International Environmental Cooperation under Fairness and Reciprocity

Costas Hadjiyiannis* Doruk İriş† and Chrysostomos Tabakis‡

January 30, 2012

Abstract

This paper explores the implications of fairness and reciprocity for self-enforcing international environmental agreements on pollution abatement. Reciprocal countries reward kind 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 failure of the Copenhagen summit and the recent success of the Cancun one. Finally, we show that these results are robust to endogenizing fair-abatement-standard perceptions.

Keywords: Reciprocity; Environmental agreements; Abatement standards; Repeated games

JEL Classification: Q50; Q58; D63

*Department of Economics, University of Cyprus, P.O. Box 20537, CY-1678 Nicosia, Cyprus; E-mail: hadjiyiannis.costas@ucy.ac.cy.
†Faculdade de Economia, Universidade Nova de Lisboa, Campus de Campolide, Lisbon 1099–032, Portugal; E-mail: dorukiris@novasbe.pt.
‡Corresponding author: Faculdade de Economia, Universidade Nova de Lisboa, Campus de Campolide, Lisbon 1099–032, Portugal; E-mail: ctampakis@novasbe.pt; Phone: +351 21 380 1639; Fax: +351 21 387 0933.
1 Introduction

In this paper, we examine the ramifications of reciprocal preferences for international environmental agreements (IEAs). A government with reciprocal preferences rewards “kind” or “fair” actions (positive reciprocity), while it punishes “unkind” or “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 question 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.

Many environmental problems are transboundary in nature and often have a global scope (e.g., climate change or marine pollution). Countries have therefore been seeking to sign IEAs in order to coordinate their environmental policies. 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 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.\(^5\)

\(^1\)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 and only if this reduces the inequity between them.

\(^3\)See http://unfccc.int/resource/docs/2009/cop15/eng/11a01.pdf.


\(^5\)For the UN Environment Programme Medium-term Strategy 2010–2013, see
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, early in the negotiations, one of India’s top negotiators expressed “concern...that [India has] been offering unilateral concessions, without obtaining any reciprocity.”6 Furthermore, 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,” while the UK’s Prime Minister Gordon Brown argued 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.”7 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 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

---

7 See http://www.guardian.co.uk/environment/2009/dec/19/copenhagen-blame-game and http://news.bbc.co.uk/2/hi/uk_newspolitics/8423831.stm, respectively.
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.8

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. The first model we can invoke here is the median-voter model, where the government chooses policies that reflect the opinion of the median voter on the issue (in order to remain popular and stay in office).9 In such a setting, if the median voter has reciprocal preferences with regard to environmental policy, then the government’s actions will mirror these preferences. Alternatively, we could look at interest-group models. The framework that currently occupies center stage in the literature is that of Grossman and Helpman (1994), who focus on trade-policy issues. More specifically, in their paper, the incumbent government maximizes a weighted average of aggregate social welfare and political contributions by lobbies that wish to influence trade policy. Their framework has subsequently been applied to environmental policy by Fredriksson (1997) and Aidt (1998), among others. In these lobbying models, if individuals have preferences for fairness and reciprocity, these preferences will enter into the government’s objective function with some weight.

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 unilaterally deviating from its agreed-

---


9See, for example, Eriksson and Persson (2003) and McAusland (2003).
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.\textsuperscript{10} On the other hand, to model reciprocity, we follow Segal and Sobel (2007).\textsuperscript{11} 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 behave kindly by setting a higher abatement standard than the one it perceives as fair. However, if the other country in question is expected to behave unkindly by setting an unfairly low abatement standard, then a negative weight is placed on its self-interested welfare. In other words, countries are assumed to have preferences over both outcomes and strategies.

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 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 domi-

\textsuperscript{10}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).

\textsuperscript{11}See Iriş and Santos-Pinto (2008) and Hadjijianinis et al. (forthcoming) for an application of their framework to, respectively, firm collusion and multilateral tariff cooperation.
nates, 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. It could then be argued that this might be one of the plausible explanations for the failure of the 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.” The same level of ambition is expressed in the quote from the French government above. Our findings, however, demonstrate 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.

By contrast, at the 2010 UN Climate Change Conference in Cancun, 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.” 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

---

13 “Can Copenhagen Still Be Saved?” The Economist, November 17, 2009.
agreement; Cancun was expected to embarrass itself.” In summary, expectations of what can/should the negotiations achieve emerge as a key factor in our analysis, having a significant effect on their final outcome. Actually, to the best of our knowledge, our paper is the first to identify such a role for expectations in multilateral environmental negotiations. The policy implications are then immediate. The careful management of expectations is critical for the success of international environmental negotiations, and most importantly, 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 non-cooperative 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, some authors explore in different contexts whether international lobbying can help governments internalize cross-national externalities (e.g., Conconi, 2003; Gawande et al., 2006; Aidt and Hwang, 2008) and therefore be an alternative to international agreements, which are the focus of this paper.

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, whereas Section 4 analyzes the dynamic game. Section 5 discusses the robustness of our main results to relaxing some of the model’s key assumptions. Finally, Section 6 offers some concluding remarks. 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 \( q^J \in [0, q_{\text{max}}] \) so as to maximize its individual welfare, where \( J \in \{A, B\} \) and \( q_{\text{max}} \) 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 \( \Psi^J \) country \( J \) faces is a function of the level of its own emissions and those of country \(-J\). Assuming that both \( b^J (q^J) \) and \( \Psi^J (q^J, q^{-J}) \) are twice continuously differentiable functions, we then have: \( \frac{\partial b^J}{\partial q^J} < 0 \), \( \frac{\partial \Psi^J}{\partial q^J} < 0 \), and \( \frac{\partial \Psi^J}{\partial q^{-J}} < 0 \). We further maintain the assumption that \( \frac{\partial^2 \Psi^J}{\partial q^J \partial q^{-J}} > 0 \), 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 assumption 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:

\[
RW^J (q^J, q^{-J}, q_{\text{f}}^{-J}) = SW^J (q^J, q^{-J}) + \gamma w^J (q^{-J}, q_{\text{f}}^{-J}) SW^{-J} (q^J, q^{-J}).
\]

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

\[
SW^J (q^J, q^{-J}) = b^J (q^J) - \Psi^J (q^J, q^{-J}).
\]

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

\[
w^J (q^{-J}, q_{\text{f}}^{-J}) = \begin{cases} 
   0 & \text{if } q^{-J} = q_{\text{f}}^{-J} \\
   \in (0, 1) & \text{if } q^{-J} > q_{\text{f}}^{-J} \\
   \in [-1, 0) & \text{otherwise}
\end{cases}
\]

with \( q_{\text{f}}^{-J} \) being the \( q^{-J} \) country \( J \) deems “fair.” We maintain the assumptions that country \( J \)'s weight function \( w^J (q^{-J}, q_{\text{f}}^{-J}) \) is twice continuously differentiable in both arguments, is strictly decreasing in its own fair-abatement-level perception \( q_{\text{f}}^{-J} \), and is strictly increasing in country \(-J\)’s abatement level \( q^{-J} \). We also assume that fair-abatement-standard perceptions are common knowledge.

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 (or kind) action that should be rewarded, whereas any \(q^{-J}\) below \(q_f^{-J}\) is an unfair (or unkind) action that should be punished.

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 \(\Gamma^S\) represent the static game with self-interested countries, while \(\Gamma^R\) denotes the static game with reciprocal countries, where \(\vec{q}_f \equiv (q^J_f, q^{-J}_f)\) is the fair-abatement-standard vector. We henceforth assume that \(q^J_f = q^{-J}_f \equiv q_f\), 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 our robustness section, the qualitative nature of our findings is in general robust to introducing asymmetries in fair-abatement-level perceptions. In addition, in all that follows, we maintain the assumption that \(\gamma\) 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 \(\gamma\) is sufficiently small and \((\partial^2 \Psi^J / \partial q^J \partial q^{-J}) > 0\), 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., \((\partial^2 RW^J / \partial q^J \partial q^{-J}) < 0\)). 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., $\frac{\partial^2 RW^J}{\partial q^J \partial q_f} < 0$). To understand the sign of this derivative, note that $\frac{\partial^2 SW^J}{\partial q^J \partial q_f} = 0$ 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 $R^R(RW, w, \overline{\bar{q}} f)$ and $S^S(SW)$ admit pure Nash equilibria. In fact, if the countries’ best-response functions have a slope strictly greater than $-1$, which is hereafter assumed, then $R^R(RW, w, \overline{\bar{q}} f)$ and $S^S(SW)$ have unique pure Nash equilibria, which are symmetric: $\overline{\bar{q}} NR = (q_{NR}, q_{NR})$ and $\overline{\bar{q}} NS = (q_{NS}, q_{NS})$, respectively. Let us now 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. Moreover, let both $RW^J + RW^{-J}$ and $SW^J + SW^{-J}$ be strictly concave with respect to $q_f$. It is then straightforward to demonstrate that the Nash equilibria of $R^R(RW, w, \overline{\bar{q}} f)$ and $S^S(SW)$ 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 $R^R(RW, w, \overline{\bar{q}} f)$.

**Lemma 1** The (pure) Nash equilibrium of $R^R(RW, w, \overline{\bar{q}} f)$, $\overline{\bar{q}} NR = (q_{NR}, q_{NR})$, is strictly decreasing in the fair abatement standard $q_f$, i.e., $\frac{\partial q_{NR}}{\partial q_f} < 0$.

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 abate-
ment standard, $q_{NR}$, with the self-interested one, $q_{NS}$, as well as the welfare obtained in $\Gamma^R(RW, w, \overline{q}_f)$ and $\Gamma^S(SW)$.

**Proposition 1** Under our model’s assumptions, (i) $q_{NR} \leq q_{NS}$; and (ii) for any country $J$, $RW^J(\overline{q}_{NR}, q_f) \leq SW^J(\overline{q}_{NS})$, with equality holding in either case if and only if $q_f = q_{NS}$.

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. 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 $q_f \leq q_{NS}$). Therefore, the Nash equilibrium abatement standard of $\Gamma^R(RW, w, \overline{q}_f)$ 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 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 \( \Gamma^R_\infty (RW, w, \overline{q}_f) \), and the one with self-interested countries by \( \Gamma^S_\infty (SW) \). 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 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, \( \Gamma^S_\infty (SW) \). To derive the incentive-compatibility constraint for self-interested country \( J \), we first look at its static incentive to cheat, \( \Omega^J_S (q_C) \), 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 \( q_C \):

\[
\Omega^J_S (q_C) \equiv SW^J (BR^J_S (q_C), q_C) - SW^J (\overline{q}_C) \equiv SW^J_D - SW^J_C, \tag{4}
\]

where \( \overline{q}_C \equiv (q_C, q_C) \) and \( BR^J_S (q_C) \) 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:

\[
\frac{\delta}{1-\delta} (SW^J (\overline{q}_C) - SW^J (\overline{q}_{NS})) \equiv \frac{\delta}{1-\delta} (SW^J_C - SW^J_N) \equiv \frac{\delta}{1-\delta} \omega^J_S (q_C), \tag{5}
\]

where \( \overline{q}_{NS} \equiv (q_{NS}, q_{NS}) \) and \( \omega^J_S \) 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 \( \Gamma^S_\infty (SW) \) is that the onetime gain from defection, \( \Omega^J_S \), does not exceed the discounted future value of cooperation, \( (\delta / (1-\delta)) \omega^J_S \):

\[
\Omega^J_S (q_C) \leq \frac{\delta}{1-\delta} \omega^J_S (q_C). \tag{6}
\]

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

\[
\delta \geq \delta^S_{q_C} \equiv \frac{SW^J_D - SW^J_C}{SW^J_D - SW^J_N}. \tag{7}
\]
Analogous relationships hold for reciprocal countries. More specifically, the incentive-compatibility constraint for a reciprocal country $J$ to uphold international environmental cooperation in $\Gamma^R_{\infty} (RW, w, \overrightarrow{q})$ is given by:

$$\Omega^J_R (q_C) \leq \frac{\delta}{1 - \delta} \omega^J_R (q_C).$$

(8)

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

$$\delta \geq \delta^R_{q_C} \equiv \frac{RW^D_J - RW^J_C}{RW^D_J - RW^J_N}.$$  \hspace{1cm} (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 $\delta^S_{q_C}$ against $\delta^R_{q_C}$.

**Proposition 2** Any cooperative abatement standard $q_C \geq q_f$ can be more easily maintained by reciprocal rather than self-interested countries (i.e., $\delta^R_{q_C} < \delta^S_{q_C}$ for any $q_C \geq q_f$). However, it is ambiguous whether reciprocal or self-interested countries can more easily sustain any given cooperative abatement standard $q_C < q_f$.

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 $\Gamma^R_{\infty} (RW, w, \overrightarrow{q})$ 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 $\Gamma^R_{\infty} (RW, w, \overrightarrow{q})$ 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 one-time 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.

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 $\Gamma^R_{\infty}(RW, w, \overline{q}_f)$ 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 $\Gamma^R_{\infty}(RW, w, \overline{q}_f)$. However, the latter effect is relatively weak for a sufficiently small $\gamma$. On the other hand, country $J$ has a stronger incentive to defect in $\Gamma^R_{\infty}(RW, w, \overline{q}_f)$ than in $\Gamma^S_{\infty}(SW)$ (see Figure 2) since in the former game (i) the defect abatement standard is lower, because the countries 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. Let $\overline{q}_{CS} \equiv (q_{CS}, q_{CS})$ denote the most cooperative equilibrium abatement-standard vector for $\Gamma^S_{\infty}(SW)$, 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 $\overline{q}_{CR} \equiv (q_{CR}, q_{CR})$ represent the most cooperative equilibrium abatement-standard vector of $\Gamma^R_{\infty}(RW, w, \overline{q}_f)$. Furthermore, let us assume in the remainder of this section that $\delta \in [\delta, \tilde{\delta}]$ so that both self-interested and reciprocal countries can maintain some environmental cooperation, but $q_{OS}$ is infeasible for either of them.

**Proposition 3** If $q_f \leq q_{CS}$, the most cooperative equilibrium abatement standard of $\Gamma^R_{\infty}(RW, w, \overline{q}_f)$ is higher than the one of $\Gamma^S_{\infty}(SW)$ (i.e., $q_{CR} > q_{CS}$ if $q_f \leq q_{CS}$). However, if $q_f > q_{CS}$, then the effect of fairness and reciprocity on the most cooperative abatement-standard equilibrium of the dynamic game is ambiguous.

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.

---

16Observe that the most cooperative abatement-standard equilibrium is the most natural focal point for either game as (i) it is the only equilibrium of the desired class that is not Pareto dominated; and (ii) nothing precludes preplay communication between the countries.
To investigate further the latter scenario, we resort to numerical analysis (available from the authors upon request). In our simulations, we set (i) $b^J (q^J) = \alpha - (q^J)^2$, where $\alpha > (q_{max})^2$; (ii) $\Psi^J (q^J, q^{-J}) = (1/2) [\beta - q^J + s (\beta - q^{-J})]^2$, where $\beta > q_{max}$ and $s \in (0, 1)$ is the degree of transboundary pollution; and (iii) $w^J(q^{-J}, q^{-J}) = ((q^{-J} - q_f) / (q^{-J} + q_f)) \in (-1, 1)$. 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 $q_f$ 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 $\delta$ 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.\footnote{To see this, simply note that $q_{CS}$ is independent of $q_f$.} 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 cooperative or “inferior” environmental agreement than in the absence of these high expectations. It could then be argued that at the Cancun summit, countries managed to make a step forward (even if a modest one) relative to Copenhagen partly due to their lowered expectations.

## 5 Robustness

In this section, we discuss the robustness of our main predictions to relaxing some of the model’s assumptions. 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 self-interested and reciprocal countries can support would quantitatively change. Nevertheless, the ramifications of fairness and reciprocity for the incentives faced by the countries would be qualitatively preserved, and thus, 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) would 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, 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 would be binding in equilibrium for both countries. Therefore, our qualitative predictions would still be valid for each country taken individually. However, 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.” Such a framework though would not be suitable for our infinitely-repeated game. The reason is that the analysis would be highly intractable as, for example, the countries could 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 re-
ducing transboundary pollution if it expects the latter to behave kindly by setting a higher abatement standard than the one deemed fair; nevertheless, the reverse is true when the latter is expected to behave unkindly by implementing 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 provides 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. Our findings therefore suggest a plausible explanation for the failure of the 2009 Copenhagen summit and the more positive outcome of the 2010 Cancun one. In particular, it is now evident that countries entered the Copenhagen negotiations with overly ambitious expectations. Our analysis demonstrates that such high expectations (represented in our framework by high fair-abatement-standard perceptions) could prove to be counterproductive, hindering the efforts for deeper international environmental cooperation. The opposite is true for the Cancun summit. Expectations were much lower, and countries managed to make a step forward (even if a modest one) relative to Copenhagen.

Appendix

Existence and Uniqueness of Static Nash Equilibria

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

Proof. We first consider $\Gamma^R(RW, w, \overline{q}_J)$. 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) $[0, q_{\text{max}}]$ is a compact interval in $\mathcal{R}^+$; (ii) $RW^J$ is twice continuously differentiable on $[0, q_{\text{max}}]$; 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 Nash equilibria.

To see that \( \Gamma^S(SW) \) also admits pure 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.

**Inefficiency of Static Nash Equilibria**

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

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

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

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

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

\[
\frac{\partial SW^J(q^J, q^{-J})}{\partial q^J}\bigg|_{q^J=q_{NS}^J} + \frac{\partial SW^{-J}(q^{-J}, q^J)}{\partial q^J}\bigg|_{q^J=q_{OS}^J} = -\frac{\partial \Psi^{-J}(q^{-J}, q^J)}{\partial q^J}\bigg|_{q^J=q_{OS}^J} > 0. \tag{12}
\]

(12) along with the strict concavity of \( SW^J + SW^{-J} \) with respect to \( q^J \) together imply that \( q_{OS}^J > q_{NS}^J \). Given the overall symmetry of our framework, it then follows trivially that \( q_{OS} > q_{NS} \).

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

\[
\frac{\partial RW^J(q^J, q^{-J})}{\partial q^J} = \frac{\partial b^J(q^J)}{\partial q^J} - \frac{\partial \Psi^J(q^J, q^{-J})}{\partial q^J} - \gamma w^J(q^{-J}, q_f) \frac{\partial \Psi^{-J}(q^{-J}, q^J)}{\partial q^J} = 0. \tag{13}
\]
Moreover, given that \( q_{OR} = q_{OR}^J = q_{OR} \) is interior, it must be a solution to:

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

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

\[
\frac{\partial RW^J(q^J, q^{-J})}{\partial q^J} \bigg|_{q^J = q_{OR}} + \frac{\partial RW^{-J}(q^{-J}, q^J)}{\partial q^J} \bigg|_{q^J = q_{OR}^J} = -\frac{\partial \psi^{-J}(q^{-J}, q^J)}{\partial q^J} \bigg|_{q^J = q_{OR}^J} + \gamma \left[ \frac{\partial w^{-J}(q^J, q_f)}{\partial q^J} \bigg|_{q^J = q_{OR}^J} SW^J(q_{OR}^J, q^{-J}) + w^{-J}(q_{OR}^J, q_f) \frac{\partial SW^J(q_{OR}^J, q^{-J})}{\partial q^J} \bigg|_{q^J = q_{OR}^J} \right]. \tag{15}
\]

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

Given now the overall symmetry of our model, it follows trivially that \( q_{OR} > q_{NR} \).

We will finally show that \( q_{OR} > q_{OS} \). Using (11), (14), and the fact that \( q_{OS}^J = q_{OS}^{-J} = q_{OS} \), we obtain:

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

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

**Proof of Lemma 1**

Given that (i) \( \Gamma^R(RW, w, \overline{q}_f) \) is a supermodular game; and (ii) \( (\partial^2 RW^J / \partial q^J \partial q_f) < 0 \) for any \( J \), the lemma follows immediately from Theorem 6 in Milgrom and Roberts (1990). Q.E.D.
Proof of Proposition 1

If \( q_f = q_{NS} \), then trivially \( q_{NR} = q_{NS} \equiv q_N \), implying that for any \( J \), \( RW^J(\overline{q}_N, q_f) = SW^J(\overline{q}_N) \) since \( w^J(q_N, q_f) = 0 \) by (3) (where \( \overline{q}_N \equiv (q_N, q_N) \)). 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 \( q_{NR} < q_{NS} \), \( SW^J(\overline{q}_{NR}) < SW^J(\overline{q}_{NS}) \) for all \( J \). But then it follows that for all \( J \), \( RW^J(\overline{q}_{NR}, q_f) < SW^J(\overline{q}_{NS}) \). Q.E.D.

Proof of Proposition 2

We want to show first that \( q_C \geq q_f \) implies that \( \delta_{q_C}^R = \left( (RW_D^J - RW_C^J) / (RW_D^J - RW_N^J) \right) < \left( (SW_D^J - SW_C^J) / (SW_D^J - SW_N^J) \right) = \delta_{q_C}^S \). To do so, we will prove:

(i) If \( q_C \geq q_f \Rightarrow RW_D^J - RW_C^J \leq SW_D^J - SW_C^J \) for any \( J \).

(ii) If \( q_C \geq q_f \Rightarrow RW_D^J - RW_N^J > SW_D^J - SW_N^J \) for any \( J \).

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

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

Therefore:

\[
RW_D^J - RW_C^J = SW^J(BR_R^J(q_C), q_C) - SW^J(\overline{q}_C) + \gamma w^J(q_C, q_f)\left( SW^{-J}(BR_R^J(q_C), q_C) - SW^{-J}(\overline{q}_C) \right) \\
\leq SW^J(BR_R^J(q_C), q_C) - SW^J(\overline{q}_C) \leq SW^J(BR_S^J(q_C), q_C) - SW^J(\overline{q}_C) = SW_D^J - SW_C^J.
\]

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., \( SW^{-J}(BR_R^J(q_C), q_C) - SW^{-J}(\overline{q}_C) \leq 0 \). 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:

\[
q_C \geq q_f \Rightarrow (RW_D^J - SW_D^J) - (RW_N^J - SW_N^J) > 0 \quad \text{for any} \quad J.
\]

By Proposition 1 we know that the Nash equilibrium abatement level of \( \Gamma_S^f(SW) \) is (weakly) higher than that of \( \Gamma_R^f(RW, w, \overline{q}_f) \), i.e., \( q_{NR} \leq q_{NS} \). Thus, \( q_f \geq q_{NS} \geq q_{NR} \), implying that
Since we will show that if 
Proof of Proposition 3

such a higher, since the weight function is negative at around sufficiently small. Taking a first-order Taylor series expansion of RW\(^J\)(BR\(_R\)(q\(_C\)), q\(_C\), q\(_f\)) around \(\gamma = 0\), we obtain:

\[
RW\(^J\)(BR\(_R\)(q\(_C\)), q\(_C\), q\(_f\)) \approx SW\(^J\)(BR\(_S\)(q\(_C\)), q\(_C\)) + \gamma w\(^J\)(q\(_C\), q\(_f\))SW\(^J\)(BR\(_S\)(q\(_C\)), q\(_C\))
\]

\[
\Leftrightarrow RW\(^J\)(BR\(_R\)(q\(_C\)), q\(_C\), q\(_f\)) - SW\(^J\)(BR\(_S\)(q\(_C\)), q\(_C\)) \approx \gamma w\(^J\)(q\(_C\), q\(_f\))SW\(^J\)(BR\(_S\)(q\(_C\)), q\(_C\)) \geq 0 \Leftrightarrow RW\(^J\)_D - SW\(^J\)_D \geq 0.
\]  

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.

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\(^J\)_N = \delta\(_S\)_{q\(_CS\)} (SW\(^J\)_D - SW\(^J\)_N) \quad \text{and} \quad RW\(^J\)_D - RW\(^J\)_N = \delta\(_R\)_{q\(_CS\)} (RW\(^J\)_D - RW\(^J\)_N).
\]

Since \(\delta\(_R\)_{q\(_CS\)} < \delta\(_S\)_{q\(_CS\)}:

\[
RW\(^J\)_D - RW\(^J\)_N < \delta\(_S\)_{q\(_CS\)} (RW\(^J\)_D - RW\(^J\)_N)
\]

\[
\Leftrightarrow (1 - \delta\(_S\)_{q\(_CS\)}) RW\(^J\)(BR\(_R\)(q\(_CS\)), q\(_CS\), q\(_f\))
\]

\[
< RW\(^J\)(\(\bar{q}\)\(_CS\), q\(_f\)) - \delta\(_S\)_{q\(_CS\)} RW\(^J\)(\(\bar{q}\)\(_NR\), q\(_f\)),
\]  

meaning that \(\Omega\(^J\)\(_R\)(q\(_CS\)) < (\(\delta\(_S\)_{q\(_CS\)} / (1 - \delta\(_S\)_{q\(_CS\)})\) \(\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 $\Gamma^N_\infty(SW)$, $q_{CS}$, 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 $q_{CS}$:

$$RW^J(\overline{q}^C_C, q_f) > RW^J(\overline{q}^C_{CS}, q_f).$$

By the continuity of $RW^J(\bullet)$, then there exists a cooperative abatement standard $\overline{q}^C_C > q_{CS}$ such that (20) still holds, or $\Omega^J_R(\overline{q}^C_C) < \left(\frac{\delta_{q_{CS}}^S}{1-\delta_{q_{CS}}^S}\right) \omega^J_R(\overline{q}^C_C)$. Since the same analysis applies to any $(q_{CS}, \delta_{q_{CS}}^S)$ pair for $\delta_{q_{CS}}^S \in [\tilde{\delta}_C, \overline{\delta}_C]$, we have that for any $\delta \in [\tilde{\delta}_C, \overline{\delta}_C]$, $q_C < q_{CR}$.

However, as we argued in the proof of Proposition 2, if $q_f > q_{CS}$, it is ambiguous by (17) and (19) whether $\delta_{q_{CS}}^R$ or $\delta_{q_{CS}}^S$ 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., $\delta_{q_{CS}}^R > \delta_{q_{CS}}^S$. 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.\(^{18}\) From (9) and (7), we have:

$$RW^J_D - RW^J_C = \delta_{q_{CS}}^S (RW^J_D - RW^J_N)$$

and

$$SW^J_D - SW^J_C = \delta_{q_{CS}}^S (SW^J_D - SW^J_N).$$

Since $\delta_{q_{CS}}^R > \delta_{q_{CS}}^S$:

$$RW^J_D - RW^J_C > \delta_{q_{CS}}^S (RW^J_D - RW^J_N) \iff \left(1 - \delta_{q_{CS}}^S\right) RW^J(\overline{q}^C_{CS}, q_{CS}, q_f) > \left(1 - \delta_{q_{CS}}^S\right) RW^J(\overline{q}^C_{CR}, q_{CS}, q_f),$$

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

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 $\Omega^J_R(\bullet)$ and the strict concavity of $\omega^J_R(\bullet)$, it is obvious that the incentive-compatibility condition for reciprocal country $J$ can only be restored at a coop-

\(^{18}\) This assumption is clearly not restrictive given the type of result we are seeking here.
ervative abatement standard $\hat{q}_C < q_{CS}$. Since the same analysis applies to any $(q_{CS}, \delta_{q_{CS}})$ pair for $\delta_{q_{CS}}^S \in [\hat{\delta}, \delta]$, we have that for any $\delta \in [\hat{\delta}, \delta]$, $q_{CS} > q_{CR}$.

Nevertheless, $\delta_{q_{CS}}^R < \delta_{q_{CS}}^S$ 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 $\delta_{q_{CS}}^R$ or $\delta_{q_{CS}}^S$ is higher. Q.E.D.

References


