means that eq. (27) cannot hold. Thus, a coalition of size k^1 is internally and externally stable.

Finally, recall that k^0 is the smallest integer weakly larger than c/b. Since k^1 is the smallest integer such that $k^1 \geq \underline{k} > c/b$ (see above), we must have $k^1 \geq k^0$. ■

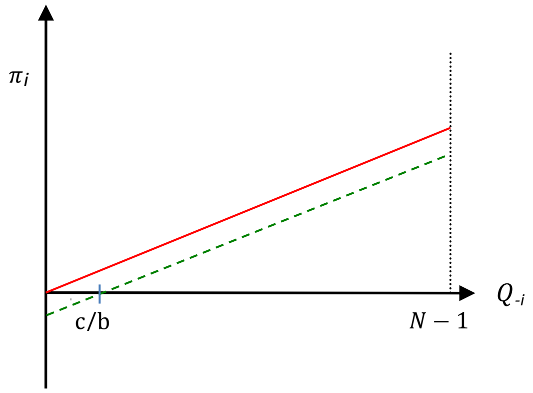

Figure 1: Payoff of country *i*, given that Q_{-i} others abate.

Red solid line: payoff if *i* pollutes.

Green dashed line: payoff if *i* abates.

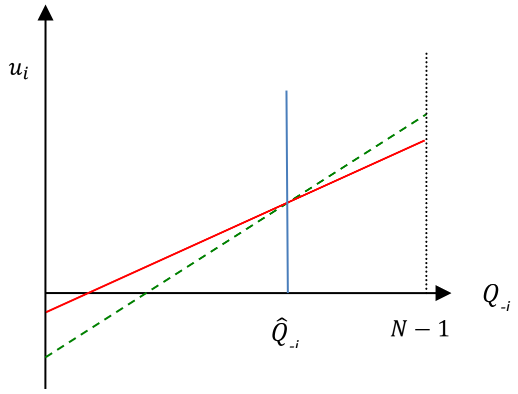

Figure 2: Utility of a reciprocal country *i*, given that Q_{-i} others abate.

Red solid line: utility if *i* pollutes.

Green dashed line: utility if *i* abates.

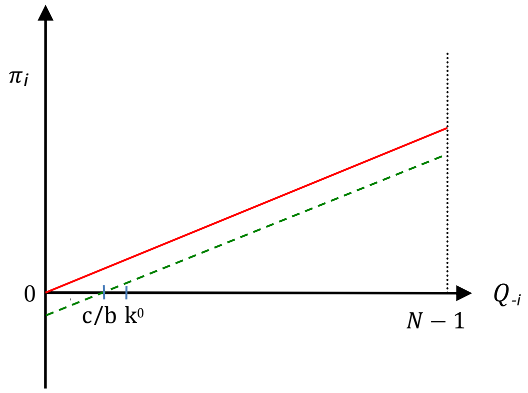

Figure 3: Payoff of country *i* (standard preferences), given that Q_{-i} others abate.

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*.

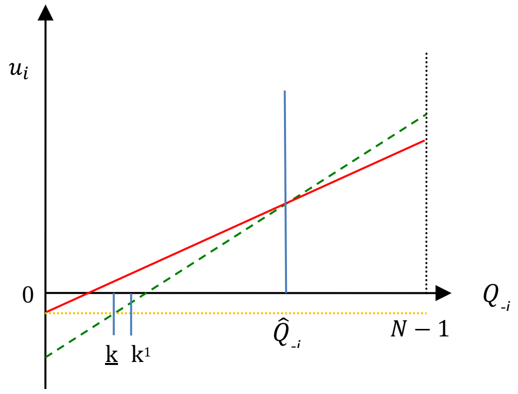

Figure 4: Utility of a reciprocal country *i*, given that Q_{-i} others abate.

Red solid line: Utility if *i* pollutes.

Green dashed line: Utility if *i* abates.

Orange dotted line: Utility if no one abates.

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

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