The B.E. Journal of Economic Analysis & Policy

Topics

Volume 12, *Issue* 1 2012 *Article* 33

International Environmental Cooperation under Fairness and Reciprocity

Costas Hadjiyiannis* Doruk İriş† Chrysostomos Tabakis‡

*University of Cyprus, hadjiyiannis.costas@ucy.ac.cy
†Universidade Nova de Lisboa, dorukiris@novasbe.pt
‡Universidade Nova de Lisboa, ctabakis@fe.unl.pt

Recommended Citation
Costas Hadjiyiannis, Doruk İriş, and Chrysostomos Tabakis (2012) "International Environmental Cooperation under Fairness and Reciprocity," *The B.E. Journal of Economic Analysis & Policy*: Vol. 12: Iss. 1 (Topics), Article 33.
DOI: 10.1515/1935-1682.2917

International Environmental Cooperation under Fairness and Reciprocity*

Costas Hadjiyiannis, Doruk İriş, and Chrysostomos Tabakis

Abstract

This paper explores the implications of fairness and reciprocity for self-enforcing international environmental agreements on pollution abatement. Reciprocal countries reward fair behavior (positive reciprocity), but retaliate against countries behaving unfairly (negative reciprocity). We demonstrate that reciprocal countries that have moderate expectations from each other with respect to their national abatement strategies can support a greater degree of environmental cooperation than self-interested ones. However, when only very high abatement standards are deemed fair, then reciprocity could have a detrimental effect on international environmental cooperation. Our model therefore provides a novel perspective on the role of expectations in environmental negotiations.

KEYWORDS: reciprocity, environmental agreements, abatement standards, repeated games

*We would like to thank Luca Opromolla, Luis Santos-Pinto, José Tavares, Anastasios Xepapadeas, and participants at CRETE 2010, ETSG 2010, IAREP/SABE/ICABEEP 2010, International Forum "Towards Global Agreements on Environmental Protection and Sustainability: Frontiers of Future Economic Research," QED Jamboree 2010, UECE Lisbon Meetings 2010, and seminars at Cardiff Business School, ISCTE, ISEG, Istanbul Bilgi University, Izmir University of Economics, and Universidade Nova de Lisboa for very helpful comments and suggestions. A previous version of the paper was written while ?ri? was visiting UC San Diego, and Tabakis was an International Faculty Fellow at the MIT Sloan School of Management. Iri? and Tabakis gratefully acknowledge financial support from the Fundação para a Ciência e a Tecnologia in Portugal (PTDC/ECO/65284/2006 and PTDC/ECO/71508/2006, respectively). Any errors are ours.

1 Introduction

Many environmental problems are transboundary in nature and often have a global scope (e.g., climate change or marine pollution). Uncoordinated environmental policy-making at the national level will not internalize the transboundary effects, and thus, countries have sought to sign international environmental agreements (IEAs).

In this paper, we examine the ramifications of reciprocal preferences for IEAs. A government with reciprocal preferences rewards "fair" actions (positive reciprocity), while it punishes "unfair" behavior (negative reciprocity), where fairness (or unfairness) is defined relative to a reference level regarding other countries' environmental policies.[1] Basically, a reciprocal government has preferences over both outcomes and strategies.[2] The significance of this approach is twofold. First, governments themselves claim to have such preferences with regard to environmental policy. Second, our analysis provides important insights into the role of expectations in international environmental negotiations.

Typically, governments stress the importance of fairness in sharing the burden of environmental protection. In fact, it is clearly stated in the Copenhagen Accord, the outcome of the December 2009 United Nations (UN) Climate Change Conference in Copenhagen, that its endorsers "shall...on the basis of equity...enhance [their] long-term cooperative action to combat climate change."[3] In a similar spirit, France at the onset of the Copenhagen summit, in response to India's commitment to reduce emissions, expressed "her determination to work with India and all her partners to put together an ambitious, just and balanced agreement in Copenhagen," while President Obama at a press conference at the end of the summit acknowledged that it is not fair to expect developing countries like China and India to be bound by the same set of legal obligations as developed countries in the fight against

[1] It is important to underscore here that fairness in our context is entirely subjective. It depends only on governments' perceptions and should not be confused with what is objectively or ethically fair. On a different note, the "reference level" is a concept widely studied in the behavioral economics literature. For more on this, see, for instance, Helson (1964) and Tversky and Kahneman (1991).

[2] We should note here that reciprocity is distinct from altruism and inequity aversion. An altruist is willing to give up own resources in order to increase the welfare of others *without* expecting reciprocation. On the other hand, an inequity averse individual is willing to forgo some material payoffs in order to achieve more equitable outcomes, i.e., an inequity averse individual wishes to benefit or hurt another person if this reduces the inequity between them.

[3] See http://unfccc.int/resource/docs/2009/cop15/eng/11a01.pdf.

climate change.[4] Moreover, fairness considerations appear to be central to the UN Environment Programme Medium-term Strategy 2010–2013 and to the official environmental policy agendas of most countries, including the European Union (EU) and the US.[5]

At the same time, governments emphasize reciprocity equally strongly. As a matter of fact, reciprocity considerations seem to have played a pivotal role in the disappointing outcome of the Copenhagen summit. In particular, most analysts agree that the principal reason for the failure of the summit was the disagreement between countries—especially between the US and the BASIC countries (i.e., Brazil, South Africa, India, and China) led by China—on how to share the burden of reducing greenhouse gas emissions. Essentially, the negotiations in Copenhagen appear to have revolved around fairness and reciprocity, with countries rejecting the various deals proposed as unfair or one-sided; and when the talks were concluded, participants started accusing each other of a total lack of willingness to compromise or reciprocate. For instance, the UK's Prime Minister Gordon Brown argued after the summit that, "If America and China were able to show they were doing more and I believe that they can, then all countries—Australia, Brazil, Japan, Korea—all those countries that have ranges [of emission cuts] would be prepared to go to their highest level of ambition."[6] In brief, in the international environmental arena, governments assert that they are not simply maximizers of their own self-interested welfare, but that they are also motivated by fairness and reciprocity considerations. One could, of course, argue that such assertions are not sincere, but are rather made for purely strategic reasons (as suggested by Lange et al., 2010). However, even though it is very hard to disentangle rhetoric from reality, it is still important to analyze the implications of reciprocal preferences

---

[4] See, respectively, http://www.ambafrance-uk.org/France-welcomes-India-s-climate.html and http://www.whitehouse.gov/the-press-office/remarks-president-during-press-availability-copenhagen.

[5] For the UN Environment Programme Medium-term Strategy 2010–2013, see http://www.unep.org/PDF/FinalMTSGCSS-X-8.pdf. For the EU and US environmental policy agendas, see http://ec.europa.eu/environment/integration/development_en.htm and http://www.usaid.gov/our_work/environment/climate/index.html, respectively.

[6] See http://news.bbc.co.uk/2/hi/uk_news/politics/8423831.stm. In addition, early in the negotiations, one of India's top negotiators expressed "concern...that [India has] been offering unilateral concessions, without obtaining any reciprocity," while after the summit, the G77 group of 130 developing nations blamed President Obama for "locking the poor into permanent poverty by refusing to reduce US emissions further" (see, respectively, http://indiatoday.intoday.in/site/Story/73850/Top%20Stories/Copenhagen+summit+braces+for+a+'no+deal'.html and http://www.guardian.co.uk/environment/2009/dec/19/copenhagen-blame-game).

for IEAs, as the existence of such preferences among governments cannot be dismissed.

It could also be argued that it is reasonable to expect governments to have reciprocal preferences towards environmental policy since such preferences are exhibited by voters. In particular, the behavioral economics literature offers extensive evidence of both positive and negative reciprocity in individual decision making. For example, experiments asking individuals to contribute to public goods typically find that their contributions far exceed what self-interested utility maximization would entail (e.g., Andreoni, 1988; Palfrey and Prisbrey, 1997; Croson, 2007), which is usually interpreted as evidence of positive reciprocity. Analogous results arise from trust or gift-exchange experiments (e.g., Berg et al., 1995; Fehr et al., 1997; Fehr et al., 1998). On the other hand, evidence for negative reciprocity is found in ultimatum-game experiments with the typical result being that people reject offers that would be accepted under the self-interested hypothesis (e.g., Güth et al., 1982; Roth et al., 1991). Further evidence of both positive and negative reciprocity among individuals is provided by Dohmen et al. (2009), who use survey data. Public opinion polls corroborate these findings, suggesting that a significant proportion of people do exhibit reciprocal preferences with respect to different global issues, including international environmental cooperation against climate change.[7] Political-economy models of environmental policy then suggest that if voters have such preferences, these preferences will be reflected in the objective function of governments. In fact, this would be the case within the context of a median-voter model (e.g., Eriksson and Persson, 2003; McAusland, 2003) as well as an interest-group model à la Grossman and Helpman (1994) (e.g., Fredriksson, 1997; Aidt, 1998).

To address the ramifications of reciprocal preferences for IEAs, we develop a dynamic game in which reciprocal countries facing a free-riding Prisoner's Dilemma problem in their dealings with one another attempt to maintain cooperation in their national pollution-abatement strategies, where pollution is assumed to be transboundary in nature. Given the lack of a supranational authority with effective enforcement mechanisms regarding environmental policy, we restrict our attention to IEAs that are self-enforcing, as in Ferrara et al. (2009). In this context, a country will choose today to adhere to the cooperative path as long as the onetime gain it could achieve by

[7] See, for instance, http://www.americans-world.org/digest/global_issues/global_warming/gw3.cfm and http://siteresources.worldbank.org/INTWDR2010/Resources/Background-report.pdf for US and global "evidence," respectively, of reciprocal preferences towards environmental policy among individuals.

unilaterally deviating from its agreed-upon abatement policies does not outweigh the discounted future welfare losses due to the ensuing breakdown in international environmental cooperation. It is important to stress here that we completely abstract from any participation considerations, which have occupied center stage in a significant part of the literature on IEAs. Instead, we look for the most cooperative equilibrium that can be supported by reciprocal countries within the context of a self-enforcing international agreement involving full participation, and we compare that with the most cooperative equilibrium that could be sustained by self-interested countries.[8] On the other hand, to model reciprocity, we assume that countries have preferences over both outcomes and strategies. More specifically, we assume that a country attaches a positive weight to the self-interested welfare of another given country if it expects the latter to set a higher abatement standard than the one it perceives as fair. However, if the other country in question is expected to set an unfairly low abatement standard, then a negative weight is placed on its self-interested welfare.[9]

We find that reciprocal countries that have moderate expectations from each other with respect to environmental policy (i.e., when the abatement standards considered fair are not too high) can support a higher degree of environmental cooperation and thus achieve higher welfare than can self-interested countries. The intuition underlying this result is straightforward. For such fair-abatement-standard perceptions, in the reciprocal game (i) the punitive Nash abatement standard is lower than in the self-interested game, which acts to make the punishment phase costlier in the former game; and (ii) countries are in a positive-reciprocity state. As a result, under the scenario in question, reciprocal countries are faced with both a weaker incentive to cheat and a stronger incentive to cooperate than self-interested ones, allowing them to support a "greener" equilibrium.

However, when reciprocal countries are highly demanding of each other regarding their environmental policies (i.e., when only very high abatement levels are perceived as fair), then the impact of reciprocity on international environmental cooperation is ambiguous. Intuitively, in such a case, reciprocal countries are in a negative-reciprocity state, meaning that they face a stronger incentive to defect than self-interested ones. On the other hand, their incentive

[8]Ferrara et al. (2009) also abstract altogether from any participation considerations with regard to IEAs. Moreover, note that self-enforcing agreements involving full participation are commonly employed in the literature on multilateral trade negotiations (e.g., Dam, 1970; Dixit, 1987; Bagwell and Staiger, 2002).

[9]As we discuss below, an axiomatic foundation to our modeling approach is provided in a seminal paper by Segal and Sobel (2007).

to cooperate remains stronger, due to the lower punitive Nash abatement standard in the reciprocal game. Simulations do confirm that for very high fair abatement standards, there indeed exist cases in which the stronger-incentive-to-cheat effect dominates, leading to less pollution abatement in the reciprocal equilibrium as compared with the self-interested one.

At a more general level, our findings provide novel insights into the role of expectations in international environmental negotiations. In particular, assuming countries have some preferences for fairness and reciprocity, our results suggest that if they arrive at the negotiations table with expectations that are greatly elevated (i.e., they have very high fair-abatement-level perceptions), this could prove counterproductive, in the sense that they might no longer be able to support very "green" environmental policies and could end up with an "inferior" agreement than in the absence of these high expectations. Actually, as we argue in the concluding section (Section 6), this might be one of the plausible explanations for the failure of the 2009 UN Climate Change Conference in Copenhagen and the more positive outcome of the 2010 one in Cancun. The policy implications of our analysis are then immediate. The careful management of expectations is critical for the success of international environmental negotiations, and, most important, the creation of a pre-negotiations high-expectations environment should be avoided.

Our paper contributes to the expanding literature on IEAs, excellent reviews of which are provided by Wagner (2001) and Barrett (2005). Lange and Vogt (2003) also study IEAs among countries with "nonstandard" preferences. However, their focus is on preferences for equity, and unlike here, they look at coalition formation. Hoel (1991) on the other hand investigates how altruistic unilateral efforts to reduce harmful emissions affect noncooperative as well as cooperative global environmental equilibria. Moreover, a number of papers look at the implications of a minimum participation clause—a clause stipulating that a given treaty will only come into force if it is ratified by at least a certain number of countries—for international environmental cooperation (e.g., Black et al., 1993; Carraro et al., 2004). Although aimed at participation, which we do not examine in this paper, minimum participation clauses contained in many IEAs are somewhat related to the ideas of fairness and reciprocity: a signatory country will not implement an IEA unless it is ratified by at least a critical number of countries because it is unfair that they should get away with free-riding. Finally, Conconi (2003) explores whether green lobbying can help governments internalize cross-national environmental externalities and therefore be an alternative to IEAs, which are the focus of this paper.

Last, in a companion paper—Hadjiyiannis et al. (forthcoming)—we explore the implications of reciprocal preferences for multilateral tariff cooperation, and find that expectations play a pivotal role in determining the final outcome of multilateral trade negotiations as well.[10] However, our current paper differs in two important ways. First, abatement-standard setting is clearly conceptually distinct from import-tariff setting. Second, countries' environmental policies emerge here as strategic substitutes, whereas in our companion paper, there is strategic complementarity between countries' commercial policies. This difference—strategic substitutability versus strategic complementarity of the choice variables—has ramifications for the theoretical analyses presented in the two papers.[11]

The remainder of the paper is organized as follows. Section 2 sets out the basics. Section 3 characterizes the static Nash equilibrium of our model, and Section 4 analyzes the dynamic game. Section 5 discusses some possible extensions of our framework. Finally, Section 6 offers some concluding remarks and relates the predictions of our model to the experiences of the Copenhagen and Cancun summits. All the proofs are relegated to the Appendix.

2 The Model

We assume that the world consists of two countries, A and B, which engage in a pollution-abatement game. In particular, in each period, country J unilaterally selects its abatement standard $q^J \in [0, q_{\max}]$ so as to maximize its individual welfare, where $J \in \{A, B\}$ and $q_{\max}$ is the highest feasible level of pollution abatement. Let b^J denote country J's surplus from economic activity. We assume that pollution abatement consumes a fraction of a country's resources and thus reduces b^J. Moreover, pollution is transboundary in nature, meaning that the aggregate environmental damage Ψ^J country J faces is a function of the level of its own emissions and those of country $-J$. Assuming that both $b^J\left(q^J\right)$ and $\Psi^J\left(q^J, q^{-J}\right)$ are twice continuously differentiable functions, we then have: $\left(\partial b^J/\partial q^J\right) < 0$, $\left(\partial \Psi^J/\partial q^J\right) < 0$, and $\left(\partial \Psi^J/\partial q^{-J}\right) < 0$. We further maintain the assumption that $\left(\partial^2 \Psi^J/\partial q^J \partial q^{-J}\right) > 0$, i.e., country J's marginal environmental benefit from its own abatement efforts is strictly decreasing in the abatement efforts of country $-J$. This is a reasonable as-

[10]See also İriş and Santos-Pinto (2008) for an investigation of the effects of fairness and reciprocity on firm collusion.

[11]Among other things, this difference has ramifications for the proofs presented in the two papers. For instance, unlike in our companion paper, in order to prove the existence of the static Nash equilibria in the present paper, the countries' strategies need to be properly redefined. See also footnotes 13 and 14.

sumption given our framework, since a higher q^{-J} reduces the aggregate environmental damage faced by country J and has, thus, a dampening effect on the latter's marginal abatement benefits.

The countries have preferences for fairness and reciprocity. More precisely, the welfare of country J is given by:

$$RW^J\left(q^J, q^{-J}, q_f^{-J}\right) = SW^J\left(q^J, q^{-J}\right) + \gamma w^J(q^{-J}, q_f^{-J}) SW^{-J}\left(q^J, q^{-J}\right). \quad (1)$$

The first term, SW^J, is the self-interested (or "standard") welfare function, i.e., economic surplus minus environmental damage:

$$SW^J\left(q^J, q^{-J}\right) = b^J\left(q^J\right) - \Psi^J\left(q^J, q^{-J}\right). \quad (2)$$

The second term, $\gamma w^J(q^{-J}, q_f^{-J}) SW^{-J}$, captures the fairness payoff for country J, where (i) $\gamma > 0$ is a scaling factor; and (ii) $w^J(q^{-J}, q_f^{-J})$ determines the (scaled) weight country J places on $-J$'s self-interested welfare SW^{-J}, and is of the following form:

$$w^J(q^{-J}, q_f^{-J}) \begin{cases} \in (0,1] \text{ if } q^{-J} > q_f^{-J} \\ = 0 \text{ if } q^{-J} = q_f^{-J} \\ \in [-1,0) \text{ otherwise} \end{cases}, \quad (3)$$

with q_f^{-J} being the q^{-J} country J deems "fair."

Intuitively, the weight function w^J reflects the fact that a reciprocal country cares about the other country's strategies. More specifically, if country J expects country $-J$ to set an abatement standard higher than the one it perceives as fair, then it wishes to reward $-J$, exhibiting positive reciprocity. If instead country J expects country $-J$ to behave unfairly by selecting an abatement standard below the one it considers fair, then it would like to punish $-J$, exhibiting negative reciprocity. Finally, if q^{-J} is expected to be exactly equal to q_f^{-J}, then RW^J collapses to SW^J, i.e., the reciprocal and self-interested welfare functions coincide for country J. In sum, equation (3) signifies that from country J's perspective, any q^{-J} in excess of q_f^{-J} is a fair action that should be rewarded, whereas any q^{-J} below q_f^{-J} is an unfair action that should be punished.

We maintain the assumptions that country J's weight function $w^J(q^{-J}, q_f^{-J})$ is twice continuously differentiable in both arguments, is strictly increasing in country $-J$'s abatement level q^{-J}, and is strictly decreasing in its own fair-abatement-level perception q_f^{-J}. The latter assumption implies that when country J's fair-abatement-level perception rises, country $-J$ needs to abate more in order to get "rewarded" with a positive weight in country J's objective function. In addition, the fair-abatement-standard perceptions are assumed to be common knowledge.

We should stress here that the reciprocal welfare function in equation (1) is consistent with the functional forms employed in the behavioral economics literature on fairness and reciprocity (e.g., Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). An axiomatic foundation to this specific representation of reciprocal preferences is actually provided in a seminal paper by Segal and Sobel (2007). In particular, they demonstrate that given a game and assuming that players' preferences satisfy the expected utility, continuity, independence, and self-interest axioms as defined in their paper, a player's preferences over strategies—including preferences for fairness and reciprocity—can be represented as a weighted average of the utilities from outcomes of the player in question and his opponents, where the weights depend on players' expected behavior (see their Theorem 1), with which our equation (1) is well in line.[12]

3 Static Game

Our aim in this section is to characterize the static Nash equilibrium of our model, and compare it with the one that would emerge in a game with self-interested countries. This equilibrium will serve as a credible punishment in the dynamic game explored in the next section, the threat of which can support international environmental cooperation in a repeated setting. Actually, the static Nash equilibrium would be the unique equilibrium for the dynamic game as well if an IEA were not feasible (e.g., due to exogenous, political reasons or because the governments were totally myopic and did not value the future at all).

Let $\Gamma^S(SW)$ represent the static game with self-interested countries, while $\Gamma^R(RW, w, \overrightarrow{q}_f)$ denotes the static game with reciprocal countries, where $\overrightarrow{q}_f \equiv \left(q_f^J, q_f^{-J}\right)$ is the fair-abatement-standard vector. We henceforth assume that $q_f^J = q_f^{-J} \equiv q_f$, i.e., the countries share a common fair-abatement-level perception. The reason for this assumption is twofold. First, it considerably simplifies our analysis. Second, as we discuss in Section 5, we expect that the qualitative nature of our findings would be, in general, preserved under asym-

[12] It is important to point out that the axioms in question are weaker than the assumptions on preferences in standard game theory. On a separate note, we could model country J's fairness payoff differently and make it independent of country $-J$'s self-interested welfare. As we demonstrate in the Online Appendix, in such a case, fairness and reciprocity would unambiguously have a positive impact on international environmental cooperation.

metric fair-abatement-level perceptions. In addition, in all that follows, we maintain the assumption that γ is sufficiently small, meaning that the relative weight of the fairness payoff in the countries' objective functions (or, equivalently, the relative weight the countries attach to each other's self-interested welfare) is not too high.

It is direct to show that given that γ is sufficiently small and $\left(\partial^2\Psi^J/\partial q^J\partial q^{-J}\right) > 0$, the cross-partial derivative of the welfare function of reciprocal country J with respect to its own abatement level and country J's abatement standard is strictly negative (i.e., $\left(\partial^2 RW^J/\partial q^J\partial q^{-J}\right) < 0$). In other words, the choice variables are (strict) strategic substitutes, reflecting the free-riding incentives the countries face in their dealings with one another. Furthermore, the cross-partial derivative of country J's welfare function with respect to its abatement level and its fair-abatement-standard perception is strictly negative (i.e., $\left(\partial^2 RW^J/\partial q^J\partial q_f\right) < 0$). To understand the sign of this derivative, note that $\left(\partial^2 SW^J/\partial q^J\partial q_f\right) = 0$ since by definition SW^J is not a function of q_f, while regarding the negative sign of the cross-partial derivative of country J's fairness payoff, simply recall that (i) pollution abatement by country J exerts a positive environmental externality on country $-J$, as it lowers the aggregate environmental damage faced by the latter; and (ii) ceteris paribus, a lower q_f results in a larger w^J.

As we show in the Appendix, both $\Gamma^R\left(RW, w, \overrightarrow{q}_f\right)$ and $\Gamma^S\left(SW\right)$ admit pure strategy Nash equilibria. In fact, if the countries' best-response functions have a slope strictly greater than -1, which is hereafter assumed, then $\Gamma^R\left(RW, w, \overrightarrow{q}_f\right)$ and $\Gamma^S\left(SW\right)$ have unique pure strategy Nash equilibria, which are symmetric: $\overrightarrow{q}_{NR} \equiv (q_{NR}, q_{NR})$ and $\overrightarrow{q}_{NS} \equiv (q_{NS}, q_{NS})$, respectively. Let now both $RW^J + RW^{-J}$ and $SW^J + SW^{-J}$ be strictly concave with respect to q^J. Moreover, let us assume that q_{NR} and q_{NS} are interior, and that the same applies to the reciprocal and self-interested fully cooperative abatement standards q_{OR} and q_{OS}, correspondingly (i.e., to the abatement standards maximizing, respectively, $RW^J + RW^{-J}$ and $SW^J + SW^{-J}$). It is then straightforward to demonstrate that the Nash equilibria of $\Gamma^R\left(RW, w, \overrightarrow{q}_f\right)$ and $\Gamma^S\left(SW\right)$ are characterized by an inefficiently low level of pollution abatement, i.e., $q_{OR} > q_{NR}$ and $q_{OS} > q_{NS}$, where actually $q_{OR} > q_{OS}$. The intuition behind the inefficiency of the static Nash equilibria runs as follows. In the presence of a positive cross-border abatement externality, pollution abatement is globally underprovided in the absence of an international environmental agreement, as the countries do not internalize the externality when unilaterally choosing their national abatement levels. However, under full international cooperation, the cross-border externality would be internalized by both countries and therefore, higher abatement standards would be implemented.

We next show how countries' fair-abatement-level perception affects the Nash equilibrium of $\Gamma^R(RW, w, \overrightarrow{q}_f)$.

Lemma 1 *The (pure strategy) Nash equilibrium of* $\Gamma^R(RW, w, \overrightarrow{q}_f)$, $\overrightarrow{q}_{NR} \equiv (q_{NR}, q_{NR})$, *is strictly decreasing in the fair abatement standard* q_f, *i.e.,* $(\partial q_{NR}/\partial q_f) < 0$.
Proof. *See the Appendix.* ■

Intuitively, as we argued above, for a given q^{-J}, a higher q_f leads to a smaller w^J. Consequently, as q_f rises, country J chooses its environmental policy with less of country $-J$'s interests in mind, resulting in a Nash equilibrium characterized by less pollution abatement worldwide.

The following assumption is now introduced: $q_f \geq q_{NS}$, i.e., too little pollution abatement is considered unfair, which is a reasonable assumption given our focus on environmental cooperation among countries. We finally compare the reciprocal static Nash abatement standard, q_{NR}, with the self-interested one, q_{NS}, as well as the welfare obtained in $\Gamma^R(RW, w, \overrightarrow{q}_f)$ and $\Gamma^S(SW)$.

Proposition 1 *Assume that the fair abatement standard* q_f *is higher than or equal to the self-interested static Nash abatement standard* q_{NS}. *Then, (i) the reciprocal static Nash abatement standard,* q_{NR}, *is lower than or equal to the self-interested one,* q_{NS}*; and (ii) in Nash equilibrium, the welfare obtained by reciprocal country* J, $RW^J(\overrightarrow{q}_{NR}, q_f)$, *is lower than or equal to the welfare obtained by self-interested country* J, $SW^J(\overrightarrow{q}_{NS})$, *with equality holding in both (i) and (ii) if and only if* $q_f = q_{NS}$.
Proof. *See the Appendix.* ■

Proposition 1 demonstrates that as compared with self-interested countries, reciprocal ones select lower Nash abatement standards, and thus, end up with lower welfare in Nash equilibrium.[13] The intuition underlying Proposition 1 is direct. At q_{NS}, reciprocal countries are in a negative-reciprocity state wishing to punish each other by lowering their abatement standards (since $q_f \geq q_{NS}$). Therefore, the Nash equilibrium abatement standard of $\Gamma^R(RW, w, \overrightarrow{q}_f)$ has to be lower than that of $\Gamma^S(SW)$.

Before proceeding with the dynamic game, a remark is in order. As we have already argued, in a world where environmental cooperation were infeasible, the static Nash equilibrium would be the only equilibrium that the

[13] In our companion paper on the implications of fairness and reciprocity for multilateral tariff cooperation (Hadjiyiannis et al., forthcoming), we find that the reciprocal static Nash tariff is *higher* than or equal to the self-interested one.

countries could support in the dynamic game as well. Proposition 1 then illustrates that in such a world, the impact of fairness and reciprocity on global pollution abatement would be negative. Of course, there is also a bright side to our static-game analysis. As Lemma 1 demonstrates, in such a noncooperative world, if reciprocal countries lowered their expectations from each other regarding their environmental policy (i.e., if they lowered their fair-abatement-standard perception), they would be able to reach a "greener" equilibrium, albeit still an "inferior" one as compared with the self-interested equilibrium.

4 Dynamic Game

We now study repeated interaction between the countries. More specifically, the dynamic game we consider is simply the static one analyzed above infinitely repeated. We assume that countries cannot make binding international commitments, but are instead limited to environmental agreements that are self-enforcing. In such a setting, countries can still maintain international environmental cooperation, whose degree depends critically on how severely they can credibly punish an offender. Our aim in this section is to investigate the ramifications of fairness and reciprocity for the most cooperative abatement-standard equilibrium that can be supported within the context of an IEA involving full participation (i.e., we totally abstract from any participation considerations). To do so, we characterize the most cooperative abatement-standard equilibrium of our dynamic game, and compare it with the one that could be sustained by self-interested countries.

To this end, denote the dynamic game with reciprocal countries by $\Gamma_{\infty}^{R}(RW, w, \overrightarrow{q}_f)$, and the one with self-interested countries by $\Gamma_{\infty}^{S}(SW)$. Moreover, let $\delta \in (0,1)$ denote the discount factor between periods. Given the overall symmetry of our framework, for both games we focus on symmetric cooperative subgame-perfect equilibria in which (i) along the equilibrium path, the countries implement a common cooperative abatement standard $q_C \in (q_{NS}, q_{OS}]$ in each period; and (ii) if at any point in the game a defection occurs, both countries revert from the following period onwards to the noncooperative Nash abatement standard of the (relevant) stage game. In other words, to enforce environmental cooperation, the countries employ grim-trigger strategies.

Let us begin our analysis with the dynamic game with self-interested countries, $\Gamma_{\infty}^{S}(SW)$. To derive the incentive-compatibility constraint for self-interested country J, we first look at its static incentive to cheat, Ω_S^J, which simply equals the onetime increase it achieves in welfare when it optimally cheats by choosing an abatement standard on its reaction curve, while country $-J$ still cooperates with q_C:

$$\Omega_S^J (q_C) \equiv SW^J \left(BR_S^J (q_C), q_C \right) - SW^J \left(\overrightarrow{q}_C \right) \equiv SW_D^J - SW_C^J, \tag{4}$$

where $\overrightarrow{q}_C \equiv (q_C, q_C)$ and $BR_S^J (q_C)$ is country J's best-response abatement standard to q_C. However, defection by any country leads to a permanent breakdown in international cooperation. Therefore, the discounted future welfare cost a defector faces is the discounted difference between the welfare under cooperation and the welfare in the punishment phase, given by:

$$\frac{\delta}{1-\delta} \left(SW^J \left(\overrightarrow{q}_C \right) - SW^J \left(\overrightarrow{q}_{NS} \right) \right) \equiv \frac{\delta}{1-\delta} \left(SW_C^J - SW_N^J \right) \equiv \frac{\delta}{1-\delta} \omega_S^J (q_C), \tag{5}$$

where ω_S^J is the per-period value of cooperation for country J (or, equivalently, its per-period incentive to cooperate). Thus, the incentive-compatibility condition for a self-interested country J to adhere to the cooperative path in $\Gamma_\infty^S (SW)$ is that the onetime gain from defection, Ω_S^J, does not exceed the discounted future value of cooperation, $(\delta / (1-\delta)) \, \omega_S^J$:

$$\Omega_S^J (q_C) \leq \frac{\delta}{1-\delta} \omega_S^J (q_C). \tag{6}$$

It follows from (6) that a given cooperative abatement standard q_C can be supported as a subgame-perfect equilibrium of $\Gamma_\infty^S (SW)$ as long as the countries are patient enough, or:

$$\delta \geq \delta_{q_C}^S \equiv \frac{SW_D^J - SW_C^J}{SW_D^J - SW_N^J}. \tag{7}$$

Analogous relationships hold for reciprocal countries. More specifically, the incentive-compatibility constraint for a reciprocal country J to uphold international environmental cooperation in $\Gamma_\infty^R (RW, w, \overrightarrow{q}_f)$ is given by:

$$\Omega_R^J (q_C) \leq \frac{\delta}{1-\delta} \omega_R^J (q_C). \tag{8}$$

Furthermore, reciprocal countries can support a given cooperative abatement standard q_C as long as they sufficiently value the future, or:

$$\delta \geq \delta_{q_C}^R \equiv \frac{RW_D^J - RW_C^J}{RW_D^J - RW_N^J}. \tag{9}$$

We are at this point ready to state our first result regarding the impact of fairness and reciprocity on international environmental cooperation. Using (7) and (9), we now compare $\delta_{q_C}^S$ against $\delta_{q_C}^R$.

Proposition 2 *1. Any cooperative abatement standard q_C higher than or equal to the fair abatement standard q_f can be more easily maintained by reciprocal rather than self-interested countries (i.e., $\delta^R_{q_C} < \delta^S_{q_C}$ for any $q_C \geq q_f$).*

2. However, it is ambiguous whether reciprocal or self-interested countries can more easily sustain any given cooperative abatement standard q_C lower than the fair abatement standard q_f.

Proof. *See the Appendix.* ■

To understand Proposition 2, recall first that for any cooperative abatement standard q_C above the fair abatement level q_f, the countries attach a positive weight to each other's self-interested welfare, i.e., they are in a positive-reciprocity state. In such a case, two reinforcing forces are at work. On the one hand, for any country J, the value of cooperation at q_C is higher in $\Gamma^R_\infty(RW, w, \overrightarrow{q}_f)$ than in $\Gamma^S_\infty(SW)$ (see Figure 1) because in the former game (i) the noncooperative (punitive) Nash abatement standard is lower, raising the severity of punishment; and (ii) infinite Nash reversion would also be costly for country $-J$, which acts to heighten the cost of the punishment phase for country J itself. On the other hand, the static incentive country J has to deviate from q_C is weaker in $\Gamma^R_\infty(RW, w, \overrightarrow{q}_f)$ than in $\Gamma^S_\infty(SW)$ (see Figure 2) since in the former game (i) the defect abatement standard is higher, as the countries are in a positive-reciprocity state; and (ii) defection would hurt $-J$, mitigating J's potential onetime gains from cheating. It, then, follows that reciprocal countries can more easily support any given cooperative abatement standard above the fair one than can self-interested countries.[14]

However, for any cooperative abatement standard q_C below the fair abatement level q_f, the countries attach a negative weight to each other's self-interested welfare, i.e., they are in a negative-reciprocity state. Two observations can then be readily made for any such $q_C < q_f$. On the one hand, for any country J, the value of cooperation at q_C is higher in $\Gamma^R_\infty(RW, w, \overrightarrow{q}_f)$ than in $\Gamma^S_\infty(SW)$ (see Figure 1) since the punitive Nash abatement standard is lower in the former game. Of course, infinite Nash reversion would be costly for country $-J$ as well, which acts to lower the cost of the punishment phase for country J in $\Gamma^R_\infty(RW, w, \overrightarrow{q}_f)$. However, the latter effect is relatively weak for a sufficiently small γ. On the other hand, country J has a stronger incentive to defect in $\Gamma^R_\infty(RW, w, \overrightarrow{q}_f)$ than in $\Gamma^S_\infty(SW)$ (see Figure 2) since in the former game (i) the defect abatement standard is lower, because the countries

[14] In Hadjiyiannis et al. (forthcoming), we find that reciprocal countries can more easily sustain any given cooperative tariff *below* the fair one than can self-interested countries.

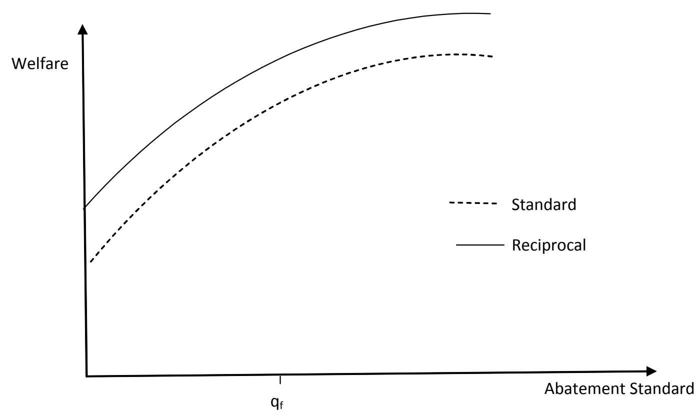

Figure 1: The Incentive to Cooperate

are in a negative-reciprocity state, wishing to punish each other; and (ii) defection would hurt country $-J$, raising the gains from cheating for country J. It is therefore ambiguous whether reciprocal or self-interested countries can more easily sustain any given cooperative abatement standard below the fair abatement standard.

Our final result is a restatement of Proposition 2 in terms of the most cooperative equilibria that can be sustained by reciprocal and self-interested countries.[15] Let $\overrightarrow{q}_{CS} \equiv (q_{CS}, q_{CS})$ denote the most cooperative equilibrium abatement-standard vector for $\Gamma_{\infty}^{S}(SW)$, i.e., q_{CS} is the highest pollution-abatement standard that does not invite cheating in the dynamic game with self-interested countries. Similarly, let $\overrightarrow{q}_{CR} \equiv (q_{CR}, q_{CR})$ represent the most cooperative equilibrium abatement-standard vector of $\Gamma_{\infty}^{R}(RW, w, \overrightarrow{q}_f)$. Furthermore, let us assume in the remainder of this section that $\delta \in \left[\underline{\delta}, \overline{\delta}\right]$ so that both reciprocal and self-interested countries can maintain some environmental cooperation, but q_{OS} is infeasible for either of them.

[15]Note that the most cooperative abatement-standard equilibrium is the most natural focal point for either game, since (i) it is the only equilibrium of the desired class that is not Pareto dominated; and (ii) nothing precludes preplay communication between the countries.

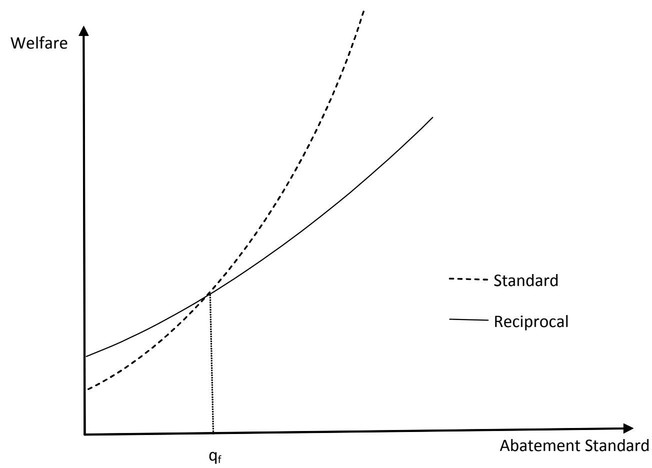

Figure 2: The Incentive to Deviate

Proposition 3 *1. If the fair abatement standard q_f is lower than or equal to the self-interested most cooperative equilibrium standard q_{CS}, the most cooperative equilibrium abatement standard of the dynamic game with reciprocal countries is higher than the one of the dynamic game with self-interested countries (i.e., $q_{CR} > q_{CS}$ if $q_f \leq q_{CS}$).*

2. However, if the fair abatement standard q_f exceeds the self-interested most cooperative equilibrium standard q_{CS}, then the effect of fairness and reciprocity on the most cooperative abatement-standard equilibrium of the dynamic game is ambiguous.

Proof. *See the Appendix.* ■

The intuition behind Proposition 3 is the same as the one underlying Proposition 2. Basically, Proposition 3 states that reciprocal countries that are moderately demanding from each other regarding their environmental policy (i.e., assuming the fair abatement standard is not too high) can support a greater degree of international environmental cooperation than self-interested ones (see Figure 3). Nevertheless, when reciprocal countries are highly demanding of each other regarding pollution abatement (i.e., when only very "green" environmental policies are considered fair), the overall effect of fairness and reciprocity on international environmental cooperation could be negative.

To investigate further the latter scenario, we resort to numerical analysis (see the Online Appendix). In our simulations, we set (i) $b^J\left(q^J\right) = \alpha -$

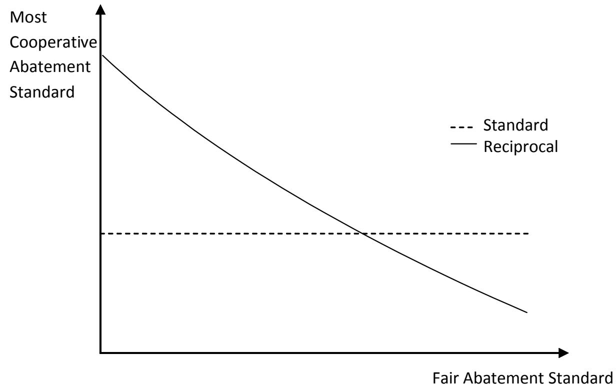

Figure 3: The Most Cooperative Abatement Standard

$\left(q^J\right)^2$, where $\alpha > (q_{\max})^2$; (ii) $\Psi^J\left(q^J, q^{-J}\right) = (1/2)\left[\beta - q^J + s\left(\beta - q^{-J}\right)\right]^2$, where $\beta > q_{\max}$ and $s \in (0,1)$ is the degree of transboundary pollution; and (iii) $w^J(q^{-J}, q_f) = \left(\left(q^{-J} - q_f\right)/\left(q^{-J} + q_f\right)\right) \in (-1,1)$. The simulations do confirm that for very high fair abatement standards and for plausible parameter values, q_{CR} does indeed lie below q_{CS} (as we depict in Figure 3). Actually, q_{CR} is more likely to lie below q_{CS} when δ is relatively low, i.e., when the countries are relatively impatient. To gain some insight into this, recall that if $q_f > q_{CS}$, reciprocal countries have around q_{CS} both a stronger incentive to cheat and a stronger incentive to cooperate than self-interested ones, making the comparison between q_{CR} and q_{CS} ambiguous. However, a lower δ weakens the relative significance of the stronger-incentive-to-cooperate force, while it leaves the stronger-incentive-to-cheat force unaffected, and hence, renders $q_{CR} < q_{CS}$ more likely.

At a more general level, Proposition 3 demonstrates that if, for whatever reason, reciprocal countries become more demanding of each other with respect to their environmental policy (i.e., if the fair abatement standard increases), a given cooperative equilibrium that could have been otherwise supported, might no longer be feasible.[16] In fact, in the simulations discussed above, we find that for plausible parameter values, q_{CR} is strictly decreasing in q_f. Our results, therefore, suggest that if reciprocal countries enter a round of international environmental negotiations with elevated expectations due to domestic and/or global political pressure, they might end up with a less co-

[16] To see this, simply note that q_{CS} is independent of q_f.

operative or "inferior" environmental agreement than in the absence of these high expectations.

5 Extensions

In this section, we discuss some possible extensions of our model. First, we could examine other strategies in the dynamic game, such as tit-for-tat or getting-even strategies. Under such strategies, the degree of environmental cooperation that reciprocal and self-interested countries can support should quantitatively change. Nevertheless, we conjecture that the ramifications of fairness and reciprocity for the incentives faced by the countries would be qualitatively preserved, and thereby, our main conclusions would still hold.

We could also introduce asymmetry in fair-abatement-level perceptions between the two countries. We could then consider two types of agreements: symmetric and asymmetric. In the former case, the incentive-compatibility condition given by (8) should bind in equilibrium only for the most demanding country, i.e., the one with the highest fair-abatement-level perception. Hence, under the symmetric-agreement scenario, we expect that the dynamic game with asymmetric countries would be equivalent to one with symmetric countries having the same fair-abatement-level perception as the most demanding country in the asymmetric-country game. On the other hand, under the asymmetric-agreement scenario, the incentive-compatibility condition should be binding in equilibrium for both countries. Therefore, our qualitative predictions should still be valid for each country taken individually. However, we expect that the overall impact of fairness and reciprocity on international environmental cooperation would be less clear-cut, as, for instance, the reciprocal most cooperative abatement standard might be higher than the self-interested one for country J, but the reverse might hold for country $-J$.

Finally, we could model reciprocity differently and allow fair-abatement-level perceptions to be formed endogenously and strategically. In particular, Rabin (1993) examines a static game where the weight a player places on an opponent's material payoff depends on his interpretation of the opponent's intentions, which are evaluated using beliefs (and beliefs about beliefs) over strategy choices, and defines a "fairness equilibrium." Dufwenberg and Kirchsteiger (2004) extend Rabin's framework to finite extensive-form games allowing players to update their beliefs about their opponents' intentions as the game unfolds, and define a "sequential reciprocity equilibrium." One should expect though that such a framework would not be suitable for our infinitely-repeated game. The reason is that the analysis would likely be highly intractable, since, for example, the countries might choose their abatement standards strategically in the early stages of the game so that they favorably affect beliefs in the later stages.

6 Conclusions

In this paper we examined the impact of fairness and reciprocity on international environmental cooperation in pollution abatement, where pollution was assumed to be transboundary in nature. More specifically, we investigated whether in the context of self-enforcing IEAs (involving full participation) reciprocal countries can support a higher degree of pollution abatement than self-interested ones. In our setting, a reciprocal country is willing to reward another given country by raising its own abatement standard and therefore reducing transboundary pollution if it expects the latter to set a higher abatement standard than the one deemed fair; nevertheless, the reverse is true when the latter is expected to implement an unfairly low level of pollution abatement. This is an important question for two reasons. First, governments themselves claim to exhibit reciprocal preferences towards environmental policy. Second, our analysis has provided a novel perspective on the role of expectations in international environmental negotiations.

We have established that reciprocal countries that have moderate expectations from each other regarding their national abatement policies (i.e., when their fair-abatement-standard perceptions are not too high) can sustain higher cooperative abatement levels than self-interested ones. On the other hand, when countries are highly demanding from each other with respect to environmental policy (i.e., when only very high abatement standards are perceived as fair), then reciprocity could have a detrimental effect on international environmental cooperation.

More generally, our results suggest that if reciprocal countries enter a round of international environmental negotiations with elevated expectations due to domestic and/or global political pressure, they might no longer be able to support very "green" environmental policies and could end up with a less cooperative or "inferior" environmental agreement than in the absence of these high expectations. It could then be argued that this might be one of the plausible explanations for the failure of the 2009 Copenhagen summit. There is little doubt that the pre-Copenhagen expectations for what could have been accomplished there were very high. For instance, according to the official UN website, more than 100 world leaders met in New York in September 2009 in order to "mobilize political will and strengthen momentum for a fair, effective, and ambitious climate deal in Copenhagen."[17] The same level of ambition is expressed in the quote from the French government presented in the introduction. Our findings, however, have demonstrated that such "ambitious" expectations (represented in our model by high fair-abatement-standard perceptions) could end up hindering the efforts for deeper international environmental cooperation. In fact, this possibility had been anticipated by a number of analysts and negotiators before the summit. For example, in November 2009, Susanne Dröge from the German Institute for International and Security Affairs said that high expectations for the summit had to be adjusted, and noted that it was important that the expectations were adjusted before the summit, because otherwise the outcome would be bad.[18]

By contrast, at the 2010 Cancun summit, expectations were much more modest. As a matter of fact, the UN Secretary General Ban Ki-moon, in his address to the opening ceremony of the high-level segment of the Cancun talks, urged nations to not be too demanding. He said, "We don't need final agreement on all the issues, but we do need progress on all the fronts. We cannot let the perfect be the enemy of the good."[19] In the light, then, of our model, it could be argued that countries managed at the Cancun summit to make a step forward (even if a modest one) relative to Copenhagen in part due to their lowered expectations. As *The Economist* concludes, "So why did Cancun succeed in making progress within the UN process where Copenhagen so spectacularly failed? One reason is low expectations. Copenhagen was meant to produce an all-encompassing agreement; Cancun was expected to embarrass itself."[20] To sum up, our analysis has demonstrated that expectations of what can/should the international environmental negotiations achieve are a key factor in determining their final outcome.

A Appendix

We first establish the existence, uniqueness, and inefficiency of both the reciprocal and the self-interested static Nash equilibria (as discussed in Section 3). We then provide the proofs for the lemma and the three propositions presented in the body of the paper.

[17] See http://un.org/wcm/content/site/climatechange/lang/en/pages/2009summit.

[18] "Can Copenhagen Still Be Saved?" *The Economist*, November 17, 2009.

[19] See http://www.guardian.co.uk/environment/2010/dec/08/ban-ki-moon-cancun-climate-deal.

[20] See http://www.economist.com/blogs/newsbook/2010/12/climate_change?page=1.

A.1 Existence, Uniqueness, and Inefficiency of Static Nash Equilibria

A.1.1 Existence and Uniqueness of Static Nash Equilibria

Lemma 2 *For both the static game with reciprocal countries* $\Gamma^R(RW, w, \overrightarrow{q}_f)$ *and the static game with self-interested countries* $\Gamma^S(SW)$*, there exist pure strategy Nash equilibria. Moreover, if the countries' best-response functions have a slope strictly greater than* -1*, then* $\Gamma^R(RW, w, \overrightarrow{q}_f)$ *and* $\Gamma^S(SW)$ *have unique pure strategy Nash equilibria, which are symmetric.*

Proof. We first consider $\Gamma^R(RW, w, \overrightarrow{q}_f)$. Let us define new strategies $a^J = q^J$ and $a^{-J} = -q^{-J}$, reversing the natural order in country $-J$'s strategy set. Then, $\left(\partial^2 RW^J / \partial a^J \partial a^{-J}\right) > 0$. Given now that the number of countries is finite and that for any country J, (i) $[0, q_{\max}]$ is a compact interval in $\mathcal{R}^+$; (ii) RW^J is twice continuously differentiable on $[0, q_{\max}]$; and (iii) $\left(\partial^2 RW^J / \partial a^J \partial a^{-J}\right) > 0$, we know from Theorem 4 in Milgrom and Roberts (1990) that $\Gamma^R(RW, w, \overrightarrow{q}_f)$ is a (smooth strictly) supermodular game. It then follows from Theorem 5 in Milgrom and Roberts (1990) that $\Gamma^R(RW, w, \overrightarrow{q}_f)$ admits pure strategy Nash equilibria.

To see that $\Gamma^S(SW)$ also admits pure strategy Nash equilibria, simply recall that $\Gamma^S(SW)$ can be obtained from $\Gamma^R(RW, w, \overrightarrow{q}_f)$ by setting $\gamma = 0$, meaning that $\Gamma^S(SW)$ is also a (smooth strictly) supermodular game. Moreover, the uniqueness of the equilibria when the countries' best-response functions have a slope strictly greater than -1 follows directly from Theorem 2.8 in Vives (1999). Last, given the overall symmetry of our setup, it follows trivially that these equilibria must be symmetric. ■

A.1.2 Inefficiency of Static Nash Equilibria

We will first establish that $q_{OS} > q_{NS}$. Given that $q^J_{NS} (= q^{-J}_{NS} \equiv q_{NS})$ is interior, it must be a solution to:

$$\frac{\partial SW^J\left(q^J, q^{-J}\right)}{\partial q^J} = \frac{\partial b^J\left(q^J\right)}{\partial q^J} - \frac{\partial \Psi^J\left(q^J, q^{-J}\right)}{\partial q^J} = 0. \tag{10}$$

Similarly, given that $q^J_{OS} (= q^{-J}_{OS} \equiv q_{OS})$ is interior, it must be a solution to:

$$\frac{\partial SW^J\left(q^J, q^{-J}\right)}{\partial q^J} + \frac{\partial SW^{-J}\left(q^J, q^{-J}\right)}{\partial q^J}$$
$$= \frac{\partial b^J\left(q^J\right)}{\partial q^J} - \frac{\partial \Psi^J\left(q^J, q^{-J}\right)}{\partial q^J} - \frac{\partial \Psi^{-J}\left(q^J, q^{-J}\right)}{\partial q^J} = 0. \tag{11}$$

Using (10) and (11), we then have:

$$\frac{\partial SW^J\left(q^J,q^{-J}\right)}{\partial q^J}\Big|_{q^J=q_{NS},q^{-J}=q_{NS}}+\frac{\partial SW^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}\Big|_{q^J=q_{NS},q^{-J}=q_{NS}} = -\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}\Big|_{q^J=q_{NS},q^{-J}=q_{NS}} > 0. \quad (12)$$

(12) along with the strict concavity of SW^J+SW^{-J} with respect to q^J together imply that $q_{OS}>q_{NS}$.

We will next demonstrate that $q_{OR}>q_{NR}$. Given that $q^J_{NR}(=q^{-J}_{NR}\equiv q_{NR})$ is interior, it must be a solution to:

$$\frac{\partial RW^J\left(q^J,q^{-J},q_f\right)}{\partial q^J}=\frac{\partial b^J\left(q^J\right)}{\partial q^J}-\frac{\partial \Psi^J\left(q^J,q^{-J}\right)}{\partial q^J} -\gamma w^J(q^{-J},q_f)\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}=0. \quad (13)$$

Moreover, given that $q^J_{OR}(=q^{-J}_{OR}\equiv q_{OR})$ is interior, it must be a solution to:

$$\frac{\partial RW^J\left(q^J,q^{-J},q_f\right)}{\partial q^J}+\frac{\partial RW^{-J}\left(q^J,q^{-J},q_f\right)}{\partial q^J}=\frac{\partial b^J\left(q^J\right)}{\partial q^J}-\frac{\partial \Psi^J\left(q^J,q^{-J}\right)}{\partial q^J} -\gamma w^J(q^{-J},q_f)\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}-\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J} +\gamma\left[\frac{\partial w^{-J}\left(q^J,q_f\right)}{\partial q^J}SW^J\left(q^J,q^{-J}\right)+w^{-J}\left(q^J,q_f\right)\frac{\partial SW^J\left(q^J,q^{-J}\right)}{\partial q^J}\right]=0. \quad (14)$$

Using (13) and (14), we then obtain:

$$\frac{\partial RW^J\left(q^J,q^{-J},q_f\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}}+\frac{\partial RW^{-J}\left(q^J,q^{-J},q_f\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}} = -\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}}+\gamma\left[\frac{\partial w^{-J}\left(q^J,q_f\right)}{\partial q^J}\Big|_{q^J=q_{NR}}SW^J\left(q_{NR},q_{NR}\right) +w^{-J}\left(q_{NR},q_f\right)\gamma w^J\left(q_{NR},q_f\right)\frac{\partial \Psi^{-J}\left(q^J,q^{-J}\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}}\right]. \quad (15)$$

For γ sufficiently small, $\frac{\partial RW^J\left(q^J,q^{-J},q_f\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}}+\frac{\partial RW^{-J}\left(q^J,q^{-J},q_f\right)}{\partial q^J}\Big|_{q^J=q_{NR},q^{-J}=q_{NR}}>0$, which along with the strict concavity of RW^J+RW^{-J} with respect to q^J together imply that $q_{OR}>q_{NR}$.

We will finally show that $q_{OR} > q_{OS}$. Using (11) and (14), we obtain:

$$\left.\frac{\partial RW^J\left(q^J, q^{-J}, q_f\right)}{\partial q^J}\right|_{q^J=q_{OS}, q^{-J}=q_{OS}} + \left.\frac{\partial RW^{-J}\left(q^J, q^{-J}, q_f\right)}{\partial q^J}\right|_{q^J=q_{OS}, q^{-J}=q_{OS}}$$
$$= \gamma \left.\frac{\partial w^{-J}\left(q^J, q_f\right)}{\partial q^J}\right|_{q^J=q_{OS}} SW^J\left(q_{OS}, q_{OS}\right) > 0. \tag{16}$$

Given (16) and the strict concavity of $RW^J + RW^{-J}$ with respect to q^J, the result then follows trivially. Q.E.D.

A.2 Proofs

A.2.1 Proof of Lemma 1

As we argued above in the proof of the existence of the static Nash equilibria, given that the number of countries is finite and that for any country J, (i) $[0, q_{\max}]$ is a compact interval in $\mathcal{R}^+$; (ii) RW^J is twice continuously differentiable on $[0, q_{\max}]$; and (iii) $\left(\partial^2 RW^J / \partial a^J \partial a^{-J}\right) > 0$, where $a^J = q^J$ and $a^{-J} = -q^{-J}$ are the countries' strategies redefined, we know from Theorem 4 in Milgrom and Roberts (1990) that $\Gamma^R(RW, w, \overrightarrow{q}_f)$ is a supermodular game. Given now the supermodularity of $\Gamma^R(RW, w, \overrightarrow{q}_f)$ and that $\left(\partial^2 RW^J / \partial q^J \partial q_f\right) < 0$ for any J, the lemma follows immediately from Theorem 6 in Milgrom and Roberts (1990). Q.E.D.

A.2.2 Proof of Proposition 1

If $q_f = q_{NS}$, then trivially $q_{NR} = q_{NS} \equiv q_N$, implying that for any J, $RW^J(\overrightarrow{q}_N, q_f) = SW^J(\overrightarrow{q}_N)$ since $w^J(q_N, q_f) = 0$ by (3) (where $\overrightarrow{q}_N \equiv (q_N, q_N)$). On the other hand, if $q_f > q_{NS}$, then $q_{NR} < q_{NS}$ by Lemma 1. These two inequalities imply that $q_{NR} < q_f$, and thus for any J, $w^J(q_{NR}, q_f) < 0$ from (3). Moreover, for $q_{NR} < q_{NS}$, $SW^J(\overrightarrow{q}_{NR}) < SW^J(\overrightarrow{q}_{NS})$ for all J. But then it follows that for all J, $RW^J(\overrightarrow{q}_{NR}, q_f) < SW^J(\overrightarrow{q}_{NS})$. Q.E.D.

A.2.3 Proof of Proposition 2

We want to show first that $q_C \geq q_f$ implies that $\delta^R_{q_C} < \delta^S_{q_C}$. To do so, we will prove:

(i) If $q_C \geq q_f \Rightarrow RW^J_D - RW^J_C \leq SW^J_D - SW^J_C$ for any J.
(ii) If $q_C \geq q_f \Rightarrow RW^J_D - RW^J_N > SW^J_D - SW^J_N$ for any J.

Let us start with (i). We have that for any J:

$$RW_C^J = SW^J(\overrightarrow{q}_C) + \gamma w^J(q_C, q_f) SW^{-J}(\overrightarrow{q}_C) \text{ and}$$
$$RW_D^J = SW^J(BR_R^J(q_C), q_C) + \gamma w^J(q_C, q_f) SW^{-J}(BR_R^J(q_C), q_C).$$

Therefore:

$$\begin{aligned} RW_D^J - RW_C^J &= SW^J(BR_R^J(q_C), q_C) - SW^J(\overrightarrow{q}_C) \\ &+\gamma w^J(q_C, q_f)\left(SW^{-J}(BR_R^J(q_C), q_C) - SW^{-J}(\overrightarrow{q}_C)\right) \\ &\leq SW^J(BR_R^J(q_C), q_C) - SW^J(\overrightarrow{q}_C) \\ &\leq SW^J(BR_S^J(q_C), q_C) - SW^J(\overrightarrow{q}_C) = SW_D^J - SW_C^J. \end{aligned} \tag{17}$$

We know from (3) that $w^J(q_C, q_f) \geq 0$ if $q_C \geq q_f$. Furthermore, the welfare of self-interested country $-J$ is (weakly) lower when country J deviates while it still cooperates than when both countries cooperate, i.e., $SW^{-J}(BR_R^J(q_C), q_C)$ $-SW^{-J}(\overrightarrow{q}_C) \leq 0$. The first inequality then follows. The second inequality stems from the fact that $BR_S^J(q_C)$ is the best reply of self-interested country J. This concludes the proof of (i).

We now turn to (ii). Let us rewrite the result we want to show:

$$q_C \geq q_f \Rightarrow \left(RW_D^J - SW_D^J\right) - \left(RW_N^J - SW_N^J\right) > 0 \text{ for any } J.$$

By Proposition 1 we know that the Nash equilibrium abatement level of $\Gamma^S(SW)$ is (weakly) higher than that of $\Gamma^R(RW, w, \overrightarrow{q}_f)$, i.e., $q_{NR} \leq q_{NS}$. Thus, $q_f \geq q_{NS} \geq q_{NR}$, implying that $w^J(q_{NR}, q_f) \leq 0$ by (3). Therefore, the following inequality holds for any J:

$$\begin{aligned} RW_N^J &= SW^J(\overrightarrow{q}_{NR}) + \gamma w^J(q_{NR}, q_f) SW^{-J}(\overrightarrow{q}_{NR}) \\ &\leq SW^J(\overrightarrow{q}_{NR}) \leq SW^J(\overrightarrow{q}_{NS}) = SW_N^J \\ &\Leftrightarrow RW_N^J - SW_N^J \leq 0. \end{aligned} \tag{18}$$

Next, we will show that $RW_D^J - SW_D^J \geq 0$ for any J. Remember that γ is assumed to be sufficiently small. Taking a first-order Taylor series expansion of $RW^J(BR_R^J(q_C), q_C, q_f)$ around $\gamma = 0$, we obtain:

$$\begin{aligned} RW^J(BR_R^J(q_C), q_C, q_f) &\approx SW^J(BR_S^J(q_C), q_C) \\ &+\gamma w^J(q_C, q_f) SW^{-J}(BR_S^J(q_C), q_C) \\ \Leftrightarrow RW^J(BR_R^J(q_C), q_C, q_f) &- SW^J(BR_S^J(q_C), q_C) \\ \approx \gamma w^J(q_C, q_f) SW^{-J}(BR_S^J(q_C), q_C) &\geq 0 \Leftrightarrow RW_D^J - SW_D^J \geq 0. \end{aligned} \tag{19}$$

The inequality holds due to $w^J(q_C, q_f) \geq 0$. By assumption, we have that $q_C > q_{NS}$, and thus q_f cannot be equal to both q_C and q_{NS} at the same time. Hence, at least one of the inequalities in (18) and (19) must be strict. This concludes the proof of part (ii). Therefore, by (i) and (ii), we finally have $\delta^R_{q_C} < \delta^S_{q_C}$.

Nevertheless, for any $q_C < q_f$, it is ambiguous by (17) and (19) whether $\delta^R_{q_C}$ or $\delta^S_{q_C}$ is higher, since the weight function is negative at q_C. Thus, it is possible that $\delta^R_{q_C} > \delta^S_{q_C}$ for such a q_C. Q.E.D.

A.2.4 Proof of Proposition 3

We will show first that if $q_f \leq q_{CS}$, then $q_{CR} > q_{CS}$. We know from Proposition 2 that if $q_{CS} \geq q_f$, then $\delta^R_{q_{CS}} < \delta^S_{q_{CS}}$. Furthermore, from (7) and (9), we have:

$$SW^J_D - SW^J_C = \delta^S_{q_{CS}} \left(SW^J_D - SW^J_N \right) \text{ and}$$
$$RW^J_D - RW^J_C = \delta^R_{q_{CS}} \left(RW^J_D - RW^J_N \right).$$

Since $\delta^R_{q_{CS}} < \delta^S_{q_{CS}}$:

$$\begin{aligned} RW^J_D - RW^J_C &< \delta^S_{q_{CS}} \left(RW^J_D - RW^J_N \right) \\ \Leftrightarrow \left(1 - \delta^S_{q_{CS}}\right) &RW^J(BR^J_R(q_{CS}), q_{CS}, q_f) \\ < RW^J(\overrightarrow{q}_{CS}, q_f) &- \delta^S_{q_{CS}} RW^J\left(\overrightarrow{q}_{NR}, q_f\right), \end{aligned} \tag{20}$$

meaning that $\Omega^J_R(q_{CS}) < \left(\delta^S_{q_{CS}} / \left(1 - \delta^S_{q_{CS}}\right)\right) \omega^J_R(q_{CS})$, or that the incentive-compatibility condition is not binding for reciprocal country J at the pair $(q_{CS}, \delta^S_{q_{CS}})$.

Note here that RW^J_N does not depend on the cooperative abatement level. Moreover, for any cooperative abatement standard q_C higher than the most cooperative equilibrium abatement level of $\Gamma^s_\infty(SW)$, q_{CS}, the welfare for reciprocal country J under defection from q_C is higher than the welfare under deviation from q_{CS}:

$$RW^J(BR^J_R(q_C), q_C, q_f) > RW^J(BR^J_R(q_{CS}), q_{CS}, q_f).$$

At the same time, for such a $q_C > q_{CS}$, country J's welfare under cooperation is also higher at q_C than at q_{CS}:

$$RW^J(\overrightarrow{q}_C, q_f) > RW^J(\overrightarrow{q}_{CS}, q_f).$$

By the continuity of $RW^J(\bullet)$, then there exists a cooperative abatement standard $\widehat{q}_C > q_{CS}$ such that (20) still holds, or $\Omega^J_R(\widehat{q}_C)$

$< \left(\delta^S_{q_{CS}} / \left(1-\delta^S_{q_{CS}}\right)\right) \omega^J_R\left(\widehat{q}_C\right)$. Since the same analysis applies to any $\left(q_{CS}, \delta^S_{q_{CS}}\right)$ pair for $\delta^S_{q_{CS}} \in [\underline{\delta}, \overline{\delta}]$, we have that for any $\delta \in [\underline{\delta}, \overline{\delta}]$, $q_{CS} < q_{CR}$.

However, as we argued in the proof of Proposition 2, if $q_f > q_{CS}$, it is ambiguous by (17) and (19) whether $\delta^R_{q_{CS}}$ or $\delta^S_{q_{CS}}$ is higher. Hence, it is possible that the minimum discount factor required for countries with reciprocal preferences to sustain cooperation at q_{CS} is higher than that for self-interested countries, i.e., $\delta^R_{q_{CS}} > \delta^S_{q_{CS}}$. Let us consider this case first. Moreover, let us make the assumption that $\Omega^J_R(\bullet)$ is a strictly convex function whereas $\omega^J_R(\bullet)$ is a strictly concave one.[21] From (9) and (7), we have:

$$\begin{gathered} RW^J_D - RW^J_C = \delta^R_{q_{CS}}\left(RW^J_D - RW^J_N\right) \text{ and} \\ SW^J_D - SW^J_C = \delta^S_{q_{CS}}\left(SW^J_D - SW^J_N\right). \end{gathered}$$

Since $\delta^R_{q_{CS}} > \delta^S_{q_{CS}}$:

$$\begin{gathered} RW^J_D - RW^J_C > \delta^S_{q_{CS}}\left(RW^J_D - RW^J_N\right) \\ \Leftrightarrow \left(1-\delta^S_{q_{CS}}\right) RW^J(BR^J_R(q_{CS}), q_{CS,} q_f) \\ > RW^J(\overrightarrow{q}_{CS}, q_f) - \delta^S_{q_{CS}} RW^J\left(\overrightarrow{q}_{NR}, q_f\right), \end{gathered}$$

meaning that $\Omega^J_R(q_{CS}) > \left(\delta^S_{q_{CS}} / \left(1-\delta^S_{q_{CS}}\right)\right) \omega^J_R(q_{CS})$, or that the incentive-compatibility condition is violated for reciprocal country J at the pair $(q_{CS}, \delta^S_{q_{CS}})$.

We know that both $\Omega^J_R(q_C)$ and $\omega^J_R(q_C)$ are strictly increasing in q_C. Therefore, given the strict convexity of $\Omega^J_R(\bullet)$, the strict concavity of $\omega^J_R(\bullet)$, and the permissible range of δ, it is obvious that the incentive-compatibility condition for reciprocal country J can only be restored at a cooperative abatement standard $\widehat{q}_C < q_{CS}$. Since the same analysis applies to any $\left(q_{CS}, \delta^S_{q_{CS}}\right)$ pair for $\delta^S_{q_{CS}} \in [\underline{\delta}, \overline{\delta}]$, we have that for any $\delta \in [\underline{\delta}, \overline{\delta}]$, $q_{CS} > q_{CR}$.

Nevertheless, $\delta^R_{q_{CS}} < \delta^S_{q_{CS}}$ is also possible by (17) and (19). In this case, as we argued above, $q_{CS} < q_{CR}$. Thus, when $q_f > q_{CS}$, it is ambiguous whether q_{CR} or q_{CS} is higher due to the ambiguity of whether $\delta^R_{q_{CS}}$ or $\delta^S_{q_{CS}}$ is higher. Q.E.D.

References

[1] Aidt, T.S., 1998. Political internalization of economic externalities and environmental policy. Journal of Public Economics 69, 1–16.

[21] This assumption is clearly not restrictive given the type of result we are seeking here.

[2] Andreoni, J., 1988. Why free ride? Strategies and learning in public good experiments. Journal of Public Economics 37, 291–304.

[3] Bagwell, K., Staiger, R.W., 2002. The Economics of the World Trading System. MIT Press, Cambridge.

[4] Barrett, S., 2005. The theory of international environmental agreements, in: Maler, K.G., Vincent, J.R. (Eds.), Handbook of Environmental Economics, Volume 3. Elsevier, Amsterdam, pp. 1457–1516.

[5] Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games and Economic Behavior 10, 122–142.

[6] Black, J., Levi, M.D., de Meza, D., 1993. Creating a good atmosphere: Minimum participation for tackling the "Greenhouse Effect." Economica 60, 281–293.

[7] Carraro, C., Marchiori, C., Oreffice, S., 2004. Endogenous minimum participation in international environmental treaties. CEPR Discussion Paper No. 4281.

[8] Conconi, P., 2003. Green lobbies and transboundary pollution in large open economies. Journal of International Economics 59, 399–422.

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

[10] Dam, K., 1970. The GATT: Law and International Economic Organization. University of Chicago Press, Chicago.

[11] Dixit, A., 1987. Strategic aspects of trade policy, in: Truman, B. (Ed.), Advances in Economic Theory: Fifth World Congress. Cambridge University Press, Cambridge, pp. 329–362.

[12] Dohmen, T., Falk, A., Huffman, D., Sunde, U., 2009. Homo reciprocans: Survey evidence on behavioural outcomes. Economic Journal 119, 592–612.

[13] Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268–298.

[14] Eriksson, C., Persson, J., 2003. Economic growth, inequality, democratization, and the environment. Environmental and Resource Economics 25, 1–16.

[15] Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293–315.

[16] Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: Experimental evidence. Econometrica 65, 833–860.

[17] Fehr, E., Kirchsteiger, G., Riedl, A., 1998. Gift exchange and reciprocity in competitive experimental markets. European Economic Review 42, 1–34.

[18] Ferrara, I., Missios, P., Yildiz, H.M., 2009. Trading rules and the environment: Does equal treatment lead to a cleaner world? Journal of Environmental Economics and Management 58, 206–225.

[19] Fredriksson, P.G., 1997. The political economy of pollution taxes in a small open economy. Journal of Environmental Economics and Management 33, 44–58.

[20] Grossman, G.M., Helpman, E., 1994. Protection for sale. American Economic Review 84, 833–850.

[21] Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.

[22] Hadjiyiannis, C., İriş, D., Tabakis, C., forthcoming. Multilateral tariff cooperation under fairness and reciprocity. Canadian Journal of Economics.

[23] Helson, H., 1964. Adaptation Level Theory: An Experimental and Systematic Approach to Behavior. Harper & Row, New York.

[24] Hoel, M., 1991. Global environmental problems: The effects of unilateral actions taken by one country. Journal of Environmental Economics and Management 20, 55–70.

[25] İriş, D., Santos-Pinto, L., 2008. Tacit collusion under fairness and reciprocity. Cahiers de Recherches Économiques 09.03, University of Lausanne.

[26] Lange, A., Vogt, C., 2003. Cooperation in international environmental negotiations due to a preference for equity. Journal of Public Economics 87, 2049–2067.

[27] Lange, A., Löschel, A., Vogt, C., Ziegler, A., 2010. On the self-interested use of equity in international climate negotiations. European Economic Review 54, 359–375.

[28] McAusland, C., 2003. Voting for pollution policy: The importance of income inequality and openness to trade. Journal of International Economics 61, 425–451.

[29] Milgrom, P., Roberts, J., 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. Econometrica 58, 1255–1277.

[30] Palfrey, T.R., Prisbrey, J.E., 1997. Anomalous behavior in public goods experiments: How much and why? American Economic Review 87, 829–846.

[31] Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281–1302.

[32] Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. American Economic Review 81, 1068–1095.

[33] Segal, U., Sobel, J., 2007. Tit for tat: Foundations of preferences for reciprocity in strategic settings. Journal of Economic Theory 136, 197–216.

[34] Tversky, A., Kahneman, D., 1991. Loss aversion in riskless choice: A reference-dependent model. Quarterly Journal of Economics 106, 1039–1061.

[35] Vives, X., 1999. Oligopoly Pricing: Old Ideas and New Tools. MIT Press, Cambridge.

[36] Wagner, U.J., 2001. The design of stable international environmental agreements: Economic theory and political economy. Journal of Economic Surveys 15, 377–411.